Universidad de Valladolid

Facultad de Ciencias

Grado en matemáticas

TRABAJO DE FIN DE GRADO:

EL MÉTODO DE K MEDIAS

AUTOR:
Carla Perucha Jurjo

TUTOR:
Carlos Matrán Bea

Julio 2022

Índice

1. Introducción **2**

2. Presentación del método de k medias **4**
2.1. Planteamiento intuitivo del problema de agrupación 4
2.2. La media como representante de los datos 5
2.3. Aplicaciones de k medias 7

3. Estudio teórico de k medias **10**
3.1. Planteamiento del modelo general de k medias 10
3.2. Propiedades matemáticas 12
3.2.1. Existencia de k medias 13
3.2.2. Unicidad del problema de k medias 16
3.2.3. Puntos frontera entre clusters 23
3.3. Consistencia del método de k-medias 25
3.3.1. Convergencia de k-ϕ-medias empíricas a k-ϕ-medias teóricas 29

4. Algoritmo de k medias **32**
4.1. Necesidad de algoritmos 32
4.2. Tres algoritmos para el problema de k medias 34
4.2.1. El algoritmo de Lloyd (1957) 35
4.2.2. El algoritmo de MacQueen (1967) 35
4.2.3. El algoritmo de Hartigan-Wong (1979) 36
4.3. Principales carencias de k medias 37
4.3.1. Conjuntos de datos apropiados para k medias 37
4.3.2. Elección de k, el número de grupos a buscar 39
4.3.3. Fuerte dependencia de la inicialización 42
4.4. Implementación 44

5. Métodos y estrategias para mejorar la inicialización **52**
5.1. Algunos métodos para la inicialización de k medias 52
5.1.1. Principales procedimientos para la inicialización 53
5.2. El procedimiento de k medias++ 54
5.2.1. Marco teórico de k medias++ 57
5.2.2. Una cota superior para el problema 58
5.2.3. Una cota inferior para el problema 67
5.2.4. Algunas generalizaciones 71
5.2.5. Ejemplos de la aplicación de k medias++ 72

6. Apéndice **75**
6.1. Algunos resultados auxiliares 75
6.2. Algunas nociones sobre complejidad computacional 78

1. Introducción

En las últimas décadas, hemos sido testigos de un avance tecnológico monumental y un crecimiento exorbitado del mundo digital. Aplicaciones como búsquedas en internet, creación de páginas web, procesamiento de imágenes, vigilancia por video... han generado conjuntos de datos de tamaño extraordinario. De igual manera, no solo se ha registrado un crecimiento en la cantidad de información sino que también se ha incrementado la variedad de tipos de datos (texto, imagen, video).

Debido a este aumento tanto de volumen como de variedad en los tipos de datos, la utilización de las técnicas propias del Análisis de Datos se ha generalizado en todo tipo de estudios e investigaciones. El Análisis Cluster o de Conglomerados ocupa un lugar destacado en esta ciencia: es la técnica multivariante encargada del estudio de métodos y algoritmos diseñados para agrupar objetos de acuerdo con sus características intrínsecas. Para realizar un Análisis Cluster, partimos de n individuos de los cuales hemos tomado medidas en d variables, en búsqueda de la agrupación natural de los objetos, buscamos asignar cada objeto a un cluster (o grupo) de tal manera que los miembros de un mismo grupo sean lo más homogéneos posibles y lo más diferentes posibles a los individuos de los otros grupos.

El Análisis Cluster se encuadra entre los métodos de clasificación no supervisada: nuestra intención es agrupar individuos en clusters homogéneos sin tener un conocimiento a priori de las categorías o tipologías de los objetos con los que trabajamos.

De las muchas técnicas de agrupamiento relativas al Análisis Cluster, en este trabajo estudiaremos detenidamente el método de k medias. La media de un conjunto de datos pretende ser el mejor representante de estos de acuerdo con el criterio de mínimos cuadrados. Partiendo de esto, el procedimiento de k medias extiende la idea natural de media y busca el mejor representante formado por no uno sino k elementos del espacio. De esta manera, una vez escogidos estos k puntos denominados centroides, se genera una partición del conjunto de datos en k grupos o clusters donde asignamos cada punto al representante más "similar" a él. Es un método muy interesante debido a su sencillez conceptual, el amplio estudio que se ha realizado en cuanto a su comportamiento asintótico, su simplicidad de implementación y el hecho de que, a pesar de ser un problema computacionalmente difícil, existen numerosas heurísticas eficientes que permiten una convergencia rápida a un óptimo local. Si bien es un procedimiento que se publica por primera vez en 1955, es a día de hoy uno de los más utilizados en el campo del Análisis Cluster.

El problema de k medias y sus diferentes aspectos tanto teóricos como computacionales siguen siendo objeto de estudio a día de hoy, por lo que encontramos artículos bastante recientes que recogen nuevos resultados del método o presentan propuestas de algoritmos más eficientes en la inicialización. Este documento tiene como objetivo introducir el procedimiento de k medias y recoger algunos de los resultados matemáticos más interesantes, incluyendo algunos menos comunes, ya que el método de k medias es tratado habitualmente desde la perspectiva algorítmica y no muy frecuentemente encontramos resultados teóricos en los manuales de k medias.

Desde el punto de vista computacional, el problema de encontrar k representantes que consigan satisfacer el criterio de mínimos cuadrados es NP-Hard. Los algoritmos existentes son iterativos y parten de una inicialización de k elementos escogidos aleatoriamente con el fin de llegar a una solución óptima localmente. Cabe entonces preguntarse si una buena inicialización proporciona mejoras

notables en la posterior agrupación de individuos utilizando el método. La respuesta es claramente afirmativa y mostraremos algunos de los procedimientos más exitosos a la hora de seleccionar centroides iniciales adecuados, provocando así no solo una mejora en la solución sino menor tiempo de ejecución del algoritmo.

Para concluir esta introducción, decir que en este documento hemos generado multitud de ejemplos visuales empleando el lenguaje de programación R, ampliamente utilizado en estadística. En este entorno de programación existen múltiples paquetes que implementan algoritmos de clustering (en particular k medias) y funciones para visualizar sus resultados. En este documento se emplean los siguientes:

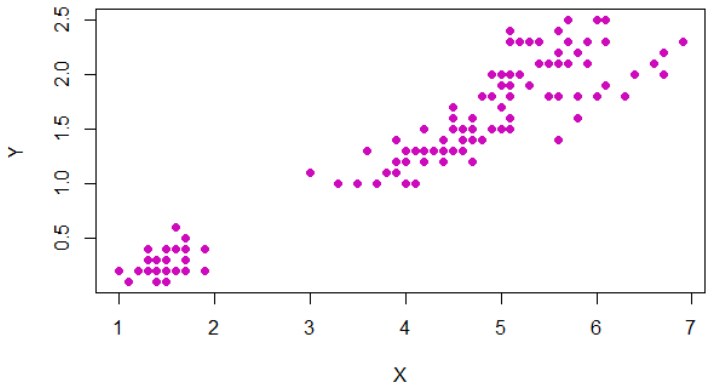
- stats: contiene las funciones $dist()$ para calcular matrices de distancias, $kmeans()$ y, a pesar de que no serán utilizados en este trabajo, funciones que nos permiten aplicar otro tipo de procedimientos como $hclust()$, $cuttree()$ para crear los clusters y $plot.hclust()$ para visualizar los resultados.
- factoextra: contiene las funciones $fviz\_cluster()$ para elaborar elegantes visualizaciones de las agrupaciones en clusters, $fviz\_nbclust()$ para determinar y visualizar el número óptimo de clusters, y otras muchas funciones relativas a otras técnicas multivariantes.
- flexclust: Contiene la función $kcca()$ que permite ejecutar el algoritmo de k medias con una variante interesante denominada k medias++ que permite una elección de centros iniciales mucho más satisfactoria cuando las condiciones son las adecuadas.

2. Presentación del método de k medias

En esta sección, expondremos un primer ejemplo que nos acercará de manera intuitiva al Análisis Cluster y al propósito de buscar agrupaciones en un conjunto de individuos. Veremos también cómo a la hora de trabajar con conjuntos de datos, a menudo estamos interesados en encontrar un representante de estos que permita describirlos de la mejor manera posible, pues resumir la información nos permite comprenderla. Las medidas de centralización tienen un papel importante en esta síntetización que tratamos de elaborar ya que nos proporcionan un representante ubicado en el "centro" del conjunto de los datos. En particular, la media resulta ser el más interesante según el criterio de mínimos cuadrados. Presentaremos también en esta sección un primer acercamiento a k medias desde la perspectiva de construir una generalización del concepto de media, cuyo papel es el de ser el representante natural de un conjunto de datos. Como ya hemos comentado, los ejemplos expuestos a lo largo de este trabajo serán procesados con el lenguaje de programación R, de uso extendido en el ámbito de la estadística.

2.1. Planteamiento intuitivo del problema de agrupación

Al inicio de este documento, comentábamos que el propósito del Análisis Cluster o de Conglomerados es buscar agrupaciones naturales que puedan servir para hallar relaciones en un conjunto de datos que sean útilies a la hora de clasificarlos. Para comenzar, presentamos un conjunto de datos constituido por 150 observaciones de individuos a los que hemos medido valores en 2 variables, X e Y, y plantearemos la manera intuitiva de buscar agrupaciones entre los individuos atendiendo a los valores que toman en estas dos variables. Estas serían algunas de las primeras observaciones del conjunto de datos y la representación de los 150 individuos si los consideramos como puntos en $\mathbb{R}^2$:



(a) Representación en el plano de los individuos

	X	Y
1	1.4	0.2
2	1.4	0.2
3	1.3	0.2
4	1.5	0.2
5	1.4	0.2
6	1.7	0.4
7	1.4	0.3
8	1.5	0.2
9	1.4	0.2
10	1.5	0.1

(b) Encabezado de los datos

Figura 1

Nuestro objetivo es encontrar una agrupación de los individuos de tal manera que aquellos que

se encuentren en el mismo grupo sean lo más homogéneos entre sí, atendiendo a los valores que toman en las dos variables que estamos considerando. Estos grupos de individuos reciben el nombre de *clusters* en la Ciencia del Análisis de Conglomerados.

Cuando uno trata de agrupar los individuos de características similares atendiendo al gráfico, se hace patente que existe un conjunto de observaciones cuyas medidas tanto en la variable X como en la variable Y son claramente menores. Parece lógico reunir todos estos individuos en un mismo grupo ya que se parecen entre sí y sus medidas son muy diferentes de las del resto de individuos.

En segundo lugar, tal vez con un poco más de esfuerzo, reparamos en que hay unas cuantos individuos con mediciones intermedias de ambos valores y otros tantos cuyas medidas en las dos variables son bastante más altas. Al contrario que antes, no es clara la separación entre estos dos grupos: parece que individuos del segundo grupo con valores relativamente altos en las variables se confunden con individuos del tercer grupo con valores relativamente bajos de las variables. Aun así, somos capaces de "cortar" este grupo y, con lo mencionado anteriormente, elaborar una clasificación intuitiva de los individuos en tres grupos.

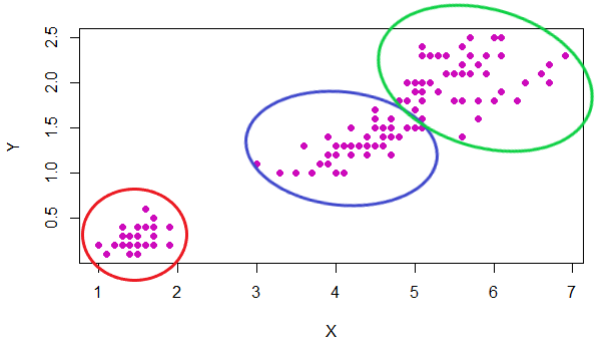


Figura 2: Agrupación intuitiva que podríamos hacer

Dos medidas muy importantes que utilizamos para medir la similaridad entre individuos son la distancia intra-cluster y la distancia inter-cluster: Nos parece natural reunir individuos cuyos valores en las variables distan poco entre sí y que se diferencian mucho de las mediciones en las variables de otros individuos. Por lo tanto, nuestro método de clustering deberá tener como objetivo minimizar las distancias entre individuos de una misma agrupación y maximizar la distancia entre individuos de diferentes grupos.

2.2. La media como representante de los datos

Como ya se ha comentado, la media de un conjunto de datos suele utilizarse como un representante natural de estos. La justificación se encuentra en el hecho de ser el elemento que minimiza la dispersión

cuadrática media:

Proposición 1. *Sea* $C = \{x_1, .., x_n\}$ *un conjunto con* n *puntos de* $\mathbb{R}^d$, $w_1, .., w_n \in \mathbb{R}$ *con* $w_i \geq 0$ $\forall i = 1, .., n$ *y* $\sum_{i=1}^n w_i = 1$, *y sea* $m = \sum_{i=1}^n x_i w_i$. *Entonces*

$$m = \arg\min_{a \in \mathbb{R}^d} \sum_{i=1}^n w_i \|x_i - a\|^2 \tag{1}$$

Demostración

Sea $F : \mathbb{R}^d \longrightarrow \mathbb{R}$ la función definida por

$$F(y) = \sum_{i=1}^n w_i \|x_i - y\|^2$$

Podemos reescribir la función F en coordenadas como

$$F(y) = \sum_{i=1}^n \left(\sum_{j=1}^d w_i (x_{i,j} - y_j)^2 \right)$$

Para estudiar los puntos críticos de F, dado que es derivable en $\mathbb{R}^d$, calculamos su gradiente $\nabla F(y) = \left(\frac{\partial F}{\partial y_1}, .., \frac{\partial F}{\partial y_d} \right)$ y lo igualamos a cero. Para $k = 1, .., d$, tenemos que

$$\frac{\partial F}{\partial y_k} = \frac{\partial}{\partial y_k} \left(\sum_{i=1}^n \sum_{j=1}^d w_i (x_{i,j} - y_j)^2 \right) = 2y_k \sum_{i=1}^n w_i - 2 \sum_{i=1}^n x_{i,k} w_i = 0 \Rightarrow y_k = \sum_{i=1}^n x_{i,k} w_i$$

Por lo tanto,

$$y = (y_1, .., y_d) = \left(\sum_{i=1}^n x_{i,1} w_i, .., \sum_{i=1}^n x_{i,d} w_i \right) = \sum_{i=1}^n (x_{i,1}, .., x_{i,d}) w_i = \sum_{i=1}^n x_i w_i = m$$

Dado que m es el único punto crítico y resulta que $\lim_{y \to \infty} F(y) = \lim_{y \to -\infty} F(y) = \infty$, debe ser un mínimo y queda probado el resultado. □

Surge entonces también de forma natural la búsqueda de un representante de nuestro conjunto de datos formado por k elementos del espacio. Un resumen en k representantes de un conjunto de puntos nos dará más información y permitirá describir mejor las observaciones sobre todo en el caso en el que existan diferencias significativas entre los valores que toman en las variables. Es decir, si denotamos $\mathcal{B}_k$ como la clase de conjuntos con k elementos de $\mathbb{R}^d$, definimos una k media de un conjunto de datos $\{x_1, .., x_n\} \subset \mathbb{R}^d$ dados pesos w_i como hemos descrito anteriormente como un conjunto $H \in \mathcal{B}_k$ que verifica

$$H = \arg\min_{G \in \mathcal{B}_k} \sum_{i=1}^n w_i \min_{g \in G} \|x_i - g\|^2 \tag{2}$$

Con esta definción, seguimos buscando la mejor descripción de acuerdo con el criterio de mínimos cuadrados, esta vez a través de k elementos. Este planteamiento parece interesante desde la perspectiva no solo de obtener un resumen en k puntos de un conjunto de datos sino de establecer un representante de cada cluster y asociar cada observación al conjunto cuyo representante es más parecido. El procedimiento que trata de formalizar esta idea natural es el método de k medias. El efecto que produciría el cálculo de $k = 3$ representantes (igual al número de grupos que habíamos formado) en el ejemplo anterior sería el siguiente, para el cual hemos utilizado la función *kmeans* de *R* cuya misión es buscar en este caso 3 representantes denominados centroides y agrupar las observaciones en función de qué centroide dista menos de cada una:

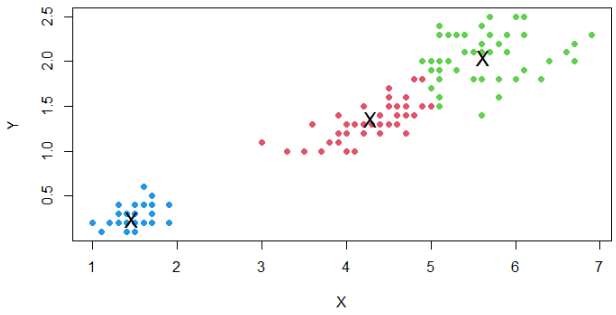


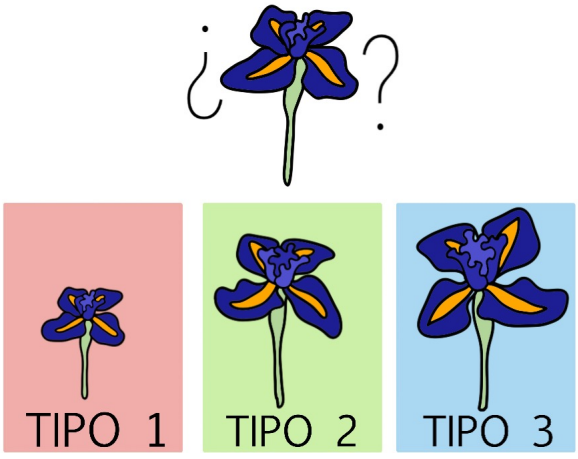
Figura 3: Tres representantes del conjunto de datos y la agrupación que provocan.

2.3. Aplicaciones de k medias

Podemos preguntarnos si el procedimiento de agrupación que hemos llevado a cabo al inicio del apartado con el fin de conseguir crear tipologías o clases de objetos similares entre sí nos ayuda a clasificar individuos cuando los datos poseen una organización subyacente o si la agrupación que hemos realizado no tiene nada que ver. En nuestro caso, el conjunto de datos con el que hemos trabajado es una reducción a dos variables del bien conocido conjunto de datos "Iris", disponible en R. En él se recogen 150 observaciones de tres especies de flores bastante similares entre sí: setosa, versicolor y virginica. Para cada tipo de flor, se ha medido la longitud y la anchura de sus pétalos. Para que la clasificación fuera no supervisada (el caso en el que desconocemos a priori a qué especie de flor pertenece cada una de las observaciones), simplemente habíamos eliminado la última columna del conjunto de datos que etiquetaba cada individuo, por lo que no teníamos conocimiento a priori ni de cuántas especies de flores habíamos podido recoger ni de cuáles eran.

Figura 4: Eliminación de etiquetas y de variables "Largura de sépalo" y "Anchura de sépalo"

Realizar una clasificación (o en nuestro caso, una agrupación) de los individuos atendiendo a sus medidas en dos variables cobra sentido ahora que conocemos la naturaleza de las observaciones, ya que parece razonable pensar que la longitud y la anchura de los pétalos de un tipo de flor pueda caracterizar una especie y diferenciarla de otros tipos de flores. De esta forma, es lógico tomar unos representantes del conjunto de individuos para realizar un resumen de qué diferentes flores hemos podido recoger y a la vez ayudarnos a agruparlas según sus similitudes entre sí. De esta manera tendríamos, de acuerdo con las dos variables, tres representantes del conjunto de flores: aquellas con petalos pequeños y estrechos, aquellas con pétalos medianos y aquellas con pétalos grandes y anchos. Esto nos permite establecer tres grupos y asociar cada observación al conjunto cuyo representante es más parecido.



Etiquetamos ahora cada observación con la especie a la que pertenece con la ayuda de la columna que habíamos eliminado en nuestros datos y comparamos con el resultado obtenido anteriormente por *kmeans*. Marcamos con un punto relleno aquellos individuos que k medias ha conseguido cla-

sificar correctamente por medio de la agrupación que realizábamos y con un punto sin rellenar las observaciones que no consigue clasificar correctamente de acuerdo con las verdaderas etiquetas de las variables. Así, vemos como en este ejemplo en concreto la búsqueda de tres representantes y la agrupación de los individuos de acuerdo con esto ha conseguido de algún modo rescatar las categorías que existían entre los objetos con bastante exactitud.

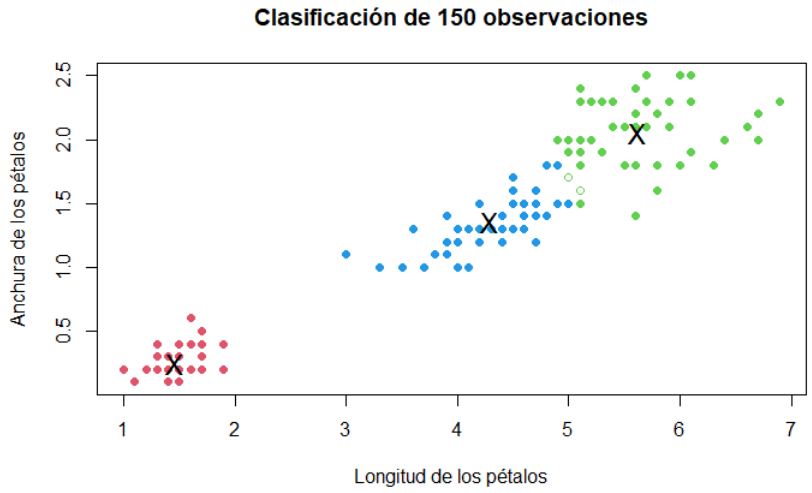


Figura 5

Después de este ejemplo, queda patente que la búsqueda de k representantes puede ser de gran utilidad bien a la hora de confirmar una partición en categorías cuando estas son conocidas, o bien para ser capaces de crear o definir nuevas clases dentro de un conjunto de individuos. Esta es la perspectiva más interesante del Análisis Cluster debido a que generalmente dado un conjunto de datos, desconocemos la tipología de los objetos e incluso el número de grupos a buscar, por lo que nuestra misión será construir categorías de individuos. Algunas de las aplicaciones del Análisis Cluster y en concreto del método de k medias son establecer agrupaciones de animales y plantas en especies distintas en la disciplina de la taxonomía, crear agrupaciones de productos que los clientes compran simultáneamente para ofrecer recomendaciones de compra en la venta online, o definir diferentes categorías de tumores atendiendo a sus propiedades y así poder aplicar posteriormente en cada caso el tratamiento más adecuado, en el campo de la medicina.

3. Estudio teórico de k medias

Una vez planteado el problema de agrupación e introducida la idea de escoger k representantes de un conjunto de individuos, buscamos plantear el problema de k medias de manera general, a través de un estudio teórico que incluya tanto conjuntos de datos como distribuciones de probabilidad. Para ello, será necesario adentrarse en el método de k medias desde una perspectiva probabilística y generalizar el concepto de media de un conjunto de datos a media de una variable aleatoria, por medio de la esperanza matemática. A su vez, extendemos el problema considerando no solo minimizar la suma de cuadrados sino otras posibles funciones que midan la disimilaridad entre los representantes y los valores que toma la variable aleatoria. De esta manera, conseguiremos formular finalmente el problema de k medias y enunciar un modelo teórico. Hablaremos a continuación sobre algunas propiedades matemáticas del método como la existencia, unicidad del problema y una característica mucho más concreta del procedimiento: La probabilidad de los puntos que se encuentran en la frontera entre dos grupos. Por último, veremos propiedades de consistencia del método que nos permitirán, en particular, asegurar que los representantes en un conjunto de datos obtenido como muestra de una distribución de probabilidad se acercan a los representantes de la distribución de probabilidad teórica. En todo este proceso, expondremos ejemplos elaborados con el lenguaje de programación R a fin de facilitar la visualización y la comprensión de los resultados planteados.

3.1. Planteamiento del modelo general de k medias

Hasta ahora, habíamos planteado la idea de buscar k elementos de $\mathbb{R}^d$ para configurar una descripción de un conjunto de puntos. Supongamos que en lugar de tener un conjunto con n datos contamos con una distribución de probabilidad. Si X es una variable aleatoria definida en un espacio probabilístico $(\Omega, \sigma, \mathbb{P})$ y P la ley de probabilidad que induce en $\mathbb{R}$, la generalización de la media de X a través del concepto de integral respecto de una probabilidad se denomina esperanza matemática y se define como

$$\mathbb{E}(X) = \int_{\mathbb{R}} x dP(x) \tag{3}$$

Para variables aleatorias multidimensionales, su esperanza o valor esperado se define componente a componente, esto es, dado un vector aleatorio $X = (X_1, .., X_d) : \Omega \longrightarrow \mathbb{R}^d$, definimos su esperanza como

$$\mathbb{E}[(X_1, .., X_d)] = [\mathbb{E}(X_1), .., \mathbb{E}(X_d)]$$

Este planeamiento del problema mediante distribuciones de probabilidad generaliza el caso de tener un conjunto con n puntos. En el caso de tener $x_1, .., x_n$ elementos de $\mathbb{R}^d$, podemos considerar P_n, la probabilidad muestral que asigna probabilidad $\frac{1}{n}$ a cada x_i, y así tendríamos que la esperanza dada en (3) no es sino el promedio de los puntos como veníamos haciendo hasta ahora:

$$\mathbb{E}(X) = \frac{1}{n} \sum_{i=1}^{n} x_i = \int x dP_n(x)$$

La esperanza de nuevo de nuevo es la mejor representación de la variable aleatoria X de acuerdo con el criterio de mínimos cuadrados:

$$\mathbb{E}(X) = \arg\min_{a\in\mathbb{R}^d} \int_{\mathbb{R}} \|x-a\|^2 dP(x) = \arg\min_{a\in\mathbb{R}^d} \int_{\Omega} \|X(\omega)-a\|^2 d\mathbb{P}(\omega) \tag{4}$$

Esto se debe a que

$$\int \|X(\omega)-a\|^2 d\mathbb{P}(\omega) = \mathbb{E}\|X-a\|^2 = \mathbb{E}\|X\|^2 + \|a\|^2 - 2\mathbb{E}(\langle X, a\rangle)$$

Utilizando que la esperanza es un operador lineal y que la esperanza de una constante es la misma constante, los mismos argumentos que utilizamos antes, garantizan (4).

En estas condiciones, de nuevo podemos pensar en tomar k representantes en lugar de sólo uno, de manera que definiríamos una k media de una variable aleatoria como un conjunto $H \in \mathcal{B}_k$ que verifica

$$H = \arg\min_{G\in\mathcal{B}_k} \int \min_{g\in G} \|x-g\|^2 dP(x) = \arg\min_{G\in\mathcal{B}_k} \int \min_{g\in G} \|X-g\|^2 d\mathbb{P} \tag{5}$$

La caracterización de la media dada en (4) invita a la consideración de otros representantes vinculados a otras medidas de disimilaridad. Podemos considerar entonces una función $\phi : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$ creciente y adecuada en cada caso con el proposito de medir la discrepancia entre puntos x e y como $\phi(\|x-y\|)$. De este modo, podemos extender el concepto de media al de ϕ-media.

Definición 1. *Sea $X : (\Omega, \sigma, \mathbb{P}) \longrightarrow \mathbb{R}^d$ un vector aleatorio y sea P la probabilidad inducida por X en $\mathbb{R}^d$. Sea $\phi : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$ una función creciente, continua y con $\phi(0) = 0$. Suponemos que $\exists a \in \mathbb{R}^d$ tal que $\int \phi(\|x-a\|)dP(x) < \infty$. El vector $m \in \mathbb{R}^d$ es una ϕ-media de X si verifica*

$$m = \arg\min_{a\in\mathbb{R}^d} \int \phi(\|x-a\|)dP(x)$$

Y finalmente, al igual que extendíamos la noción de media en la sección anterior, definimos una clase mucho más general de ϕ-medias al considerar una descripción de X en k elementos del espacio: las denominadas k-ϕ-medias.

Definición 2. *Sea $X : (\Omega, \sigma, \mathbb{P}) \longrightarrow \mathbb{R}^d$ un vector aleatorio y sea P la probabilidad inducida por X en $\mathbb{R}^d$. Sea k un entero positivo y $\phi : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$ una función creciente, continua y que verifica $\phi(0) = 0$. Suponemos que $\exists G$ de k elementos tal que $\int \min_{g\in G} \phi(\|x-g\|)dP(x) < \infty$. Dado A un subconjunto de $\mathbb{R}^d$, el problema de k medias consiste en encontrar un conjunto $H \in \mathcal{B}_k$ que minimice la expresión*

$$W_\phi^k(H, P, A) = \int_A \min_{h\in H} \phi(\|x-h\|)dP(x) = \int \min_{h\in H} \phi(\|X-h\|)d\mathbb{P} \tag{6}$$

Un conjunto que verifique esta condición se denominará k-ϕ-media de X. Denominamos k-potencial por H penalizado por ϕ de la variable aleatoria X en A a la expresión a minimizar.

Observación 1. *Para facilitar la lectura, frecuentemente descartaremos algunos de los elementos presentes en la notación, cuando el contexto los haga innecesarios. Por ejemplo, en el caso en el que $A = \mathbb{R}^d$ con la notación anterior, nos limitaremos a escribir $W_\phi^k(H, P)$.*

Del mismo modo, nos referiremos a la función a minimizar simplemente como k-ϕ-potencial por H de X cuando no exista equivocación posible y en el caso de ser ϕ la distancia euclídea al cuadrado (La función más conocida y utilizada en este problema), omitiremos ϕ de todas las notaciones.

Nuestro objetivo desde el punto de vista de crear agrupaciones es aquel de construir grupos de puntos que denominábamos clusters en función de cuál es el representante más similar a ellos. De esta manera, si $H = \{h_1, .., h_k\}$ es una k-ϕ-media de X, podemos construir una partición de $\mathbb{R}^d$ en conjuntos, $\mathscr{C}_H = \{C_1, .., C_k\}$, donde cada C_i esté constituido por los elementos de $\mathbb{R}^d$ que distan menos del elemento h_i que del elemento h_j para $i \neq j$. Es decir

$$C_i = \{x \in \mathbb{R}^d / \|x - h_i\| \leq \|x - h_j\| \text{ para } j \neq i\}$$

Estos k elementos se denominan centros de cluster o centroides en el contexto del método de k medias. Dado que pueden existir puntos que equidisten de dos o más centroides, definimos una manera única de construir la partición por ser ϕ creciente:

$$C_1 = \{x \in \mathbb{R}^d / \|x - h_1\| \leq \|x - h_j\|, j = 1, .., k\}$$
$$C_i = \{x \in \mathbb{R}^d / \|x - h_i\| \leq \|x - h_j\|, j = 1, .., k\} - C_{i-1}, \quad i = 1, .., k$$

3.2. Propiedades matemáticas

Planteado el problema de k medias, abordaremos en esta sección su estudio teórico. En primer lugar, plantearemos las hipótesis necesarias para poder garantizar la existencia de una solución al problema de k medias. A continuación, nos plantearemos si el problema tiene unicidad en sus soluciones, respuesta que es claramente negativa salvo en el caso $k = 1$ y tras aplicar unas condiciones más restrictivas a la función de disimilaridad ϕ que utilizamos para medir la discrepancia entre los centroides y los puntos. Seguido de esto, demostraremos una interesante propiedad del procedimiento de k medias: La probabilidad de los puntos que se encuentran en la frontera entre dos o más clusters es cero. Por último, estudiaremos la consistencia del método y presentaremos el resultado particular en el que trabajamos con k-ϕ-medias muestrales, especialmente interesante ya que a menudo el problema de k medias aparece en un ámbito estadístico en el que contamos con conjuntos de datos que podemos interpretar como una muestra.

