Trabajo Fin de Máster

Métodos cluster basados en la detección de modas

Alumno: Daniel Pais Romero

Tutores: Rosa M. Crujeiras, Jose Ameijeiras Alonso

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

MÁSTER EN TÉCNICAS ESTADÍSTICAS

Trabajo Fin de máster

Métodos cluster basados en la detección de modas

Área: Estadística y investigación operativa

UNIVERSIDAD DE SANTIAGO DE COMPOSTELA

Área de Conocimiento:
Estadística y investigación operativa
Tutores:
Rosa M. Crujeiras, Jose Ameijeiras Alonso
Título:
Métodos cluster basados en la detección de modas
Modalidad:
A.
Breve descripción del contenido
Se pretende explicar diferentes métodos clustering, destacando el interés en los modelos no paramétricos relacionados con la frecuencia de los datos.
Objetivos
El objetivo de este trabajo será la revisión y comparativa en escenarios simulados y sobre datos reales, de las técnicas paramétricas y no paramétricas para la agrupación de datos, donde esta agrupación se realiza basándose en la estimación de la densidad.

Resumen

Resumen

En contextos prácticos, donde necesitaremos de técnicas estadísticas para analizar datos, podemos abordar los problemas de interés desde dos perspectivas: técnicas exploratorias o descriptivas (sin un modelo pre-específico de los datos), o técnicas inferenciales o confirmatorias (donde queremos confirmar la validez de una hipótesis en nuestros datos). Ejemplos de técnicas estadísticas para análisis de datos pueden ser el análisis de componentes principales o el análisis cluster.
En lo que respecta a reconocimiento de datos en grupos con características comunes, distinguiremos dos tipos de problemas: problemas supervisados (de clasificación, para datos con niveles o categorías) o problemas no supervisados (de clustering, para datos sin niveles). Cabe mencionar el interés creciente en una mezcla de los anteriores problemas (con niveles para una porción de los datos).
El trabajo se sitúa en el marco de los problemas no supervisados y se tratarán distintas técnicas para realizar las agrupaciones.
El objetivo de este trabajo será generar grupos utilizando algoritmos basados en distribuciones donde el agrupamiento se hace en función a la probabilidad que tiene un dato de pertenecer a cierto grupo. Se pueden distinguir dos metodologías principales dependiendo del concepto de agrupación que se adopte: paramétrico y no paramétrico. Por un lado, los métodos paramétricos suponen, en general, que la densidad subyacente que genera los datos es una mixtura. En este enfoque los datos se agrupan en función de la probabilidad que tiene cada dato de pertenecer a cierta componente de la mixtura. En el caso de los métodos no paramétricos, los grupos se definen en función de donde se concentra la masa de probabilidad. Una forma de caracterizar estas regiones es en términos de las modas (máximos relativos) de la función de densidad.

Summary

In practical contents, where we will need statistical techniques to analyse data, we can approach the problems of interest from two perspectives: exploratory or descriptive techniques (without a pre-specific model of the data) and inferential or confirmatory techniques (where we want to confirm the validity of a hypothesis in our data). Examples os statistical techniques for data analysis can be the principal component analysis and the cluster analysis. When it comes to data recognition in groups with common characteristics, we will distinguish two types of problems: supervised problems (classification problems, for data with levels or categories) and unsupervised problems (clustering problems, for data without levels). It is worth to mentioning the growing interest in a mixture of the above problems (with levels for a portion of the data).

The work is located within the framework of the unsupervised problems and different techniques will be treated to carry out the groupings. The objetive of this work will be to generate groups using algorithms based on distributions where the grouping is made in function of the probability that a observation belongs to a certain group.

Two main methodologies can be distinguished depending on the concept os grouping that is adopted: parametric and non-parametric. On the one hand, parametric methods assume, in general, that the underlying density that generates the data is a mixture. In this aproach, the data is clustered according to the probability that each obsrvation has of belonging to a certain component of the mixture. On the other hand, in non-parametric methods, groups are defined according to where the probability mass is concentrated. One way to characterize these regions in terms of the modes (relative maximums) of the density functions.

Índice general

1. Introducción a los métodos clustering **7**

1.1. Introducción a los métodos 7

1.1.1. Fundamentos clustering 8

1.1.2. Tipos de clustering 8

1.1.3. Clustering basado en distancias 9

1.2. El algoritmo k-medias 10

1.2.1. Funcionamiento del algoritmo k-medias 10

1.2.2. Limitaciones del clustering empleando k-medias 11

1.3. Introducción a los datos reales 12

1.4. Organización del trabajo 15

2. Clustering basado en modelos paramétricos **17**

2.1. Modelos de mixturas 17

2.2. Algoritmo de esperanza-maximización (EM) 19

2.2.1. Caso particular de mixturas de distribuciones normales 20

2.2.2. Relación entre el algoritmo k-medias y el algoritmo EM 20

2.3. Elección del número de clusters 21

2.4. Estrategia para realizar clustering paramétrico 22

2.5. Limitaciones del clustering paramétrico 23

3. Clustering basado en modelos no paramétricos **25**

3.1. Clustering modal 26

3.2. Algoritmo de cambio medio 27

3.3. Elección del clustering modal 28

4. Datos reales **31**

4.1. Base de datos de críminalidad en USA 31

4.2. Datos de glucosa 34

4.2.1. Comparación de los tres enfoques con el conjunto de datos de glucosa 36
4.2.2. Segunda comparación de los tres enfoques con el conjunto de datos de glucosa 38
4.3. Datos de tiros en la liga americana de baloncesto 39

A. Anexo de código **43**

Bibliografía **53**

Capítulo 1

Introducción a los métodos clustering

1.1. Introducción a los métodos

Los avances tecnológicos y el crecimiento en aplicaciones como búsquedas en Internet provocaron la necesidad de trabajar con bases de datos de gran tamaño y dimensión. Junto con el crecimiento de la cantidad de datos, también aumenta la variabilidad de estos, complicando la tarea de analizarlos.

Para abordar este problema, se puede pensar en agrupar los datos con el fin de aprovechar las similitudes entre ellos para que sean más sencillos de analizar. Estos problemas de clasificar datos en grupos, se conocen como problemas de clustering. Además tener como objetivo el análisis datos, otros propósitos de agrupar los datos serán los siguientes: para obtener información sobre los datos, generar hipótesis, detectar anomalías e identificar características destacadas; para identificar el grado de similitud entre organismos o como método para organizar los datos y resumirlos a través de sus características cluster.

Una cualidad de la metodología clustering es que resulta interdisciplinar, esto quiere decir que se usa en múltiples ámbitos, por lo tanto es sencillo encontrar ejemplos donde se usará clustering. Para obtener rentabilidad económica en tareas de marketing, puede ser útil encontrar grupos de clientes con un comportamiento similar que puedan ser consumidores de los productos que se lanzan al mercado. En campos como la medicina, será provechoso clasificar personas de acuerdo a un grado de enfermedad para proporcionar los cuidados necesarios a cada paciente. En marcos deportivos, puede ser ventajoso para un equipo incorporar un jugador realizando una selección previa [1]. Para llegar a esta recopilación, será fructífero agrupar jugadores de acuerdo a características de juego.

Al elegir un método clustering, hay que decidir cuantos grupos coger o que criterio utilizar para agrupar datos en un cluster u otro. En esta sección, se explicará que características comunes presenta un método cluster. Luego se explicarán varios métodos clustering,

comparándolos, y finalmente se añadirán datos reales para ilustrar estos métodos.

1.1.1. Fundamentos clustering

El análisis de clusters tiene por objeto agrupar elementos en grupos homogéneos en función de las similitudes entre ellos. Los fundamentos del análisis clustering son:

- Que cada elemento pertenezca a uno y solo uno de los grupos;
- que todo elemento quede clasificado;
- que cada grupo sea internamente homogéneo. Esto es no puede haber datos que dentro del grupo tenga distintas características.
- Al emplear ciertos métodos, será importante prefijar el número de clusters para llegar a una clasificación adecuada.
- Se pretende que los algoritmos clustering sean estables, esto es, al utilizar los algoritmos con diferentes valores iniciales o con diferentes subconjuntos de los datos, se deberían obtener los mismos clusters.

1.1.2. Tipos de clustering

En esta sección se hará una breve introducción a los procedimientos clustering. Lo primero que se destaca es la amplitud y diversidad de estos métodos, tanto en sus objetivos como en los procedimientos que emplean. El aspecto común a todos ellos es el propósito de formar grupos que todavía no están definidos. Podemos plantear la siguiente clasificación de las técnicas clustering. Los métodos jerárquicos, también conocidos como métodos de taxonomía numérica, pues tienen mucha semejanza con la taxonomía de los seres vivos en biología. Partiendo de una matriz de distancias entre individuos, el objetivo es construir un árbol de clasificación en distintos niveles. En el nivel más bajo estarán grupos de individuos muy próximos o semejantes. A niveles superiores se crean grupos como resultado de la agregación de los grupos de nivel inferior. Por ejemplo, dada una muestra de animales, pretendemos estructurarlos según su categoría taxonómica, de forma que se empezarán agrupando en los taxones menos inclusivos (como pueden ser el reino o el dominio) y finalmente se agruparán en los taxones más inclusivos (como la especie).
Los métodos de particionamiento. Al contrario de los métodos jerárquicos, los métodos de particionamiento sólo pretenden una división de los individuos en un conjunto de grupos, todos al mismo nivel, que constituyen una partición de la población original. Un ejemplo de estos métodos es el algoritmo k-medias (algoritmo basado en la distancia a centroides).

Los métodos antes descritos no establecen suposiciones sobre el modelo estadístico que generó los datos, y en este sentido se deben interpretar como métodos descriptivos que ayudan a la comprensión sobre las propiedades que presenta un conjunto de datos multivariantes. Cabe destacar que se basean en el concepto de distancia y que generalmente se utiliza la distancia euclidiana.
Sin embargo, con bastante frecuencia, se agrupan los datos mediante métodos basados en distribuciones. El agrupamiento se hace en función de la probabilidad que tiene un dato de pertenecer a cierto grupo. Se distinguen:

- Técnicas paramétricas. Se asume un modelo paramétrico, se necesitan por lo tanto unos parámetros para conocer la distribución de los datos. A partir de la distribución se realiza el clustering. Un ejemplo de procedimiento que permite estimar los parámetros en una distribución paramétrica en contextos de información incompleta es el algoritmo EM (de esperanza-maximización).