Estaremos interesados un poco más adelante en probar convergencias entre conjuntos del mismo número de puntos. Decimos que $H_n = \{h_1^n, .., h_k^n\}$ converge a $H = \{h_1, .., h_k\}$ si para cada n, existe un reordenamiento de $\{h_1^n, .., h_2^n\}$, digamos $(h_{(1)}^n, .., h_{(k)}^n)$, de tal manera que $h_{(j)}^n$ converge a h_j^0. Dado que esta notación se vuelve complicada debido al renombramiento de los puntos, usaremos por comodidad la convergencia en el sentido de Hausdorff (que no es sino una manera más sencilla de escribir lo que estamos diciendo) y veremos que son equivalentes.

Definición 3. *Sea $(\mathbb{R}^d, D)$ un espacio métrico donde D es la distancia Euclídea y sean $A, B \subset \mathbb{R}^d$. Sea $\rho(A, B) = \sup\{D(a, B) : a \in A\}$ y $\rho(B, A) = \sup\{D(b, A) : b \in B\}$. Si $\mathcal{C}$ es la clase de los conjuntos compactos de $\mathbb{R}^d$, definimos la distancia de Hausdorff $\mathcal{H} : \mathcal{C} \times \mathcal{C} \longrightarrow [0, \infty)$ como*

$$\mathcal{H}(A, B) = \text{máx}\{\rho(A, B), \rho(B, A)\}$$

Para el caso que nos ocupa, bastará con considerar la clase de conjuntos $\mathcal{B}_k$ que habíamos definido anteriormente.

Proposición 2. *La convergencia en el sentido de la métrica de Hausdorff en los conjuntos de $\mathcal{B}_k$ es equivalente a la convergencia en $\mathbb{R}^d$ punto a punto.*

La prueba de este resultado está disponible en el Apéndice (4). Una vez demostrado esto, se utilizará indistíntamente cualquiera de las dos convergencias para hablar de $H_n \longrightarrow H_0$, ya que se entenderán como equivalentes por ser $H_n, H_0 \in \mathcal{B}_k$.

El siguiente lema nos será de gran utilidad para probar la convergencia de H_n a H_0, permitiéndonos trabajar con subsucesiones, siendo una propiedad universal para todas las convergencias asociadas a métricas. De nuevo, encontramos su demostración en el Apéndie (9).

Lema 1. *Sea $\{H_n\}_{n=0}^{\infty}$ una sucesión de elementos de $\mathcal{B}_k$. Entonces H_n converge a H_0 si y solo si toda subsucesión de $\{H_n\}_n$ admite una nueva subsucesión convergente a H_0.*

Con el fin de facilitar o ilustrar ciertos resultados que demostraremos a continuación, fluctuaremos entre escritura en términos de variables aleatorias y escritura como puntos de $\mathbb{R}^d$ de las k medias según resulte una redacción más clara en una situación u otra.

3.2.1. Existencia de k medias

En este apartado, abordaremos el problema de existencia de conjuntos de k puntos que minimicen la expresión (6). La existencia de k-ϕ-medias no está en absoluto garantizada sin imponer condiciones sobre la variable X; simplemente pensando en la media de una variable aleatoria nos damos cuenta de que puede tomar valores infinitos. Consideremos el siguiente ejemplo: Sea X la variable aleatoria que toma los valores $\frac{(-1)^j 3^j}{j}$ con probabilidad $\mathbb{P}(X = \frac{(-1)^j 3^j}{j}) = \frac{2}{3^j}$, para $j = 1, 2, ...$ Sabemos que hemos definido una probabilidad dado que la suma de la serie geométrica $\sum_{j=1}^{\infty} \frac{2}{3^j} = 1$. Sin embargo, dado que

$$\sum_{j=1}^{\infty} |x_j| p_{x_j} = \sum_{j=1}^{\infty} \frac{|(-1)|^j 3^j}{j} \frac{2}{3^j} = \sum_{j=1}^{\infty} \frac{2}{j} = \infty$$

por ser una serie divergente Riemann, no existe la esperanza de la variable X.

Los resultados presentes en esta sección son una adaptación de los que encontramos en el artículo [4], donde se trabaja con una clase más general de k-ϕ-medias denominadas k-ϕ-medias recortadas. En él, se presenta una técnica que consiste en eliminar parte de los datos de manera no arbitraria con el fin de hacer más robusto el procedimiento de k medias.

Nuestro principal objetivo será demostrar el siguiente resultado, que impone ciertas condiciones sobre la variable X con el fin de asegurar la existencia de sus k-ϕ-medias.

Teorema 1. *Sea $X : \Omega \longrightarrow \mathbb{R}^d$ un vector aleatorio con $\mathscr{L}(X) = P$, k entero positivo y $\phi : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$ una función creciente, continua, tal que $\phi(0) = 0$ y que verifica $\phi(x) < \phi(\infty)$ $\forall x$, donde entendemos $\phi(\infty) = \lim_{x\to\infty} \phi(x)$. Supongamos además que se verifica $\int \inf_{g\in G_0} \phi(\|x - g\|) dP(x) < \infty$ para algún $G_0 \in \mathcal{B}_k$. En estas condiciones, podemos afirmar que existe una k-ϕ-media de X.*

Observación 2. *Denotamos por $V_\phi^k(P)$ el mínimo k-ϕ-potencial de X que podamos obtener eligiendo k puntos, es decir, $V_\phi^k(P) = \inf_{H\in\mathcal{B}_k} W_\phi^k(H, P)$.*

Bajo las hipótesis del teorema anterior, sabemos que $V_\phi^k(P)$ es finito:

$$V_\phi^k(P) = \inf_{G \in \mathcal{B}_k} W_\phi^k(G, P) \leq W_\phi^k(G_0, P) < \infty$$

Probaremos previamente un lema que nos ayudará a demostrar el teorema principal.

Lema 2. *Sea $H = \{h_1, .., h_k\}$ un subconjunto de $\mathbb{R}^d$. Son equivalentes*

1. *$W_\phi^k(H, P) > 0$*
2. *Existe $h_0 \in \mathbb{R}^d$ tal que $W_\phi^{k+1}(H \cup h_0, P) < W_\phi^k(H, P)$ Es decir, a no ser que el k-ϕ-potencial por H de X sea exactamente cero, siempre podemos añadir centroides para reducirlo.*

Demostración

Es claro que 2) $\Rightarrow$ 1), ya que por definición el k-ϕ-potencial de X es mayor o igual que cero. Demostramos entonces 1) $\Rightarrow$ 2). Denotamos por $Sop(P)$ el soporte de la probabilidad P. Consideramos el caso en el que este es acotado y el caso en el que no.

- Si $Sop(P)$ es acotado: Sea $r = \text{máx}_x\{\text{ínf}_{h \in H}\|x - h\|\}$, es decir, la distancia máxima entre uno de los valores que toma la variable X y su centroide asociado. Sabemos que este valor es finito ya que $\exists M > 0$ tal que $Sop(P) \subset \overline{B}(0, M)$ y los puntos con $\|x\| > M$ tienen probabilidad cero. Construimos entonces la bola $B_0 = B(h_0, r_0)$ en $\mathbb{R}^d$ que verifique
  1. $\int_{B_0} dP(x) > 0$
  2. El centro de la bola h_0 dista al menos $\frac{2}{3}r$ del resto de centroides. Es decir, $\text{ínf}_{i=1,..,k}\|h_0 - h_i\| > \frac{2}{3}r$.
  3. El radio debe ser $r_0 < \frac{1}{3}r$
- Si $Sop(P)$ no es acotado: Para todo $r > 0$, $\exists h_0 \notin B(0, r)$ tal que $P(B(h_0, r_0)) > 0$ para todo $r_0 > 0$. De esta manera, tomamos r suficientemente grande y r_0 suficientemente pequeño para que los puntos $x \in B(h_0, r_0)$ verifiquen $\|x - h_0\| < \|x - h_i\|$. Construimos así $B_0 = B(h_0, r_0)$

De esta manera, si tomamos $x \in B_0$, tenemos que para $i = 1, .., k$,

$$\|x - h_i\| \geq \|h_0 - h_i\| - \|x - h_0\| > \frac{2}{3}r - \frac{1}{3}r = \frac{1}{3}r$$

Mientras que para h_0 se da

$$\|x - h_0\| < r_0 < \frac{1}{3}r$$

Así, resulta que $\text{mín}_{h \in H \cup h_0}\|x - h\| = \|x - h_0\|$ cuando $x \in B_0$.

De este modo, escribiendo las integrales en términos de B_0 y B_0^c, llegamos a

$$W_\phi^k(H, P) = \int \inf_{h \in H} \phi(\|x - h\|) dP(x) = \int_{B_0} \inf_{h \in H} \phi(\|x - h\|) d\mathbb{P} + \int_{B_0^c} \inf_{h \in H} \phi(\|x - h\|) dP(x) >$$

$$> \int_{B_0} \phi(\|x-h_0\|)dP(x) + \int_{B_0^c} \inf_{h\in H} \phi(\|x-h\|)dP(x)$$

Si agrupamos ambos trozos, llegamos a la siguiente desigualdad:

$$W_\phi^k(H,P) > \int \min\{\phi(\|x-h_0\|), \inf_{h\in H}(\phi(\|x-h\|)\}dP(x) = \int \inf_{h\in H\cup h_0} \phi(\|x-h\|)dP(x) = W_\phi^{k+1}(H\cup h_0, P)$$

Con lo que concluimos la demostración.

□

Demostración

Demostraremos el Teorema (1). Buscamos probar que en la expresión $V_\phi^k(P) = \inf_{G\in\mathcal{B}_k} W_\phi^k(G,P)$, este inferior realmente es un mínimo. Consideremos $H_n = \{h_1^n, .., h_k^n\} \in \mathcal{B}_k$ tal que $W_\phi^k(H_n,P) \longrightarrow \inf_{G\in\mathcal{B}_k} W_\phi(G)$. Sea $I = \{i \in \{1,..,k\} / \liminf_n \|h_i^n\| < \infty\}$.
Se tiene que I es no vacío: Si esto no ocurriera, $\forall i \in \{1,..,k\}$ resultaría que $\liminf_n \|h_i^n\| = \infty$. Veamos que si esto fuera así, $V_\phi^k(P)$ no sería el inferior. Tenemos, aplicando el Lema de Fatou

$$V_\phi^k(P) = \liminf_n W_\phi^k(H_n,P) = \liminf_n \int \inf_{i=1,..,k} \phi(\|x-h_i^n\|)dP(x) \geq \int \liminf_n \inf_{i=1,..,k} \phi(\|x-h_i^n\|)dP(x)$$

Dado que $f(y) = \inf_{i=1,..,k} \phi(y)$ es una función continua, se tiene que

$$\int \liminf_n \inf_{i=1,..,k} \phi(\|x-h_i^n\|)dP(x) = \int \inf_{i=1,..,k} \phi(\liminf_n \|x-h_i^n\|)dP(x)$$

Aplicamos a continuación la segunda desigualdad triangular y el hecho de que $I = \emptyset$:

$$V_\phi^k(P) \geq \int \phi(\inf_{i=1,..,k} \liminf_n \|x-h_i^n\|)dP(x) \geq \int \phi(\inf_{i=1,..,k} (\liminf_n \|h_i^n\| - \|x\|))dP(x) = \int \phi(\infty)dP(x)$$

Pero esto es absurdo, pues $V_\phi^k(P)$ no sería el inferior.

Por lo tanto, sabemos que I no es vacío: Existe al menos un índice i con $\liminf_n \|h_i\| < \infty$. Es decir, existe $M > 0$ y una subsucesión $\{h_i^{n_k}\}_k$ tal que $\|h_i^{n_k}\| < M$ para $k \geq k_0$. De esta subsucesión acotada, por el Teorema de Bolzano-Weierstrass, podemos extraer una subsucesión convergente. Para simplificar la notación, seguiremos denominándola $\{h_i^n\}_n$. Definamos $J = \{i \in \{1,..,k\} / h_i^n \longrightarrow h_0$ para un cierto $h_0\} = \{1,..,s\}$ si los ordenamos adecuadamente. Por lo que acabamos de decir, J es no vacío dado que I no lo es.

Para los índices $i \in J$, ya que $h_i^n \longrightarrow h_i^0$, realizando un razonamiento similar al anterior por medio del Lema de Fatou

$$V_\phi^k(P) = \liminf_n W_\phi^k(H_n,P) = \liminf_n \int \inf_{i=1,..,k} \phi(\|x-h_i^n\|)dP(x) \geq$$

$$\geq \int \liminf_n \inf_{i=1,..,k} \phi(\|x-h_i^n\|)dP(x) = \int \inf_{i=1,..,s} \phi(\|x-h_i^0\|)dP(x) \geq V_\phi^s(P)$$

Por el lema (2), sabemos que tener k puntos disminuye el potencial penalizado por ϕ frente a tener s puntos , por lo que necesariamente se tiene $V_\phi^k(P) = V_\phi^s(P)$ y así

$$W_\phi^s(\{h_1^0, .., h_s^0\}, P) = \lim_{n\to\infty} W_\phi^s(\{h_1^n, .., h_s^n\}, P) = V_\phi^s(P)$$

Tenemos dos opciones:

- Si $k = s$, el conjunto $H_0 = \{h_1^0, .., h_k^0\}$ verifica $W_\phi^k(H_0, P) = V_\phi^k(P)$ y por lo tanto es una k-ϕ-media de X.
- Si $s < k$, dado que $V_\phi^k(P) = V_\phi^s(P)$, por el lema (2) necesariamente $V_\phi^s(P) = 0$ y podemos añadir más puntos a los s que ya teníamos hasta llegar hasta k, con lo que seguirmos teniendo garantizada la existencia de una k-ϕ-media.

□

3.2.2. Unicidad del problema de k medias

Planteadas las hipótesis bajo las cuales existen las k-ϕ-medias de un vector aleatorio X, nuestra siguiente pregunta será si el problema de encontrar un conjunto H de k puntos que minimice el potencial penalizado por ϕ de nuestro vector aleatorio (bajo las condiciones que aseguran su existencia) tendrá una única solución o si por el contrario existen varios conjuntos de estas características con los que se alcanza el mínimo. El interés que tenemos al respecto no es solo teórico, sino que también será interesante desde el punto de vista computacional debido a que la multiplicidad de soluciones puede provocar un mayor tiempo de ejecución. En el marco práctico nos preguntaremos si, dependiendo de la inicialización que tomemos, somos capaces de encontrar más de una solución óptima. La respuesta es claramente afirmativa, tanto si consideramos una muestra de n puntos como si pensamos en la k-ϕ-media de una distribución de probabilidad.

Algunos ejemplos

Consideremos en primer lugar el caso en el que tenemos un conjunto de datos. Supongamos que queremos hallar una 2-media del conjunto de puntos $\{(-1, 0), (0, 0), (1, 0)\}$, cuya representación en el plano es la siguiente:

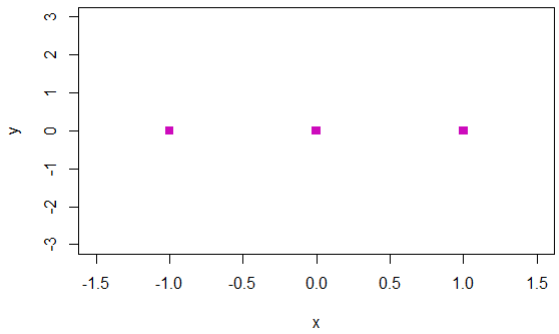
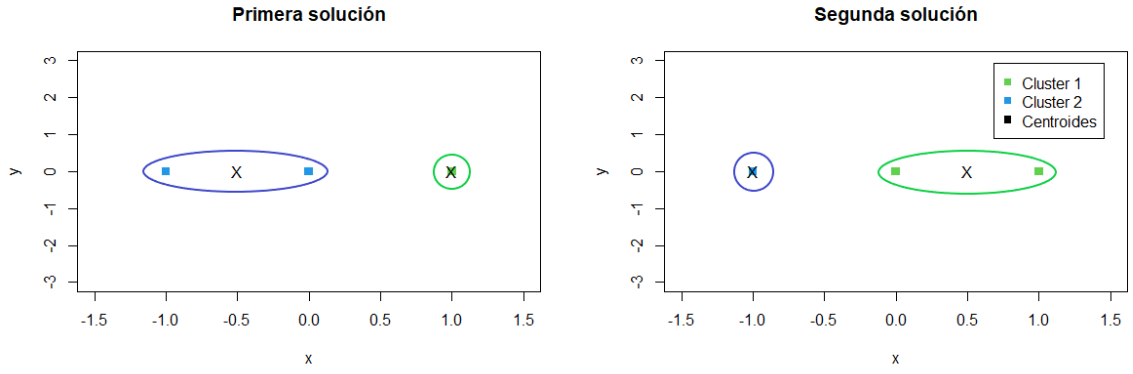


Figura 6

Supongamos que la medida de disimilaridad que utilizamos es la distancia Euclídea. Dado que los puntos estan colocados de manera que uno equidista de dos de ellos, más alejados entre sí, es claro que existen dos soluciones óptimas:

1. $C_1 = \{(-1, 0)\}$ y $C_2 = \{(0, 0), (1, 0)\}$ y centros $h_1 = (-1, 0), h_2 = (0'5, 0)$
2. $C_1 = \{(-1, 0), (0, 0)\}$ y $C_2 = \{(1, 0)\}$ y centros $h_1 = (-0'5, 0), h_2 = (1, 0)$

dado que, en ambos casos, $\sum_{i=1}^{3} \text{mín}_{j=1,2} \|x_i - h_j\|^2 = 1$ es mínima.
Al buscar una agrupación y dos centroides con *kmeans* de R obtenemos los dos resultados ejecutando el programa dos veces:



(a) Primera Solución

(b) Segunda Solución

Figura 7

En un caso tan sencillo como este vemos claro que la unicidad del problema no se da y existen dos soluciones que nos permiten llegar a un potencial mínimo. La obtención de dos soluciones parece consecuencia de la presencia de un punto que equidista de dos. Uno podría pensar en clasificar ese individuo del medio en dos grupos simultáneamente. Es decir, colocar los centroides en los puntos $(\frac{-1}{2}, 0)$ y $(\frac{1}{2}, 0)$ y dejar ese punto en la frontera de ambos clusters:

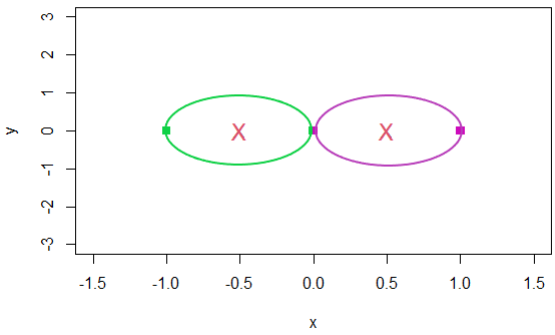


Figura 8

Sin embargo, el método de k medias evita situaciones así: Veremos más adelante cómo la frontera de un grupo tendrá siempre probabilidad cero si calculamos los clusters utilizando el algoritmo.

Esto que acabamos de observar no solo ocurre al calcular las k medias de un conjunto de datos sino cuando nos planteamos si las k medias de una distribución de probabilidad son únicas. Presentamos el siguiente ejemplo para ilustrar cómo la respuesta a esta cuestión a menudo es negativa y además dejar patente que, salvo en casos muy concretos, no tenemos resultados sobre la unicidad del problema.

Consideramos el experimiento consistente en seleccionar un punto al azar en el círculo de radio unidad centrado en el origen de coordenadas (denomimamos B a esta bola).

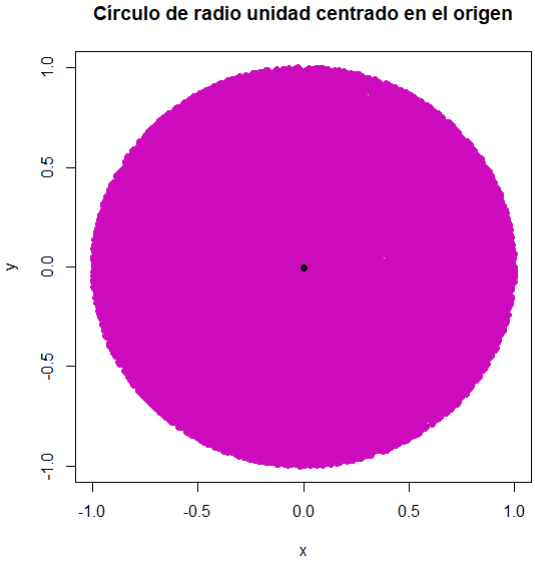


Figura 9

Sean X e Y la abscisa y ordenada respectivamente del punto elegido. La elección completamente al azar de un punto en el círculo sin que haya zonas con más preferencia que otras puede modelizarse a traves de la función de densidad conjunta

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{\pi} & si\ x^2+y^2 \leq 1 \\ 0 & si\ x^2+y^2 > 1 \end{cases}$$

Sea $\vec{x} = (x,y)$. Nuestro objetivo es encontrar una 2-media de (X,Y) (tomamos como media de disimilaridad la distancia Euclídea al cuadrado), es decir, hallar M_1 y M_2 en B de tal manera que minimicen $\int \text{mín}_{i=1,2}\|\vec{x}-M_i\|^2 dP(\vec{x})$.

Calculamos en primer lugar las densidades marginales de X y de Y:

$$f_X(x) = \frac{2}{\pi}\sqrt{1-x^2}, \quad -1 < x < 1$$

$$f_Y(y) = \frac{2}{\pi}\sqrt{1-y^2}, \quad -1 < y < 1$$

A continuación, hallamos la esperanza de (X,Y), sabiendo que $\mathbb{E}(X,Y) = (\mathbb{E}(X),\mathbb{E}(X))$. Así pues

$$\mathbb{E}(X) = \int_{-1}^{1} x dP(x) = \int_{-1}^{1} x f_X(x) dx = \int_{-1}^{1} \frac{2}{\pi} x\sqrt{1-x^2} dx = 0$$

De manera análoga, se tiene $\mathbb{E}(Y) = 0$ y con ello, $\mathbb{E}(X,Y) = (0,0)$.

Es claro que estos dos puntos M_1 y M_2 dividen a B en dos subconjuntos:

$$A_1 = \{\vec{x} \in B / \|\vec{x} - M_1\|^2 \leq \|\vec{x} - M_2\|^2\},\ A_2 = \{\vec{x} \in B / \|\vec{x} - M_1\|^2 \geq \|\vec{x} - M_2\|^2\}$$

Donde M_i sería la media del conjunto A_i para $i = 1, 2$, es decir, $\mathbb{E}((X,Y)/A_i) = M_i$. Además, el conjunto $A_1 \cap A_2$ tiene probabilidad cero. De este modo, utilizando esperanzas condicionadas, podemos expresar $\mathbb{E}(X,Y)$ como combinación lineal convexa de M_1 y M_2, ya que $P(A_1) = 1 - P(A_2)$:

$$\mathbb{E}(X,Y) = \mathbb{E}((X,Y)|A_1)P(A_1) + \mathbb{E}((X,Y)|A_2)P(A_2)$$

Esto supone que ambos puntos deben estar en un diámetro de la circunferencia. Sin pérdida de generalidad, supondremos que es el correspondiente al eje de abscisas. Así, los puntos M_1 y M_2 son de la forma

$$M_1 = (m_1, 0),\ M_2 = (m_2, 0),\ m_1, m_2 \in \mathbb{R}$$

A su vez, gracias a esta consideración podemos reescribir los conjuntos A_1 y A_2:

$$A_1 = \{\vec{x} \in B / x \leq \frac{m_1 + m_2}{2}\},\ A_2 = \{\vec{x} \in B / x \geq \frac{m_1 + m_2}{2}\}$$

Podemos reescribir la expresión a minimizar de este modo:

$$\int_B \min_{i=1,2} \|\vec{x} - M_i\|^2 dP(\vec{x}) = \int_{A_1} \|\vec{x} - M_1\|^2 dP(x) + \int_{A_2} \|\vec{x} - M_2\|^2 dP(\vec{x}) \tag{7}$$

Sin embargo, observamos que solo depende de la densidad de la primera coordenada y coincide, salvo por una constante, con la expresión correspondiente a la 2-media de la distribución marginal correspondiente. De esta manera, hemos logrado plantear el problema en términos de una sola dimensión. En estas condiciones, existen resultados interesantes que nos garantizan la unicidad de la 2-media de (X,Y) y más generalmente de la k media. El que vamos a utilizar aparece en el artículo [16] en el que gracias a resultados previamente probados por B. Flury, se establece el siguiente lema:

Lema 3. *Sea X una variable aleatoria univariante y f su función de densidad. Suponemos que f es simétrica en torno al origen y log-cóncava. Entonces, f tiene dos únicos puntos principales (denominados 2-media en nuestro caso) que son simétricos respecto del origen.*

Dado que $f_X(x)$ satisface estas hipótesis, podemos aplicar el lema y afirmar que existen dos únicos puntos m_1 y m_2 que minimizan la expresión (7) y que verifican además que $m_1 = -m_2$.
Desarrollando la expresión a minimizar y realizando algunos cálculos engorrosos pero sencillos, se prueba que en este problema en concreto $m_1 = \frac{-4}{3\pi}$ y $m_2 = \frac{4}{3\pi}$.

El hecho de que $m_1 = -m_2$ y con ello $\frac{m_1+m_2}{2} = 0$, implica que los dos grupos obtenidos A_1 y A_2 serían exactamente los semicírculos de B delimitados por el eje de ordenadas. Habíamos supuesto que el diámetro que unía M_1 y M_2 era precisamente el eje de abscisas, pero aplicando una rotación a M_1, M_2 y al diámetro, existen infinitos pares de puntos que consiguen minimizar la expresión (7). Por lo tanto, es claro que existen infinitas soluciones al problema. Para ilustrarlo, generamos 50000 puntos en B uniformemente y aplicamos *kmeans* de R reiteradamente. Obtenemos tanto centroides como particiones de B completamente diferentes en cada caso.

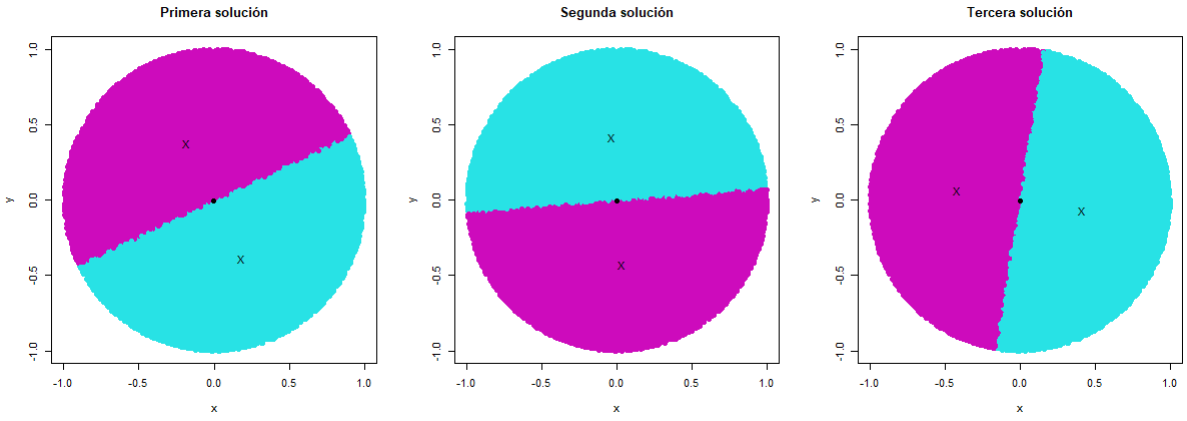


Figura 10

Existen pocos resultados relativos a la unicidad de k-ϕ-medias. Algunos de los existentes exigen condiciones más fuertes sobre la función de densidad de probabilidad. Otros, como el que presentamos a continuación, demandan propiedades más fuertes de la función ϕ y reducen sus resultados al caso $k = 1$.

Proposición 3. *Si la función ϕ es estrictamente convexa, entonces la ϕ-media de P es única.*

Demostración

Supongamos que tenemos m y $m^* \in \mathbb{R}^d$ dos ϕ-medias de X diferentes y tomemos $\lambda \in (0, 1)$. Creamos una combinación lineal convexa de m y m^*, $m_\lambda = \lambda m + (1 - \lambda)m^*$. Recordemos que la definición de que m sea ϕ-media de X, en este caso media, es

$$\int \phi(\|x - m\|)dP(x) = \int \phi(\|X - m\|)d\mathbb{P} \leq \int \phi(\|X - g\|)d\mathbb{P}$$

para cualquier elemento $g \in \mathbb{R}$.

Dado que la función ϕ es estrictamente convexa y creciente, sumando y restando λx, y aplicando la desigualdad triangular, tenemos que

$$\int \phi(\|x-m_\lambda\|)dP(x) = \int \phi(\|x-(\lambda m+(1-\lambda)m^*)\|)dP(x) = \int \phi(\|x-\lambda m-m^*+\lambda m^*+\lambda x-\lambda x\|)dP(x) \leq$$

$$\leq \int \phi(\|\lambda x - \lambda m\| + \|x - m^* + \lambda m^* - \lambda x\|)dP(x) = \int \phi(\lambda\|x - m\| + (1 - \lambda)\|x - m^*\|)dP(x)$$

Como ϕ es estrictamente convexa, se tiene que

$$\int \phi(\lambda\|x - m\| + (1 - \lambda)\|x - m^*\|)dP(x) \leq \lambda \int \phi(\|x - m\|)dP(x) + (1 - \lambda) \int \phi(\|x - m^*\|)dP(x)$$

Ya que tanto m como m^* son ϕ-medias, resulta que

$$\lambda \int \phi(\|x-m\|)dP(x) + (1-\lambda)\int \phi(\|x-m^*\|)dP(x) = \int \phi(\|x-m\|)dP(x)$$

Y, por lo tanto, podemos afirmar que

$$\int \phi(\|x-m_\lambda\|)dP(x) \leq \int \phi(\|x-m\|)dP(x)$$

Dado que ϕ es estrictamente convexa, salvo que $\|x-m\| = \|x-m^*\|$ $P-c.s.$, la desigualdad es estricta. Como m es una ϕ-media, necesariamente debe cumplirse $\|x-m\| = \|x-m^*\|$ con probabilidad 1: si no, habríamos encontrado un valor que minimiza aún más la expresión. Definimos el conjunto

$$A = \{y \in \mathbb{R}^d / \|y-m\| = \|y-m^*\| \quad P-c.s.\}$$

Pero podemos reescribirlo como

$$B = \{y \in \mathbb{R}^d / \exists x \in \mathbb{R}^d \ \ con \ \ \langle x, m-m^*\rangle = 0 \ \ e \ \ y = x + \frac{m+m^*}{2} \quad P-c.s.\}$$

Donde $\langle \cdot, \cdot \rangle$ denota el producto escalar euclideo (véase la demostración en el Apéndice, (5)).

Pero entonces, resulta que si $\|y-m\| = \|y-m^*\| \ \ P-c.s.$, podemos escribir $y = x + \frac{m+m^*}{2} \ \ P-c.s.$ con $\langle x, m-m^*\rangle = 0$. De este modo, desarrollando de nuevo el producto escalar

$$\|x + \frac{m+m^*}{2} - m_\lambda\|^2 = \|(x + \frac{m+m^*}{2} - m^*) - \lambda(m-m^*)\|^2 =$$

$$= \|x + \frac{m+m^*}{2} - m^*\|^2 + 2\lambda\langle (x + \frac{m+m^*}{2} - m^*), m^*-m\rangle + \lambda^2\|m^*-m\|^2$$

Dado que $\|y-m\| = \|y-m^*\| \ \ P-c.s.$, podemos reescribir la expresión anterior como

$$\|x + \frac{m+m^*}{2} - m\|^2 + \lambda^2\|m^*-m\|^2 + 2\lambda\langle x, m-m^*\rangle - \lambda\|m^*-m\|^2$$

Sabemos que $\langle x, m-m^*\rangle = 0 \ P-c.s.$, por lo que cancelamos el término. Como $\lambda \in (0,1)$, se tiene que $\lambda^2 < \lambda$, por lo que

$$\|x + \frac{m+m^*}{2} - m\|^2 + \lambda^2\|m^*-m\|^2 - \lambda\|m^*-m\|^2 < \|x + \frac{m+m^*}{2} - m\|^2$$

Es decir, hemos llegado a que la siguiente desigualdad se verifica con probabilidad 1. Dado que ϕ es estrictamente convexa, podemos afirmar

$$\|x + \frac{m+m^*}{2} - m_\lambda\|^2 < \|x + \frac{m+m^*}{2} - m\|^2 \Rightarrow \phi(\|x + \frac{m+m^*}{2} - m_\lambda\|^2) < \phi(\|x + \frac{m+m^*}{2} - m\|^2)$$

De este modo, llegamos a un absurdo ya que m era ϕ-media de X, por lo que debe ser finalmente $m = m^*$. □

3.2.3. Puntos frontera entre clusters

Hemos comentado que la solución del problema de k medias no es necesariamente única y que a pesar de que podría parecer posible encontrar puntos a la misma distancia de dos centros de cluster y de algún modo "repartir" su masa de probabilidad entre ambos grupos, el método evita estas situaciones y la frontera de un cluster siempre tendrá probabilidad cero. En este apartado pretendemos demostrar esto último planteando qué condiciones debemos imponer sobre ϕ para poder asegurar esto. Para ello, nos apoyaremos en la proposición demostrada en el apartado anterior que asegura la unicidad de la ϕ-media en el caso de ser ϕ convexa, y a su vez usaremos que la ϕ-media de un conjunto A si existe, siendo ϕ convexa, siempre está en $\overline{A}$ (Ver proposición (6) en el Apéndice). Este resultado sobre la probabilidad de los puntos en la frontera de dos clusters se encuentra en [3] en el caso de k-ϕ-medias recortadas.

Teorema 2. *Sea X vector aleatorio, k entero positivo y $\phi : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$ una función creciente, continua, tal que $\phi(0) = 0$ y convexa. Supongamos además que se verifica $\int \text{ínf}_{g\in G}\, \phi(\|x-g\|)dP(x) < \infty$ para algún $G \in \mathcal{B}_k$ con el fin de asegurar la existencia de la k-ϕ-media. Sea $H = \{h_1, .., h_k\} \in \mathcal{B}_k$ una k-ϕ-media de X y $\mathscr{C} = \{C_1, .., C_k\}$ la partición de $\mathbb{R}^d$ inducida por H. Supongamos además que la derivada de ϕ existe, es continua y que $h_i \neq h_j$ si $i \neq j$.*

Entonces, $P(\{x \in \mathbb{R}^d / \|x - h_i\| = \|x - h_j\|\}) = 0$ si $i \neq j$. Es decir, los puntos en la frontera de dos o varios clusters tienen probabilidad 0.

Demostración

Sea C_i el cluster correspondiente al centro h_i y suponemos que $\int_{C_i} dP(x) > 0$ para evitar el caso trivial, al igual que $h_i \neq h_j$ si $i \neq j$. Sea $B_{i,j} = \{x \in \mathbb{R}^d / \|x - h_i\| = \|x - h_j\|\}$, es decir, la frontera entre los clusters C_i y C_j. Si consideramos qué elemento de H minimiza la integral en $C_i \cup C_j$, llegamos a que dependiendo de que x cojamos, elegiremos bien h_i o bien h_j:

$$\int_{C_i\cup C_j} \min_{l=1,..,k}(\phi(\|x-h_l\|)dP(x) = \int_{C_i\cup C_j} \text{mín}(\phi(\|x-h_i\|, \phi(\|x-h_j\|))dP(x)$$

Podemos escribir $C_i \cup C_j = (C_j \cup B_{i,j}) \sqcup (C_i - B_{i,j})$, donde $\sqcup$ denota la unión disjunta. De esta manera escribiríamos

$$\int_{C_i\cup C_j} \text{mín}(\phi(\|x-h_j\|), \phi(\|x-h_i\|))dP(x) = \int_{C_j\cup B_{i,j}} \phi(\|x-h_j\|)dP(x) + \int_{C_i-B_{i,j}} \phi(\|x-h_i\|)dP(x)$$

Se tiene que la ϕ-media de $C_i \cup B_{i,j}$, la de $C_i - B_{i,j}$ y h_i deben coincidir (De manera análoga para h_j). Si esto no fuera así, podríamos cambiar h_i y h_j en H por las respectivas ϕ-medias de $C_j \cup B_{i,j}$ y $C_i - B_{i,j}$, denotando ese nuevo conjunto de $\mathcal{B}_k$ por H^*. De esta manera, dado que la ϕ-media de un conjunto es única, H no sería una k-ϕ-media:

$$\int_{C_i\cup C_j} \min_{h\in H}(\phi(\|x-h_l\|)dP(x) > \int_{C_i\cup C_j} \min_{h\in H^*}(\phi(\|x-h_l\|)dP(x)$$

Sea h^0 la ϕ-media de $C_i - B_{i,j}$. Vamos a comprobar que si la probabilidad de los puntos frontera entre clusters C_i y C_j es estrictamente positiva, entonces h^0 no será ϕ-media de $C_i \cup B_{i,j}$ y por lo

tanto llegaríamos a un absurdo. Denotamos por h^1 la ϕ-media de $B_{i,j}$, que también necesitaremos para las demostraciones.

En primer lugar, comprobamos que $\int_{C_i-B_{i,j}} dP(x) > 0$. Recordemos que $h_i \in \overline{C}_i$ por ser ϕ-media de C_i y notemos que $\overline{B_{i,j}} = B_{i,j}$ por ser cerrado. Si fuera $\int_{C_i-B_{i,j}} dP(x) = 0$, necesariamente $h_i \in B_{i,j}$. De lo contrario, $h_i \in C_i - B_{i,j}$ pero $\int_{C_i-B_{i,j}} \phi(\|x-h\|)dP(x) = 0$ para cualquier $h \in \mathbb{R}^d$, por lo que existirían infinitas ϕ-medias y esto es absurdo.
Pero si $h_i \in B_{i,j}$, entonces h_i cumpliría $\|h_i - h_j\| = \|h_i - h_i\| = 0$ y resultaría $h_i = h_j$, por lo que llegaríamos a un absurdo. Es decir, necesariamente $\int_{C_i-B_{i,j}} dP(x) > 0$.

Además, sabemos que $h^0 \neq h^1$. Si no, resultaría que

$$\int_{C_i-B_{i,j}} \phi(\|x-h^0\|)dP(x) + \int_{B_{i,j}} \phi(\|x-h^1\|)dP(x) = \int_{C_i} \phi(\|x-h^0\|)dP(x) \leq \int_{C_i} \phi(\|x-h_i\|)dP(x)$$

Donde la desigualdad es estricta a no ser que $h_i = h^0$ por ser la ϕ-media de un conjunto única. Dado que h_i es ϕ-media, se tiene que verificar $h^0 = h^1 = h_i$. Pero entonces como $h^1 \in \overline{B_{i,j}}$, necesariamente $h_i \in \overline{B_{i,j}}$ y razonando como antes llegaríamos a un absurdo.

Como $h^1 \neq h^0$, podemos escoger una base ortogonal $\{e_1, .., e_d\}$ de tal manera que las coordenadas en esta base de h^0 sean $(0, 0, .., 0)$ y $e_1 = h^1 - h^0 = (1, 0, .., 0)$.
A continuación, construimos la función $f : \mathbb{R} \longrightarrow \mathbb{R}^d$ definida como $f(t) = (t, 0, .., 0) = h^t$ y escribimos

$$\phi'(\|x - f(t_0))\|) = \frac{d}{dt}\phi(\|x - h^t)\|)|_{t=t_0}$$

Definimos las siguientes funciones

$$H_1(t) = \int_{C_i-B_{i,j}} \phi(\|x-h^t\|)dP(x), \quad H_2(t) = \int_{B_{i,j}} \phi(\|x-h^t\|)dP(x)$$

Además, dado que ϕ' existe y es continua, las derivadas de $H_1(t)$ y de $H_2(t)$ también existen y son continuas

$$H_1'(t) = \int_{C_i-B_{i,j}} \phi'(\|x-h^t\|)dP(x), \quad H_2'(t) = \int_{B_{i,j}} \phi'(\|x-h^t\|)dP(x)$$

Como h^0 es ϕ-media de $C_i - B_{i,j}$, minimiza la función $\phi(\|x - h^t\|)$ en $C_i - B_{i,j}$. Por lo tanto, podemos asegurar que $H_1'(0) = 0$.

Veamos que $H_2(t)$ es estríctamente convexa en $[0, 1]$. Dados $t, r \in [0, 1]$, queremos ver que

$$H_2(\lambda t + (1-\lambda)r) < \lambda H_2(t) + (1-\lambda)H_2(r)$$

Como $h^t = (t, 0, .., 0)$, entonces $h^{\lambda t+(1-\lambda)r} = (\lambda t + (1-\lambda)r, 0, .., 0) = \lambda(t, 0, .., 0) + (1-\lambda)(r, 0, .., 0) = \lambda h^t + (1-\lambda)h^r$. De este modo, utilizando que ϕ es estrictamente convexa para establecer la desigualdad, se tiene

$$H_2(\lambda t + (1-\lambda)r) = \int_{B_{i,j}} \phi(\|x - h^{\lambda t+(1-\lambda)r}\|)dP(x) = \int_{B_{i,j}} \phi(\|x - \lambda h^t + (1-\lambda)h^r\|)dP(x) \leq$$

$$\leq \lambda \int_{B_{i,j}} \phi(\|x - h^t\|) dP(x) + (1-\lambda) \int_{B_{i,j}} \phi(\|x - h^r\|) dP(x) = \lambda H_2(t) + (1-\lambda) H_2(r)$$

Por lo que queda probado que $H_2(t)$ es estrictamente convexa en $[0, 1]$.

Dado que h^1 es ϕ-media de $B_{i,j}$, minimiza $\phi(\|x - h^t\|)$ en $B_{i,j}$. Por lo tanto, sabemos que $H_2'(1) = 0$. Al ser estrictamente convexa, $H_2'(t) < 0$ para cualquier $t < 1$. En concreto, se tiene $H_2'(0) < 0$.
Si consideramos $H(t) = H_1(t) + H_2(t)$, podemos afirmar que $H'(t) < 0$ por lo que es decreciente en $[0, t_0]$ para $t_0 > 0$. Por lo tanto, $H(0) \geq H(t)$, $t \in [0, t_0]$. Así, se tiene

$$H(0) = \int_{C_i \cup B_{i,j}} \phi(\|x - h^0\|) dP(x) \geq \int_{C_i \cup B_{i,j}} \phi(\|x - h^{t_0}\|) dP(x) = H(t_0)$$

Pero de esta manera h^0 no sería ϕ-media de $C_i \cup B_{i,j}$ pero sí de $C_i - B_{i,j}$, y hemos dicho que ambos valores debían coincidir para no llegar a un absurdo. Por lo tanto, podemos concluir que $P(B_{i,j}) = 0$.
□

3.3. Consistencia del método de k-medias

A menudo, nuestro marco de trabajo consiste en un conjunto de n individuos que toman medidas en d variables. Los consideramos como n puntos de $\mathbb{R}^d$, $\{x_1, .., x_n\}$, y tratamos de agruparlos en función de las similaridades que presentan entre ellos, encontrando k representantes para tener un resumen en k puntos de nuestro conjunto. En particular, cuando las observaciones se obtienen como una muestra en un modelo probabilístico, se plantea el problema de "consistencia estadística": Nos preguntamos si las k-ϕ-medias que hallamos para una muestra se acercan a las k-ϕ-medias de la distribución de probabilidad teórica.

Una k-ϕ-media de una variable aleatoria son k puntos concretos (no dependen de la aleatoriedad), mientras que una k-ϕ-media empírica que hallamos por medio de una muestra depende del ω escogido, por lo que obtenemos valores distintos de los centroides al generar muestras nuevas. Comprobémoslo en un ejemplo.

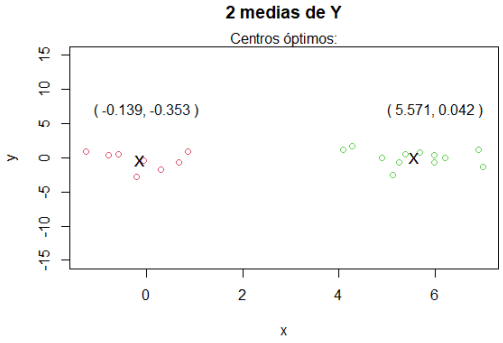
Supongamos que tenemos $X_1 \sim N((0,0), \Sigma)$ y $X_2 \sim N((6,0), \Sigma)$, normales multivariantes con matríz de covarianza común $\Sigma = \left(\begin{smallmatrix} 2 & 0 \\ 0 & 2 \end{smallmatrix}\right)$. Supongamos también que Z sigue una distribución de Bernoulli de parámetro $p = \frac{1}{2}$, esto es $Y \sim B(\frac{1}{2})$, y definamos

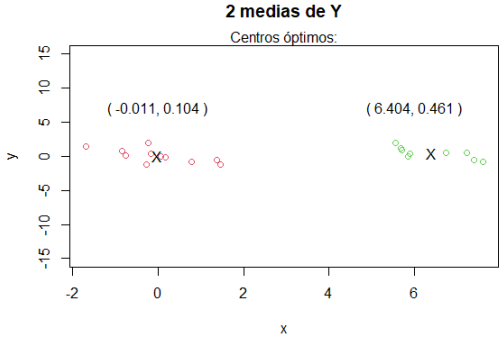
$$Y = X_1 Z + (1 - Z) X_2$$

Y es una variable que toma el valor de X_1 o de X_2 con igual probabilidad. Generamos dos muestras diferentes de 20 individuos de la variable Y y buscamos, utilizando como función de disimilaridad la distancia Euclídea, la 2-media de cada muestra:

Los dos centros óptimos que hallamos para esta primera muestra son $h_{1,1} = (-0{,}139, -0{,}353)$ y $h_{2,1} = (5{,}571, 0{,}042)$.

En el caso de la segunda muestra, los centros óptimos que hallamos son $h_{1,2} = (-0{,}011, -0{,}104)$ y $h_{2,2} = (6, 404, 0{,}461)$. Para el primer centroide, parece que los resultados son parecidos mientras





que para el segundo podríamos dudar de la fiabilidad del método, ya que los resultados son considerablemente diferentes. ¿Tenemos la garantía de estar aproximando verdaderamente la 2 medias de Y?

Para generar la primera muestra, hemos tomado $\omega \in \Omega$, $Y_1, Y_2, .., Y_n$ v.a.i.i.d con $\mathscr{L}(Y_i) = \mathscr{L}(Y)$ y hemos aplicado el método de 2 medias a la muestra $y_1, \ldots, y_n$ con $Y_i(\omega) = y_i$, $i = 1, .., n$. En el caso de la segunda, hemos seleccionado otro ω, llamémoslo $\omega' \in \Omega$, y hemos aplicado el método a la muestra (distinta) $y_1{}', \ldots, y_n{}'$ con $X_i(\omega') = y_i'$. Puntos diferentes generan centroides diferentes. A lo largo de este trabajo, enunciaremos y demostraremos un Teorema que nos permitirá garantizar la convergencia de la k-ϕ-media empírica hacia la k-ϕ-media teórica de una variable aleatoria. Esto es, podremos asegurar que a pesar de obtener centroides distintos y que a priori no se parecen, al aumentar el tamaño de la muestra tenemos garantizado que los centroides obtenidos se aproximan hacia la k-ϕ-media teórica de la variable.

La demostración de este resultado de convergencia será muy similar a la que llevábamos a cabo en apartado de existencia de k-ϕ-medias. Una formulación en términos de variables aleatorias será más satisfactoria ahora para poder aplicar resultados como la Ley Fuerte de los Grandes números o el Teorema de Representación de Skorohod. Cabe decir también que como llevábamos haciendo hasta

ahora, supondremos que las variables con las que trabajamos llegan a $\mathbb{R}^d$ para mayor simplicidad, es decir, son vectores aleatorios. Sin embargo, todas las conclusiones descritas en esta sección pueden generalizarse a un espacio de Banach B uniformemente convexo sustituyendo la convergencia por la convergencia débil en el sentido funcional y recurriendo a una generalización sobre el Teorema de Glivenko-Cantelli debida a V.S.Varadarajan [15]. Encontramos demostraciones de consistencia del método en $\mathbb{R}^d$ en el artículo escrito por D. Pollard [12] y para espacios más generales en [5].

Observación 3. *Sea $Z : \Omega \longrightarrow \mathbb{R}^d$ una variable aleatoria con $\mathscr{L}(Z) = P$, $\phi : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$ una función continua, creciente y con $\phi(0) = 0$. Sea H un conjunto de $\mathcal{B}_k$. Para no tener que recurrir a la notación $\mathscr{L}(X)$, escribiremos lo siguiente en un abuso de notación:*

$$W_\phi^k(H, Z) = \int \inf_{h \in H} \phi(\|Z - h\|) d\mathbb{P}$$

Consideramos $\{Z_n\}_{n=0}^\infty$ variables aleatorias con valores en $\mathbb{R}^d$ definidas en el espacio probabilístico $(\Omega, \sigma, \mathbb{P})$. Sea $H_n = \{h_1^n, ..., h_k^n\}$ una k media de Z_n. Buscamos probar, bajo ciertas condiciones, que si $Z_n \longrightarrow Z_0$, está garantizado que H_n converge a H_0.

Teorema 3. *Sean $\{Z_n\}_{n=0}^\infty$ vectores aleatorios definidos en el espacio probabilístico $(\Omega, \sigma, \mathbb{P})$, $\phi : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$ una función continua, creciente y con $\phi(0) = 0$. Suponemos que se cumple $W_\phi^k(G_n, Z_n) < \infty$ con G_n un conjunto de $\mathcal{B}_k$ adecuado para cada n y sea H_n una k-ϕ-media de Z_n. Supongamos que*

1. *$Z_n \longrightarrow Z_0$ $\mathbb{P} - c.s$*
2. *El conjunto H_0 es único*
3. *$W_\phi^k(H_0, Z_n) = \int \phi(\text{ínf}_{h \in H_0} \|Z_n - h\|) d\mathbb{P} \longrightarrow \int \phi(\text{ínf}_{h \in H_0} \|Z_0 - h\|) d\mathbb{P} = W_\phi^k(H_0, Z_0)$*

Entonces, se tiene que $H_n \longrightarrow H_0$

Demostración

Por el lema (9) visto anteriormente, sabemos que para probar la convergencia $H_n \longrightarrow H_0$ es equivalente probar que de toda subsucesión de $\{H_n\}_n$ es posible extraer una subsucesión convergente. Tomamos una subsucesión que seguiremos denotando por $\{H_n\}_n$ para simplificar la notación.

Sea $I = \{i \in \{1, .., k\} / \liminf_n \|h_i^n\| < \infty\}$ y veamos que I es no vacío. Si esto no ocurriera, $\forall i \in \{1, .., k\}$ resultaría que $\liminf_n \|h_i^n\| = \infty$. Veamos que si esto fuera así, cualquier conjunto de k elementos sería una k media de Z_0. Dado que H_n es k-ϕ-media de Z_n, se tiene

$$W_\phi^k(H_0, Z_0) = \liminf_n W_\phi^k(H_n, Z_0) \geq \liminf_n W_\phi^k(H_n, Z_n)$$

Si aplicamos el Lema de Fatou a la expresión anterior llegamos a que

$$\liminf_n W_\phi^k(H_n, Z_n) = \liminf_n \int \phi(\inf_{i=1,..,k} \|Z_n - h_i^n\|) d\mathbb{P} \geq \int \liminf_n \phi(\inf_{i=1,..,k} \|Z_n - h_i^n\|) d\mathbb{P}$$

Utilizando que la función dentro de la integral es continua y aplicando la segunda desigualdad triangular llegamos a

$$W_\phi^k(H_0, Z_0) \geq \int \phi(\inf_{i=1,..,k}(\liminf_n(\|h_i^n\| - \|Z_n\|))d\mathbb{P} = \int \phi(\infty)d\mathbb{P}$$

Dado que estamos suponiendo que $I = \emptyset$. Entonces existirían inifinos valores que minimizan el potencial penalizado por ϕ, pero esto es absurdo ya que habíamos supuesto que la k-ϕ-media H_0 era única.

Por lo tanto, sabemos que I no es vacío: Existe al menos un índice i con $\liminf_n \|h_i\| < \infty$. Es decir, existe $M > 0$ y una subsucesión $\{h_i^{n_k}\}_k$ tal que $\|h_i^{n_k}\| < M$ para $k \geq k_0$. De esta subsucesión acotada, por el Teorema de Bolzano-Weierstrass, podemos extraer una subsucesión convergente. Para simplificar la notación, seguiremos denominándola $\{h_i^n\}_n$. Definamos $J = \{i \in \{1, .., k\} \ / h_i^n \longrightarrow g_i$ para un cierto $g_i\}$. Por lo que acabamos de decir, J es no vacío dado que I no lo es.

Nuestro objetivo será comprobar que $I = J = \{1, .., k\}$ y, además, que los puntos hacia los que convergen los h_i son los correspondientes a la k-ϕ-media de Z_0, esto es, $\{g_1, .., g_k\} = H_0$.

Dado $i \in J$, ya que $Z_n \xrightarrow{c.s.} Z_0$, podemos afirmar que $Z_n - h_i^n \xrightarrow{c.s.} Z_0 - g_i$. Además, como $\|\cdot\| : \mathbb{R}^d \longrightarrow \mathbb{R}$ es una función continua,

$$\lim_{n\to\infty} \|Z_n - h_i^n\| = \|Z_0 - g_i\| \quad \mathbb{P} - c.s.$$

Realizamos un razonamiento similar al anterior por medio del Lema de Fatou

$$W_\phi^k(H_0, Z_0) \geq \liminf_n \int \phi(\inf_{i=1,..,k} \|Z_n - h_i\|)d\mathbb{P} = \int \phi(\inf_{i=1,..,k}(\liminf_n \|Z_n - h_i^n\|))d\mathbb{P}$$

Y, dado que $\|Z_n - h_i^n\| \xrightarrow{c.s.} \|Z_0 - g_i\|$, se tiene que

$$\int \phi(\inf_{i=1,..,k}(\liminf_n \|Z_n - h_i^n\|))d\mathbb{P} = \int \phi(\inf_{i=1,..,k}(\|Z_0 - g_i\|))d\mathbb{P}$$

Por lo tanto, resulta que el conjunto de k puntos $\{g_1, .., g_k\}$ verifica

$$W_\phi^k(H_0, Z_0) \geq \int \phi(\inf_{i=1,..,k} \|Z_0 - g_i\|)d\mathbb{P}$$

Dado que H_0 era único, necesariamente los conjuntos deben coincidir: $H_0 = \{g_1, .., g_k\}$ y es claro además que $J = \{1, .., k\}$ (para todos los índices podemos asegurar la convergencia). Como habíamos comenzado tomando una subsucesión y hemos conseguido extraer de ella una nueva subsucesión convergente, por el lema anterior tenemos el resultado buscado.

□

3.3.1. Convergencia de k-ϕ-medias empíricas a k-ϕ-medias teóricas

El resultado que acabamos de demostrar nos permite corroborar la idea intuitiva de que si Z_n converge a Z_0 $\mathbb{P}-c.s.$, también lo harán sus k-ϕ-medias. Un caso particular que resulta especialmente interesante es aquel que mencionábamos al inicio, en el que queremos garantizar la convergencia de las k-ϕ-medias muestrales a las teóricas. Definimos de manera más rigurosa el concepto de k-ϕ-media muestral, haciendo patente la importancia del ω escogido.

Definición 4. *Sea X_0 un vector aleatorio definido en el espacio probabilístico $(\Omega, \sigma, \mathbb{P})$ con llegada en $\mathbb{R}^d$. Sean $X_1, ..., X_n, ...$ vectores aleatorios i.i.d. definidas en el mismo espacio con $\mathscr{L}(X) = \mathscr{L}(X_i) = P$, y sea P_n^w la distribución de probabilidad empírica que otorga masa $\frac{1}{n}$ a cada $X_i(\omega)$, $i = 1, .., n$. Una k-ϕ-media empírica o muestral es un conjunto $H_n^\omega \in \mathcal{B}_k$ que verifica*

$$\frac{1}{n}\sum_{i=1}^{n} \min_{h \in H_n^\omega} \phi(\|X_i(\omega) - h)\|) \leq \sum_{i=1}^{n} \min_{g \in G} \phi(\|X_i(\omega) - g)\|) \tag{8}$$

Para cualquier conjunto $G \in \mathcal{B}_k$

Es claro que, dado que P_n^w depende del $\omega \in \Omega$ elegido, también lo hará H_n^ω. Gracias a los resultados previamente probados, conseguimos asegurar que a medida que el tamaño de la muestra crece, los conjuntos H_n^ω se parecen cada vez más a H_0, la k-ϕ-media teórica. Para conseguirlo, utilizaremos el Teorema de representación de Skorohod (ver Teorema 25.6 en [2]) y el Teorema de Glivenko-Cantelli (ver Teorema en 20.6 [2]).

Teorema 4 (Skorohod)**.** *Sea $(\mathcal{W}, \alpha, \lambda)$ el espacio probabilístico donde $\mathcal{W} = (0, 1)$, α consiste en la σ álgebra de Borel en $(0, 1)$ y λ es la medida de Lebesgue. Sea $\{Pn\}_{n=0}^{\infty}$ una sucesión de probabilidades. Si $P_n \xrightarrow{d} P$ (convergencia en distribución), existen $Y_0, Y_1, Y_2, ...$ variables reales definidas en $(\mathcal{W}, \alpha, \lambda)$ tales que*

1. *$P_{Y_n} = P_n \ \forall n \in \mathbb{N}$*
2. *$Y_n \longrightarrow Y_0 \ \lambda - c.s.$*

Teorema 5 (Glivenko-Cantelli)**.** *Sean $X_1, ..., X_n, ...$ v.a.i.i.d. definidas en un mismo espacio probabilístico con función de distribución común $F(x)$. Definimos la función de distribución empírica como*

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{[X_i, \infty)}(x)$$

Donde I_A es la función indicadora de A. Entonces, se tiene convergencia uniforme de F_n a F. Esto es:

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \longrightarrow 0 \quad \textit{casi seguro}$$

Además, dado que se tiene $F_n \xrightarrow{d} F$, podemos afirmar que $P_n \xrightarrow{d} P$

En el caso que estamos tratando, dado que trabajamos con vectores aleatorios en lugar de con variables aleatorias, sería necesario utilizar la versión multivariante del teorema de Glivenko-Cantelli. La demostración para variables aleatorias (llegada en $\mathbb{R}$) basa su argumento en el carácter creciente de la función F. En el caso de la versión de Glivenko-Cantelli para variables aleatorias con valores en $\mathbb{R}^d$, es frecuente recurrir a desiguadades de tipo exponencial, como la de Kiefer (ver p.ej. Teorema 2.1.3 B y Teorema 2.1.4 A en [13])

Teorema 6. *En el marco descrito anteriormente, sea $\{H_n^\omega\}_{n=1}^\infty$ una sucesión de k-ϕ-medias empíricas. Suponemos que X_0 vector aleatorio definido en el espacio probabilístico $(\Omega, \sigma, \mathbb{P})$ con llegada a $\mathbb{R}^d$ cuya única k-ϕ-media es H_0. Entonces, se tiene que $H_n^\omega \longrightarrow H_0$ para $\omega \in \Omega_0 \subset \Omega$ con $\mathbb{P}(\Omega_0) = 1$.*

Demostración

Sean $X_1, ..., X_n, ...$ vectores aleatorios i.i.d. definidos en el mismo espacio con $\mathscr{L}(X) = \mathscr{L}(X_i) = P$, y sea P_n^w la distribución de probabilidad empírica para $X_1(\omega), ..., X_n(\omega)$. Por el teorema de Glivenko-Cantelli, sabemos que $P_n^\omega \xrightarrow{d} P$; Estamos entonces en condiciones de aplicar el teorema de Skorohod: existen $Y_0^\omega, Y_1^\omega, Y_2^\omega, ...$ variables reales definidas en $(\mathcal{W}, \alpha, \lambda)$ y $\Omega_0 \subset \Omega$ con $\mathbb{P}(\Omega_0) = 1$ tales que $P_{Y_n^\omega} = P_n$ y $Y_n^\omega \longrightarrow Y_0^\omega$ $\lambda - c.s.$ si $\omega \in \Omega_0$.
Veamos que se verifican las hipótesis del teorema que nos asegurará que $H_n^\omega \longrightarrow H_0$:

1. Por Skorohod, se tiene que $Y_n^\omega \longrightarrow Y_0^\omega$ $\lambda - c.s.$ si $\omega \in \Omega_0$, por lo que se verifica la primera hipótesis

2. La unicidad de H_0 se verifica ya que hemos exigido que X_0 tenga una única k-ϕ-media.

3. Nos gustaría ver que

$$\int \phi(\inf_{h \in H_0} \|Y_n^\omega - h\|) d\lambda \longrightarrow \int \phi(\inf_{h \in H_0} \|Y_0^\omega - h\|) d\lambda$$

Dado que Y_n^ω es un vector aleatorio para cada $n = 1, 2, ...$, tenemos que $\phi(\text{ínf}_{h \in H_0} \|Y_n^\omega - h\|)$ también lo es por ser ϕ continua. Reescribimos utilizando el Teorema del Transfer

$$\int \phi(\inf_{h \in H_0} \|Y_n^\omega - h\|) d\lambda = \int \phi(\inf_{h \in H_0} \|y - h\|) dP_{Y_n^\omega}(y), \quad \forall n = 0, 1, 2, ...$$

Comprobaremos en primer lugar que si H_0 es la k-ϕ-media de X_0, también lo es de Y_0^ω. Como se tiene que $P_{Y_0^\omega} = P_0 = \mathscr{L}(X_0)$ y sabemos que H_0 es k-ϕ-media de X_0, resulta que

$$\int \phi(\inf_{h \in H_0} \|y-h\|) dP_{Y_0^\omega}(y) = \int \phi(\inf_{h \in H_0} \|x-h\|) dP_0^\omega(x) \leq \int \phi(\inf_{g \in G} \|x-g\|) dP_0^\omega(x) = \int \phi(\inf_{g \in G} \|y-h\|) dP_{Y_0^\omega}(y)$$

para cualquier conjunto $G \in \mathcal{B}_k$.
Por lo tanto, H_0 verifica la condición de ser k-ϕ-media de Y_0^ω:

$$\int \phi(\inf_{h \in H_0} \|y - h\|) dP_{Y_0^\omega}(y) \leq \int \phi(\inf_{g \in G} \|y - h\|) dP_{Y_0^\omega}(y)$$

A continuación escribiremos P_n^ω en lugar de $P_{Y_n^\omega}$ ya que hemos afirmado que son iguales. Denotamos por $\phi(\text{ínf}_{h \in H_0} \|Y_n^\omega - h\|) = f(Y_n^\omega)$ y escribimos

$$\int \phi(\inf_{h \in H_0} \|y - h\|) dP_n^\omega(y) = \int f(y) dP_n^\omega(x) = \frac{1}{n} \sum_{i=1}^n f(X_i(\omega))$$

Si se diera que $\mathbb{E}(f(Y_0^\omega)) < \infty$, dado que H_0 es fijo, la Ley Fuerte de los Grandes Números aseguraría que

$$\int f(x)dP_n^\omega(x) \xrightarrow{c.s.} \int f(x)dP_0(x) = \int \phi(\inf_{h\in H_0} \|x-h\|)dP_0(x) = \int \phi(\inf_{h\in H_0} \|Y_0^\omega-h\|)d\lambda, \quad \omega \in \Omega_0$$

Y con ello, tendríamos la condición requerida en (3). Veamos que efectivamente $\mathbb{E}(f(Y_0^\omega)) < \infty$. Si no lo fuera, dado que H_0 es la k-ϕ-media de Y_0^ω , se verificaría

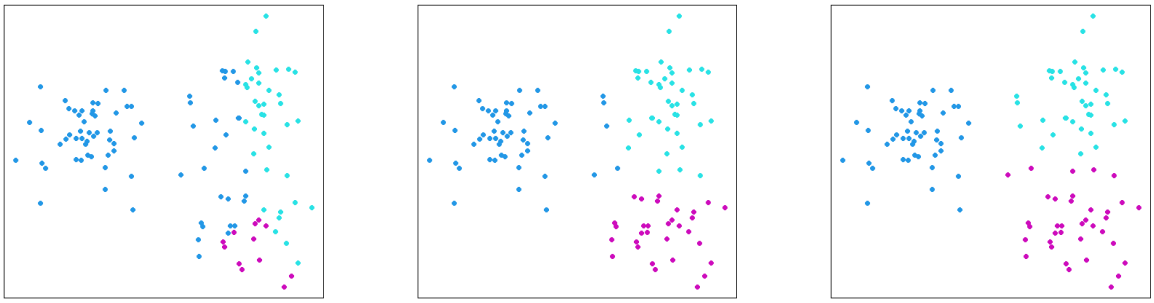
$$\infty = \mathbb{E}(f(Y_0^\omega)) = \int f(x)dP(x) = \int \phi(\inf_{h\in H_0} \|Y_0^\omega - h\|)d\lambda \leq \int \phi(\inf_{g\in G} \|Y_0^\omega - g\|)d\lambda$$

para cualquier conjunto $G \in \mathcal{B}_k$ por definición de k-ϕ-media. Esto es absurdo, ya que cualquier conjunto de k elementos sería una k-ϕ-media de Y_0^ω y habíamos supuesto la unicidad de H_0. Por lo tanto, debe darse $\mathbb{E}(f(Y_0^\omega)) < \infty$ y con ello logramos satisfacer la tercera hipótesis del Teorema.