- Técnicas no paramétricas. No se asume un modelo paramétrico, y a partir de la distribución estimada se aplica el clustering. En los modelos no paramétricos puede ser difícil distinguir densidades de los datos, sobretodo con datos multidimensionales. Un ejemplo sería el algoritmo mean-shift (de cambio medio).

1.1.3. Clustering basado en distancias

En esta sección se extenderán los métodos jerárquicos. Los algoritmos que se emplean para crear esta jerarquía pueden ser de dos tipos. Los métodos aglomerativos, los cuales partiendo de los individuos, se construyen grupos formados por individuos, y en niveles superiores se van construyendo otros grupos con la agregación de grupos formados en etapas anteriores. Los métodos divisivos: se vuelve a partir de un grupo total formado por todos los individuos y se genera una división en subgrupos, que más adelante serán subdivididos. Los más comunes son los aglomerativos, que constan de los siguientes pasos.

Algoritmo 1.1.

Se definen los grupos $C_1, ..., C_n$ formados cada uno por un individuo.

Se buscan los dos grupos C_j y C_l que estén más próximos y se juntan.

Se recalculan las distancias de todos los demás grupos al nuevo. Según como se determinan estas distancias, obtendremos distintos métodos. Por ejemplo, si la distancia entre grupos fuese la mínima distancia entre los elementos de un grupo y otro, se estaría empleando el método del mínimo.

Si el número de grupos es uno, se detiene el algoritmo. En caso contrario, se vuelve al segundo paso de búsqueda del par de grupos más cercanos.

Con respecto a los métodos de particionamiento, se formarán grupos por proximidad en el espacio d-dimensional. Un criterio natural para la formación de grupos, consistiría en elegir la partición que haga mínima la variabilidad dentro de cada grupo, medida por la suma de cuadrados intra-grupo. Sin embargo, dependiendo del número de individuos y de grupos, será imposible recorrer todas las particiones que se pueden formar. Se echa mano entonces de algoritmos que, partiendo de una solución razonable, aporten mejoras sucesivas. Como ya se dijo un ejemplo es el algoritmo k-medias.

1.2. El algoritmo k-medias

Uno de los métodos clustering más populares es el algoritmo k-medias. Como se recoge en *Jain*, (2010) pesar de ser propuesto hace más de 50 años, y de las miles de variantes que surgieron desde entonces, k-medias se sigue utilizando frecuentemente. Destaca por ser fácil de implementar, simple y eficiente. Además es facilmente extensible en diferentes direcciones. Por ejemplo, para la cuantificación de imágenes (técnica de compresión utilizada para reducir el número de colores necesarios para representar una imagen, y por lo tanto el tamaño de la misma).

1.2.1. Funcionamiento del algoritmo k-medias

En esta sección se describirá y se verá como se puede implementar el algoritmo k-medias. Supongamos una muestra de n elementos con p variables. El objetivo es dividir esta muestra en un número de grupos prefijado k, de acuerdo a minimizar el error entre la media empírica de un cluster y los puntos del cluster. Requiere las siguientes etapas.
Para comenzar, se deben seleccionar k puntos como centros de los grupos iniciales. Esto puede hacerse: asignando aleatoriamente los objetos a los grupos y tomando los centros de los grupos así formados. También tomando como centros los k puntos más alejados entre sí. O bien si se dispone de información a priori, seleccionando los centros de acuerdo a esa información. Luego, se deben calcular las distancias (generalmente euclídeas) de cada elemento al centro de los k grupos, y asignar cada elemento al grupo más próximo. La asignación se realiza secuencialmente y al introducir un nuevo elemento en un grupo se recalculan las coordenadas de la nueva media de grupo. Finalmente, una vez se tenga definido un criterio de optimalidad, se comprobará si reasignando uno a uno cada elemento de un grupo a otro mejora el criterio. Si no es posible mejorar, terminarél proceso.

Algoritmo 1.2. *Para implementar el algoritmo en la práctica [2], consideremos* $X = \{x_i\}$, $i = 1, ..., n$ *el conjunto de* n *puntos* p*-dimensionales que queremos agrupar en* k *grupos, que denominaremos* $C_j, j \in 1, .., k$*. Como ya se indicó, el algoritmo* k*-means encuentra una partición que minimice el error cuadrático medio entre la media de un grupo (centroide), la cuál denotaremos por* μ_j *y los puntos del grupo. Se define como:*

$$J(C_j) = \Sigma_{x_i \in C_j} ||x_i - \mu_j||^2$$

El objetivo del algoritmo k*-medias es minimizar la siguiente suma (el error cuadrático medio usando la norma euclidiana):*

$$J(C) = \Sigma_{j=1}^{k} \Sigma_{x_i \in C_j} ||x_i - \mu_j||^2$$

La solución obtenida al minimizar es un mínimo de forma local (aunque con probabilidad grande, está demostrado que converge a un mínimo global con grupos bien separados). Dado que el error cuadrático se reduce con el incremento del número de clusters (con $J(C) = 0$ cuando $k = n$), se puede minimizar solamente para un número fijo de clusters.

1.2.2. Limitaciones del clustering empleando k-medias

En multitud de casos, los métodos clustering basados en distancias, son una herramienta simple que proporciona buenos resultados. Sin embargo, ya se indicó que hay gran variedad de métodos para realizar tareas de agrupamiento de datos. Esto se debe a intentos de mejorar los recursos clustering con el fin de superar las limitaciones de los métodos más antiguos como puede ser k-medias. Las restricciones más importantes del algoritmo k-medias están relacionadas con las probabilidades de que las observaciones estén en un grupo u otro. Habrá datos que aunque se clasifiquen en un grupo, sea con menos seguridad que otras y el algoritmo k-medias no cuantifica como de seguro es la clasificación de una observación individualmente. Un aspecto interesante a la hora de elegir un algoritmo antes que otro es si es aplicable sin tener que proporcionar a priori valores que puedan ser mal elegidos provocando que el algoritmo falle. Al utilizar k-medias, se necesita un número de grupos de partida.
Otro inconveniente que va a estar presente en todas las formas de agrupar que se discutirán en este trabajo será la presencia de datos atípicos. La disposición de los grupos se puede graficar y los grupos se corresponden con zonas de datos cercanos, entonces un dato que no se encuentre en estas zonas (atípico) no estará bien clasificado.

1.3. Introducción a los datos reales

Base de datos de crímenes

La primera base de datos está relacionada con los crímenes cometidos en 50 estados de USA en el año 1973 y se puede encontrar en $McNeil, D.R.(1977)$. Se consideran las variables asaltos y asesinatos, que indican la frecuencia con la que suceden tanto los asaltos como asesinatos (vienen dados en cien miles). No se proporciona una clasificación real de zonas con más o menos criminalidad, por lo tanto con esta base de datos se implementará el algoritmo k-means y los métodos basados en modelos, y se interpretarán los resultados. Los datos se distribuyen como se muestra en la figura 1.1.

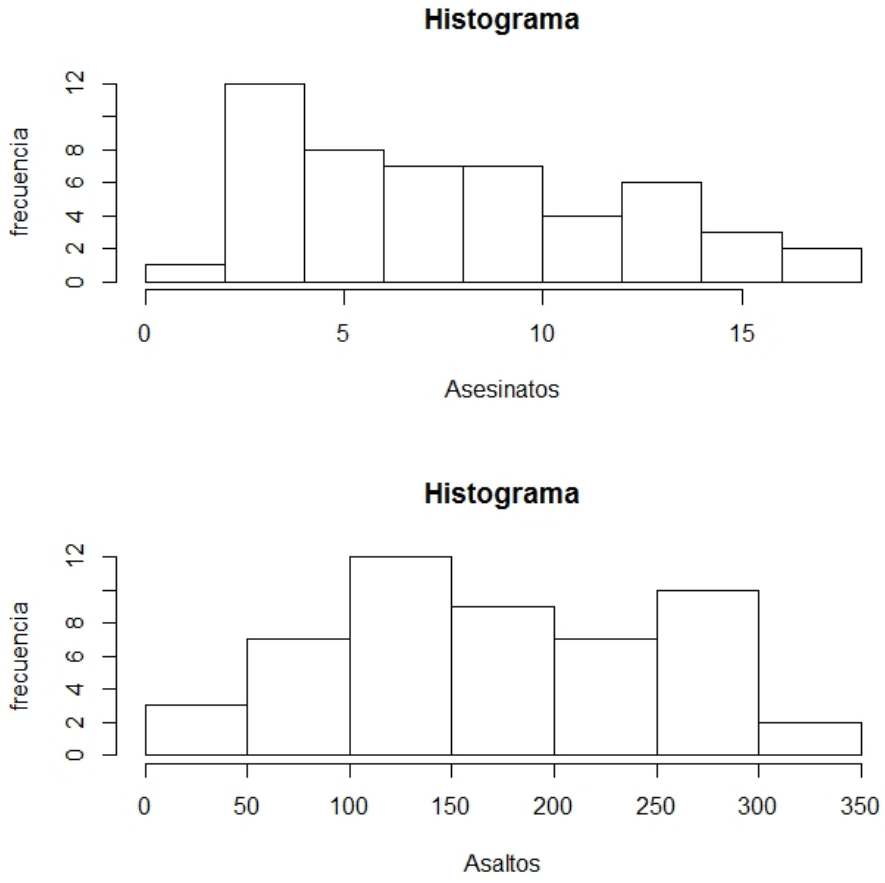

Figura 1.1: Histogramas de las variables asesinatos y asaltos. Panel superior: variable asesinatos. Panel inferior: variable asaltos.

Base de datos de glucosa

La segunda base de datos usada fue propuesta por *Reaven y Miller*, (1979) Se recoge una muestra de 145 adultos que no padecen obesidad. Las variables que aparecen son: la medida de tolerancia a la glucosa (glutest), la medida de resistencia a la insulina (sspg), el nivel de glucosa en una persona en ayunas (glufast) y el peso relativo expresado como el cociente entre el peso de la persona a la que se le hace el estudio y el peso esperado que debería tener una persona con la misma altura.