□

4. Algoritmo de k medias

Ahora que hemos presentado el modelo teórico de k medias y estudiado sus propiedades matemáticas, nos preguntamos cómo hallar de manera práctica una solución al problema de k medias. Introduciremos diferentes algoritmos iterativos que consiguen encontrar soluciones óptimas localmente y expondremos algunas funciones que encontramos en el lenguaje de programación R a la hora de llevar a cabo esta labor. Hablaremos por último de las principales dificultades propias del procedimiento de k medias.



(a) (b) (c)

Figura 11: Evolución de los grupos en un conjunto de datos al buscar los centros de cluster mediante un algoritmo iterativo

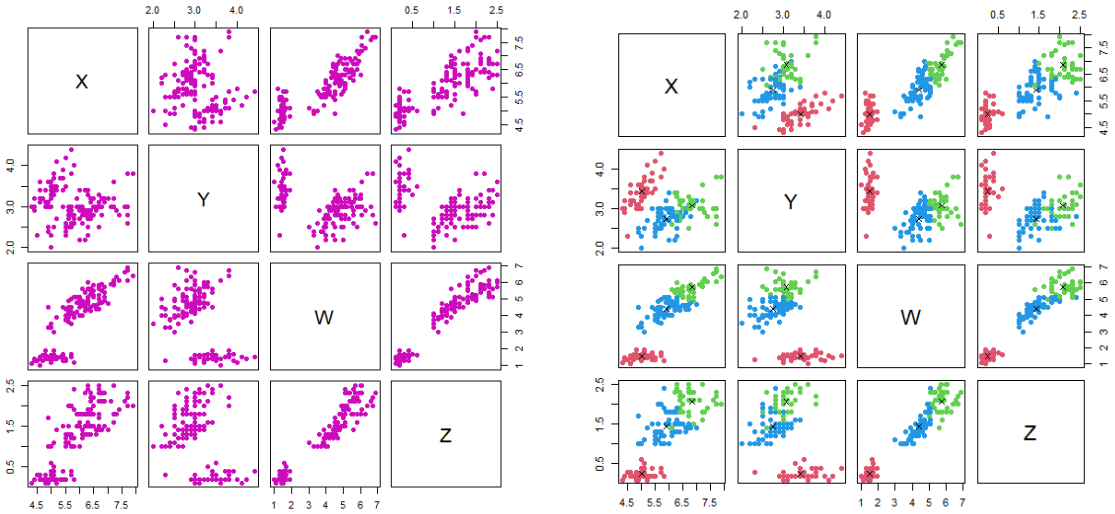
4.1. Necesidad de algoritmos

Consideremos de nuevo el conjunto de datos con el que trabajábamos al inicio de este documento y veamos qué problemas encontramos a la hora de buscar de manera práctica tres representantes del conjunto de observaciones y la partición en tres grupos de los individuos. Principalemnte, existen dos razones por las que resulta necesario realizar el proceso de agrupación de un conjunto de datos por el método de k medias de forma automática:

1. **Imposibilidad de encontrar agrupaciones visualmente si trabajamos con más de dos variables**

   Habíamos considerado anteriormente que los 150 individuos del conjunto de datos tomaban medidas en X e Y. En ese caso, conseguíamos realizar una agrupación intuitiva gracias a la visualización en dos dimensiones de estos individuos. Sin embargo, en el momento en el que aumentamos el número de variables, la búsqueda de patrones visuales se hace impensable. Si añadimos al conjunto de datos las mediciones en las variables que faltan (esto es, ahora trabajamos con las variables X, Y, W, Z) y representamos las observaciones por pares de variables, nos encontramos con la tarea prácticamente imposible de buscar agrupaciones visualmente

(12a). En la figura de la derecha, utilizamos el procedimiento de 3-medias para hallar tres representantes (marcados con una X) y la división en tres grupos que inducen:



(a) Observaciones sin agrupar

(b) Una agrupación dada por k medias

Figura 12: Individuos con valores en 4 variables en lugar de 2

Esto nos da una idea de cuánto necesitamos procedimientos automáticos de agrupamiento o clustering ya que, llegados a cierto punto de complejidad, somos incapaces de buscar esos patrones visuales que nos permiten crear grupos de individuos.

2. Cantidad desorbitada de posibilidades para elegir k representantes

Una primera idea ingenua que podríamos tener para conseguir k representantes sería realizar todas las posibles combinaciones de agrupaciones en k grupos del conjunto de datos y quedarnos con aquellas que minimizan el k-ϕ-potencial del conjunto de datos. Sin embargo, esta perspectiva no es en absoluto realista ya que el número de posibilidades se dispara desorbitadamente. Analicemos qué ocurre en el ejemplo. Debemos hallar todas las posibles combinaciones de 150 elementos en 3 grupos. Con fijar las posibilidades para los dos primeros grupos, el tercero queda ya elegido. Sean C_1 y C_2 dos de los grupos, sea k_1 el número de puntos de C_1 y k_2 el número de puntos de C_2. En primer, lugar escogeremos k_1 puntos de los 150, cosa que podemos hacer de $\binom{150}{k_1}$ formas. A continuación, de los $150 - k_1$ puntos restantes, elegimos k_2: tenemos $\binom{150-k_1}{k_2}$ posibilidades. Dado que solo necesitamos que $k_1, k_2 > 0$ y que $k_1 + k_2 < n$ para construir los

tres grupos, el número total de combinaciones para formar los tres clusters sería

$$\sum_{\substack{k_1,k_2>0 \\ k_1+k_2<n}} \binom{150}{k_1}\binom{150-k_1}{k_2}$$

Es claro que resulta imposible calcular todas las opciones y comprobar cuál resulta la mejor por el criterio de mínimos cuadrados, por lo que necesitamos confeccionar algoritmos que nos den una aproximación de la solución o consigan una solución óptima localmente.

4.2. Tres algoritmos para el problema de k medias

Consideremos a partir de ahora $\mathcal{X} = \{x_1, .., x_n\}$ un conjunto de puntos de $\mathbb{R}^d$, k un entero positivo y fijemos como función de disimilaridad ϕ la norma euclídea al cuadrado. Asignamos peso $\frac{1}{n}$ a cada uno de los puntos del conjunto de datos $\mathcal{X}$. Sea $H = \{h_1, .., h_k\} \subset \mathbb{R}$ y sea $\mathscr{C} = \{C_1, .., C_k\}$ partición asociada a H de $\mathcal{X}$.

Una vez descrito el marco de trabajo con el que continuaremos de ahora en adelante, si P_n denota la probabilidad muestral que asigna peso $\frac{1}{n}$ a cada $x_i \in \mathcal{X}$, escribiremos $W^k(H) = W^k_\phi(H, P_n)$ a fin de facilitar la lectura. Llamamos k-potencial de $\mathcal{X}$ por H a $W^k(H)$.

Encontrar H que minimice el k-potencial de $\mathcal{X}$ es equivalente a encontrar H que minimice la suma de distancias al cuadrado en cada uno de los clusters C_j:

$$H = \arg\min_{G\in\mathcal{B}_k} \sum_{i=1}^{n} \min_{j=1,..,k} \|x_i - g_j\|^2 = \arg\min_{G\in\mathcal{B}_k} \sum_{j=1}^{k} \sum_{x\in C_j} \|x - g_j\|^2 \tag{9}$$

Demostramos en la proposición (1) que la media aritmética de los datos, $m = \frac{1}{n}\sum_{i=1}^{n} x_i$, verificaba

$$m = \arg\min_{a\in\mathbb{R}^d} \sum_{i=1}^{n} \|x_i - a\|^2$$

Por lo tanto, si queremos minimizar (9), cada centro de cluster h_j deberá corresponderse con el promedio de los puntos de C_j, esto es:

$$h_j = \frac{1}{|C_j|} \sum_{x\in C_j} x$$

Esto nos permite plantear el problema desde la perspectiva de encontrar una partición del conjunto $\mathcal{X}$ de tal manera que se minimice la función objetivo, donde los k centroides se corresponderán con el promedio de los puntos asignados a cada cluster. Los tres algoritmos más conocidos de k medias que presentaremos a continuación tienen como común denominador el cálculo iterativo de los centroides como media de los puntos de cada respectivo cluster y la reasignación de los datos a los centroides una vez actualizados.

4.2.1. El algoritmo de Lloyd (1957)

Este algoritmo es el procedimiento de cálculo de centroides más conocido y extendido. Los pasos para encontrar los centros de cluster son los siguientes:

1. Escoger aleatoriamente k centroides iniciales $H = \{h_1, .., h_k\}$.
2. Para cada punto x_i, $i = 1, .., n$
   - Para cada $j = 1, .., k$, calcular la distancia $d_{i,j}$ de x_i a cada centro h_j.
   - Asignar x_i al cluster C_j cuyo centro h_j es más cercano
3. Para cada centro h_j, $j = 1, .., k$
   - Recalcular el centro de masa de cada cluster y asignar a h_j ese valor. Esto es

$$h_j := \frac{1}{|C_j|} \sum_{x \in C_j} x$$

4. Repetir los pasos 2) y 3) hasta que H se estabilice.

4.2.2. El algoritmo de MacQueen (1967)

El algoritmo de Lloyd tiene la inconveniencia de que cuando actualiza la asignación de los datos no actualiza los centroides, por lo que podemos llegar a asociar observaciones incorrectamente a un centroide simplemente porque este no estaba actualizado.

El algoritmo de MacQueen (1967) solventa esta carencia iterando sobre los puntos del conjunto de datos y corrigiendo en ese mismo momento el centroide correspondiente: Una vez reasignamos un punto a un cluster, recalculamos su respectivo centroide. En este caso, la iteración se repite hasta que ningún punto cambia de grupo. Cabe observar también que la versión de MacQueen conlleva un número más elevado de operaciones que el algoritmo de Lloyd. Los pasos para calcular los centros de cluster según MacQueen son los siguientes:

1. Escoger aleatoriamente k centroides iniciales $H = \{h_1, .., h_k\}$.
2. Para cada punto x_i, $i = 1, .., n$
   - Para cada $j = 1, .., k$, calcular la distancia $d_{i,j}$ de x_i a cada centro h_j.
   - Asignar x_i al cluster C_j cuyo centro h_j es más cercano
   - Recalcular el centro de masa del cluster C_j y asignar a h_j ese valor. Esto es

$$h_j := \frac{1}{|C_j|} \sum_{x \in C_j} x$$

3. Repetir 2) hasta que ningún punto cambie de grupo.

4.2.3. El algoritmo de Hartigan-Wong (1979)

Tanto el método de Lloyd como el de MacQueen eligen los centroides iniciales escogiendo k observaciones aleatoriamente. El algoritmo de Hartigan-Wong propone asociar un número del 1 a k a cada punto del conjunto de datos y tomar como centroide inicial h_i la media de los puntos a los que habíamos asociado la etiqueta i: de esta manera, casi todos los centroides se sitúan por la zona central de la nube de datos y el algoritmo es menos propenso a converger a un óptimo local. Al igual que MacQueen, actualiza los centroides en el momento en el que cada punto es reasignado.

El algoritmo de Hartigan-Wong incluye otra modificación que hace al método más flexible frente a los otros dos algoritmos. Recordemos que la función objetivo que tratábamos de minimizar es la suma de distancias al cuadrado desde cada punto hasta su centroide asignado. Hartigan-Wong permite que un punto se pueda asignar a un centroide incluso si este no es el más cercano si disminuye la función objetivo en mayor medida. De esta manera, resulta más sofisticado al considerar el impacto global que tendría asignar una observación a un centro de cluster.

1. Para cada $i = 1, .., n$, asignar aleatoriamente un número desde 1 hasta k a cada punto x_i.
2. Para cada $j = 1, .., k$, construimos C_j el conjunto de puntos a los que hemos asignado el número j. Calculamos el centroide h_j como

$$h_j := \frac{1}{|C_j|} \sum_{x \in C_j} x$$

3. Para cada punto x_i, $i = 1, .., n$
   - Para cada centro h_j, $j = 1, .., k$
     - Asignar x_i al cluster C_j.
     - Calcular el valor de la función objetivo, $W_{i,j}$, dada esta asignación.
   - Asignar x_i al cluster C_j cuyo $W_{i,j}$ es menor.
   - Recalcular el centro de masa del cluster C_j y asignar a h_j ese valor. Esto es

$$h_j := \frac{1}{|C_j|} \sum_{x \in C_j} x$$

4. Repetir 3) hasta que ningún punto cambie de grupo.

Al ser el más refinado de los tres algoritmos que hemos presentado, acarrea un número mayor de operaciones. En el caso de contar con un conjunto de datos en el que los grupos son fáciles de separar, el algoritmo de Lloyd será probablemente más rápido a la hora de encontrar una agrupación. Sin embargo, Hartigan-Wong es más flexible y complejo por lo que resulta más recomendable utilizarlo cuando no tenemos información sobre el conjunto de datos.

4.3. Principales carencias de k medias

El método de k medias es un procedimiento muy útil a la hora de buscar agrupaciones en numerosos conjuntos de datos, pero debido a sus características intrínsecas presenta ciertas deficiencias y debilidades con determinados conjuntos de datos. En este apartado, estudiamos cuales son los factores más importantes que deterioran el funcionamiento del algoritmo. A su vez, dejaremos el camino preparado para hablar de métodos relativos a la inicialización que consiguen subsanar en numerosas ocasiones estas deficiencias, un asunto que trataremos en secciones posteriores. Supondremos en todo momento que la medida de disimilaridad utilizada es la distancia euclídea al cuadrado.

4.3.1. Conjuntos de datos apropiados para k medias

Para entender un poco mejor qué estructura intrínseca deben tener los datos que tratamos de agrupar para que k medias consiga su cometido, debemos tratar de entender cómo hemos construido el algoritmo y qué disposiciones tiende a favorecer por ello. Una vez asignamos puntos a un centro de cluster en particular, dado que el criterio elegido ha sido aquel de minimizar la distancia euclídea al cuadrado y sabemos que los subconjuntos de la forma $\{x \in \mathbb{R}^d / \|x - C\|^2 \leq R^2\}$ son bolas en el espacio, la apariencia de los grupos que forma k medias es esférica. Por ejemplo, consideramos el conjunto de datos con 200 observaciones de $X_1 \sim N((6,0), \Sigma)$ y otras 200 de $X_2 \sim N((0,0), \Sigma)$, con matriz de covarianza común $\Sigma = \left(\begin{smallmatrix} 2 & 0 \\ 0 & 2 \end{smallmatrix}\right)$. En este caso, el algoritmo clasifica correctamente debido a la apariencia esférica de los grupos.

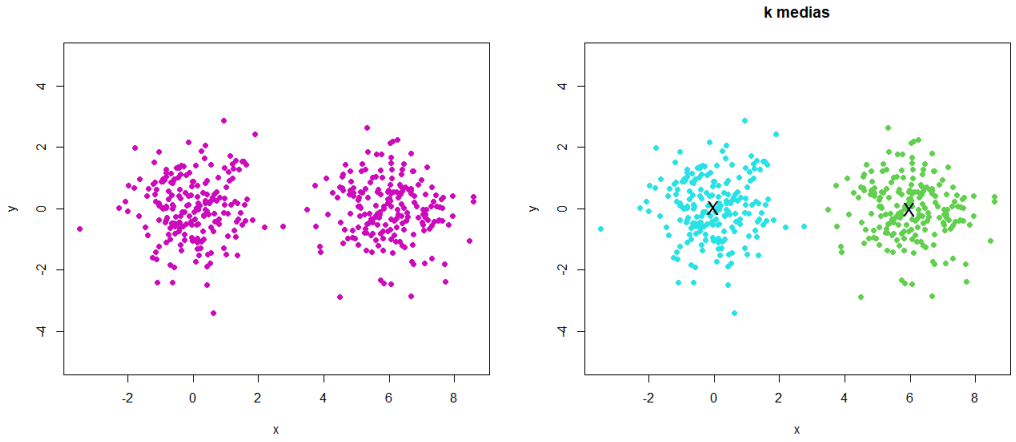


Figura 13

Debido a esta predilección por estructuras esféricas, en el momento en el que nos enfrentamos a un conjunto de datos con una forma "complicada", k medias suele fallar en la agrupación (Ver *Shaped Sets, Spiral N=312, k=3, D=2* en [8]):

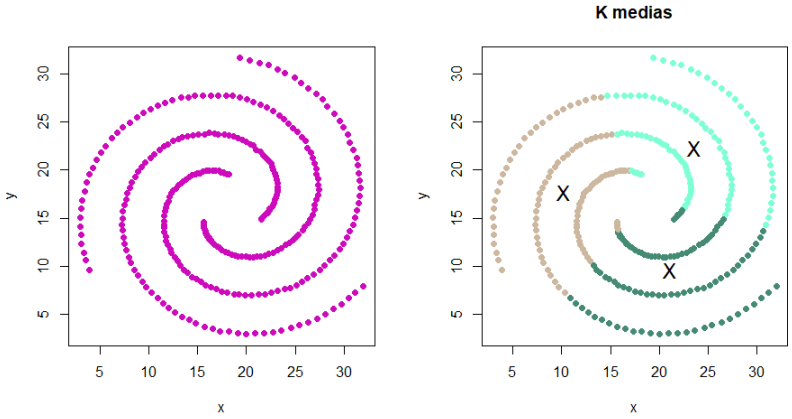


Figura 14

Sin embargo, no hace falta buscar conjuntos de datos con formas disparatadas: Supongamos que contamos con 200 observaciones de $X_1 \sim N((6,0),\Sigma)$ y otras 200 de $X_2 \sim N((0,0),\Sigma)$, con matriz de covarianza común $\Sigma = \begin{pmatrix} 0{,}5 & 0 \\ 0 & 20 \end{pmatrix}$. Dado que Σ nos indica una variabilidad mucho más alta en la segunda coordenada de los datos que en la primera, sabemos a priori que van a ser grupos con apariencia "alargada". Debido a que no son esféricos, k medias simplemente opta por partirlos a la mitad alejándose por completo de la solución:

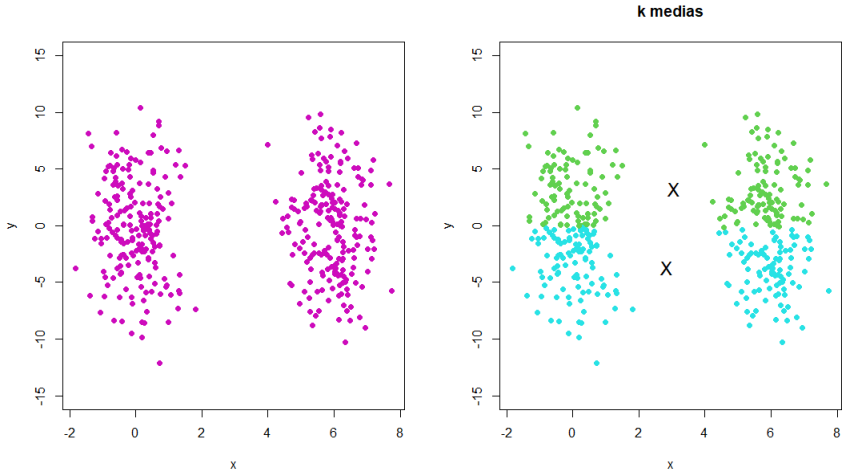


Figura 15

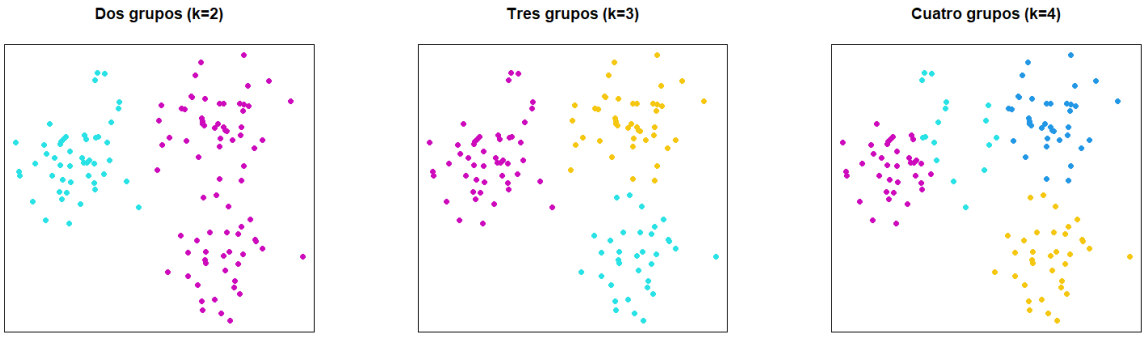
Para esta carencia del método, se han introducido modelos más ambiciosos que permiten formas

elipsoidales diversas para cada agrupamiento.

A su vez, es un método muy sensible ante contaminaciones o pequeñas desviaciones del modelo, ya que si existen puntos aislados del resto, su aportación al potencial será muy grande y para disminuirlo, el método de k medias tenderá a desplazar sus centros en esa dirección a fin de reducirlo. Este problema se puede solventar a través de procedimientos de recorte, que podemos encontrar en artículos como [4].

4.3.2. Elección de k, el número de grupos a buscar

A la hora de buscar agrupamientos con el algoritmo de k medias en un conjunto de datos, hemos supuesto que tenemos un número entero positivo k prefijado que se corresponde con el número de grupos a buscar. En algunas ocasiones, desearemos particionar un conjunto de observaciones en un número fijo de grupos con alguna finalidad, pero el caso más común es aquel en el que el número de clusters viene determinado por la naturaleza de los datos y las relaciones que albergan los individuos entre sí. Por ello, necesitamos disponer de algún criterio que nos permita establecer en cuántos grupos parece más creíble particionar un conjunto de datos.



(a) (b) (c)

Figura 16: Posibles elecciones de k para un mismo conjunto de datos.

La elección de k es un problema intrínsecamente imposible de resolver si consideramos un conjunto de datos arbitrario, ya que depende de la estructura del conjunto de datos y el número de grupos a determinar deberá ser elegido a medida en cada caso: no existe un método general. Es claro que este número no puede estimarse con el fin de minimizar el k-potencial, ya que la forma de hacerlo sería tomar tantos grupos como observaciones consiguiendo siempre un k-potencial de 0.

Para tratar de solventar esto, existen varias soluciones *ad hoc* que permiten establecer un criterio para la elección de k. No proporcionan siempre una solución del todo satisfactoria ya que su buen funcionamiento está supeditado a la estructura intrínseca del conjunto de datos. Presentamos tres

de las más utilizadas:

1. **Criterio del *Codo***
   Buscamos realizar un test que compare la mejora al elegir $k+1$ grupos frente a elegir k comparando la distancia intra-cluster obtenida variando el número de grupos. Para ello, ejecutamos el algoritmo de k medias para para diferentes valores de k, calculamos la distancia intra-cluster obtenida en cada caso y representamos ambos datos en un gráfico. Para el conjunto de datos de la imagen de arriba, se tendría la siguiente representación:

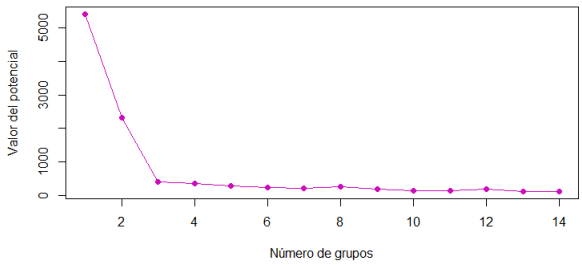


Figura 17

   En esta gráfica se debería de apreciar un cambio brusco en la evolución del potencial, teniendo la línea representada una forma similar a la de un brazo y su codo. El punto en el que se observa ese cambio brusco será aquel que seleccionaremos como número de clusters a buscar en el conjunto de datos, en este caso $k = 3$. Es recomendable ejecutar este procedimiento repetidas veces, dado que una mala inicialización para algún valor de k puede alterar el resultado. Podemos encontrar más información sobre este criterio p.ej. en [1].

2. **Método de la Silueta**

   El Método de la Silueta define para cada punto del conjunto de datos una medida de similaridad entre él y los demás individuos del mismo cluster en comparación con aquellos de clusters diferentes. A cada punto se le asigna un valor entre -1 y 1 denominado índice de silueta, donde un valor alto equivale a una mejor coincidencia con el cluster en el que debe clasificarse y menor disposición a pertenecer a otro grupo. La media del índice de silueta de todos los puntos se utiliza para indicar la calidad de la agrupación.

   Dado un conjunto de datos $\{x_1, .., x_n\}$, si para un punto x_i denotamos por a_i la media de las distancias de x_i al resto de puntos del cluster al que ha sido asignado, y por b_i la media de las distancias de x_i a puntos de grupos diferentes al suyo, el índice de silueta de x_i se calcula del siguiente modo, donde k es el número de grupos:

$$s(x_i, k) = \frac{b_i - a_i}{\text{máx}\{b_i, a_i\}} \tag{10}$$

De esta manera, un valor cercano a cero denota que la observación es cercana a un cluster vecino, un índice cercano a 1 el punto está muy integrado en el cluster y está lejos del resto de grupos, y un valor próximo a -1 indica que los clusters no están asignados correctamente. Si realizamos la media de índices de silueta de todos los puntos, obtenemos el índice de silueta para k grupos:

$$S(k) = \frac{1}{n}\sum_{i=1}^{n} s(x_i, k) = \frac{1}{n}\sum_{i=1}^{n} \frac{b_i - a_i}{\text{máx}\{b_i, a_i\}}$$

La técnica para elegir k consistirá entonces en:

a) Ejecutar el algoritmo para diferentes valores de k

b) Para cada agrupación en k clusters, calular el índice de silueta $S(k)$

c) Escoger como número de clusters óptimo aquel que maximiza el índice de silueta.

Podemos encontrar más información relativa a este procedimiento para elegir el número óptimo de grupos en [7]. Si aplicamos este criterio al conjunto de datos anterior, obtendríamos lo siguiente:

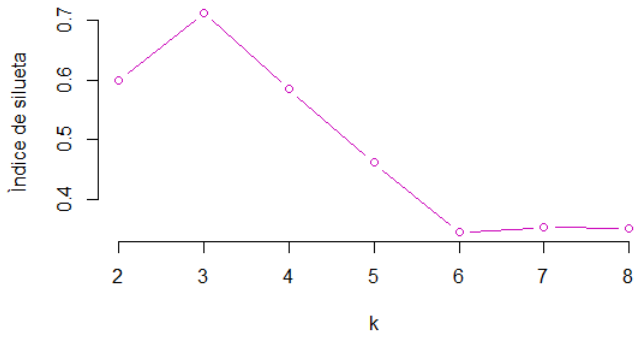


Figura 18

Dado que el índice de silueta más alto se alcanza para $k = 3$, este sería el número de clusters óptimo con este criterio.