Con el fin de evaluar la eficacia de una clasificación resultante de aplicar los métodos que se describen en los tres primeros capítulos, se tiene una clasificación de antemano como sigue. Se distingue diabetes subclínica, diabetes manifiesta y normal. La diabetes manifiesta es el estado más avanzado y se caracteriza por elevada concentración de glucosa en sangre. La diabetes de tipo química no tiene síntomas de diabetes pero presenta anormalidad en la tolerancia a la glucosa.

En la Figura 1.2 se describen tres diagramas de caja (por cada estado de enfermedad) para la variable tolerancia a la glucosa. Los datos tienen más dispersión para el grupo diabetes alta. Para los grupos diabetes química y diabetes alta los datos presentan asimétría.

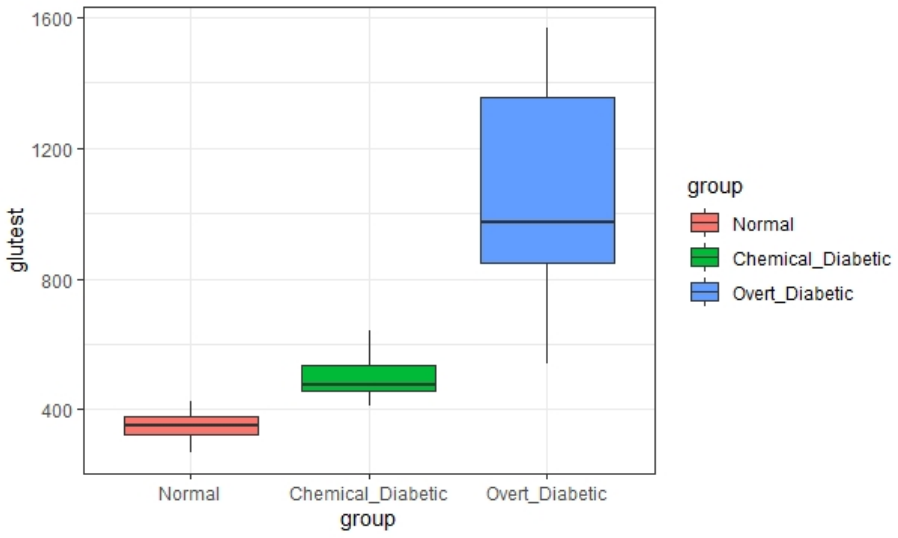

Figura 1.2: Para la variable tolerancia a la glucosa, aparecen tres boxplots, uno por cada estado de enfermedad

En la figura 1.3 se puede ver un boxplot para la variable resistencia a la insulina. La dispersión de las cajas para los distintos grados de diabetes se parecen. Para el grupo

diabetes normal, se observan datos atípicos por encima de la caja.

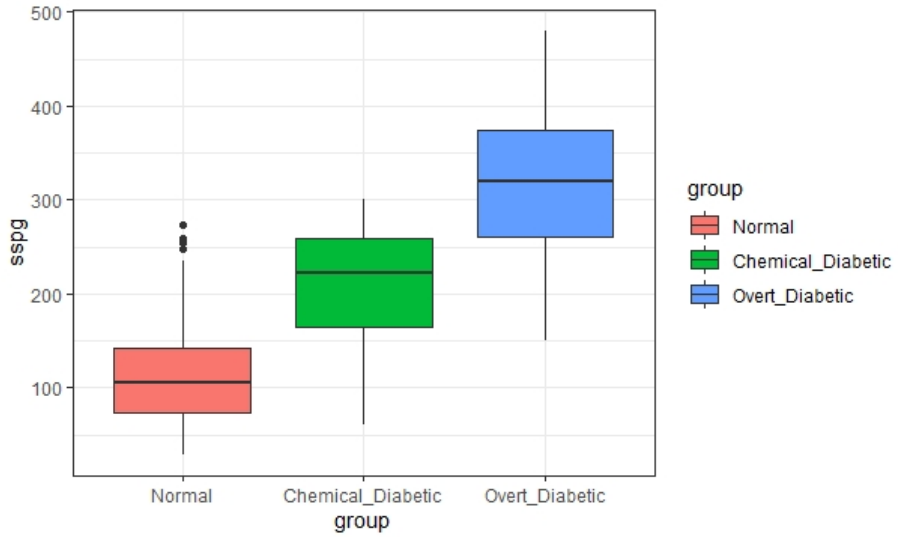

Figura 1.3: Para la variable resistencia a la insulina, aparecen tres boxplots, uno por cada estado de enfermedad

En la figura 1.4 se recogen los boxplot de las restantes variables.

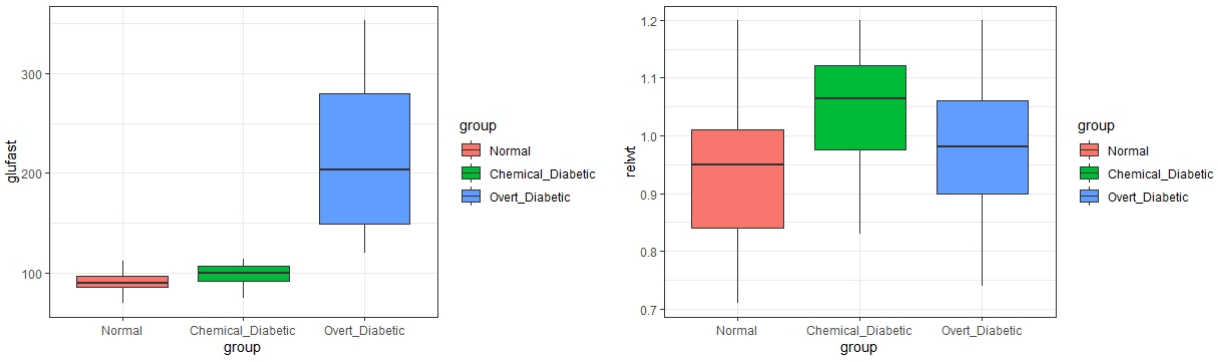

Figura 1.4: A la izquierda se representan los diagramas de caja para la variable nivel de glucosa para una persona en ayunas y a la derecha para el peso relativo.

Base de datos de baloncesto

La tercera base de datos [5] recoge los datos todos los jugadores en una temporada regular en la liga americana de baloncesto. Las variables con las que analizarán las observaciones serán los tiros hechos en la temporada, el porcentaje en tiros de campo, los puntos por partido que se promedian y los rebotes por partido que se promedian.
A través del clustering, para algunos pares de variables, se hará una interpretación acerca de si se pueden agrupar los jugadores por su posición en el campo, y se discutirá si la posición se relaciona con mejores valores en alguna de las características de juego. Por ejemplo, un jugador alto que juega en el puesto de pívot (cerca del aro), podría pensarse que tiene mayor número de rebotes y mayor efectividad en tiros que un base (que suele jugar más lejos del aro).

1.4. Organización del trabajo

En este capítulo, el objetivo era introducir la estrategia para hacer clustering y empezar a explicar los posibles caminos que son viables para abordar la tarea de agrupar observaciones. Como ayuda para resumir todos los medios para hacer clustering, las principales fuentes utilizadas fueron las siguientes. En *Hartigan*, (1975) se detallan ejemplos de disciplinas donde es posible ejecutar clustering. En la siguiente parte, de recopilación de los métodos clustering, se sigue un esquema como el que se muestra en *Peña*, (2002). En la parte focalizada en el algoritmo k-medias, la bibliografía tomada como apoyo se recoge en *Jain*, (2009).
En el segundo capítulo, la vía elegida para llevar a cabo clustering será apoyarse en modelos de probabilidad. Será necesario calcular los parámetros que determinan las componentes del modelo. Esta tarea puede desempeñarse de varias formas pero la más habitual será echar mano de la función de verosimilitud. Un algoritmo de estimación de parámetros que se describirá será el algoritmo EM. Como tercera forma de efectuar clustering, y con el fin de compararlas usando conjuntos de datos reales en el último capítulo, se explicará el clustering basado en modelos no paramétricos. Este camino consiste en ver la frecuencia con que aparecen las observaciones y formar grupos según zonas con alta concentración de los datos. Como en los casos anteriores, se elegirá un algoritmo para representar este enfoque clustering, el algoritmo del cambio-medio.

Capítulo 2

Clustering basado en modelos paramétricos

Los modelos de probabilidad han sido propuestos desde hace mucho tiempo como una ayuda para realizar análisis clustering. Uno de los primeros análisis empleando modelos de mixturas fué propuesto por Karl Pearson hace algo más de 140 años. En este tipo de clustering, en general, se asume que los datos vienen de una mixtura de distribuciones de probabilidad (se supondrán mixturas normales), donde cada mixtura representa a un cluster diferente. Dado que todo gira en torno a un modelo, se necesita saber el número de componentes que tiene y cómo elegir si dos o más particiones están relativamente cerca. Para dar respuesta a este número, se utilizará el criterio de información de Bayes (BIC). Por lo tanto, se describe una metodología clustering basada en mixturas de normales multivariantes en las que el BIC es usado para comparar distintos modelos.
Las particiones se determinarán mediante una combinación de clustering jerárquico y un algoritmo denomidado algoritmo de esperanza máxima (EM). Así, además de obtener una clasificación, se tendrá una medida de incertidumbre acerca de la clasificación resultante.
Este capítulo se organiza de la siguiente manera: se hará un repaso de los modelos de mixturas, luego se explicará el algoritmo EM. En la Sección 2.3 se tratará el número de clusters que debemos elegir. En la última parte del capítulo se extenderá el clustering paramétrico y se darán las limitaciones de este tipo de clustering.(Véase *Fraley*, 2002) .