3. **Remuestreo**

Si un conjunto de datos tiene una estructura intrínseca cuyo número adecuado de grupos es k, entendemos que al quitar parte de los puntos del conjunto arbitrariamente este sigue manteniendo esa misma organización. El método de remuestreo consiste en tomar subconjuntos

de la muestra y estimar para ellos el número de grupos óptimo, ya que entendemos que la estructura de los datos de un subconjunto de la muestra será parecida a la estructura global del conjunto de datos. De esta manera conseguimos intuir el número de clusters del conjunto de datos sin que puntos "conflictivos" (aquellos que no parecen pertenecer a ningún grupo) nos alejen de la solución. Para estimar el número de clusters óptimo de los subconjuntos, podemos analizar la estabilidad de las soluciones obtenidas con diferentes remuestreos y con diferentes valores de k.

4.3.3. Fuerte dependencia de la inicialización

El algoritmo de k medias tiene un primer paso que consiste en elegir k centroides a partir de los cuales, realizando operaciones, llegaremos a los centros de cluster definitivos. La manera de elegir estos k puntos es completamente decisiva en el resultado final: una mala elección de los centroides iniciales puede suponer el fracaso absoluto del procedimiento a la hora de agrupar a los individuos. La razón por la que esto sucede es la gran dificultad que tienen los centroides para moverse de un cluster a otro si la distancia que los separa es demasiado grande, por lo que es necesario colocar desde el principio un buen candidato cerca de cada grupo. Observamos esto con un ejemplo: Consideramos el conjunto de 1351 observaciones agrupadas en 9 clusters bastante separados entre sí (conjunto de datos extraido de [8]: *Synthetic data with Gaussian clusters. N=1351 vectors in k=9 clusters in 2 dimensional space*). En el momento en el que k medias no escoge un centro en cada uno de estos clusters con su inicialización aleatoria, el algoritmo está abocado al fracaso ya que los centros de cluster no se desplazan tal y como querríamos. Realizamos un par de iteraciones para ver como responde k medias ante el conjunto de datos:

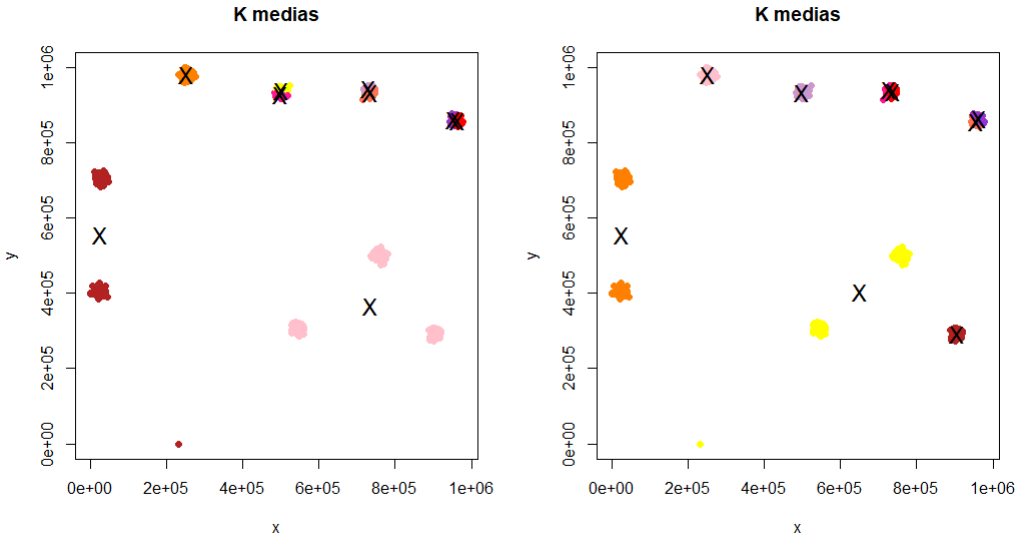


Figura 19

Dado que escoge varios centros muy próximos entre sí, varios clusters se quedan en un principio descubiertos y los centroides de otros grupos se ven obligados a desplazarse en su dirección para cubrir esa zona de puntos.

Otra situación muy dependiente de la inicialización del método es aquella en la que trabajamos con un conjunto de datos con grupos "desbalanceados" o desequilibrados. Esto ocurre cuando existe mayor densidad de puntos en unos grupos que en otros, por lo que es mucho más probable escoger centros pertenecientes a estos clusters que al resto. En el ejemplo que mostramos a continuación, se han considerado 8 grupos, tres de ellos cuentan con 2000 observaciones cada uno mientras que los cinco restantes solo con 100 por grupo (conjunto de datos en [8]: *Unbalanced, Synthetic 2-d data with N=6500 vectors and k=8 Gaussian clusters*). De esta manera, dado que es más factible elegir centros en estos tres clusters densos, el algoritmo falla en su ejecución numerosas veces:

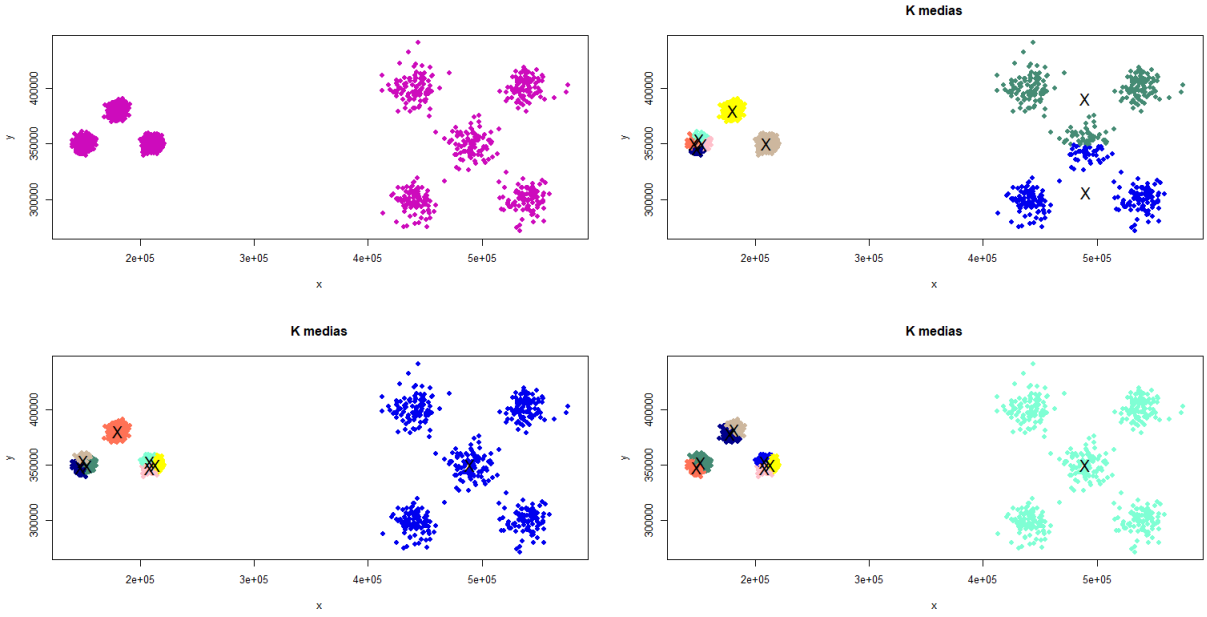


Figura 20

Nos centraremos en particular en esta desventaja del método de k medias y describiremos en la siguiente sección los procedimientos utilizados para enmendar esta debilidad del algoritmo. Además, profundizaremos en un método concreto denominado k medias++ que mejora notablemente situaciones como las que hemos descrito hasta ahora, todo ello motivado por el hecho de que el éxito del algoritmo de k medias, cuando el conjunto de datos admite una buena clasificación en grupos esféricos, depende mayoritariamente de realizar una buena inicialización.

4.4. Implementación

Presentaremos algunas funciones presentes en el entorno de programación R que nos resultarán de gran ayuda para buscar agrupaciones en un conjunto de datos siguiendo el procedimiento de k medias.

En primer lugar, decir que podemos implementar fácilmente el algoritmo de k medias (en sus distintas variantes) en cualquier lenguaje de programación, ya que tan solo necesitaremos, además de poder trabajar con matrices y vectores, medir distancias entre puntos y calcular la media de un conjunto de observaciones. En nuestro caso, hemos construido una función *kmeans_lloyd* para implementar el algoritmo de Lloyd en R y poder mostrar visualmente el proceso de agrupación de los datos en cada una de las iteraciones. El código que utilizaremos será el siguiente:

```
kmedias_lloyd <- function(k,datos,N){

 # k: número de grupos que buscamos
 # datos: conjunto de datos en el que buscar grupos
 # N: número de iteraciones máximas permitido

 n<- nrow(datos)                    #Cuántas observaciones tenemos
 index <- sample(n, k)
 centroides <- datos[index,]        #Realizamos un sorteo para elegir k centros

 A<- matrix(NA,nrow=nrow(datos),ncol=ncol(datos))
 H<- matrix(NA,nrow=k,ncol=ncol(datos))
 #escribimos los datos en una matriz en lugar de trabajar con dataframes:
 for (i in 1:k){H[i,] <- unlist(as.vector(centroides[i,])) }
 for (i in 1:n){A[i,]<-unlist(as.vector(datos[i,])) }
 todos<-rbind(A,H) #Matriz con los datos y los centroides
 #cluster al que pertenece cada punto:
 cluster<- matrix(NA,nrow=nrow(datos),ncol=1)

 d<- matrix(NA,nrow=k,ncol=1)  #distancias entre centros y observaciones

 Cota<- sum(dist(todos))  #cota para comparar distancias
 #COMIENZA EL ALGORITMO
 for (i in 1:N){     #Número máximo de veces que repetimos el algoritmo.
   todos<-rbind(A,H) #Actualizamos la matriz con los nuevos centroides
   for(s in 1:n){
     cota<-Cota      #Volvemos a tomar la cota
   for (j in 1:k){
     #Para cada punto, calculamos la distancia a cada centro de cluster
     d[j]<- dist(todos[c(s,n+j),])
     if(d[j]<cota){ #Si es mejor, cambio el cluster y la cota

       cluster[s]<- j
```

```
        cota <- d[j]
      }
    }
  } #Cada punto ya está asociado a un cluster. Recalculamos los centroides:
    for (j in 1:k){
      suma<- 0
      cuantos<-0
      for (s in 1:n){
        if(cluster[s]==j){
          cuantos<- cuantos+1
          suma=suma+todos[s,]
        } #Reescribimos los centroides: Les asignamos la media del cluster
        H[j,]<- 1/cuantos * suma
      }
  }
}
  return(H)
}
```

Este programa en concreto devuelve el valor de los centroides de cada grupo (Podríamos devolver en su lugar el vector *cluster* en el cual vienen clasificadas las observaciones). Añadiendo este fragmento de código, conseguiremos ilustrar gráficamente con qué grupos contamos coloreando cada cluster de un color y dónde están ubicados los centroides en cada iteración con una "X":

```
plot(datos, col=cluster+3, pch=19, xlab= "X",
ylab= "Y")
points(H, pch= "X", cex= 2)
```

Consideremos de nuevo el conjunto de datos de 150 observaciones medidas en las variables X e Y. Utilizamos la función *kmeans_lloyd* para buscar tres grupos y los respectivos centroides. Estas serían las cuatro primeras iteraciones al ejecutar la función una vez:

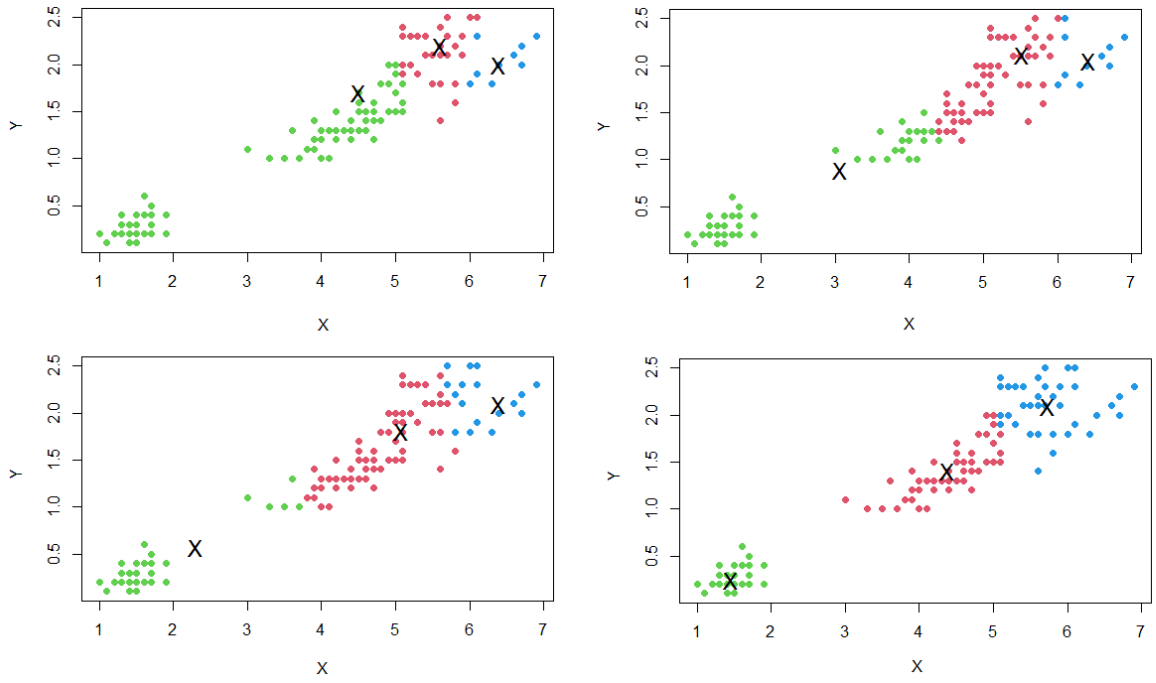


Figura 21: Cuatro primeras iteraciones del algoritmo

Observamos como a pesar de haber escogido tres centros iniciales alejados de los que luego serán los adecuados, el algoritmo permite que se desplacen hacia otras zonas con observaciones desprovistas de centroides y así esparcir los centros de cluster mejor. Al realizar diferentes iteraciones, encontramos disposiciones de los centroides que hacen que el algoritmo llegue a un óptimo local más rápido y otras que en cambio lo entorpecen. Por ejemplo, para esta muy mala elección de los centroides iniciales, las agrupaciones realizadas en las primeras cuatro iteraciones del método distan mucho de la solución óptima:

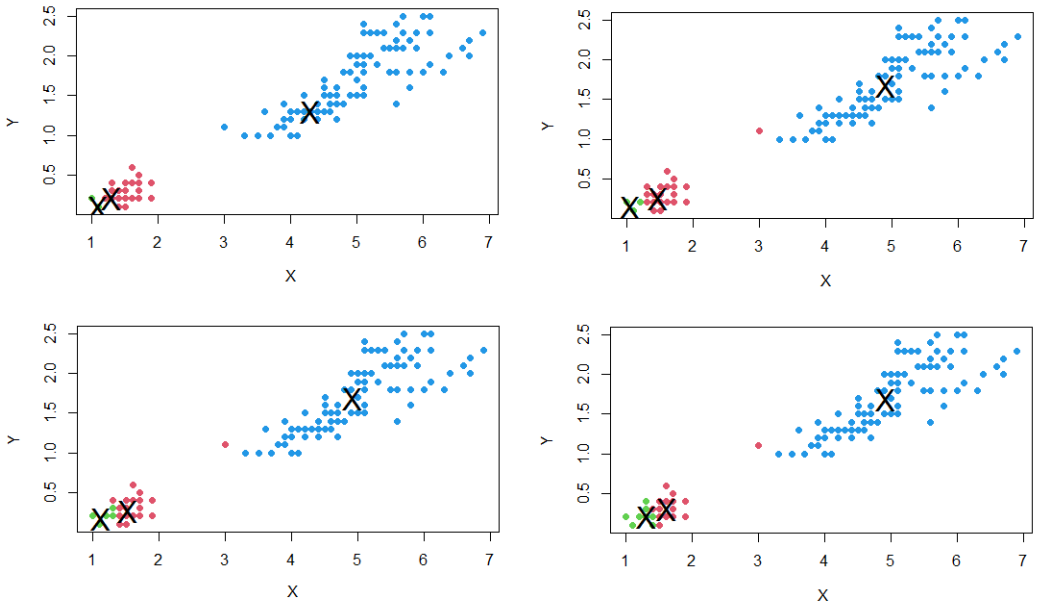


Figura 22: Elección desfavorable de los centros

Si buscamos una función que realice este proceso sin necesidad de programarla nosotros desde cero, podemos acudir al paquete **stats** de R que contiene, entre otras, la función llamada *kmeans*.

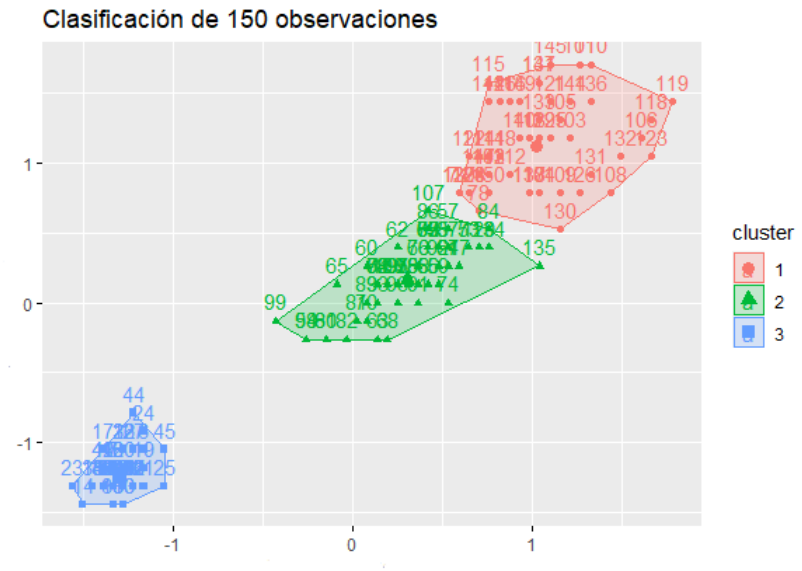
La función *kmeans* particiona un conjunto de datos dado en k grupos. Su uso es sencillo y además es muy versátil, ya que nos proporciona la opción de escoger los centros iniciales, el número máximo de iteraciones que queremos realizar y el algoritmo concreto para buscar los centroides (Los descritos anteriormente:"Lloyd", "MacQueen" y "Hartigan-Wong", junto con otro debido a Forgy). Esta función devuelve un dataframe en el que están presentes

- Un vector que indica a qué cluster enviamos cada punto.
- Una matriz con los centroides finales.
- El potencial correspondiente a esa agrupación en clusters.
- La distancia intra-cluster, para cada cluster.
- La distancia inter-cluster, para cada cluster.
- Un vector con el número de observaciones asignadas a cada cluster.
- Número de iteraciones llevadas a cabo.

Por ejemplo, si aplicamos *kmeans* de R en la versión de Lloyd al conjunto de datos con el que estábamos trabajando, obtenemos lo siguiente:

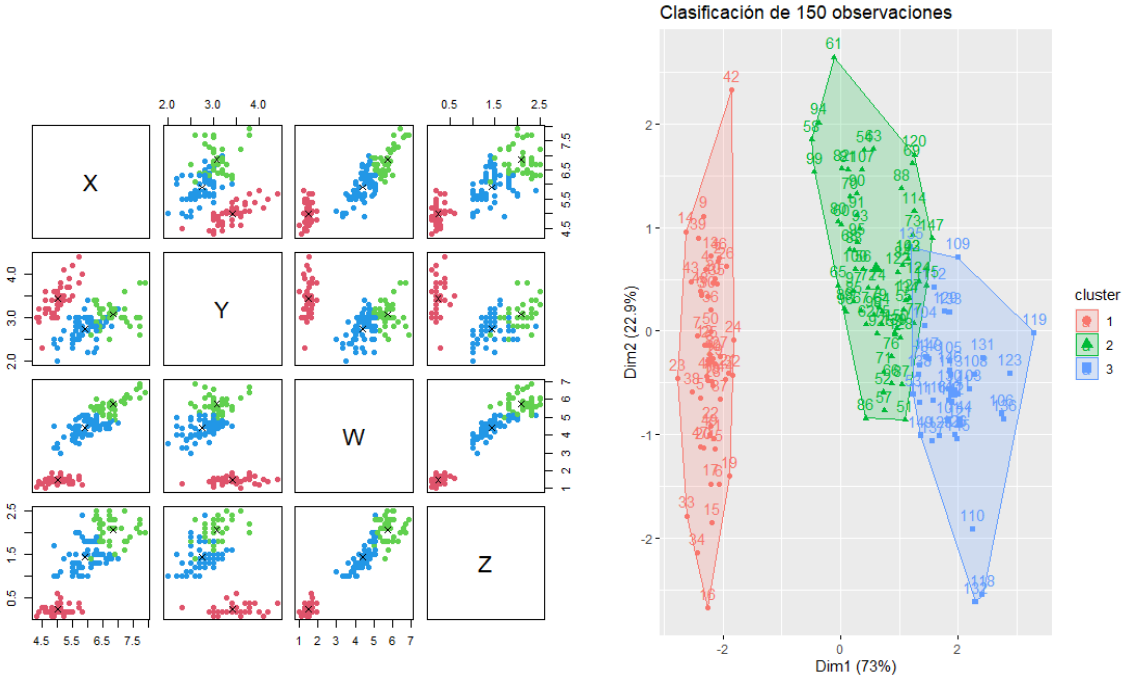
```
K-means clustering with 3 clusters of sizes 50, 46, 54

Cluster means:
         X        Y
1 1.462000 0.246000
2 5.626087 2.047826
3 4.292593 1.359259

Clustering vector:
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [37] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [73] 3 3 3 3 3 2 3 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 2 2 2 2 3 2
[109] 2 2 2 2 2 2 2 2 2 2 3 2 2 2 3 2 2 3 3 2 2 2 2 2 2 2 2 2 3 2 2 2 2 2 2
[145] 2 2 2 2 2 2

Within cluster sum of squares by cluster:
[1]  2.02200 15.16348 14.22741
 (between_SS / total_SS =  94.3 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"
[5] "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"
```

Otro paquete interesante que proporciona unos gráficos y visualizaciones muy intuitivas es **factoextra**. En él, encontramos la función $fviz\_cluster$ que permite ver las agrupaciones más claramente, etiquetando el número de observación a la hora de efectuar el gráfico:



Si aplicamos el procedimiento de k medias al conjunto de datos original, aquel con cuatro variables X, Y, W y Z, y tratamos de ilustrar gráficamente lo que ocurre, obtenemos una representación visual dada por pares de variables que puede resultar menos intuitiva a la hora de pensar qué está pasando con nuestros datos.

La función $fviz\_cluster$ que mencionábamos anteriormente nos ofrece una solución interesante para observar en un solo gráfico la agrupación efectuada por k medias utilizando Análisis en Componentes Principales (PCA). Es un procedimiento multivariante cuyo objetivo fundamental es reducir la dimensión del conjunto de datos minimizando la pérdida de información para conseguir una descripción más sencilla de los mismos. Para ello, se construye un nuevo conjunto de variables a partir de combinaciones lineales de las variables originales que recojan la mayor cantidad de información posible y se escogen tantas direcciones o componentes sean necesarias para expresar la información y representarla. En nuestro caso para poder ver una representación en dos dimensiones, buscamos extraer las dos primeras componentes principales. Aplicando la función $fviz\_cluster$ conseguimos el resultado. Observamos que en el caso del gráfico con PCA queda recogida el 86.7 % de la información, por lo que puede ser interesante utilizarlo cuando necesitemos una visualización sencilla de las agrupaciones.

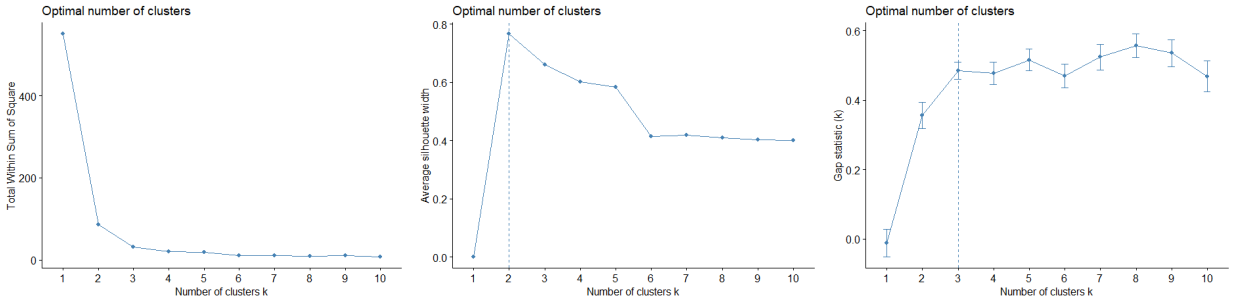


(a) Gráfico de pares de variables

(b) Gráfico con PCA

Figura 23: Una forma más cómoda de visualizar los resultados

Por otro lado, en este mismo paquete, encontramos la función *fviz_nbclust* que elabora de manera automática un diagrama para elegir k. Nos permite elegir el método para determinar el número óptimo de clusters: *Distancia Intra-cluster*, *Método de la Silueta* o *Método de Estadística de Brecha* (Esta última compara la variación intra-grupo total para diferentes valores de k con sus valores esperados bajo una distribución sin agrupamiento obvio; Encontramos detallado el procedimiento en [14]). En el ejemplo que nos ocupa, para el caso de dos variables X, Y, el resultado sería el siguiente:



(a) Método de distancia Intra-cluster (b) Método de la Silueta (c) Método de Estadística de Brecha

Figura 24: Diferentes métodos en R para elegir el número de grupos

Por lo tanto, en este caso parece adecuado buscar 2 o 3 grupos en el conjunto de datos.

5. Métodos y estrategias para mejorar la inicialización

El objetivo de esta sección es describir someramente algunos de los métodos de inicialización más comunes a la hora de elegir los centroides iniciales para ejecutar el algoritmo de k medias. Después, profundizaremos en el procedimiento de k medias++, presentando los pasos para llevarlo a cabo e implementarlo. A continuación, expondremos los principales resultados matemáticos respecto a la acotación que podemos conseguir del valor medio del potencial hallado con el algoritmo y, para terminar, comentaremos algunos ejemplos con diferentes conjuntos de datos que nos permitirán observar las ventajas de k medias++ frente a k medias "standard". Dado que en ocasiones haremos referencia a cuestiones del campo de la complejidad computacional, se ha incluido en el anexo un apartado donde se recopilan aquellas nociones más importantes con las que trabajaremos.

5.1. Algunos métodos para la inicialización de k medias

Es sabido que k medias presenta deficiencias con ciertos conjuntos de datos, como pueden ser los grupos no esféricos, grupos desequilibrados... Sin embargo, no es el único algoritmo al cual le sucede esto. A la hora de preguntarnos si existe un método de clustering que sea mejor que todos los demás, la respuesta es negativa: debido a cómo se construyen los procedimientos de clustering y en qué estrategia basan su agrupación de individuos, siempre podemos encontrar o construir un conjunto de datos que sea completamente desfavorable para hallar una agrupación buena utilizando un algoritmo en particular. Existen otros algoritmos distintos de k medias bien conocidos que en muchos casos proporcionan mejores resultados que k medias. Sin embargo, este es muy popular por varias razones:

1. Su implementación es muy sencilla y es posible construir una función *kmedias* en cualquier lenguaje de programación.

2. Es uno de los algoritmos más estudiados, tanto de manera práctica con diferentes conjuntos de datos como de manera teórica. Existen pocos procedimientos de clustering para los cuales se hayan probado tantos resultados de convergencia, existencia, probabilidad de puntos frontera entre clusters... Es lógico preferir usar un procedimiento estudiado exhaustivamente que presenta alguna limitación antes que un algoritmo potencialemnte bueno que puede presentar algún inconveniente "oculto" o todavía no examinado.

3. La actuación de k medias mejora notablemente con una buena elección de centros iniciales, además de realizando el procedimiento repetidas veces para poder elegir la opción que más minimiza nuestro potencial.

Es por ello por lo que ponemos empeño en buscar técnicas para elegir centroides iniciales de manera inteligente en lugar de simplemente realizar un sorteo y escoger k puntos como habíamos hecho hasta ahora. Existen muchos procedimientos para la inicialización y trataremos de recoger algunos de los principales. Sus descripciones, principales ventajas y algunas de sus insuficiencias han sido extraidas del artículo *How much can k-means be improved by using better initialization and repeats?* [8], donde se recoge mucha más información al ensayar el porcentaje de éxito que tiene cada algoritmo de inicialización con cada tipo de conjunto de datos.

5.1.1. Principales procedimientos para la inicialización

Comentaremos algunos de los métodos más conocidos para elegir los centroides iniciales. Como hemos comentado antes, también es interesante repetir k medias para obtener soluciones diversas con diferentes inicializaciones.

Los requisitos que le pedimos a un algoritmo de inicialización son, entre otros, que sea simple de implementar, que su complejidad sea como mucho la de k medias y que no sea necesario ningún parámetro adicional que los que ya tenemos para poder ejecutarlo. Comentar que la técnica que estábamos llevando a cabo hasta ahora se denomina **Random Centroids** y consiste en reordenar aleatoriamente los datos y escoger los k primeros.

- **Random Partition**
  Consiste en generar una partición aleatoria del conjunto de datos y calcular sus centroides. Su ventaja es que evita coger puntos no representativos de la muestra (outliers), pero provoca una aglomeración de centroides en la zona central del conjunto de datos.

- **Furthest Point Heuristic**
  Este procedimiento elige un punto aleatoriamente de la muestra y lo escoge como centroide y va añadiendo el resto. El siguiente centro de custer que elijamos será el punto que más dista de los centroides ya elegidos. De esta manera conseguimos esparcir los centros, pero es un procedimiento con tendencia a coger outliers.

- **Sorting Heuristic**
  Consiste en ordenar los puntos de la muestra de acuerdo con un criterio (Distancia al centro, densidad, centralidad, mayor varianza...) y escoger bien los k primeros, bien uniformemente (puntos en la posición $\frac{N}{k}$)... Funciona adecuadamente cuando los clusters están bien separados y tienen valor diferente respecto al criterio de orden que utilizamos.

- **Projection Based Heuristics**
  Consiste en proyectar los puntos sobre un eje, como podría ser por ejemplo una dirección principal (Uitilizando Análisis en Componentes Principales) y particionar este eje en k segmentos del mismo tamaño. A continuación, se calculan los centroides de los puntos asociados a cada segmento del eje. Es un método interesante cuando los datos tienen una estructura "Unidimensional" (Se ven bien representados en una sola dimensión).

- **Density Based Heuristics**
  Calcular la "densidad" de cada punto (Por ejemplo, según cuántos otros puntos de la muestra hay en la bola de radio ϵ, dividir en celdas los datos y calcular la cantidad de puntos que hay, calcular la distancia media de los puntos a sus k vecinos más próximos...). No resulta del todo interesante ya que en dimensiones altas requiere un número muy elevado de operaciones.

- **Inicialización k means++**

  El procedimiento busca elegir los centros de forma que puntos más alejados de los centroides ya escogidos tengan probabilidad más alta de ser seleccionados. Para ello se otorga una ponderación a cada punto en función de cuanto dista del centroide más cercano. Por lo general, mejora notablemente

5.2. El procedimiento de k medias++

Una buena inicialización del algoritmo de k medias permite no solo una convergencia más rápida del método hacia una solución aproximada sino una mejora considerable del error cometido. El algoritmo de inicialización que presentamos a continuación fue propuesto en 2007 por David Arthur y Sergei Vassilvitskii como una solución a las carencias que presenta el algoritmo de k medias. El procedimiento de k medias++ especifica un procedimiento para la elección de centros de cluster iniciales antes de ejecutar el método de k medias como tal, permitiendo así encontrar una solución "O(log k) competeitiva" a la solución óptima de k medias.

Intuitivamente, veíamos que la expansión de los centroides a la hora de inicializar el algoritmo parecía una buena idea a fin de ser capaz de "cubrir" todas las zonas y asignar un centro de cluster posteriormente a cada grupo. Consideremos el siguiente conjunto de datos con 15 grupos y 600 individuos en total (conjunto de datos presente en [8]: *Shaped Sets, R15 N=600, k=15, D=2*). A la hora de hallar agrupaciones en el este conjunto de datos, a pesar de ser todos los grupos esféricos y bien diferenciados, k medias lo encuentra muy dificil si inicialmente no sitúa un centroide en cada uno de los grupos más alejados, pues el algoritmo alcanza un mínimo local y no desplaza los centros hacia esas zonas:

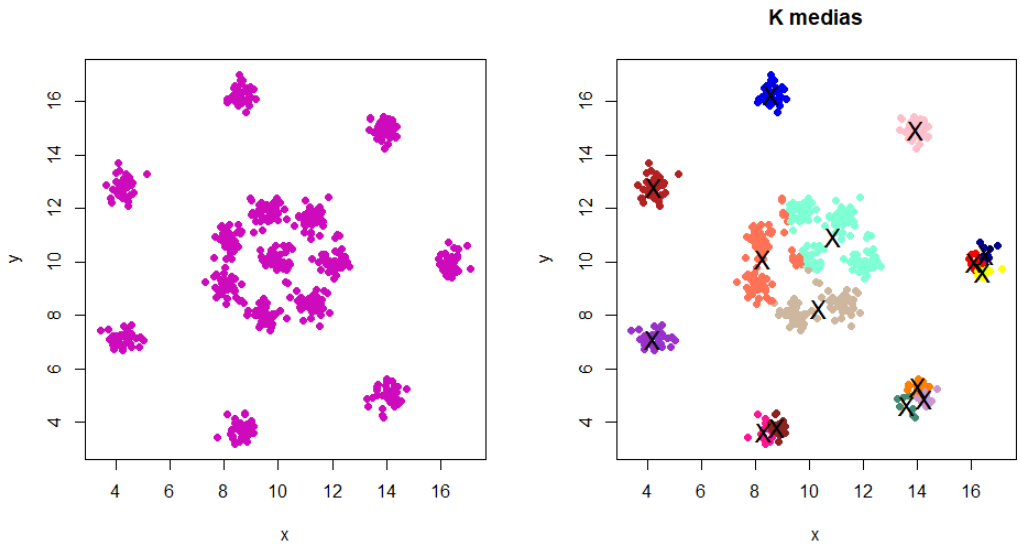


Figura 25: Actuación de k medias ante 15 clusters diferenciados pero muy separados

La pregunta radica entonces en cómo lograr implementar esta idea de esparcir los centroides de manera que sea un procedimiento aleatorio, consigamos cubrir todas las zonas y no caigamos (o al menos no siempre) en elegir outliers que seguramente formen grupos por si solos al estar mucho más alejados del resto de puntos. La filosofía de k medias ++ se basa en la idea de elegir los centros de tal manera que la probabilidad de coger nuevos centroides alejados de los puntos ya elegidos sea

mayor. El algoritmo tiene por lo tanto las siguientes etapas:

1. Utilizando una variable aleatoria X con distribución uniforme, escogemos un centro entre los puntos de la muestra.
2. Dada una observación x de la muestra, se define la distancia $D(x) = \text{ínf}_{c \in \mathscr{C}}\{d(x,c)\}$: es decir, la distancia entre un punto x y el centroide más cercano de los que ya hemos seleccionado.
3. Construimos la variable aleatoria Y que otorga probabilidad proporcional a $D(x)^2$ a cada punto x aún no escogido.
4. Se repiten los pasos 2 y 3 hasta que se hayan seleccionado k centros.
5. Una vez tenemos los centros iniciales, se procede con el método de k medias tal y como lo conocemos.

Es un procedimiento interesante ya que el algoritmo de k medias converge muy rápidamente después de la selección de puntos iniciales y además con considerables mejoras en las soluciones obtenidas, ya que vimos que k medias es capaz de generar soluciones arbitrariamente peores que la solución óptima si la elección de centros iniciales es desafortunada (Véase el ejemplo de los cuatro puntos).

Al igual que vimos que era fácil implementar *kmedias* en el entorno estadístico R, de igual manera será sencillo construir una función *kmedias_plus* que nos proporcione k centros elegidos siguiendo los pasos descritos anteriormente. En este caso, utilizamos la función *sample* bien conocida en R que nos permite realizar un muestreo con una distribución personalizada, en nuestro caso, aquella ponderación que hemos denominado $\mathcal{D}^2$. Un posible código para conseguirlo sería el siguiente:

```
kmedias_plus <- function (datos,k,N){

 n<- nrow(datos)  # Cuántas observaciones tenemos
 index <- sample(n, size=1)
 centroide <- datos[index,] # Sorteo para elegir el primer centro

 A<- matrix(NA,nrow=nrow(datos),ncol=ncol(datos))
 H<- matrix(NA,nrow=k,ncol=ncol(datos))
 dx2<- matrix(NA,nrow=nrow(datos),ncol=1)

 #escribimos los datos en una matriz en lugar de trabajar con dataframes:
 H[1,] <- unlist(as.vector(centroide))
 distancia <- matrix(NA,nrow=n, ncol=k) #distancias punto-centroide

 for (i in 1:n){A[i,]<-unlist(as.vector(datos[i,])) }

 for (i in 2:k){    #Para los k-1 centros restantes
   todos<-rbind(A,H)
   for(s in 1:n){
     for(j in 1:k-1){
```

```
      distancia[s,j]<- dist(todos[c(s,n+j),])
      }
    dx2[s]<- min(distancia[s,],na.rm=TRUE)^2  # Nos quedamos con la mínima
    }
    denominador <- sum(dx2^2)
    ponderacion<- dx2/denominador
    # Muestreo con distribución personalizada:
    nuevo_centro<- sample(n, size=1, prob=ponderacion)
   H[i,]<- A[nuevo_centro,] # Tomamos nuestro siguiente centro
  }
  return(H)
}
```

Tras ejecutar esta función y conseguir los centros, símplemente deberíamos utilizar *kmeans* o la versión que habíamos elaborado para conseguir aplicar el procedimiento entero. En el ejemplo anterior, si utilizamos la función que acabamos de presentar para elegir los centros (Triángulos negros en el gráfico de la izquierda), observamos como existe una tendencia muy favorable a coger centros de grupos alejados y, con ello, mejorar claramente el resultado:

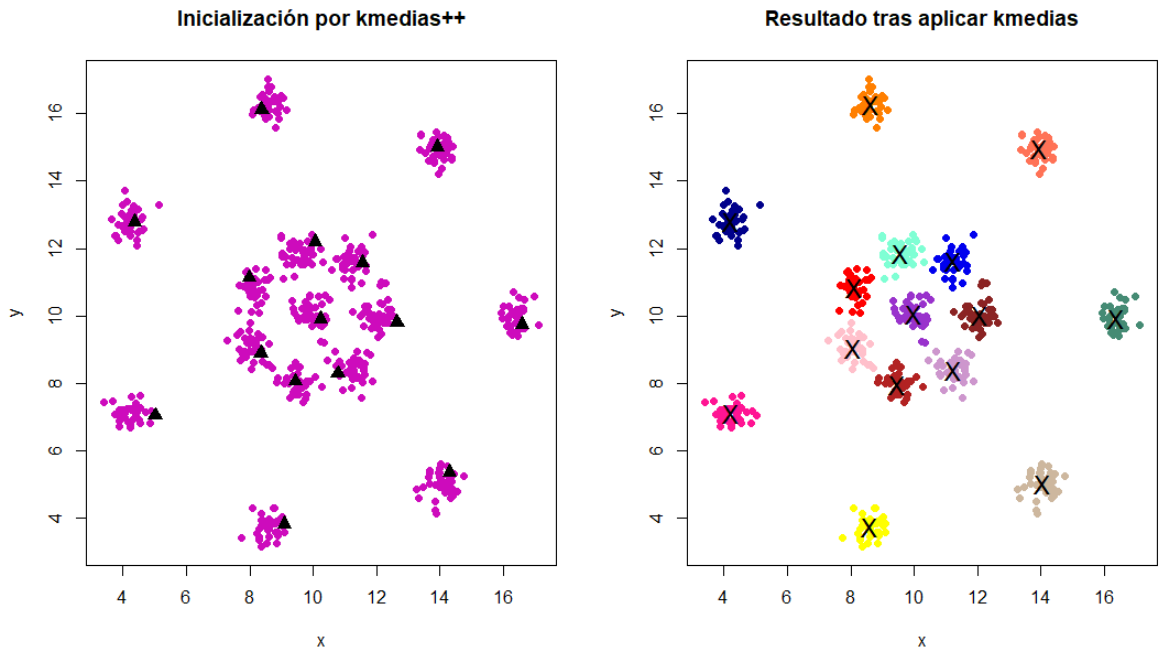


Figura 26

De nuevo, en R existen funciones encargadas de realizar el procedimiento de k medias++. En este caso, será necesario instalar la librería **flexclust**, en la cual encontramos la función *kcca*. Su

cometido es, como veníamos haciendo, hallar los centroides de un conjunto de datos y particionarlo en clusters. Sin embargo, podemos añadir la instrucción $control = list(initcent = "kmeanspp")$ con el fin de que los centros iniciales sean escogidos con k medias++. Esto será de gran ayuda para mejorar la agrupación en muchos conjuntos de datos. Al igual que con el comando *kmeans* de R ya utilizado, podemos obtener la agrupación de los datos en forma de vector donde encontramos el número de cluster al que ha sido asignado cada dato, además del número de iteraciones, el conjunto de centroides finales, distancia inter-cluster e intra-cluster... Si utilizamos la función *kcca* directamente en el ejemplo anterior, logramos resultados similares:

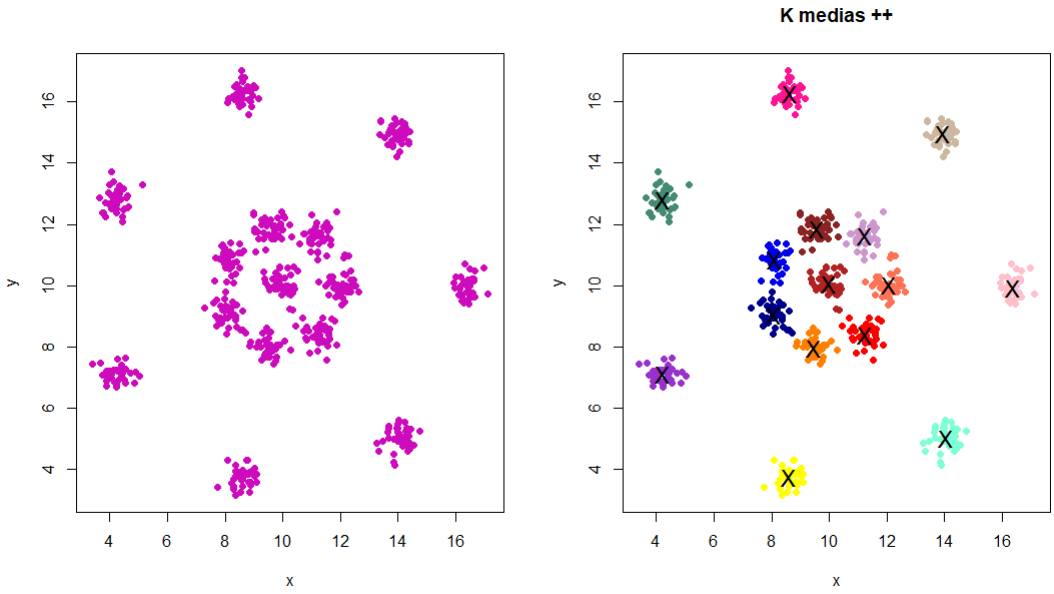


Figura 27: Mejora notable de la agrupación gracias a la actuación de k medias++

Dado que las mejoras son claramente palpables, es lógica la aspiración de estudiar el marco teórico de k medias++ e indagar un poco más en sus características.

5.2.1. Marco teórico de k medias++

Esta sección y las dos siguientes son una versión adaptada del trabajo original de [6]. Se han justificado detalladamente los argumentos correspondientes.

Seguimos trabajando con el marco teórico descrito anteriormente, con la salvedad de que ahora estaremos interesados en trabajar con subconjuntos A de la muestra: Sea k un entero positivo, $\mathcal{X} = \{x_1, .., x_n\} \in \mathbb{R}^d$ un conjunto de n puntos, $A \subset \mathcal{X}$ y $H = \{h_1, \ldots, h_k\} \subset \mathbb{R}^d$ un conjunto de k elementos de $\mathbb{R}^d$ que minimizan la función

$$W^k(H, A) = \sum_{x \in A} \min_{h \in H} \|x - h\|^2 \tag{11}$$

Sea $\mathscr{C} = \{C_1, \ldots, C_k\}$ la partición de $\mathcal{X}$ inducida por H. Para facilitar los resultados, hemos supuesto que la medida de disimilaridad utilizada es la norma euclidea al cuadrado en lugar de una función ϕ como hacíamos anteriormente. Una vez demostrados los resultados, esbozaremos el camino a seguir si queremos llegar a las mismas conclusiones con otra familia de funciones de disimilaridad.

Dada una instancia del problema de k medias, denotamos por H_{OPT} el conjunto de centros óptimos, $\mathscr{C}_{OPT}$ la partición óptima en clusters y $W^k_{OPT}(A)$ el valor óptimo del potencial.

El algoritmo de k medias++ que habíamos descrito anteriormente se reescribiría de este modo con los nuevos elementos introducidos:

1. Se elige h_1 de manera uniforme en $\mathcal{X}$
2. Se escoge el siguiente centroide inicial h_i con probabilidad
$$P(Y = h_i) = \frac{D(h_i)^2}{\sum_{x \in \mathcal{X}} D(x)^2}$$
Donde $D(h_i) = \text{mín}_{l=1,..,i-1}\, d(h_l, h_i)$, distancia de h_i al centroide más cercano. Denominamos a este peso $\mathcal{D}^2$.
3. Repetimos (2) hasta haber elegido k centros.
4. Para cada $i = \{1, .., k\}$, $C_i \in \mathscr{C}$ será el conjunto de puntos que están más cerca de h_i que de h_j para $i \neq j$.
5. Para cada $i = \{1, .., k\}$, recalculabamos el centroide h_i asignándole el valor el centro de masa del cluster, es decir
$$h_i := \frac{1}{|C_i|} \sum_{x \in C_i} x$$
6. Repetimos el proceso hasta que $\mathcal{C}$ y $W^k(H)$ permanecen estables.

5.2.2. Una cota superior para el problema

Nuestro objetivo será demostrar el siguiente resultado que nos permitirá tener un control sobre el valor del potencial que tratamos de minimizar en el problema de k medias.

Teorema 7. *Si $\mathscr{C}$ y H se construyen utilizando el procedimiento de k medias++, la correspondiente función potencial $W^k(H)$ verifica que $\mathbb{E}(W^k(H)) \leq 8(\ln k + 2)W^k_{OPT}$*

Para poder llegar a esta conclusión, demostraremos unos lemas previos. El primero nos dará una relación entre las distancias de un conjunto de puntos y su centro de masa.

Lema 4. *Sea S un conjunto de puntos cuya media se denota por $c(S)$. Sea z un punto arbitrario. Entonces*
$$\sum_{x \in S} \|x - z\|^2 - \sum_{x \in S} \|x - c(S)\|^2 = |S| \cdot \|c(S) - z\|^2$$

Demostración

Sabemos que $c(S)$ es el promedio, por lo que se calcula como $c(S) = \frac{1}{|S|}\sum_{x\in S} x$. Utilizando la identidad notable $\|x\|^2 - \|y\|^2 = \langle x-y, x+y\rangle$ y efectuando operaciones, podemos reescribir la expresión y llegar al resultado deseado.

□

Observación 4. *Sea H el conjunto de centros. La probabilidad de escoger un centro en el cluster $A \in \mathscr{C}_{OPT}$ es $\frac{W^k(H,A)}{W^k(H)}$.*

Esto se debe a que la probabilidad de escoger un punto $a \in A$ es

$$\frac{D(a)^2}{\sum_{x\in A} D(x)^2} = \frac{\text{mín}_{h\in H}\|a-h\|^2}{\sum_{x\in A}\text{mín}_{h\in H}\|x-h\|^2} = \frac{\text{mín}_{h\in H}\|a-h\|^2}{W^k(H)}$$

Y, sumando para todos los posibles valores de a, obtenemos que la probabilidad de escoger un centro en A es

$$\sum_{a\in A}\frac{\text{mín}_{h\in H}\|a-h\|^2}{\sum_{x\in A}\text{mín}_{h\in H}\|x-h\|^2} = \frac{\sum_{a\in A}\text{mín}_{h\in H}\|a-h\|^2}{W^k(H)} = \frac{W^k(H,A)}{W^k(H)}$$

A continuación, presentaremos un lema que nos permitirá acotar el valor medio de $W^1(\{h\})$ cuando añadimos el primer centro.

Lema 5. *Sea A un cluster arbitrario de $\mathscr{C}_{OPT}$ y Z la variable aleatoria que elige un cluster del que tomar un centro. Sea H el conjunto de centros de cluster con un solo centro elegido uniformemente en A sabiendo que $Z = A$, es decir, $H = \{a_0\}$ con $a_0 \in A$. Entonces $\mathbb{E}(W^1(H)|Z = A) = 2 \cdot W^1_{OPT}(A)$*

Demostración

Sea U la variable aleatoria que elige como centro $a_0 \in A$ uniformemente sabiendo que $Z = A$. Esto es, $P(U = a_0|Z = A) = \frac{1}{|A|}$. De este modo

$$\mathbb{E}(W^1(H,A)|Z=A) = \mathbb{E}_u(\mathbb{E}(W^1(H,A)|Z=A)|U) = \sum_{a_0\in A} P(U=a_0|Z=A)\cdot\mathbb{E}(W^1(H,A)|U=a_0) =$$

$$= \sum_{a_0\in A}\frac{1}{|A|}\mathbb{E}\left(\sum_{a\in A}\min_{h\in H}\|a-h\|^2|U=a_0\right) = \sum_{a_0\in A}\frac{1}{|A|}\mathbb{E}\left(\sum_{a\in A}\|a-c\|^2|\ \ c=a_0\right)$$

Dado que ya hemos elegido $c = a_0$ en la última igualdad, el sumatorio toma un valor concreto y no tiene componente aleatorio. Escribimos entonces

$$\mathbb{E}(W^1(H,A)|Z=A) = \sum_{a_0\in A}\frac{1}{|A|}\sum_{a\in A}\|a-a_0\|^2$$

Sea $h(A)$ el verdadero centro de cluster óptimo para A, es decir, $h(A) \in H_{OPT}$. Al aplicar el lema anterior, llegamos a que

$$\sum_{a_0\in A}\frac{1}{|A|}\cdot\sum_{a\in A}\|a-a_0\|^2 = \sum_{a_0\in A}\frac{1}{|A|}\left(\sum_{a\in A}\|a-h(A)\|^2 + |A|\cdot\|a_0-c(A)\|^2\right) =$$

$$= \sum_{a_0 \in A} \|a_0 - c(A)\|^2 + \sum_{a \in A} \frac{1}{|A|} \sum_{a_0 \in A} \|a - h(A)\|^2$$

El último sumatorio en $a_0 \in A$ tiene un argumento que no depende de a_0, por lo tanto lo sumamos simplemente $|A|$ veces:

$$\sum_{a_0 \in A} \|a_0 - c(A)\|^2 + \sum_{a \in A} \frac{1}{|A|} \sum_{a_0 \in A} \|a - h(A)\|^2 = \sum_{a_0 \in A} \|a_0 - c(A)\|^2 + \sum_{a \in A} \|a - h(A)\|^2 = 2 \cdot \sum_{a \in A} \|a - h(A)\|^2$$

Como $h(A)$ es el centro de A que optimiza $W^1(H, A)$, llegamos a la igualdad deseada:

$$\mathbb{E}(W^1(H, A)|Z = A) = 2 \cdot \sum_{a \in A} \|a - h(A)\|^2 = 2 \cdot W^1_{OPT}(A)$$

□

En el siguiente lema probaremos un resultado análogo cuando añadimos otro centro más a H.

Lema 6. *Sea A un cluster arbitrario de $\mathscr{C}_{OPT}$ y Z la variable aleatoria que elige el cluster del que tomamos el nuevo centro. Sea $\mathscr{C}$ una partición arbitraria de $\mathcal{X}$. Si añadimos un centro $a_0 \in A$ a $H = \{h_1, .., h_s\}$ con peso $\mathcal{D}^2$, entonces $\mathbb{E}(W^{s+1}(H \cup \{a_0\}, A)|Z = A) \leq 8 \cdot W^s_{OPT}(A)$.*

Demostración

Sea V la variable aleatoria que elige como centro $a_0 \in A$ con ponderación $\mathcal{D}^2$ sabiendo que $Z = A$. Esto es, $P(V = a_0|Z = A) = \frac{D(a_0)^2}{\sum_{a \in A} D(a)^2}$. Sea $H' = H \cup \{a_0\}$. De esta manera

$$\mathbb{E}(W^{s+1}(H', A)|Z = A) = \mathbb{E}_u(\mathbb{E}(W^{s+1}(H', A)|Z = A)|V) = \sum_{a_0 \in A} P(V = a_0|Z = A) \cdot \mathbb{E}(W^{s+1}(H', A)|V = a_0)$$

$$= \sum_{a_0 \in A} \frac{D(a_0)^2}{\sum_{a \in A} D(a)^2} \cdot \mathbb{E}\left(\sum_{a \in A} \min_{h \in H \cup \{c\}} \|a - h\|^2 | c = a_0\right)$$

Dado que suponemos que los centros elegidos anteriormente están fijos, el valor dentro de la esperanza no es aleatorio sino un valor concreto. Dado que o bien asociamos a al nuevo centro de cluster a_0 o bien a uno de los ya elegidos, la expresión anterior se reescribe

$$\mathbb{E}(W^{s+1}(H', A)|Z = A) = \sum_{a_0 \in A} \frac{D(a_0)^2}{\sum_{a \in A} D(a)^2} \sum_{a \in A} \min\{D(a), \|a - a_0\|^2\}$$

Sabemos por la desigualdad triangular que $D(a_0) = D(a_0 - a + a) \leq D(a) + \|a - a_0\|$. Utilizando la proposicion (7) del Anexo, podemos afirmar que

$$D(a_0)^2 \leq 2D(a)^2 + 2\|a - a_0\|^2$$

Por último, si sumamos en $a \in A$ y lo aplicamos a la expresión anterior:

$$\sum_{a_0 \in A} \frac{1}{|A|} \frac{\sum_{a \in A} D(a_0)^2}{\sum_{a \in A} D(a)^2} \sum_{a \in A} \min\{D(a), \|a - a_0\|^2\} \leq$$

$$\leq \frac{2}{|A|}\sum_{a_0\in A}\frac{\sum_{a\in A}D(a)^2}{\sum_{a\in A}D(a)^2}\sum_{a\in A}\text{mín}\{D(a),\|a-a_0\|^2\}+\frac{2}{|A|}\sum_{a_0\in A}\frac{\sum_{a\in A}\|a-a_0\|^2}{\sum_{a\in A}D(a)^2}\sum_{a\in A}\text{mín}\{D(a),\|a-a_0\|^2\}=$$

$$=\frac{2}{|A|}\sum_{a_0\in A}\sum_{a\in A}\text{mín}\{D(a),\|a-a_0\|^2\}+\frac{2}{|A|}\sum_{a_0\in A}\frac{\sum_{a\in A}\|a-a_0\|^2}{\sum_{a\in A}D(a)^2}\sum_{a\in A}\text{mín}\{D(a),\|a-a_0\|^2\}$$

Si en el primer sumando acotamos $\text{mín}\{D(a),\|a-a_0\|^2\}\leq\|a-a_0\|^2$, y en el segundo $\text{mín}\{D(a),\|a-a_0\|^2\}\leq D(a)$, ambos dos sumandos quedan iguales y llegamos a que

$$\mathbb{E}(W^{s+1}(H',A)|Z=A)\leq\frac{4}{|A|}\sum_{a_0\in A}\sum_{a\in A}\|a-a_0\|^2=8\cdot W^s_{OPT}(A)$$

Siendo la última igualdad consecuencia del lema (5). □

Tras haber estudiado qué ocurre con el valor medio del potencial al escoger el primer centro o al añadir otro cento de un cluster A presente en la solución óptima, veremos qué ocurre en el caso general en el que añadimos varios centros simultáneamente.

Lema 7. *Sea $\mathscr{C}$ una partición arbitraria de $\mathcal{X}$ en clusters y H el conjunto con los correspondientes centroides. Si consideramos un cluster $A\in\mathscr{C}_{OPT}$, decimos que A es **cubierto** si hemos elegido algún centro en A, y decimos que es **no cubierto** si por el contrario ningúno de los centros de cluster ya elegidos pertenece a A. Sea $u>0$ el número de clusters de $\mathscr{C}_{OPT}$ cubiertos y denotamos por $\mathcal{X}_u$ el conjunto de puntos de estos clusters. Del mismo modo, sea $\mathcal{X}_c=\mathcal{X}-\mathcal{X}_u$ el conjunto de puntos correspondientes a los clusters cubiertos. Supongamos que añadimos $t\leq u$ centros $\{c_1,..,c_t\}$ aleatoriamente a $H=\{h_1,..,h_s\}$ utilizando la ponderación $\mathcal{D}^2$. Si $H'=H\cup\{c_1,..,c_t\}$ son los centros resultantes tras este proceso y $W^{s+t}(H')$ su potencial, se tiene que*

$$\mathbb{E}(W^{s+t}(H'))\leq(W^s(H,\mathcal{X}_c)+8\cdot W^s_{OPT}(\mathcal{X}_u))(1+H_t)+\frac{u-t}{u}\cdot W^s(H,\mathcal{X}_u)\tag{12}$$

Donde $H_t=1+\frac{1}{2}+\frac{1}{3}+..+\frac{1}{t}$.

Demostración

Para probar este resultado, realizaremos un proceso inductivo. Veremos que si se verifica la desigualdad para $(t-1,u-1)$ y para $(t-1,u)$, entonces se tiene el resultado para (t,u).
En primer lugar, comprobaremos que se verifica para los primeros valores de u y de t:

- Cuando $t=0$ y $u>0$

  Si $t=0$, buscamos probar que $\mathbb{E}(W^s(H')\leq W^s(H,\mathcal{X}_c)+8\cdot W^s_{OPT}(\mathcal{X}_u)+W^s(H,\mathcal{X}_u)$. Dado que $t=0$, no añadimos ningún centro nuevo a H, por lo que $H'=H$ y con ello se mantiene $W^s(H)$. Estamos considerando que $x_1,..,x_n$ son los resultados de un sorteo en particular y con ello están fijos, por lo que $W^s(H)$ también lo está. Es decir, $\mathbb{E}(W^s(H))=W^s(H)$. De este modo, tendríamos que

  $$\mathbb{E}(W^s(H))=W^s(H,\mathcal{X}_u)+W^s(H,\mathcal{X}_c)\leq W^s(H,\mathcal{X}_u)+W^s(H,\mathcal{X}_c)+8\cdot W^s_{OPT}(\mathcal{X}_u)$$

  con lo que queda demostrado

- Cuando $u = t = 1$

  En este caso, necesitamos comprobar que $\mathbb{E}(W^{s+1}(H')) \leq 2 \cdot (W^s(H, \mathcal{X}_c) + 8 \cdot W^s_{OPT}(H, \mathcal{X}_u))$. Dado que $t = 1$, se ha seleccionado solo un nuevo centro. Hemos podido escogerlo en un cluster ya cubierto o en uno que aún no tuviese centroides. Denotamos por h' el nuevo centro que añadimos a H.

  Como hemos comentado antes, la probabilidad de escoger un centro en el conjunto $\mathcal{X}_u$ es

  $$P_u = \frac{W^{s+1}(H', \mathcal{X}_u)}{W^{s+1}(H')}$$

  Análogamente podemos calcular la probabilidad de escoger un centro en el conjunto $\mathcal{X}_c$, P_c.

  Si asociamos con 1 el suceso "Tomar el nuevo centro en un cluster no cubierto" y por 0 el suceso "Tomar el nuevo centro en un cluster ya cubierto", podemos definir la variable aleatoria Z como aquella que toma el valor 0 con probabilidad P_c y el valor 1 con probabilidad P_u.

  Podemos reescribir entonces

  $$\mathbb{E}(W^{s+1}(H')) = \mathbb{E}_Z(\mathbb{E}(W^{s+1}(H')|Z)) = P(Z=0)\cdot\mathbb{E}(W^{s+1}(H')|Z=0) + P(Z=1)\cdot\mathbb{E}(W^{s+1}(H')|Z=1)$$

  a) Si elegimos h' en un cluster descubierto, resulta

  $$\mathbb{E}(W^{s+1}(H)|Z=1) = \mathbb{E}\left(\sum_{x\in\mathcal{X}} \min_{h\in H'} \|x-h\|^2 | Z=1\right) = \mathbb{E}\left(\sum_{x\in\mathcal{X}} \min_{h\in H\cup\{h'\}} \|x-h\|^2 | Z=1\right) =$$

  $$= \mathbb{E}\left(\sum_{x\in\mathcal{X}_c} \min_{h\in H\cup\{h'\}} \|x-h\|^2 | Z=1\right) + \mathbb{E}\left(\sum_{x\in\mathcal{X}_u} \min_{h\in H\cup\{h'\}} \|x-h\|^2 | Z=1\right)$$

  Analizamos qué ocurre con estos dos sumandos. En primero de ellos, dado que estamos cogiendo $x \in \mathcal{X}_c$, los puntos estarán más cerca de un centro de cluster ya elegido y añadir h' no cambia el potencial. De este modo, seguimos obteniendo el potencial de los puntos pertenecientes a clusters cubiertos:

  $$\mathbb{E}\left(\sum_{x\in\mathcal{X}_c} \min_{h\in H\cup\{h'\}} \|x-h\|^2 | Z=1\right) = \mathbb{E}(W^s(H, \mathcal{X}_c)|Z=1) = W^s(H, \mathcal{X}_c)$$

  Ya que $W^s(H, \mathcal{X}_c)$ depende de los centros que teníamos en H y no de dónde cojamos h'. Respecto al segundo sumando, habíamos probado en el lema anterior qué ocurría al tomar centros en clusters que no estaban cubiertos y añadirlos a H, donde ya hay centroides. Por lo tanto

  $$\mathbb{E}\left(\sum_{x\in\mathcal{X}_u} \min_{h\in H\cup\{h'\}} \|x-h\|^2 | Z=1\right) \leq 8 \cdot W^s_{OPT}(\mathcal{X}_u)$$

Así, tendríamos que en el caso de coger el centro en un cluster sin cubrir, el valor medio del nuevo potencial estaría acotado de la siguiente manera:

$$\mathbb{E}(W^{s+1}(H')|Z=1) \leq 8 \cdot W^s_{OPT}(\mathcal{X}_u) + W^s(H, \mathcal{X}_c)$$

b) Si escogemos h' cogemos en un cluster ya cubierto, dado que $H \subset H'$, tendríamos que

$$\mathbb{E}(W^s(H')|Z=0) \leq \mathbb{E}\left(\sum_{x\in\mathcal{X}} \min_{h\in H} \|x-h\|^2 | Z=0\right) = \mathbb{E}(W^s(H)|Z=0) = W^s(H)$$

Ya que quitar un centro puede dejar W igual o empeorarla. De este modo, si valoramos ambos dos casos, el valor medio del nuevo potencial queda acotado como sigue

$$\mathbb{E}(W^{s+1}(H') = P(Z=0) \cdot \mathbb{E}(W^{s+1}(H')|Z=0) + P(Z=1)\mathbb{E} \cdot (W^{s+1}(H')|Z=1) \leq$$

$$\leq P_c \cdot W^s(H) + P_u(8 \cdot W^s_{OPT}(\mathcal{X}_u) + W^s(H, \mathcal{X}_c)) \leq W^s(H, \mathcal{X}_c) + (8 \cdot W^s_{OPT}(\mathcal{X}_u) + W^s(H, \mathcal{X}_c))$$

Por lo que queda probada la desigualdad.