2.1. Modelos de mixturas

En el clustering basado en modelos paramétricos se asume que los datos son generados de una mixtura de distribuciones de probabilidad donde cada componente representa un cluster. Se define una mixtura finita de G componentes como sigue:

Definición 2.1. $f(x_j|\rho) = \sum_{i=1}^{G} \pi_i f_i(y_j|\theta_i)$

siendo $\rho = (\tau_1, ..., \tau_G, \theta_1, ..., \theta_G)$ los parámetros del modelo.
Dadas las observaciones $x = (x_1, ..., x_n)$, dada $f_k(x_i|\theta_k)$ la densidad asociada a la componente k-ésima evaluada en la observación x_i, siendo θ_k sus parámetros. En el clustering basado en modelos de mixturas se tratarán de estimar los parámetros y a que cluster pertenece cada observación. Esto puede ser abordado de dos maneras como se indica en *Fraley*, (1998).

- El enfoque de clasificación por verosimilitud maximiza:

$$L_C(\theta_1, ..., \theta_G; \rho_1, ..., \rho_n|x) = \prod_{i=1}^{n} f_{\rho_i}(x_i|\theta_{\rho_i}) \tag{2.1}$$

  donde ρ_i son valores discretos etiquetando la clasificación: $\rho_i = k$ si x_i pertenece a la k-ésima componente (grupo).

- El enfoque de verosimilitud con mixturas maximiza:

$$L_M(\theta_1, ..., \theta_G; \tau_1, ..., \tau_n|x) = \prod_{i=1}^{n} \sum_{k=1}^{G} \tau_k f_k(x_i|\theta_k) \tag{2.2}$$

  donde τ_k es la probabilidad de que una observación pertenezca al k-ésimo grupo.

Siguiendo la literatura (*Fraley*, (1998)) se emplearán como f_k la densidad de una normal multivariante. Por lo tanto, los parámetros con los que se trabaja serán el vector de medias μ_k y la matriz de covarianzas Σ_k y la densidad tendrá la forma:

$$f_k(x_i|\mu_k, \Sigma_k) = \frac{\exp\{-(x_i - \mu_k)^T * \Sigma_k^{-1} * (x_i - \mu_k)/2\}}{(2\pi)^{p/2}| * \Sigma_k * |^{1/2}} \tag{2.3}$$

Los clusters van a ser elipsoides centrados en las medias μ_k. La matriz de covarianzas Σ_k determina las restantes características geométricas. Así, todas las matrices de covarianzas fuesen diagonales (con el mismo valor en todos los coeficientes), entonces todos los clusters serían esféricos con el mismo tamaño. En cambio si todas las matrices de covarianzas fuesen constantes, todos los clusters tendrían la misma geometría pero no necesariamente tendrían que ser esféricos.
Cabe destacar que se puede descomponer esta matriz de covarianzas en un producto de matrices más sencillas de estimar, que se calcularán utilizando los autovalores de la matriz de covarianzas. Cada una de ellas indicará una característica geométrica del cluster. (Véase *Fraley*, (1998)).

2.2. Algoritmo de esperanza-maximización (EM)

El algoritmo EM es un método iterativo que se utiliza para encontrar el estimador de máxima verosimilitud de un parámetro θ que proviene de una distribución paramétrica de probabilidad. Se engloba en el enfoque de verosimilitud con mixturas. Como se indica en *Fraley*, (1998), el algoritmo alterna iterativamente entre hacer conjeturas sobre los datos completos y encontrar el θ que maximiza la función de verosimilitud con los datos observados.
Se definen como y_i las observaciones compuestas por la parte observada x_i y z_i) que será la parte no observada y se define como:

$$z_{ik} = 1 \text{ si } x_i \text{ pertenece al grupo } k$$
$$z_{ik} = 0 \text{ en otro caso.}$$

Sea la densidad de x_i dada z_i:

$$\prod_{k=1}^{G} f_k(x_i|\theta_k)^{z_{ik}},$$

y cada z_i es indepediente e identicamente distribuida con distribución multinomial con G categorías y probabilidades $\tau_1, ..., \tau_G$.
La expresión del logaritmo de la función de verosimilitud para un modelo de mixturas con datos observados y no observados como los definidos anteriormente es:

$$l(\theta_k, \tau_k, z_{ik}|x) = \sum_{i=1}^{n} \sum_{k=1}^{G} z_{ik}[\log(\tau_k f_k(x_i|\theta_k))]$$

Teniendo en cuenta los datos que se tienen, los parámetros iniciales y la función de verosimilitud antes descrita, el algoritmo EM funciona de la siguiente forma:

- Paso de esperanza. Antes de buscar los parámetros que maximizan la función de verosimilitud, se deben estimar las cantidades z_{ik}. Para ello, con los datos observados, los pesos y los parámetros iniciales, se calcula la probabilidad de que los datos observados vengan de la i-ésima componente de la mixtura.

$$\hat{z}_{ik} = \frac{\hat{\tau}_k f_k(y_i|\hat{\theta}_k)}{\sum_{j=1}^{G} \hat{\tau}_j f_j(y_i|\hat{\theta}_j)}$$

- Paso de maximización. En este paso, utilizando los parámetros calculados en el paso anterior, sustituimos en el logaritmo de la función de verosimilitud para un modelo de mixturas con datos observados y no observados, y maximizamos obteniendo: nuevos parámetros correspondientes a las componentes de la mixtura y nuevos pesos.
  Actualizados los parámetros y pesos, repetiriamos el paso de esperanza hasta verificarse un criterio de parada.

En resumen, el algoritmo EM itera entre el paso de esperanza en el cual calcula los valores de $\hat{z}_{ik}$ (se parte de parámetros iniciales) y entre el paso de maximización, donde se maximiza una función obteniendo los parámetros que definen la mixtura.

2.2.1. Caso particular de mixturas de distribuciones normales

En el caso de las mixturas de distribuciones normales, el algoritmo EM funciona satisfactoriamente en la mayoría de los casos, sin embargo presenta algunos inconvenientes que motivan la búsqueda de otros enfoques clustering como el clustering basado en modelos no paramétricos. Se indicó como son las densidades que están presentes en la mixtura (se componen de la matriz de covarianzas). Pues bien, si la matriz de covarianzas está mal condicionada, el algoritmo EM no arrojará resultados correctos. Otra desvantaja del algoritmo EM es su radio de convergencia, el cual es bastante lento, lo que significa que el algoritmo puede ser muy costoso computacionalmente. Como se destacaba con el algoritmo de las k-medias, el algoritmo EM no agrupa de forma correcta datos atípicos.

2.2.2. Relación entre el algoritmo k-medias y el algoritmo EM

La primera coincidencia entre el algoritmo k-medias y el algoritmo EM está en el procedimiento que siguen ambos algoritmos. Para obtener las particiones de los datos, se buscan cantidades mínimas, ya sean distancias mínimas entre observaciones y centroides (k-medias) o funciones de verosimilitud (en el algoritmo EM). Otra similitud entre ambos algoritmos es que interesa una distribución de los datos particular. El algoritmo k-medias proporciona resultados muy buenos cuando se tienen datos bien separados en clusters, considerando la distancia euclidiana generalmente para cuantificar la proximidad de los clusters. El algoritmo EM realiza un clustering óptimo si los datos se distribuyen de acuerdo a mixturas de normales (con varianza especificada para no complicar la tarea de hacer clustering).
Además en el caso normal multivariante con matriz de covarianzas normal esférica dentro de los clusters; los valores que hacen mínima la verosimilitud para calcular los parámetros de las componentes de la mixtura (en el algoritmo EM) coinciden con la suma de cuadrados de las distancias de las observaciones a los centroides (en el algoritmo k-medias).
Se recoge en *Hartigan*, (1975) un caso particular unidimensional:

$$\text{Dados } x_1, ..., x_j, x_{j+1}, ..., x_n$$

Se asume que los datos están ordenados de la siguiente forma:

$$x_1 \leq ... \leq x_j \leq x_{j+1} \leq x_n$$

Si se cumple:

$$\frac{x_j - x_1}{x_{j+1} - x_j} \text{ y } \frac{x_n - x_{j+1}}{x_{j+1} - x_j} \text{ son suficientemente pequeños;}$$

entonces el algoritmo k-medias y el algoritmo EM asumiendo una mixtura con igual varianza, arroja como resultado clustering:

$$\{x_1, ..., x_j\}, \{x_{j+1}, ..., x_n\}.$$

2.3. Elección del número de clusters

Como vimos en la sección anterior, los parámetros de cada componente se pueden estimar utilizando el algoritmo EM, pero primero se necesitará saber el número de componentes de la mixtura. Esta cantidad está relacionada con la selección del modelo de forma que si el modelo es más sencillo, entonces el número de grupos necesarios para hacer una buena agrupación de los datos aumentará. Por ejemplo, si los datos se distribuyen como un elipsoide alargado y se ajusta un modelo esférico, se necesitarán varios modelos esféricos para aproximar el modelo elipsoidal.

Una ventaja del enfoque clustering basado en modelos de mixturas es que se puede usar el factor de Bayes para comparar modelos. Esto proporciona unas medidas no sólo para la parametrización del modelo (proporciona un equilibrio entre la bondad de ajuste del modelo y la simplicidad del mismo), sino también para el número de clusters que serán apropiados.

La idea básica recogida en $Fraley, (2002)$ es que dados varios modelos $M_1, ..., M_K$ con probabilidades $P(M_i), i = 1, .., K$ (a menudo cogidas iguales); la probabilidad a posterior del modelo M_k dados unos datos x, es proporcional a la probabilidad de los datos dado el modelo M_k.

$$P(M_k|x) \approx P(x|M_k)P(M_k)$$

Cuando los parámetros del modelo son desconocidos, la probabilidad de los datos dado el modelo M_k se calcula (por la ley de las probabilidades totales) de la forma:

$$P(x|M_k) = \int P(x|\theta_k, M_k)P(\theta_k, M_k)d\theta_k$$

donde $P(x|\theta_k, M_k)$ es la distribución conjunta a priori de θ_k (el vector de parámetros del modelo k). La cantidad resultante de calcular la anterior integral se denomina verosimilitud integrada.

La principal dificultad al usar el factor de Bayes es el cálculo de la integral que define la verosimilitud integrada. Para modelos regulares, se aproxima mediante el BIC:

$$2\log P(x|M_k) \approx 2\log P(x|\hat{\theta}_k, M_k) - \nu_k \log(n) = BIC_k$$

donde ν_k es el número de parámetros independientes estimados en el modelo M_k.