Con esto, estamos en disposición de probar el paso de inducción. Supongamos que se cumple la desiguadad para $(t-1, u)$ y para $(t-1, u-1)$, y veamos que es cierto para (t, u).

Tenemos u clusters sin cubrir y escogemos t centros. Sean $A_1, .., A_u$ los clusters sin cubrir, es decir $\mathcal{X}_u = \cup_{i=1}^u A_i$. De manera similar al paso anterior, denotamos con 0 el suceso "Tomar el primero de los t centros en un cluster cubierto" y con i para $i = 1, .., u$ el suceso "Tomar el primero de los t centros en el cluster sin cubrir A_i". Definimos entonces la variable aleatoria Z como aquella que toma el valor 0 con probabilidad P_c y el valor i con probabilidad $P_{A_i} = \frac{W^{s+t}(H', A_i)}{W^{s+t}(H')}$.

De este modo, de manera análoga, escribimos

$$\mathbb{E}(W^{s+t}(H')) = \mathbb{E}_Z(\mathbb{E}(W^{s+t}(H'))|Z)) = P(Z=0)\cdot\mathbb{E}(W^{s+t}(H')|Z=0) + \sum_{i=1}^{u} P(Z=i)\cdot\mathbb{E}(W^{s+t}(H')|Z=i) =$$

$$= P_c\cdot\mathbb{E}(W^{s+t}(H')|Z=0) + \sum_{i=1}^{u} P_{A_i}\cdot\mathbb{E}(W^{s+t}(H')|Z=i) = P_c\mathbb{E}(W^{s+t}(H')|Z=0) + P_u \sum_{i=1}^{u} \mathbb{E}(W^{s+t}(H')|Z=i)$$

Nuestra misión será acotar $\mathbb{E}(W^{s+t}(H')|Z=i)$ para $i = 0, .., u$.

- Supongamos que hemos tomado nuestro primer centro, h', en un cluster ya cubierto (Esto quiere decir que, $Z = i$ con $i = 1, .., u$). Denotamos por $H'' = H - \{h'\}$.
  Resulta que

$$W^{s+t}(H') = \sum_{x\in\mathcal{X}} \min_{h\in H'} \|x-h\|^2 \leq \sum_{x\in\mathcal{X}} \min_{h\in H'-\{h'\}} \|x-h\|^2 = W^{s+t-1}(H'')$$

De este modo, se tiene que

$$\mathbb{E}(W^{s+t}(H')|Z=0) \leq \mathbb{E}(W^{s+t-1}(H'')|Z=0)$$

Pero hemos visto que $W^{s+t-1}(H'')$ no depende de h', por lo que

$$\mathbb{E}(W^{s+t-1}(H'')|Z=0) = \mathbb{E}(W^{s+t-1}(H'')|\text{“}Elijo\ \ h'\ \ en\ \ \mathcal{X}_c\text{ ”}) = \mathbb{E}(W^{s+t-1}(H''))$$

Pero $W^{s+t-1}(H'')$ es el potencial correspondiente a añadir $t-1$ centro y tener u clusters sin cubrir. Por lo tanto, nos encontramos en la situación $(t-1,u)$ y podemos aplicar la hipótesis de inducción:

$$\mathbb{E}(W^{s+t}(H')|Z=0) \leq (W^s(H,\mathcal{X}_c) + 8W^s_{OPT}(\mathcal{X}_u))(1+H_{t-1}) + \frac{u-t+1}{u}\cdot W^s(H,\mathcal{X}_u)$$

- Supongamos que hemos tomado nuestro primer centro en un cluster no cubierto, digamos A_i (Esto es, $Z=i$ con $i=1,..,u$). Sea Y la variable aleatoria que elige un punto $a\in A_i$ con probabilidad p_a como primer centro, es decir, $P(Y=a|Z=i)$. Por lo tanto, $c_1=a$ y podemos expresar el valor medio de $W^{s+t}(H')$ sabiendo que elegimos el primer centro en el conjunto A_i, como

$$\mathbb{E}(W^{s+t}(H')|Z=i) = \mathbb{E}_Y(\mathbb{E}((W^{s+t}(H')|Y)|Z=i)) = \sum_{a\in A_i} p_a\cdot\mathbb{E}(W^{s+t}(H')|Y=a)$$

Estudiamos cómo acotar $\mathbb{E}(W^{s+t}(H')|Y=a)$. Denotamos por H' el conjunto de centros que ya teníamos añadiendo los t nuevos centros, $H_t = H\cup\{a,c_2,..,c_t\}$ junto con los t nuevos centros añadidos. Es decir, en este caso el nuevo potencial W' es

$$W^{s+t} = \sum_{x\in\mathcal{X}} \min_{h\in H\cup\{a,c_2,..,c_t\}} \|x-h\|^2$$

Para poder aplicar la hipótesis de inducción, consideramos A_i cubierto y añadimos a al conjunto de centros de cluster fijos, es decir $H''=H\cup\{a\}$. Por lo tanto, $H'=H''\cup\{c_2,..,c_t\}$. Denotamos $W^{s+t}_a(H') = \sum_{x\in A_i}\min_{h\in H''\cup\{c_2,..,c_t\}}\|x-h\|^2$, es decir, la aportación de A_i al potencial tras haber añadido a al conjunto de centros. De este modo, si además partimos la suma en A_i, $\mathcal{X}_c$ y $\mathcal{X}_u$, tenemos

$$W^{s+t}(H') = \sum_{x\in\mathcal{X}_c}\min_{h\in H''\cup\{c_2,..,c_t\}}\|x-h\|^2 + \sum_{x\in A_i}\min_{h\in H''\cup\{c_2,..,c_t\}}\|x-h\|^2 + \sum_{x\in\mathcal{X}_u-A}\min_{h\in H''\cup\{c_2,..,c_t\}}\|x-h\|^2$$

$$= W^{(s+1)+(t-1)}(H',\mathcal{X}_c) + W^{(s+1)+(t-1)}_a(H') + W^{(s+1)+(t-1)}(H',\mathcal{X}_u - A_i)$$

De esta manera, resulta que el número centros fijos es $s+1$, de clusters sin cubrir es $u-1$ y el número de centros que hemos añadido es $t-1$, por lo que podemos aplicar la hipótesis de inducción. Si lo incluimos en la expresión anterior, tendríamos que

$$\sum_{a\in A_i} p_a\cdot\mathbb{E}(W^{s+t}(H')|Y=a) \leq \sum_{a\in A_i} p_a\cdot[(W^s(H,\mathcal{X}_c) + W^{s+t}_a(H') + 8W^s_{OPT}(\mathcal{X}_u)] -$$

$$- \sum_{a \in A_i} p_a[8 \cdot W^s_{OPT}(A)(1 + H_{t-1})] + \cdot[\frac{u-t}{u-1} \cdot (W^s(H, \mathcal{X}_u) - W^s(H, A_i))]$$

Al aplicar la propiedad distributiva y utilizar que $\sum_{a \in A_i} p_a = 1$, solo falta ver qué ocurre con el término $\sum_{a \in A_i} p_a W_a^{s+t}(H')$. Pero resulta que es exactamente

$$\sum_{a \in A_i} p_a W^s_a(H) = \sum_{a \in A_i} p_a \cdot \mathbb{E}(W_a^{s+t}(H')) = \sum_{a \in A_i} P(Y = a | Z = i) \cdot \mathbb{E}(W_a^{s+t}(H') | Y = a)$$

$$= \mathbb{E}_Y(\mathbb{E}(W_a^{s+t}(H')|Y)|Z = i) = \mathbb{E}(W(A_i W_a^{s+t}(H')|Z = i) \leq 8 W^s_{OPT}(A_i)$$

Siendo la última desigualdad consecuencia del lema anterior. Así, si recopilamos las expresiones, se tiene

$$\sum_{a \in A_i} p_a \cdot \mathbb{E}(W^{s+t}(H')|Y = a) \leq (W^s(H, \mathcal{X}_c) + 8 \cdot W^s_{OPT}(\mathcal{X}_u))(1 + H_{t-1}) + \frac{u-t}{u-1} \cdot (W^s(H, \mathcal{X}_u) - W^s(H, A_i))$$

Así, en la expresión anterior, se tiene

$$\sum_{i=1}^{u} \frac{W^s(H, A_i)}{W^s(H)} \cdot \mathbb{E}(W^{s+t}(H')|Z = i) \leq \sum_{i=1}^{u} \frac{W^s(H, A_i)}{W^s(H)} (W^s(H, \mathcal{X}_c) + 8 \cdot W^s_{OPT}(\mathcal{X}_u))(1 + H_{t-1}) +$$

$$+ \sum_{i=1}^{u} \frac{W^s(H, A_i)}{W^s(H)} \frac{u-t}{u-1} \cdot (W^s(H, \mathcal{X}_u) - W^s(H, A_i))$$

Dado que la suma de $W^s(H, A_i)$ en todos los clusters A_i no cubiertos es $W^s(H, \mathcal{X}_u)$, la línea anterior se escribe equivalentemente como

$$\frac{W^s(H, \mathcal{X}_u)}{W^s(H)} (W^s(H, \mathcal{X}_c) + 8 \cdot W^s_{OPT}(\mathcal{X}_u))(1 + H_{t-1}) + \frac{u-t}{u-1} \frac{1}{W} \cdot (W^s(H, \mathcal{X}_u)^2 - \sum_{i=1}^{u} W^s(H, A_i)^2)$$

Por la proposición (7), se tiene que $\sum_{i=1}^{u} W(A_i)^2 \geq \frac{1}{u} W(\mathcal{X}_u)^2$. Aplicada a esta última expresión, tendríamos que esta sería igual a

$$= \frac{W^s(H, \mathcal{X}_u)}{W^s(H)} (W^s(H, \mathcal{X}_c) + 8 \cdot W^s_{OPT}(\mathcal{X}_u))(1 + H_{t-1}) + \frac{u-t}{u-1} \frac{1}{W^s(H)} \left(\frac{(u-1) W^s(H, \mathcal{X}_u)^2}{u} \right)$$

Simplificando el denominador y sacando factor común a $\frac{W^s(H, \mathcal{X}_u)}{W^s(H)}$, lo podemos reescribir finalmente así

$$\frac{W^s(H, \mathcal{X}_u)}{W^s(H)} \left(W^s(H, \mathcal{X}_c) + 8 \cdot W^s_{OPT}(\mathcal{X}_u))(1 + H_{t-1}) + \frac{u-t}{u} W^s(H, \mathcal{X}_u) \right)$$

Hemos conseguido acotar por lo tanto, suponiendo que el primer centro se escogía en un cluster cubierto, la expresión

$$\frac{W^s(H,\mathcal{X}_c)}{W^s(H)}\cdot\mathbb{E}(W^{s+t}(H')|Z=0) \leq \frac{W^s(H,\mathcal{X}_c)}{W^(H)}(W^s(H,\mathcal{X}_c)+8\cdot W^s_{OPT}(\mathcal{X}_u))(1+H_{t-1})+\frac{u-t+1}{u}\cdot W^s(H,\mathcal{X}_u)$$

Y si por el contrario suponemos que el primer centro se escoge en un cluster A_i no cubierto, habíamos visto que

$$\sum_{i=1}^{u}\frac{W^{s+t}(H',A_i)}{W}\cdot\mathbb{E}(W^{s+t}(H',|Z=i) \leq \frac{W^s(H,\mathcal{X}_u)}{W^s(H)}\left(W^s(H,\mathcal{X}_c)+8\cdot W^s_{OPT}(\mathcal{X}_u))(1+H_{t-1})+\frac{u-t}{u}W^s(H,\mathcal{X}_u)\right)$$

Utilizando que $P_u + P_c = 1$, agrupando términos y aplicando la propiedad distributiva, llegamos a

$$\mathbb{E}(W^{s+t}(H')) \leq \frac{u-t}{u}W^s(H,\mathcal{X}_u)+\frac{W^s(H,\mathcal{X}_c)}{W^s(H)}\left(\frac{W^s(H,\mathcal{X}_u)}{u}\right)$$

De nuevo, si nos centramos en nuestro objetivo principal, habríamos conseguido acotar el valor medio de $W^{s+t}(H')$ por

$$\mathbb{E}(W^{s+t}(H')) \leq (W^s(H,\mathcal{X}_c)+8\cdot W^s_{OPT}(\mathcal{X}_u))(1+H_{t-1})+\frac{u-t}{u}W^s(H,\mathcal{X}_u)+\frac{W^s(H,\mathcal{X}_c)}{W^s(H)}\frac{W^s(H,\mathcal{X}_u)}{u}$$

Pero resulta que

$$\frac{W^s(H,\mathcal{X}_c)}{W^s(H)}W^s(H,\mathcal{X}_u) \leq W^s(H,\mathcal{X}_c) \leq W^s(H,\mathcal{X}_c)+8\cdot W^s_{OPT}(\mathcal{X}_u)$$

Por lo que, utilizando que $\frac{1}{u} \leq \frac{1}{t}$, obtenemos el resultado deseado agrupando $\frac{1}{t}$ y H_{t-1}:

$$\mathbb{E}(W^{s+t}(H') \leq (W^s(H,\mathcal{X}_c)+8\cdot W^s_{OPT}(\mathcal{X}_u))(1+H_t)+\frac{u-t}{u}W^s(H,\mathcal{X}_u)$$

□

Gracias a este lema, podemos finalmente demostrar el teorema principal: El método de k medias++ es $O-(\log k)$ competitivo.

Teorema 8. *Si $\mathscr{C}$ se construye utilizando el procedimiento de k medias++, la correspondiente función potencial $W^k(H)$ verifica que $\mathbb{E}(W^k(H)) \leq 8(\ln k+2)W^k_{OPT}$*

Demostración

Supongamos que solo hemos escogido el primer centro de manera uniforme, de acuerdo con el primer paso de k medias++. Sea A el cluster en $\mathscr{C}_{OPT}$ en el cual hemos escogido el primer centro. En esta situación, tenemos $k-1$ clusters de $\mathscr{C}_{OPT}$ sin cubrir. Aplicamos el lema anterior para $t = u = k-1$. De este modo, como sabemos, la esperanza del nuevo potencial se acota del siguiente modo

$$\mathbb{E}(W^k(H')) \leq (W^k(H,\mathcal{X}_c)+8\cdot W^k_{OPT}(\mathcal{X}_u))(1+H_t) = (W^k(H,A)+8\cdot W^k_{OPT}(\mathcal{X}-A))(1+H_{k-1})$$
$$= (\mathbb{E}(W^k(H,A))+8\cdot W^k_{OPT}(\mathcal{X})-8\cdot W^k_{OPT}(A))$$

Utilizando el lema (5) , escribimos

$$\mathbb{E}(W^k(H') \leq (8 \cdot W^k_{OPT}(A) + 8 \cdot W^k_{OPT}(\mathcal{X}) - 8 \cdot W^k_{OPT}(A))(1 + H_{k-1}) = (8 \cdot W^k_{OPT}(\mathcal{X}))(1 + H_{k-1})$$

Como $H_{k-1} = 1 + \frac{1}{2} + \frac{1}{3} + .. + \frac{1}{k-1} \leq 1 + \ln k$, llegamos finalmente a

$$\mathbb{E}(W^k(H')) \leq (8 \cdot W^k_{OPT}(\mathcal{X}))(2 + \ln k)$$

Que es lo que queríamos probar. □

5.2.3. Una cota inferior para el problema

Acabamos de encontrar una cota superior para el valor medio del potencial aportado por $\{x_1, .., x_n\}$ en relación al potencial óptimo. Nuestra siguiente pregunta será si somos capaces de encontrar una cota inferior. Para ello, diseñaremos un problema particular en el que el valor medio del potencial se verá acotado inferiormente.
Nuestro problema será el siguiente: Vamos a construir $\mathcal{X} = \{x_1, .., x_n\}$ un conjunto de n puntos, contando con k un entero positivo y δ, Δ dos números positivos que verifican $n \gg k$ y $\Delta \gg \delta$. Seguimos los siguientes pasos

1. Elegimos en primer lugar $h_1, .., h_k$ centros de cluster tales que $\|h_i - h_j\|^2 = \Delta^2 - \left(\frac{n-k}{n}\right)\delta^2$ para $i \neq j$.
2. Para cada centro h_i, añadimos $\frac{n}{k}$ puntos, $x_{i,1}, .., x_{i,\frac{n}{k}}$, que formen un simplex regular de lado δ, centro h_i y radio $\sqrt{\frac{n-k}{2n}}$.
3. Si consideramos una dimensión suficientemente grande como para que los vectores sean ortogonales entre sí, tenemos que

$$\|x_{i,i'} - x_{j,j'}\| = \begin{cases} \delta & si\ i = j \\ \Delta & si\ i \neq j \end{cases}$$

Veremos como bajo estas hipótesis, el algoritmo de k medias++ no es mejor que $2(\ln k)-$ competitivo.

Observación 5. *El potencial óptimo de este problema es exactamente $W_{OPT} = \frac{n-k}{2}\delta^2$, ya que podemos calcularlo como*

$$W^k_{OPT} = \sum_{x \in \mathcal{X}} \min_{h \in H_{OPT}} \|x - h\|^2 = \sum_{x \in \mathcal{X}} \min_{i=1,..,k} \|x - h_i\|^2 = \sum_{x \in \mathcal{X}} \frac{n-k}{2n}\delta^2 = n \cdot \frac{n-k}{2n}\delta^2 = \frac{n-k}{2}\delta^2$$

Ya que los n puntos distan $\frac{n-k}{2n}\delta^2$ de su respectivo centro de cluster.

Lema 8. *Sea H un conjunto con $k - t \geq 1$ centroides asociados a $\mathcal{X}$. Sea $u > 0$ el número de clusters de $\mathscr{C}_{OPT}$ sin cubrir (No hemos elegido ningún punto de H en ellos). Supongamos que añadimos t centros $\{c_1, .., c_t\}$ aleatoriamente a $H = \{h_1, .., h_{k-t}\}$, con $H' = H \cup \{c_1, .., c_t\}$. Sea $W^{k-t}(H')$ el potencial de $\mathcal{X}$ tras este proceso.*
Denotamos por

$$\alpha = \frac{n - k^2}{n}, \quad \beta = \frac{\Delta^2 - 2k\delta^2}{\Delta^2}, \quad H'_u = \sum_{i=1}^{u} \frac{k-i}{ki}$$

Bajo estas condiciones, tenemos una cota inferior para la esperanza del potencial:

$$\mathbb{E}(W^{k-t}(H')) \geq \alpha^{t+1}\left(n\delta^2 \cdot (1 + H'_u) \cdot \beta + \left(\frac{n}{k}\Delta^2 - 2n\delta^2\right)(u - t)\right) \tag{13}$$

Demostración

Al igual que hacíamos con el lema de la sección anterior, probaremos el resultado por inducción. Para el caso $t = 0$, tenemos escogidos k centros, todos ellos entre los puntos de $\mathcal{X}$. Además, no añadimos más centros. De este modo, podemos escribir

$$\mathbb{E}(W^k(H)) = W^k(H) = \sum_{x\in\mathcal{X}} \min_{h\in H} \|x-h\|^2$$

Observemos que hemos seleccionado los centros h_i entre los puntos de $\mathcal{X}$. Esto quiere decir que existen k puntos $x_{(1)}, .., x_{(k)}$ tales que $h_i = x_{(i)}$. De este modo podemos reescribir el sumatorio como

$$\sum_{x\in\mathcal{X}} \min_{h\in H} \|x-h\|^2 = \sum_{x\in\mathcal{X}} \min_{i=1,..,k} \|x-x_{(i)}\|^2 = \sum_{x\in\mathcal{X}_u} \min_{i=1,..,k} \|x-x_{(i)}\|^2 + \sum_{x\in\mathcal{X}_c} \min_{i=1,..,k} \|x-x_{(i)}\|^2$$

Los puntos de $\mathcal{X}_u$ son aquellos pertenecientes a clusters no cubiertos, por lo que no hemos cogido ningún centro ahí. Es decir, $\|x-x_{(i)}\|^2 = \Delta^2$ si $x \in \mathcal{X}_u$ $\forall i = 1, .., k$. Como tenemos u clusters sin cubrir con $\frac{n}{k}$ puntos cada uno, resulta que $|\mathcal{X}_u| = u \cdot \frac{n}{k}$, por lo que la contribución de $\mathcal{X}_u$ al potencial es exactamente

$$\sum_{x\in\mathcal{X}_u} \min_{i=1,..,k} \|x-x_{(i)}\|^2 = u \cdot \frac{n}{k} \cdot \Delta^2$$

Por otro lado, los puntos en $\mathcal{X}_c$ son aquellos pertenecientes a clusters donde hemos colocado algún centro. Por ello, $\|x-x_{(i)}\|^2 = \delta^2$ si $x \in \mathcal{X}_c$ y elegimos el i apropiado. El número total de puntos en los clusters cubiertos es $|\mathcal{X}_c| = (k-u) \cdot \frac{n}{k}$, ya que tenemos $(k-u)$ clusters cubiertos con $\frac{n}{k}$ puntos cada uno. Sin embargo, como los centros $x_{(i)}$ están en $\mathcal{X}_c$ también y distan 0 de sí mismos, para calcular la contribución de $\mathcal{X}_c$ al potencial debemos restar el valor que hemos añadido al considerar estos k centros, es decir

$$\sum_{x\in\mathcal{X}_c} \min_{i=1,..,k} \|x-x_{(i)}\|^2 = (k-u) \cdot \frac{n}{k} \cdot \delta^2 - k\delta^2$$

Así, si recopilamos todo esto, podemos afirmar que

$$W^k(H) = u \cdot \frac{n}{k} \cdot \Delta^2 + (k-u) \cdot \frac{n}{k} \cdot \delta^2 - k\delta^2 = \left(n - u \cdot \frac{n}{k} - k\right)\delta^2 + u\frac{n}{k}\Delta^2$$

Si hemos supuesto que $u > 0$, se cumple que $k \geq u+1$, y encadenando desigualdades conseguimos acotar α:

$$\frac{n - u\cdot\frac{n}{k} - k}{n - \frac{n}{k}} \geq \frac{\frac{n}{k} - k}{\frac{n}{k}} = \frac{\frac{n-k^2}{k}}{\frac{n}{k}} = \frac{n-k^2}{n} = \alpha$$

De esta manera, como $\left(n - u\cdot\frac{n}{k} - k\right) \geq \alpha\left(n - \frac{n}{k}u\right)$, se tiene

$$W^k(H) = \left(n - u\cdot\frac{n}{k} - k\right)\delta^2 + u\frac{n}{k}\Delta^2 \geq \alpha\left(n - \frac{n}{k}u\right)\delta^2 + u\frac{n}{k}\Delta^2 \geq \alpha\left(n\delta^2\beta - \frac{n}{k}u\beta\delta^2 + u\frac{n}{k}\Delta^2\right)$$

Siendo esta última desigualdad consecuencia de $\alpha, \beta \leq 1$. Como sabemos que $n \cdot u \geq \frac{u}{k}u$, tenemos que $u\delta^2 n \geq \frac{n}{k}u\beta\delta^2$. Aplicando esta desigualdad y viendo que

$$H'_u = \sum_{i=1}^{u} \frac{k-i}{ki} = \frac{k-1}{k} + \frac{k-2}{2k} .. + \frac{k-u}{2u} \leq u\left(\frac{k-1}{k}\right) \leq u$$

Se tiene

$$\alpha\left(n\delta^2\beta + u\delta^2 n - 2u\delta^2 n + u\frac{n}{k}\Delta^2\right) \geq \alpha\left(n\delta^2\beta(1+H'_u) + \left(\frac{n}{k}\Delta^2 - 2\delta^2 n\right)u\right)$$

Y con ello, hemos llegado a la desigualdad deseada, comprobando que es cierta cuando $t = 0$:

$$\mathbb{E}(W^k(H)) \geq \alpha\left(n\delta^2\beta(1+H'_u) + \left(\frac{n}{k}\Delta^2 - 2\delta^2 n\right)u\right)$$

Procedemos entonces al paso de inducción. Supongamos que tenemos u clusters descubiertos y que añadimos t nuevos centros, por lo que tenemos $k-t$ centros en H. Sean $A_1, .., A_u$ los clusters sin cubrir, es decir $\mathcal{X}_u = \cup_{i=1}^{u} A_i$. De manera similar a la demostración del lema anterior, denotamos con 0 el suceso "Tomar el primero de los t centros en un cluster cubierto" y con i para $i = 1, .., u$ el suceso "Tomar el primero de los t centros en el cluster sin cubrir A_i". Definimos entonces la variable aleatoria Z como aquella que toma el valor 0 con probabilidad P_c y el valor i con probabilidad $\frac{W^{k-t}(H',A_i)}{W^{k-t}(H')} = P_{A_i}$.

De este modo, de manera análoga, escribimos

$$\mathbb{E}(W^{k-t}(H')) = \mathbb{E}_Z(\mathbb{E}(W^{k-t}(H')|Z)) = P(Z=0)\cdot\mathbb{E}(W^{k-t}(H')|Z=0) + \sum_{i=1}^{u} P(Z=i)\cdot\mathbb{E}(W^{k-t}(H')|Z=i) =$$

$$= P_c\cdot\mathbb{E}(W^{k-t}(H')|Z=0) + \sum_{i=1}^{u} P_{A_i}\cdot\mathbb{E}(W^{k-t}|Z=i) = P_c\cdot\mathbb{E}(W^{k-t}(H')|Z=0) + P_u\sum_{i=1}^{u}\mathbb{E}(W^{k-t}(H')|Z=i)$$

Nuestra misión será acotar $\mathbb{E}(W'|Z=i)$ para $i = 0, .., u$.

- Supongamos que hemos tomado nuestro primer centro, h', en un cluster ya cubierto (Esto quiere decir que, $Z = 0$). Si $H'' = H - \{h'\}$, resulta que

  $$W^{k-t}(H') = \sum_{x\in\mathcal{X}} \min_{h\in H'} \|x-h\|^2 = \sum_{x\in\mathcal{X}} \min_{h\in H'-\{h'\}} \|x-h\|^2 = W^{k-t+1}(H'')$$

  Ya que hemos añadido un centro perteneciente a los clusters cubiertos: los puntos que distaban δ de los centroides siguen haciéndolo pues es la mínima distancia, mientras que los puntos que distaban Δ siguen estando a esta distancia del nuevo centro ya que este no ha sido escogido en su cluster. De este modo, utilizando que W^{k-t+1} no depende de h', tenemos

  $$\mathbb{E}(W^{k-t}(H')|Z=0) = \mathbb{E}(W^{k-t+1}(H'')|Z=0) = \mathbb{E}(W^{k-t+1}(H''))$$

  Pero es el potencial correspondiente a añadir $t-1$ centro y tener u clusters sin cubrir. Por lo tanto, nos encontramos en la situación $(t-1, u)$ y podemos aplicar la hipótesis de inducción:

  $$\mathbb{E}(W^{k-t}(H')|Z=0) = \mathbb{E}(W^{k-t+1}(H'')) \geq \alpha^t\left(n\delta^2\cdot(1+H'_u)\cdot\beta + \left(\frac{n}{k}\Delta^2 - 2n\delta^2\right)(u-t+1)\right)$$

  Si consideramos la probabilidad de escoger un centro en $\mathcal{X}_c$, multiplicamos y dividimos por $(k-t)\delta^2$ y vemos que gracias a la cota establecida anteriormente para α, se tiene que

  $$\frac{n - u\cdot\frac{n}{k} - k}{n - \frac{n}{k}} + \frac{t}{n-\frac{n}{k}} \geq \alpha$$

Podemos llegar a que

$$\frac{W^{k-t}(H', \mathcal{X}_c)}{W^{k-t}(H')} = \frac{(k-u)\cdot\frac{n}{k}\cdot\delta^2 - (k-t)\delta^2}{u\cdot\frac{n}{k}\cdot\Delta^2 + (k-u)\cdot\frac{n}{k}\cdot\delta^2 - (k-t)\delta^2} \geq \alpha\frac{(k-t)\delta^2}{u\cdot\Delta^2 + (k-u)\cdot\delta^2}$$

- Supongamos que hemos tomado nuestro primer centro, h', en un cluster no cubierto, digamos A_i (Esto es, $Z = i$ con $i = 1, .., u$). Sea Y la variable aleatoria que elige un punto $a \in A_i$ con probabilidad p_a como primer centro, es decir, $P(Y = a|Z = i)$. Por lo tanto, $c_1 = a$ y podemos expresar el valor medio de $W^{k-t}(H')$ sabiendo que elegimos el primer centro en el conjunto A_i, como

$$\mathbb{E}(W^{k-t}(H')|Z = i) = \mathbb{E}_Y(\mathbb{E}((W^{k-t}(H')|Y)|Z = i)) = \sum_{a\in A_i} p_a \cdot \mathbb{E}(W^{k-t}(H')|Y = a)$$

En primer lugar, observamos que $p_a = P(Y = a|z = i)$ es la misma para todos los puntos de A_i, ya que todos distan Δ de los centros de cluster ya elegidos. Como en cada A_i hay $\frac{n}{k}$ puntos, la probabilidad de elegir uno de ellos será $\frac{1}{\frac{n}{k}} = \frac{k}{n}$.