El convenio usado para calibrar diferencias entre el BIC se describe en $Fraley$, (2002). Si las diferencias son menos que dos unidades, entonces se concluye evidencia débil de disparidad, diferencias en el intervalo entre dos y seis significan evidencia positiva de disparidad, y a partir de seis se habla de evidencia fuerte. Además, el elegiremos el modelo de mixturas con k componentes que tenga menor BIC_k

2.4. Estrategia para realizar clustering paramétrico

El clustering jerárquico de tipo aglomerativo, con términos gaussianos, a menudo proporciona buenas particiones aunque no las más óptimas. El algoritmo EM puede refinar esas particiones cuando arranca con valores suficientemente cercanos al valor óptimo. Una estrategia tomada de $Fraley$, (1998) para realizar clustering es:

- Determinar el número máximo de clusters T y un conjunto de candidatos a parametrizaciones del modelo gaussiano a considerar (si por ejemplo se considera una matriz de varianzas constante o de otra forma). En general T debería de ser lo más pequeña posible.

- Hacer un clustering jerárquico para el modelo gaussiano sin restricciones (2.3), obteniendo las clasificaciones hasta los T grupos.

- Aplicar EM para cada número de clusters y cada parametrización, empezando con la clasificación del clustering jerárquico.

- Calcular el BIC de cada modelo con un grupo con las diferentes parametrizaciones. Calcular también el BIC a partir de la verosimilitud de mixturas para $2, ..., T$ grupos. Esto proporciona una matriz de valores BIC correspondientes a cada posible combinación de parametrización y número de clusters.

- Graficar los valores del BIC para cada modelo y elegir el adecuado. Queda así descrita la parametrización del modelo y el número de clusters que resulta.

Es importante aplicar este procedimiento en un número de clusters tan grande como sea necesario.

Nota En este capítulo, se suponen modelos paramétricos de mixturas. Se componen por modelos más sencillos con parámetros que se han estimado por máxima verosimilitud. Sin embargo, se pueden estimar de otras formas, por ejemplo mediante el método de los momentos como se puede ver en *Pearson K.*, (1894). Aunque con los avances en computación, el algoritmo EM desplazó al ajuste por el método de los momentos, sigue teniendo interés en ciertos casos como en mixturas de normales con misma varianza. Véase *Furman W.D.*, *Lindsey B.G.*, (1994).
Se distinguirá también el clustering paramétrico extendido a mixturas de distribuciones no normales. Un caso importante son las t-distribuciones. Para este tipo de mixturas *McLachlan* y *Peel*, (1998) implementaron el algoritmo EM y su variante ECM (esperanza-condicional maximización) para realizar la estimación por verosimilitud de los parámetros. Otro algoritmo igualmente válido fue propuesto por Merg y Rubin, reemplazando el paso de maximizar del algoritmo EM por un número computacionalmente simple de etapas de maximización condicional.

2.5. Limitaciones del clustering paramétrico

Los métodos clustering basados en modelos de mixturas normales multivariantes tienen mucho éxito en diversas aplicaciones, sin embargo, su uso está limitado ante datos con alta-dimensionalidad o datos masivos.

Datos de alta dimensión Una limitación del clustering basado en modelos con datos multidimensionales es el número de parámetros por componente en las mixturas normales multivariantes. Además, si la dimensión de los datos es grande comparada con el número de observaciones, la covarianza estimada en modelos elipsoidales será a menudo singular, provocando que el algoritmo EM falle.
Una estrategia ligada con datos multidimensionales es la reducción de la dimensión. Si las correlaciones son evidentes entre las variables, se puede seleccionar un subconjunto de ellas para solucionar el problema de la alta dimensión. Las componentes principales son usadas a menudo para reducir la dimensión, pero a veces, transformar los datos teniendo en cuenta sus componentes principales puede dificultar la tarea de hacer clustering.
Otro enfoque de los datos multidimensionales es reemplazar los datos por distancias o diferencias entre los datos. Esto se aplica en recuperación de información, donde cada dimensión se corresponde con una palabra o término que puede aparecer o no en el documento.

Grandes conjuntos de datos Una de las razones por las que el interés en agrupar datos es cada vez mayor es el deseo de usarlo en conjuntos de datos muy grandes. El clustering basado en modelos no es una herramienta aconsejable frente a estos datos. La mayor

limitación es que el tiempo eficiente en los métodos jerárquicos basados en modelos tiene requisitos de memoria inicial proporcionales al número de grupos en la partición inicial, que en general, asocia cada observación con un grupo donde hay un sólo elemento asignado. A veces se reduce esta memoria agrupando ciertas observaciones juntas por adelantado.
Cuando el tamaño muestral es moderadamente grande, una solución propuesta en *Fraley*, (2002), es habitual coger una muestra aleatoria de los datos y después aplicar clustering basado en modelos a la muestra. Los resultados serán extendidos a toda la muestra de datos.
En este contexto, como se puede ver en *Bradley.P.S.*,*Fayyad.U.*, *Reina.C.*(1998) surgieron técnicas computacionales para hacer el algoritmo EM más eficiente cuando tenemos grandes conjuntos de datos. Un manejo es el desarrollo de los métodos de un paso, llamados así dado que los datos necesitan ser cargados en memoria una sola vez.

Capítulo 3

Clustering basado en modelos no paramétricos

Como se introduce en *Chacon*, (2020); la media, mediana y la moda son las medidas más frecuentes a la hora de determinar la tendencia de un archivo de datos. Es más frecuente emplear la mediana o la media, aunque hay situaciones en que la moda (o modas en el caso de haber varias) puede ser la medida más recomendable. Un ejemplo donde la moda sería una medida más precisa de localización que la media o la mediana sería en datos que se agrupan en gran parte al principio y los restantes se reparten en un intervalo grande (distribución asimétrica de los datos, que provoca que la media se infle).
La definición de moda, en el sentido clásico, es la observación que más se repite. Pero cuando las observaciones provienen de una variable aleatoria continua, dado que todos los valores observados son diferentes con probabilidad una, esta definición no debe de ser empleada. En este caso, se define la moda como el valor o valores donde la función de densidad alcanza un máximo.
En la tarea de realizar clustering, se distinguen además de los métodos paramétricos descritos en el capítulo anterior, los métodos basados en modelos no paramétricos. A partir de la función de densidad o generalmente de una estimación (ya que la densidad no suele ser conocida), habrá una correspondencia entre zonas con alta densidad de puntos rodeadas de zonas con menos densidad y clusters. Se define una mixtura de densidades de la siguiente manera:

$$f = \sum_{k=1}^{L} \pi_k f_k \tag{3.1}$$

El capítulo se dividirá de la siguiente forma: se introducirá la estimación modal (de una única moda o varias), la cual se divide según la forma de calcular los estimadores de las modas en estimación modal directa o indirecta. Luego se hará un repaso de uno de los

métodos de clustering modal no paramétrica, el algoritmo de mean-shift. Finalmente se enunciarán algunas ventajas de este tipo de clustering, que expliquen el interés de su uso.

3.1. Clustering modal

Cualquier distribución cuya densidad tenga más de un máximo local se llamará multimodal. Las distribuciones multimodales se usan para reflejar la existencia de varias subpoblaciones dentro de la distribución. Pueden ser modeladas como una mixtura de densidades definida anteriormente en 3.1. Además, los pesos de las mixturas ($\pi_l > 0$) suman 1 y cada subpoblación se representa con las densidades f_l. Cabe destacar que el número de componentes de la mixtura y el número de modas no tiene porque coincidir, a no ser que las componentes estén suficientemente separadas.
A menudo, la estimación de las modas se abordará de la misma manera que en el caso unimodal, esto es, considerando los máximos relativos de la densidad estimada $\hat{f}$. Una cuestión de mucha importancia es el número de modas que se deben usar. En general, se suele contrastar la importancia de una moda para saber si se debe añadir o no o se puede ver con herramientas gráficas, como por ejemplo el paquete multimode desarrollado por *Ameijeiras − Alonso J., Crujeiras R.M. y Rodriguez − Casal*, 2016.
El objetivo del análisis cluster será obtener una partición de la muestra completa, donde cada conjunto de la partición se asocia con los dominios de atracción de las modas. Por ejemplo, dada una muestra unidimensional, con función de densidad f. Los valores donde la función de densidad alcance sus máximos serán representativos para seleccionar los clusters y los puntos donde haya mínimos serán de utilidad para delimitar los clusters. Con problemas de suavidad de la densidad, el clustering modal puede fallar. Se distinguen dos métodos para efectuar clustering modal:

1. Métodos de caza de modas (mode hunting): se calculan las modas de la densidad y después se asocian las observaciones a alguna moda. Un ejemplo de estos métodos es el algoritmo de cambio medio (en el siguiente capítulo se explicará este algoritmo). *Li*, 2007 propuso un método que mezcla el clustering paramétrico (suponiendo una estructura de mixturas) con métodos no paramétricos (estimación de la densidad tipo núcleo). Este método extiende al algoritmo EM descrito en el capítulo anterior, y basicamente encuentra máximos locales en la mixtura de tipo núcleo dada.

2. Métodos basados en conjuntos de nivel: se asocian modas a regiones con alta densidad (o estimación de la densidad). Para explicar como funcionan estos métodos es necesario definir los conjuntos de nivel $N(\lambda)$.

$$N(\lambda) = \{x \in \mathbb{R}^d : f(x) \geq \lambda\},\ 0 \leq \lambda \leq \text{máx}\, f.$$

Entonces para cualquier elección λ, se evalúa la función anterior obteniendo una partición de clusters. Cuando no se tiene la función de densidad, se sustituye la misma por la densidad estimada.

El problema está en saber si este valor es conocido o si existe, y aún existiendo, si puede agrupar correctamente todos los clusters. Además, no siempre es sencillo encontrar las componentes conectadas, lo cual si pasa en el caso unidimensional donde los conjuntos conectados son intervalos. Este problema se puede solucionar utilizando herramientas gráficas como los árboles de cluster, que es una estructura jerárquica que cuenta el número de componentes conectadas para un nivel de confianza λ que va cambiando.

En *Menardi*, (2016) hai un ejemplo donde se ve como funcionan los árboles cluster. Dados unos datos, donde su función de densidad tiene tres modas m_1, m_2 y m_3, ordenadas de menor a mayor densidad. Empezando con $\lambda = \lambda_0 = 0$, se obtiene un conjunto de nivel y por lo tanto el árbol cluster muestra una rama. Se aumenta el valor de λ, considerando ahora λ_1, obteniendo dos componentes principales. Una de las componentes (C_1) está asociada con la moda con menor frecuencia, que representa un núcleo del cluster, pero seguimos aumentando el valor de λ hasta tener tantas modas como componentes. Con $\lambda = \lambda_2 = f(m_1)$, la componente asociada al la moda m_1 desaparece, en cambio la otra se mantiene. Cogiendo un λ_3 mayor, la componente C_2 se divide en dos componentes que se asocian con las modas m_2 y m_3. En el árbol cluster aparecen dos ramas con esta última elección de λ. Se tienen entonces los cluster, que serán las componentes resultantes. Si hubiese alguna observación fuera de las componentes, se incluiría en uno de los clusters de forma jerárquica.

3.2. Algoritmo de cambio medio

El algoritmo de cambio medio (o algoritmo mean-shift), se localiza entre los procedimientos de caza de modas. Dada una observación, se clasificará en un grupo u otro dependiendo de la moda más cercana moviéndose en la dirección del gradiente de una aproximación de la función de densidad. Siguiendo lo descrito en *Menardi*, (2016), se considera el estimador tipo núcleo:

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K_H(x - x_i) \tag{3.2}$$

H será la matriz ventana, que es simétrica definida positiva. Además $K_H(x) = |H|^{-1/2} K(H^{-1/2}x)$. Sea $x^{(0)}$ un punto genérico, sea $w_i(x^{(s)})$ el vector de pesos de las componentes de x_i en la

etapa s. El algoritmo de cambio medio funciona de la siguiente manera:

Algoritmo 3.1. $x^{s+1} = \sum_{i=1}^{n} w_i(x^{(s)})x_i = x^{(s)} + [\sum_{i=1}^{n} w_i(x^{(s)})x_i - x^{(s)}] = x^{(s)} + M(x^{(s)})$

Entonces $M(x^{(s)})$ se denota por cambio medio, es lo que varía un punto con respecto al anterior. El algoritmo finaliza cuando el cambio medio es menor que una cantidad dada en un criterio de parada.
En conclusión, el algoritmo consiste en el desplazamiento desde un punto inicial x a otro (punto de mayor densidad) que resulta del promedio de los pesos de los datos dentro de una vecindad determinada por una región centrada en x. Los pesos están relacionados con las distancias al origen y sus valores quedan determinados por el núcleo definido en 3.2.

Algoritmo de cambio medio como un método gradiente

Como se dijo anteriormente, el cambio medio es una estimación intuitiva del gradiente de la densidad. En el caso de tener un núcleo gaussiano, se describe en $Menardi$, (2016) el algoritmo:

$$x^{s+1} = x^{(s)} + a\frac{\nabla \hat{f}(x^{(s)})}{\hat{f}(x^{(s)})}$$

Suele elegirse el gradiente normalizado para aumentar la velocidad de convergencia.
El algoritmo de cambio medio tiene sus ventajas e inconvenientes para elegirlo por delante de otros métodos como pueden ser los que se discutieron en capítulos anteriores. Es interesante dado que es un algoritmo simple de entender e implementar. Otra ventaja es que no necesita un número de grupos especificado de antemano, como si sucedía con otros algoritmos como k-medias. Sí se especifica un ancho de ventana, que va a estar directamente relacionado con el número de clusters que aparezcan como resultado. Si cogemos una ventana pequeña en la estimación de la densidad, el número de clusters obtenidos será mayor.

3.3. Elección del clustering modal

La virtud principal del clustering modal radica en el procedimiento usado. Asumiendo una verdadera estructura en la población, se puede representar una partición ideal que los métodos clustering tratan de aproximar. La clasificación inicial será usada como punto de referencia para evaluar un clustering o comparar alternativas. El número de clusters en el clustering modal pasa a ser una propiedad intrínseca del mecanismo generador de datos, por lo tanto su determinación es parte del proceso de estimación. Cabe destacar que los clusters obtenidos mediante clustering modal no se limitan a una forma particular. Esto hace que se vean los clusters como agrupaciones más naturales de los datos. Esto no pasa en modelos basados en mixturas, cuyos clusters tienen una forma predeterminada. Otro

motivo para la elección del clustering modal por delante del clustering basado en distancias es que para cada observación, proporciona una probabilidad de pertenecer a los clusters (como sucedía en el clustering basado en modelos paramétricos). La confianza máxima se asocia a las observaciones cercanas a las modas.
Sin embargo, el clustering modal también presenta ciertas limitaciones. Hay muchas ocasiones donde no se puede garantizar tener una función de densidad en los dominios de atracción de las modas, por ejemplo cuando no se tienen condiciones de regularidad en la función de densidad o cuando los datos no son continuos (con variables discretas o categóricas). Además es muy complejo computacionalmente, por el número de operaciones que se requieren y por el número de iteraciones que son necesarias para llegar a la convergencia del algoritmo.

Capítulo 4

Datos reales

En este capítulo, en cada sección se utilizará una base de datos de las descritas en el primer capítulo. Como punto de partida se trabajará con los datos de crímenes en USA. Con el fin de ver claramente como agrupa y como se explican esos grupos, se trabajará con dos dimensiones. Los enfoques utilizados en el primer conjunto de datos serán el enfoque basado en distancias y el enfoque basado en modelos paramétricos. En el segundo capítulo, se realizará el clustering no paramétrico al conjunto de datos de glucosa. Se hará una comparativa, en términos de datos bien clasificados disponiendo de una clasificación inicial, de los tres métodos. En la última sección, se hará una interpretación con el enfoque basado en modelos no paramétricos, de los datos relativos a la eficacia de lanzamientos en la liga americana de baloncesto.

4.1. Base de datos de críminalidad en USA

Para esta base de datos, el objetivo será clasificar conjuntamente los estados de USA según similitud en el número de asesinatos (murder) y número de asaltos (assault). El algoritmo k-medias realiza la clasificación para 3 clusters descrita en la Figura 4.1. En la misma figura se arroja la clasificación con el enfoque clustering basado en modelos paramétricos (mixtura de 3 distribuciones gaussianas).

La clasificación real de los estados según la zona geográfica donde se encuentren será la siguiente. Los estados costeros del sur son: Alaska, California, Arizona, Nuevo Méjico, Texas, Louisiana, Missisipi, Alabama, Florida, Hawai, Georgia, Carolina del Sur, Carolina del Norte, Islas Vírgenes.

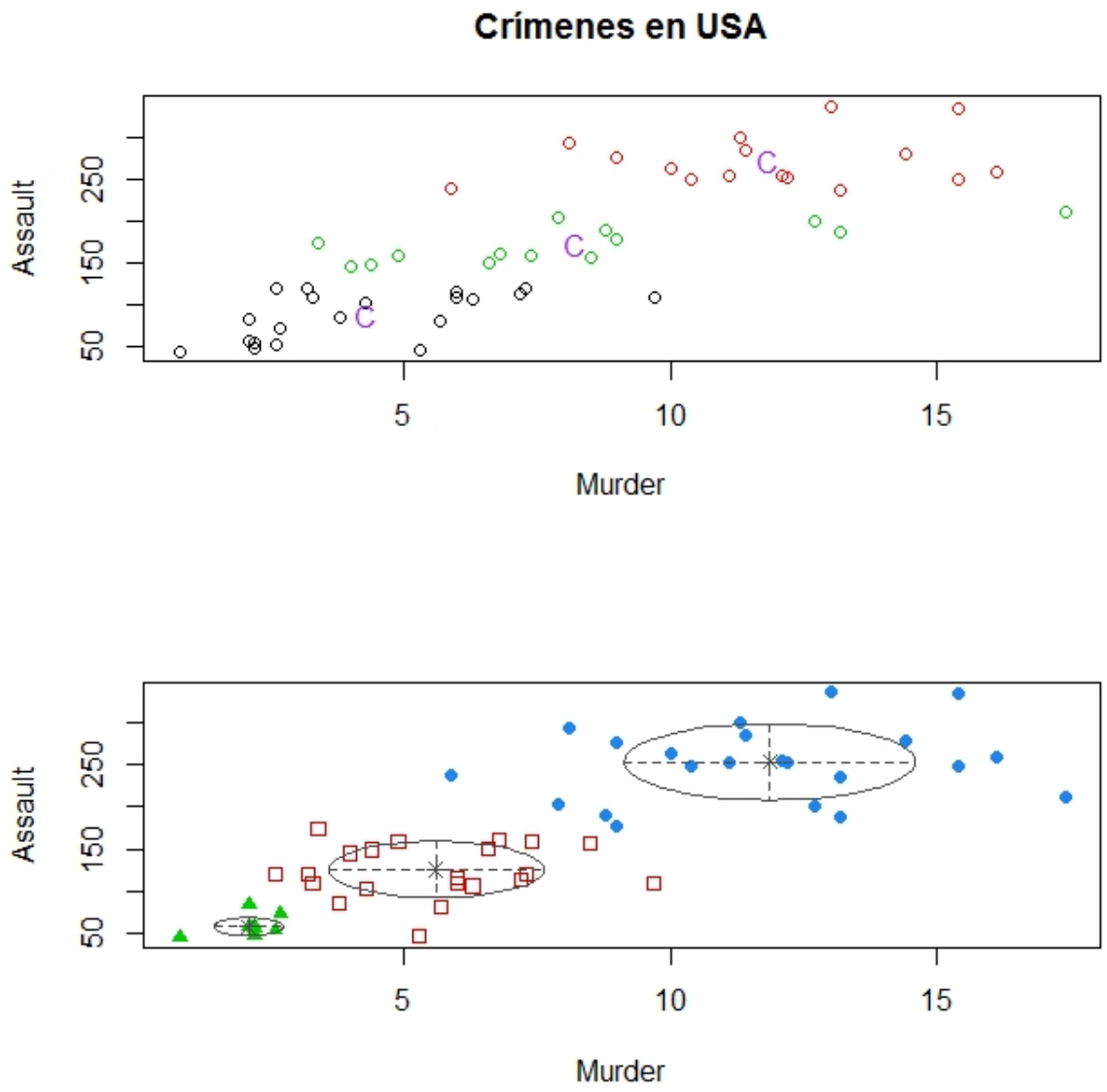

Figura 4.1: Resultado del análisis cluster. Panel superior: algoritmo k-medias. Panel inferior: algoritmo EM.