Estudiamos entonces cómo acotar $\mathbb{E}(W^{k-t}(H')|Y = a)$. Consideramos A_i cubierto y añadimos a al conjunto de centros de cluster fijos: $H'' = H' \cup \{a\}$. Tenemos que el potencial sería

$$W^{k-t}(H') = \sum_{x\in\mathcal{X}} \min_{h\in H\cup\{a,c_2,..,c_t\}} \|x - h\|^2 = u\cdot\frac{n}{k}\cdot\Delta^2 + (k-u)\cdot\frac{n}{k}\cdot\delta^2 - (k-t)\delta^2 \geq$$

$$\geq (u-1)\cdot\frac{n}{k}\cdot\Delta^2 + (k-u+1)\cdot\frac{n}{k}\cdot\delta^2 - (k-t+1)\delta^2$$

En esta situación, podemos aplicar la hipótesis de inducción (caso $(u-1, t-1)$) y afirmar que

$$\mathbb{E}(W^{k-t}(H')|Y = a) \geq \alpha^t\left(n\delta^2\cdot(1 + H'_{u-1})\cdot\beta + \left(\frac{n}{k}\Delta^2 - 2n\delta^2\right)(u-t)\right)$$

Así, si introducimos esto en la expresión anterior

$$\sum_{a\in A_i} p_a\cdot\mathbb{E}(W^{k-t}(H')|Y = a) = \sum_{a\in A_i}\frac{k}{n}\cdot\mathbb{E}(W^{k-t}(H')|Y = a) \geq \alpha^t\left(n\delta^2\cdot(1 + H'_{u-1})\cdot\beta + \left(\frac{n}{k}\Delta^2 - 2n\delta^2\right)(u-t)\right)$$

Ya que A_i tiene $\frac{n}{k}$ puntos, lo que provoca que el sumatorio $\sum_{a\in A_i}\frac{k}{n}$ sea 1.
Si estudiamos cual es la probabilidad de escoger el centro entre los puntos $\mathcal{X}_u$, llegamos a que

$$\frac{W^{k-t}(H', \mathcal{X}_u)}{W^{k-t}(H')} = \frac{u\cdot\frac{n}{k}\cdot\Delta^2}{u\cdot\frac{n}{k}\cdot\Delta^2 + (k-u)\cdot\frac{n}{k}\cdot\delta^2 - (k-t)\delta^2} \geq \alpha\left(\frac{u\cdot\Delta^2}{u\cdot\Delta^2 + (k-u)\cdot\delta^2}\right)$$

Tras el estudio de ambos casos, recopilamos los resultados obtenidos en la expresión principal:

$$\mathbb{E}(W^{k-t}(H')) = \frac{(k-t)\delta^2}{u\cdot\Delta^2 + (k-u)\cdot\delta^2}\cdot\alpha^{t+1}\left(n\delta^2\cdot(1 + H'_u)\cdot\beta + \left(\frac{n}{k}\Delta^2 - 2n\delta^2\right)(u-t+1)\right) +$$

$$+\frac{u\Delta^2}{u\Delta^2+(k-u)\delta^2}\cdot\alpha^{t+1}\left(n\delta^2\cdot(1+H'_{u-1})\cdot\beta+\left(\frac{n}{k}\Delta^2-2n\delta^2\right)(u-t)\right)$$

Separando los términos del primer sumando para poder agrupar y utilizando que $H'_u - H'_{u-1} = \frac{k-u}{uk}$, tras unas cuentas algo pesadas pero simples, logramos ver que

$$\mathbb{E}(W^{k-t}(H')) \geq \alpha^{t+1}\left(n\beta\delta^2(1+H'_u)+\left(\frac{n}{k}\Delta^2-2n\delta^2\right)(u-t)\right)$$

Que coincide con el resultado deseado (13), por lo que hemos concluido.

□

De esta manera, logramos probar el resultado deseado:

Teorema 9. *La ponderación $\mathcal{D}^2$ no es mejor que $2(\ln k)-$competitiva.*

Supongamos que construimos una partición en clusters $\mathscr{C}$ como hemos descrito anteriormente. Vamos a aplicar el lema con $u = t = k-1$ tras elegir el primer centro. Dado que $1+H'_{k-1} = H_k > \ln k$, se tiene que

$$\mathbb{E}(W^k(H')) \geq \alpha^k\beta\delta^2\ln k$$

Fijamos todos los términos salvo Δ y n, que hacemos que tiendan a infinito. Resulta que

$$\lim_{\Delta\to\infty}\alpha = \lim_{\Delta\to\infty}\frac{n-k^2}{n} = 1, \quad \lim_{\Delta\to\infty}\beta = \lim_{\Delta\to\infty}\frac{\Delta^2-2k\delta^2}{\Delta^2} = 1$$

Por esta razon, haciendo tender $\Delta, n \to \infty$ en ambos lados y recordando que $W^k_{OPT} = \frac{n-k}{2}\delta^2$, se tiene que

$$\mathbb{E}(W^k(H')) \geq n\delta^2\ln k = \frac{n}{n-k}2W^k_{OPT} \geq 2(\ln k)W^k_{OPT}$$

Con lo que conseguimos el resultado deseado.

5.2.4. Algunas generalizaciones

Como hemos comentado, en el estudio particular de k medias++, hemos utilziado $\|x-h\|^2$ como medida de disimilariad entre puntos. Los resultados probados en las dos subsecciones anteriores pueden ser generalizados en el caso de minimizar un potencial $W^{k,[l]}(H) = \sum_{x\in\mathcal{X}}\text{mín}_{h\in H}\|x-h\|^l$, con $l \geq 1$. Lo único que debemos hacer en este cso es utilizar la ponderación $\mathcal{D}^l$, es decir, la probabilidad de elegir x_0 como nuevo centro sería $\frac{D(x_0)^l}{\sum_{x\in\mathcal{X}}D(x)^l}$. Se pueden probar entonces unos lemas análogos a los que habíamos presentado previamente, obteniendo:

- El valor esperado del potencial al escoger el primer centro en $A \in \mathscr{C}_{OPT}$ se acota como

$$\mathbb{E}(W^{1,[l]}(H), A) \leq 2^l W^{1,[l]}_{OPT}(A)$$

- Si $\mathscr{C}$ se construye con la ponderación $\mathcal{D}^l$, el correspondiente potencial satisface

$$\mathbb{E}(W^{k,[l]}(H')) \leq 2^{2l}(\ln k + 2)W^{k,[l]}_{OPT}$$

5.2.5. Ejemplos de la aplicación de k medias++

Ahora que quedan estudiadas las bases teóricas que nos permiten acotar el valor medio del potencial que obtenemos inicializando los centros con k medias++, abordaremos la agrupación de los conjuntos de datos que resultaban problemáticos para k medias y veremos hasta que punto k medias++ logra enmendar estos errores.

- **Carencias no resueltas por k medias++**

  El algoritmo que hemos descrito solo busca una forma más coherente de elegir los centros, por lo que no podemos esperar mejoras muy notables en conjuntos de datos cuya dificultad para ser clasificados radicaba en la naturaleza no esférica de sus grupos. Por ejemplo, si observamos qué ocurre con conjuntos de datos como "Espiral":

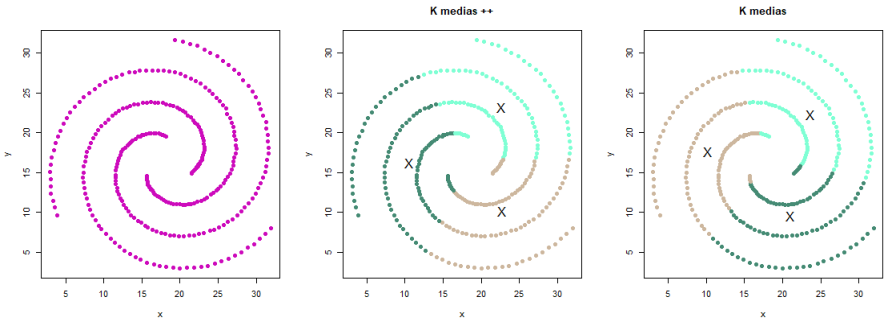


Figura 28: No existe una mejora perceptible en la agrupación de este conjunto de datos por k medias++

O como aquellos con grupos en apariencia "alargados" que observábamos antes, nos damos cuenta de que siguen sin ser agrupados correctamente con k medias++ (ver [8]: *Synthetic 2-d Gaussian clusters to test skewness, N=1000, k=6*):

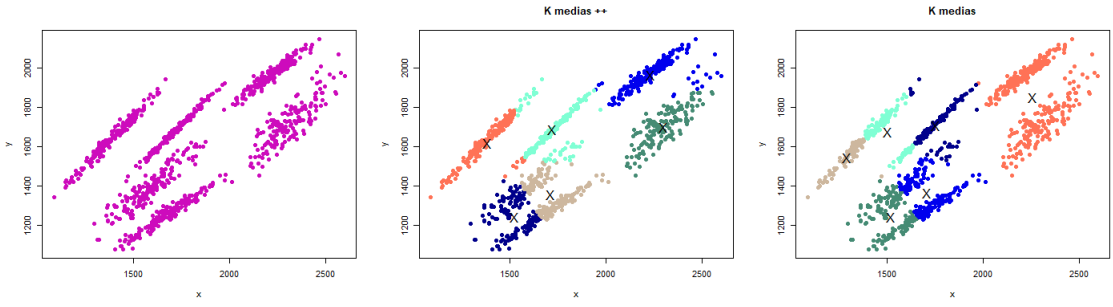


Figura 29: No existe una mejora perceptible en la agrupación de este conjunto de datos por k medias++

Esto se debe a que el principal problema no reside en la inicialización sino en la forma de los grupos, por lo que en estos casos sería más conveniente buscar otros métodos de clustering más apropiados para esto ya que ofrecerán soluciones más satisfactorias.

- **Situaciones resueltas con k medias++**

  Los problemas enmendados por k medias++ son aquellos cuyo principal obstáculo radicaba en el esparcimiento adecuado de centroides. Veíamos dos ejemplos, aquel correspondiente a grupos muy separados entre sí y aquel con grupos mucho más densos que otros, siendo en ambos los grupos con forma esférica. En los dos casos el principal problema era lograr escoger un centroide dentro de cada grupo, cosa que k medias++ consigue la mayor parte de las veces gracias a la probabilidad que asigna a cada punto de ser elegido como nuevo centroide, favoreciendo aquellos que distan más y por lo tanto alejándonos de los centros de cluster ya tomados. En este caso, es claro que k medias++ logra una agrupación mucho mejor del conjunto de datos:

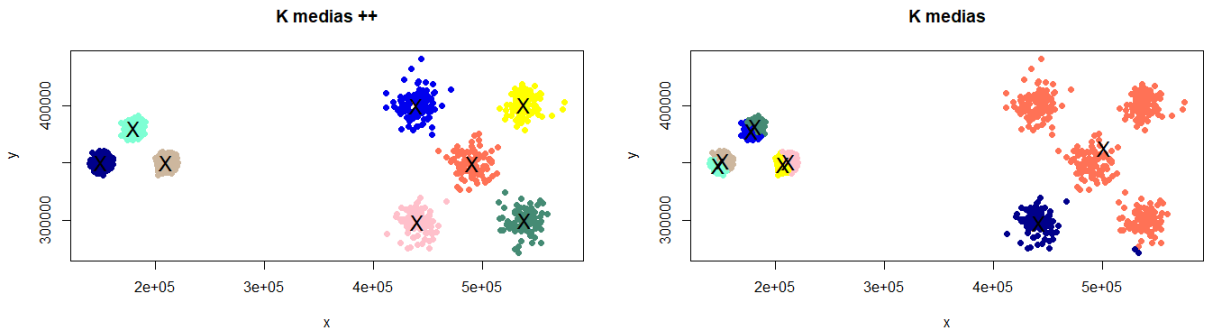


Figura 30: Existe una mejora clara en la agrupación de este conjunto de datos por k medias++

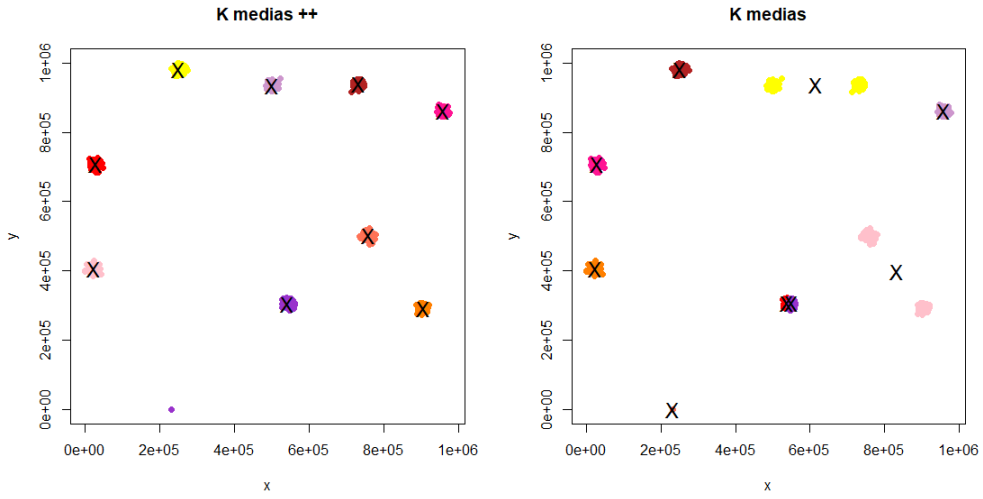


Figura 31: Existe una mejora clara en la agrupación de este conjunto de datos por k medias++

6. Apéndice

Esta sección tiene la finalidad de reunir algunas definiciones o demostraciones de carácter secundario necesarias para el desarrollo y mejor entendimiento del resto del documento.

6.1. Algunos resultados auxiliares

En este apartado recogeremos los resultados necesarios para la realización de las demostraciones de las secciones principales.

Proposición 4. *La convergencia en el sentido de la métrica de Hausdorff en los conjuntos de $\mathcal{B}$ es equivalente a la convergencia en $\mathbb{R}^d$ punto a punto.*

Demostración
Sea $\{A_n\}_{n=1}^{\infty}$ una sucesión de conjuntos de $\mathcal{B}$ y $A \in \mathcal{B}$. Si A_n converge a A en el sentido de la métrica de Hausdorff, significa que $\forall \delta > 0$, $\exists n_0 \in \mathbb{N}$ tal que si $n \geq n_0$, entonces $\mathcal{H}(A_n, A) < \delta$. Esto quiere decir que $\mathcal{H}(A_n, A) = \text{máx}\{\rho(A, A_n), \rho(A_n, A)\} < \delta$. Como los conjuntos de $\mathcal{B}$ son finitos, realmente se tiene $\rho(A, A_n) = \text{máx}\{D(a, A_n) : a \in A\} < \delta$. Dado que $A_n = \{a_1^n, .., a_k^n\}$ y $A = \{a_1, .., a_k\}$ tienen el mismo número de puntos, existe un reordenamiento de los elementos de A_n, $\{a_{m_1}^n, .., a_{m_k}^n\}$, tal que $a_{m_j}^n$ es el punto de A_n que menos dista de a_j (Emparejamos cada punto de A_n al punto más cercano de A). Para simplificar la notación, seguimos denotando al reordenamiento de A_n igual que al propio A_n. De este modo, está claro que

$$\text{máx}\{D(a, A_n) : a \in A\} = \underset{1 \leq j \leq k}{\text{máx}} D(a_j^n, a_j) = \underset{1 \leq j \leq k}{\text{máx}} \|a_j^n - a_j\| < \delta$$

Esto implica que $\forall j = 1, .., k$, se tiene que $\|a_j^n - a_j\| < \delta$ si $n \geq n_0$, y queda probada la convergencia punto a punto.

Recíprocamente, si $\forall \delta > 0$ y para todos los índices $j \in \{1, .., k\}$ $\exists n_0 \in \mathbb{N}$ tal que si $n \geq n_0$, se tiene que $\|a_j^n - a_j\| < \delta$ si $n \geq n_0$, también se verifica que $\text{máx}\{D(a, A_n) : a \in A\} = \text{máx}_{1 \leq j \leq k} D(a_j^n, a_j) = \text{máx}_{1 \leq j \leq k} \|a_j^n - a_j\| < \delta$. Con ello, llegamos a que $\mathcal{H}(A_n, A) < \delta$ si $n \geq n_0$.

□

Lema 9. *Sea $\{H_n\}_{n=0}^{\infty}$ una sucesión de elementos de $\mathcal{B}$. Entonces H_n converge a H_0 si y solo si toda subsucesión de $\{H_n\}_n$ admite una nueva subsucesión convergente a H_0.*

Demostración
La condición necesaria es evidente, ya que si $H_n \longrightarrow H_0$, sabemos que cualquier subsucesión de H_n también converge a H_0.
Veamos la condición suficiente. Supongamos entonces que de toda subsucesión de $\{H_n\}_n$ podemos extraer una nueva subsucesión que converge a H_0. Razonemos por reducción al absurdo suponiendo que H_n no converge a H_0: es decir, $\exists \delta > 0$ de manera que $\mathcal{H}(H_n, H_0) > \delta$ infinitas veces. Realizando un ordenamiento y renombramiento de índices similar al ya comentado antes, esto implicaría que existe al menos un $j \in \{1, .., k\}$ tal que $\|h_j^n - h_j\| > \delta$ para infinitos términos.

Podemos construir entonces la subsucesión $\{H_{n_k}\}_k$ de términos que verifican $\|h_j^{n_k} - h_j\| > \delta$ $\forall k$. Por hipótesis, dada H_{n_k}, podemos extraer una nueva subsucesión de esta que sea convergente a H_0.

Pero esto es absurdo, ya que todos los términos de $\{H_{n_k}\}_k$ distan al menos δ de su correspondiente en H_0. Por lo tanto, debe darse $H_n \longrightarrow H_0$. □

Proposición 5. *Dado el siguiente conjunto*

$$A = \{y \in \mathbb{R}^d / \|y - m\| = \|y - m^*\| \quad P - c.s.\}$$

Se puede reescribir como

$$B = \{y \in \mathbb{R}^d / \exists x \in \mathbb{R}^d \ \ con \ \ \langle x, m - m^* \rangle = 0 \ \ e \ \ y = x + \frac{m + m^*}{2} \quad P - c.s.\}$$

Donde $\langle \cdot, \cdot \rangle$ denota el producto escalar euclideo.

Demostración

El conjunto A de los puntos de $\mathbb{R}^d$ que equidistan de m y m^* se corresponde con el hiperplano que pasa por el punto medio de m y m^*, $\frac{m^*+m}{2}$, y que es ortogonal a la dirección $m - m^*$. De este modo, si un punto $y \in \mathbb{R}^d$ equidista de m y m^*, estará en este hiperplano y existirá un vector x que verifique $\langle x, m - m^* \rangle = 0$ de tal manera que y se escriba como $y = \frac{m+m^*}{2} + x$.

□

Proposición 6. *Sea $A \subset \mathbb{R}^d$ un conjunto convexo y $\phi : \mathbb{R}^d \longrightarrow \mathbb{R}$ una función estrictamente convexa y creciente. Entonces, si la media de A existe, pertenece a la adherencia de A.*

Demostración

Denotamos por $\overline{A}$ la adherencia de A. Sea m la media del conjunto A, es decir

$$\int_A \phi(\|x - m\|) dP(x) \leq \int_A \phi(\|x - g\|) dP(x)$$

para cualquier $g \in A$. Si suponemos que $m \notin \overline{A}$, entonces pertenece a $\in \mathbb{R} - A$ abierto, por lo que existe $r > 0$ tal que $B(m, r) \cap \overline{A} = \emptyset$. Denotamos por a_m el punto de $\overline{A}$ que verifica $\|m - a_m\| \leq \|m - a\|$ para $a \in \overline{A}$, es decir, el punto de $\overline{A}$ más cercano a m. Sabemos que $\|m - a_m\| \geq r > 0$. Además, dado que $\overline{A}$ es cerrado y convexo, a_m corresponde con la única proyección de m sobre el conjunto $\overline{A}$. Dado que ϕ es una función creciente y la integral verifica la propiedad de monotonía, bastará con ver que para cualquier $a \in \overline{A}$

$$d(a, a_m) = \|a - a_m\| \leq \|a - m\| = d(a, m)$$

Y con ello llegaríamos a un absurdo por la proposición anterior, ya que la media de un conjunto es única por ser ϕ creciente y estrictamente convexa, y en este caso a_m estaría más cerca de los puntos de A que el propio m. Dado un punto $a \in A$, si consideramos el triángulo de vértices a, a_m y m, tenemos tres posibilidades:

- La hipotenusa del triángulo la forman los vértices m y a_m, por lo que $d(m, a_m) > d(a, m)$. Esto es absurdo ya que a_m era el punto más cercano a m.

- La hipotenusa del triángulo la forman los vértices a y a_m, por lo que $d(a, a_m) > d(a, m)$. Dado que $\overline{A}$ es cerrado y convexo, el segmento $\lambda a + (1-\lambda)a_m$, para $\lambda \in [0,1]$, está enteramente contenido en $\overline{A}$, por lo que seríamos capaces de encontrar un $\lambda \in (0,1)$ y con ello a_λ tal que $d(a_\lambda, m) < d(a, m)$ (Es decir, construir un triángulo cuya hipotenusa mide $d(a, m)$). Pero hemos afirmado que a_m es único y habríamos encontrado otro candidato a_λ para ser el punto que menos dista de m, lo cual es absurdo.
- La hipotenusa del triángulo la forman los vértices m y a, por lo que $d(m, a) > d(a_m, a)$. Debe ser esta opción e implicaría que m no es media del conjunto A, por lo que concluimos que m debe pertenecer a $\overline{A}$.

□

Proposición 7. *Sean $z_1, .., z_n$ números reales. Se verifica la siguiente desigualdad*

$$n \cdot \sum_{i=1}^{n} z_i^2 \geq \left(\sum_{i=1}^{n} z_i \right)^2$$

Demostración

Si desarrollamos ambos sumatorios, estamos tratando de probar

$$n(z_1^2 + ..z_n^2) \geq (z_1 + .. + z_n)^2$$

Para ello, prestamos atención a las siguientes observaciones:

- Para $n = 2$ la desigualdad es cierta, ya que $(z_1 - z_2)^2 \geq 0$, por lo tanto $z_1^2 + z_2^2 \geq 2z_1z_2$ y con ello, sumando z_1^2 y z_2^2 en ambos lados, obtenemos que $2(z_1^2+z_2^2) = 2z_1+2z_2 \geq z_1^2+z_2^2+2z_1z_2 = (z_1 + z_2)^2$.
  Es decir, si logramos remitirnos al caso $n = 2$ de algún modo, sabremos que se verifica.
- Nos damos cuenta de que

  $$(z_1 + z_2 + .. + z_n)^2 = (z_1 + (z_2 + .. + z_n))^2 = z_1^2 + 2z_1(z_2 + .. + z_n) + (z_2 + .. + z_n)^2$$

  Y, a su vez, este último término se puede tratar del mismo modo

  $$(z_2 + z_3 + .. + z_n)^2 = (z_2 + (z_3 + .. + z_n))^2 = z_2^2 + 2z_2(3 + .. + z_n) + (z_3 + .. + z_n)^2 = ...$$

  Por lo tanto, podemos reescribir de forma compacta

  $$(z_1 + z_2 + .. + z_n)^2 = \sum_{i=1}^{n}(z_i^2 + \sum_{j=i+1}^{n} 2z_iz_j) = (z_1^2 + .. + z_n^2) + \sum_{i=1}^{n}\sum_{j=i+1}^{n} 2z_iz_j$$

- Puedo tratar de añadir términos para crear identidades notables del tipo $(z_i + z_j)^2$ para todas las combinaciones de índices dado que tengo términos de la forma $2z_iz_j$:

  - Uso z_1^2 para formar $(z_1 + z_2)^2$. Debo añadir $(n-2)$ veces z_1^2 para formar $(z_1 + z_i)^2$ con $i \in \{3, 4, .., n\}$.

- Uso z_i^2 para formar $(z_i + z_{i+1})^2$. Debo añadir $(n-2)$ veces z_{n-1}^2 para formar $(z_i + z_j)^2$ con $i \in \{1, 2, .., n\} - \{i, i+1\}$.

- Uso z_n^2 para formar $(z_n + z_1)^2$. Debo añadir $(n-2)$ veces z_n^2 para formar $(z_n + z_i)^2$ con $i \in \{2, 3, .., n-1\}$.

De este modo, llegamos a que añadiendo $(n-2)(z_1^2 + .. + z_n^2)$ podríamos formar todas las combinaciones de $(z_i + z_j)^2$.

Por lo tanto, en lugar de preguntarnos si es cierto que

$$n \cdot \sum_{i=1}^{n} z_i^2 \geq \left(\sum_{i=1}^{n} z_i \right)^2$$

Trataremos de ver si se verifica

$$n(z_1^2 + .. + z_n^2) + (n-2)(z_1^2 + .. + z_n^2) \geq (z_1^2 + .. + z_n^2) + \sum_{i=1}^{n} \sum_{j=i+1}^{n} 2z_i z_j + (n-2)(z_1^2 + .. + z_n^2)$$

Que, agrupando como hemos comentado anteriormente, quedaría de esta forma

$$(2n-2)(z_1^2 + .. + z_n^2) \geq \sum_{i=1}^{n} \sum_{j=i+1}^{n} (z_i + z_j)^2$$

Comprobamos que efectivamente para el caso $n = 2$ se cumplía la desigualdad. Es decir, para cada $(z_i + z_j)^2$ del lado izquierdo necesito $2z_i^2 + 2z_j^2$ al lado derecho para que se verifique. Por lo tanto, para cada z_i^2 del lado izquierdo, necesito $2z_i^2$ al otro lado. Cada z_i^2 del lado izquierdo aparece $(n-1)$ veces ya que se combina con todos en $(z_i + z_j)^2$ salvo consigo mismo, por lo que para que la desigualdad fuera cierta, necesitaría tener $2(n-1)$ veces z_i^2 al otro lado de la desigualdad. Pero esto es justo lo que tenemos, por lo cual se verifica la desigualdad.

□

6.2. Algunas nociones sobre complejidad computacional

La teoría de la complejidad computacional es la parte de la teoría de la computación encargada de estudiar los recursos requeridos para resolver un problema (tiempo de ejecución y espacio de memoria) y clasificar estos en aquellos cuyo proceso de resolución es tratable y aquellos donde el consumo de recursos se dispara rápidamente hasta llegar a ser prohibitivo.

Su objetivo es establecer una métrica abstracta de la cantidad de recursos necesarios para calcular una solución. Denominamos tiempo de ejecución de un algoritmo a la medida del número de pasos o de operaciones elementales necesarias para conseguir un resultado. El espacio de memoria es la medida del número de posiciones de memoria necesarios para almacenar los cálculos y el resultado final.

Distinguimos dos categorías de algoritmos: Aquellos con complejidad polinómica y aquellos con complejidad exponencial. Esta última corresponde con los algoritmos cuyo costo de cálculo se vuelve inasumible rápidamente a medida que los datos de entrada crecen. Más concretamente:

- Algoritmo con complejidad polinómica: La tasa de crecimiento de su coste respecto al tamaño de la entrada (n) es del orden de $O(nk)$ para alguna constante k.
- Algoritmo con complejidad exponencial: La tasa de crecimiento de su coste respecto al tamaño de la entrada (n) es del orden de $2^{O(nk)}$ para alguna constante k.

De esta forma, somos capaces de agrupar los problemas en clases de complejidad. Denominamos clase P al conjunto de los problemas resolubles en tiempo polinómico. Denominamos clase NP al conjunto de todos los problemas verificables en tiempo polinómico (es decir, dada una posible solución, esta se puede verificar en un tiempo polinómico por una máquina de Turing). La clase NP-Hard es aquella que contiene a los problemas de decisió que son como mínimo tan dificiles como un problema de NP. El algoritmo de k medias al que hacemos referencia en el documento pertenece a esta clase.

Por otro lado, el Análisis Competitivo es la disciplina encargada del estudio de la actuación de un algoritmo en una instancia de un problema en comparación con el comportamiendo del algoritmo en el óptimo. Decimos que un algoritmo es "competitivo" si la razón entre su comportamiento en un caso cualquiera y en el caso óptimo está acotada. En este documento afirmaremos que "El método de k means++ es $O(\log k) - competitivo$ a la solución óptima de agrupamiento en clusters". Esta noción no trata la complejidad del algoritmo, sino que trata de ver cómo de efectivo es y cuánto puede alejarse como mucho del valor óptimo de una instancia. Sabemos que el objetivo de k medias es minimizar W^k. La afirmación anterior por lo tanto quiere decir que el k-potencial esperado de una solución proporcionada por k means++ es como mucho $8(\ln k + 2)$ veces el potencial de la mejor solución posible.

Referencias

[1] Bholowalia, P., and Kumar, A. (2014). *EBK-means: A clustering technique based on elbow method and k-means in WSN.* International Journal of Computer Applications, 105(9).

[2] Billingsley, P. (2013). *Convergence of probability measures.* John Wiley and Sons.

[3] Cuesta-Albertos, J. A., Gordaliza, A., Matrán, C. (1997) *Trimmed fc-means and the Cauchy mean value property.* New Trends in Probability and Statistics, 3, 247-265.

[4] Cuesta-Albertos, J. A., Gordaliza, A., Matrán, C. (1997). *Trimmed k-means: an attempt to robustify quantizers.* The Annals of Statistics, 25(2), 553-576.

[5] Cuesta, J. A., Matrán, C. (1988). *The strong law of large numbers for k-means and best possible nets of Banach valued random variables.* Probability theory and related fields, 78(4), 523-534.

[6] D. Arthur, S. Vassilvitskii (2006). *k-means++: The Advantages of Careful Seeding.* Stanford, 2006.

[7] Dudek, A. (2019, September). *Silhouette index as clustering evaluation tool.* In Conference of the Section on Classification and Data Analysis of the Polish Statistical Association (pp. 19-33). Springer, Cham.

[8] Fränti, P., Sieranoja, S. (2019). *How much can k-means be improved by using better initialization and repeats?.* Pattern Recognition, 93, 95-112.

[9] Hartigan, J. A. and Wong, M. A. (1979). *Algorithm AS 136: A k-means clustering algorithm..* Journal of the royal statistical society. series c (applied statistics), 28(1), 100-108.

[10] Jain, A. K. (2010). *Data clustering: 50 years beyond K-means.* Pattern recognition letters, 31(8), 651-666.

[11] MacQueen, J. (1967, June). *Some methods for classification and analysis of multivariate observations.* In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297).

[12] Pollard, D. (1981). *Strong consistency of k-means clustering.*The Annals of Statistics, 135-140.

[13] Serfling, R. J. (2009). *Approximation theorems of mathematical statistics.* John Wiley and Sons.

[14] Tibshirani, R., Walther, G., and Hastie, T . (2001).*Estimating the number of clusters in a data set via the gap statistic.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(2), 411-423.

[15] Varadarajan, V. S. (1958). *On the convergence of sample probability distributions.* The Indian Journal of Statistics (1933-1960), 19(1/2), 23-26.

[16] Yamamoto, W., Shinozaki, N. (2000). *On uniqueness of two principal points for univariate location mixtures.* Statistics and probability letters, 46(1), 33-42.