Los estados costeros del norte son: Oregon, Washington, Virginia, Delaware, Maryland, New Jersey, Connecticut, Rhode Island, Massachusetts,New Hampshire, Maine, New York. Los estados no costeros son: Idaho, Nevada, Montana, Wyoming, Colorado, Dakota del Norte, Dakota del Sur, Nebraska, Kansas, Oklahoma, Minnesota, Iowa, Missouri, Arkansas, Wisconsin, Illinois, Michigan, Indiana, Kentucky, Tenessee, Ohio, Virginia del Este, Pennsylvania, Vermont.

La clasificación empleando k-medias de los estados según la criminalidad es la siguiente. Los estados con menor criminalidad son: Connecticut, Hawai, Idaho, Indiana, Iowa, Kansas, Kentucky, Maine, Minnesota, Montana, Nebraska, New Hampshire, Dakota del Norte, Ohio, Pennsylvania, Dakota del Sur, Utah, Vermont y Wisconsin. Los estados con criminalidad intermedia son: Alabama, Alaska, Arizona, California, Delaware, Florida, Illinois, Louisiana, Maryland, Michigan, Missisipi, Nevada, Nuevo Méjico, Nueva York, Carolina del Norte, Carolina del Sur. Los estados con mayor criminalidad son: Arkansas, Colorado, Georgia, Massachusetts, Missouri, Nueva Jersey, Oklahoma, Rhode Island, Tennessee, Texas, Virginia, Washington, Wyoming.

Como interpretación de la partición resultante al usar el algoritmo k-medias (teniendo en cuenta en qué cluster se agrupa cada estado y donde se sitúa geograficamente), los estados costeros situados más al sur son los que tienen mayor criminalidad (mayor número de asesinatos y asaltos). Los estados costeros situados hacia el norte y estados no costeros situados al sur son los que menor criminalidad arrojan. Como último grupo, los estados no costeros situados más al norte, tienen un número de asesinatos y asaltos que está entre los dos primeros. Se obtiene por tanto una clasificación con coincidencias según características geográficas como presencia de costa en el estado.

Entre el clustering basado en distancias y el clustering suponiendo modelos de mixturas, se puede observar por ejemplo que el grupo con menor criminalidad tiene menor cantidad de datos al agrupar con el algoritmo basado en modelos paramétricos. Pasa lo contrario con el grupo de mayor criminalidad.

4.2. Datos de glucosa

En la sección anterior, con un conjunto de datos reales, los métodos que se trabajaron fueron los propuestos en los dos primeros capítulos. Con el conjunto de datos de glucosa, que recoge medidas de glucosa e insulina en personas, el objetivo será trabajar con el algoritmo de cambio medio, que se explicaba en el Capítulo 3. Además se hará una comparación de los tres enfoques clustering.
La clasificación basada en modelos no paramétricos agrupa observaciones según la distancia a las modas tomando como dirección el gradiente de la densidad o de una estimación de la misma. Entonces puede ser de utilidad analizar el número de modas antes de aplicar cualquier algoritmo clustering basado en modelos no paramétricos. En la Figura 4.2 se muestra una representación de la densidad de los datos de glucosa para las variables resistencia a la insulina y tolerancia a la glucosa.

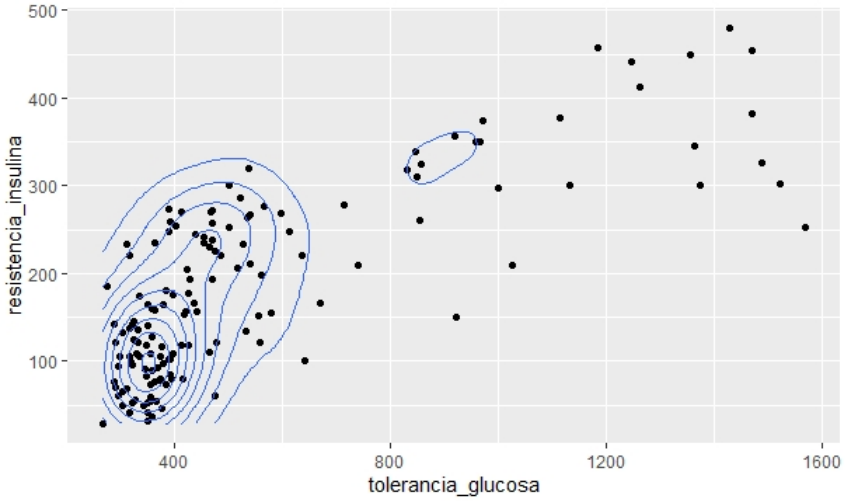

Figura 4.2: Gráfica para estudiar la frecuencia con que aparecen las observaciones. En el eje horizontal se encuentra la tolerancia a la glucosa y en el vertical la resistencia a la insulina

Se observan dos concentraciones de datos diferenciadas que aparecen entre elipsoides que indican la presencia de dos modas. Así, el número de grupos al realizar el clustering no paramétrico será el mismo que el número de modas y por lo tanto será dos. A continuación, se verá como agrupa el algoritmo mean-shift cambiando la ventana en la Figura 4.3.

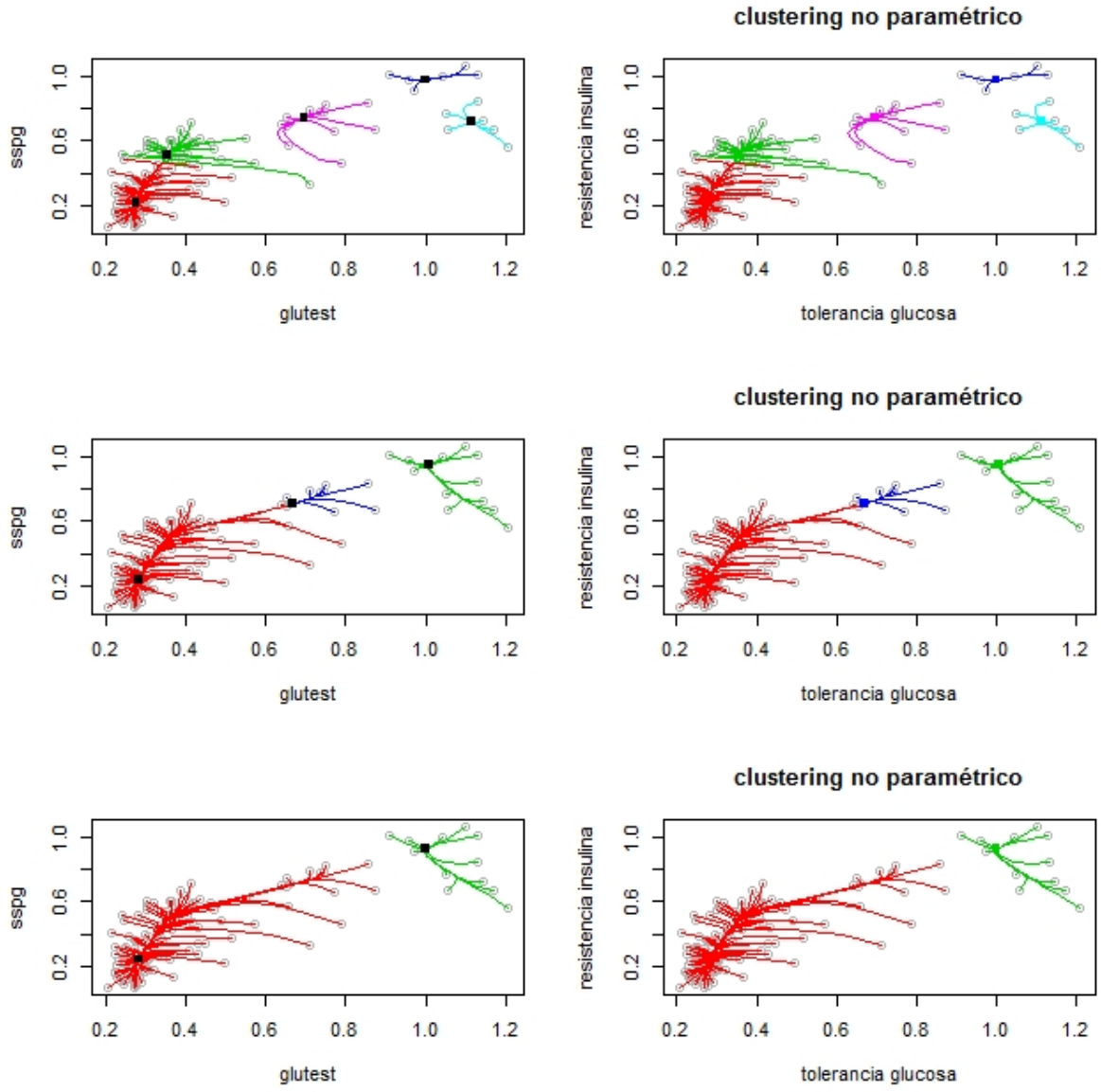

Figura 4.3: En la Figura se muestran tres particiones diferentes obtenidas al realizar clustering no paramétrico, en concreto el algoritmo del cambio medio. La diferencia entre las gráficas de la izquierda y derecha simplemente está en como se destacan los centros de los grupos. En la primera fila el selector de ventana es $h = 0{,}08$, en la segunda $h = 0{,}111$ y en la tercera $h = 0{,}12$.

Conforme se aumenta el selector de ventana, el número de grupos disminuye. Se puede

analizar el porcentaje de observaciones bien agrupadas tomando los selectores anteriores, y con el aumento del tamaño de la ventana se obtiene un peor porcentaje de acierto al agrupar. Como conclusión, es importante fijar el selector de ventana para el caso no paramétrico, dado que se los datos se dividen en tres grupos (clasificación real) y nuestra partición es de 10 grupos, habrá más datos mal clasificados.

4.2.1. Comparación de los tres enfoques con el conjunto de datos de glucosa

En esta parte, se hará una comparación de los clusters que resultan al trabajar con los distintos enfoques clustering (la cual se recoge en la Figura 4.4). Cabe destacar que esta comparación será para los datos de glucosa considerando el par de variables resistencia a la insulina y tolerancia a la glucosa. Como ayuda para comparar las vías para agrupar datos, utilizará la siguiente tabla.

	Enfoque distancias	Enfoque modelos de mixturas	Enfoque no paramétrico
Número de datos bien clasificados	110	127	85
Error (%)	24	12	41

Cuadro 4.1: Tabla de observaciones bien clasificadas y del error cometido en cada enfoque

Mediante el clustering no paramétrico, como se puede ver en el Cuadro 4.1, el error es el más alto de los tres, en el 41 por ciento de los casos, las observaciones se agrupan en el cluster equivocado. El enfoque basado en modelos de mixturas agrupa bien 127 de las 145 observaciones de las que se compone nuestra muestra, por lo tanto en este caso es el método que proporciona mejores resultados.
Analizando los clusterings resultantes que aparecen en la Figura 4.4, lo primero que destaca es el número de clusters en cada enfoque. En el clustering no paramétrico, no se fija de antemano un número de clusters, en cambio el algoritmo mean-shift agrupa las observaciones en 3 grupos distintos por la elección que se hizo de la ventana. Sin embargo, a la hora de implementar los algoritmos que se engloban en los marcos del clustering basado en distancias o en el clustering basado en mixturas, el número de clusters es un parámetro que se fija de antemano.

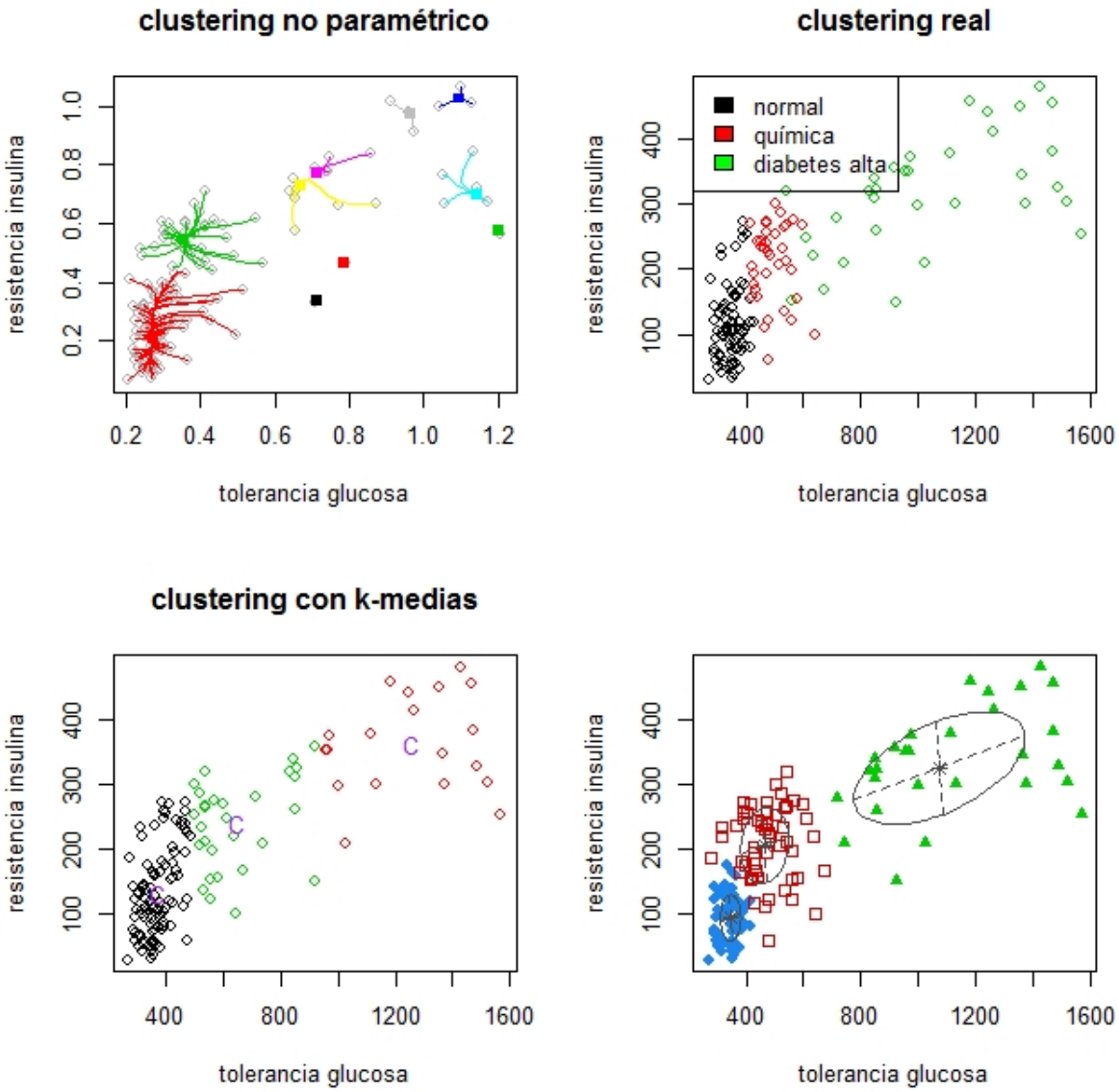

Figura 4.4: En primer lugar se muestra la partición resultante al emplear el algoritmo del cambio medio. A su derecha aparece la clasificación verdadera dada de antemano. En primer lugar, en la segunda fila, se puede ver como se agrupan las observaciones utilizando el algoritmo k-medias. En último lugar, está el resultado de usar el algoritmo EM (clustering paramétrico).

Otro aspecto interesante es la forma de los clusters con cada enfoque. En el clustering no paramétrico, las obseraciones se distribuyen en relación a las modas. En el clustering basado en distancias, las observaciones se reparten de acuerdo a cercanía respecto a los

centroides. En el enfoque basado en modelos de mixturas, los clusters tienen una forma elipsoidal, con tantos elipsoides como componentes hay en la mixtura.

4.2.2. Segunda comparación de los tres enfoques con el conjunto de datos de glucosa

En la anterior sección y en la primera parte de esta, se realizaba una clasificación para datos bidimensionales. Así era sencillo ver graficamente como se repartían los puntos en un diagrama de dispersión y diferenciarlos por grupo. En esta sección, la dimensión con la que se trabajará será 4. Se repetirá una tabla que aparece en el Cuadro 4.2 para interpretar cual de los enfoques clustering funciona mejor con los datos de glucosa considerando 4 variables.

	Enfoque distancias	Enfoque modelos de mixturas	Enfoque no paramétrico
Número de datos bien clasificados	119	126	88
Error (%)	18	13	39

Cuadro 4.2: Tabla de observaciones bien clasificadas y del error cometido en cada enfoque

Los resultados obtenidos son parecidos a los del caso bidimensional. En el enfoque basado en distancias, el número de datos bien clasificados sube ligeramente y pasa lo mismo con el enfoque no paramétrico. También sube el error en el caso del enfoque paramétrico, luego al realizar clustering basado en modelos de mixturas, hay más observaciones bien clasificadas con menor dimensión. La elección de un método por delante de otro con estos datos y dimensión es la misma que en el caso bidimensional. El enfoque más adecuado es el paramétrico seguido del enfoque basado en distancias.

4.3. Datos de tiros en la liga americana de baloncesto

En esta sección, el objetivo no va a ser comparar enfoques cluster para elegir un método antes que otro con una base de datos dada. Lo que interesa es interpretar las particiones que resultan y analizar si las posiciones de un jugador están relacionadas con mejores números en algunas de las características del juego.
En la Figura 4.5 se puede ver el clustering resultante al implementar el algoritmo de cambio medio, y la dispersión de los datos. Fijándose en la dispersión, los datos se concentran entre 0 y 50 tiros a canasta. Implicará la presencia de una moda y de un cluster. Hay observaciones en el rango de tiro de 50 a 200, pero no se distinguen zonas con alta densidad de datos, por lo tanto en ese rango aparecen distintos clusters pequeños.
A la vista de la dispersión, es cierto que los jugadores que tiran a canasta más veces juegan en las posiciones de alero y base. En el clustering no paramétrico, queda reflejado en la presencia de dos grupos distintos en color azul y rosa con pocas observaciones en cada uno. Por otra parte, entre 0 y 50 tiros, cualquier jugador (independientemente de la posición), realiza ese número de tiros. Se refleja en un grupo grande en el clustering donde hay jugadores de las cinco posiciones de campo.
Como idea general, para las características: tiros a canasta y porcentaje de tiros, los jugadores que más tiran con un porcentaje de acierto del 60 por ciento, juegan de aleros o bases. Además, tirando menos tiros con buenos o malos porcentajes, no quedan determinados grupos claros de jugadores según su posición en el campo.

En la figura 4.6 se recoge la dispersión de las variables porcentaje de acierto en tiros y rebotes por partido, y el clustering no paramétrico. Como en el caso anterior, interesa ver se los grupos que resultan se relacionan con posiciones del campo en concreto. Por ejemplo, es lógico pensar que un jugador alto que juega en la posición de pívot, tenga buenos números en rebote y porcentaje de acierto en tiros dado que juega cerca del aro.
Viendo el diagrama de dispersión, hay jugadores que tienen un porcentaje de tiros del 0 por ciento, independientemente de la posición. Esto se debe a que los datos son recogidos de todos los jugadores de la liga, aunque tuvieran un papel residual en sus equipos, pudiendo haber jugado menos de un minuto en toda la temporada regular. En el clustering se refleja en los grupos en color verde o rosa dependiendo del promedio de rebotes que tengan. Con un porcentaje del 70 por ciento en tiros aproximadamente y un número de rebotes por partido superior a 10, aparecen varios datos de pívots. En el clustering no paramétrico se refleja en un cluster azul. Como pasaba con las anteriores variables, entre el 60 y 90 por ciento (con

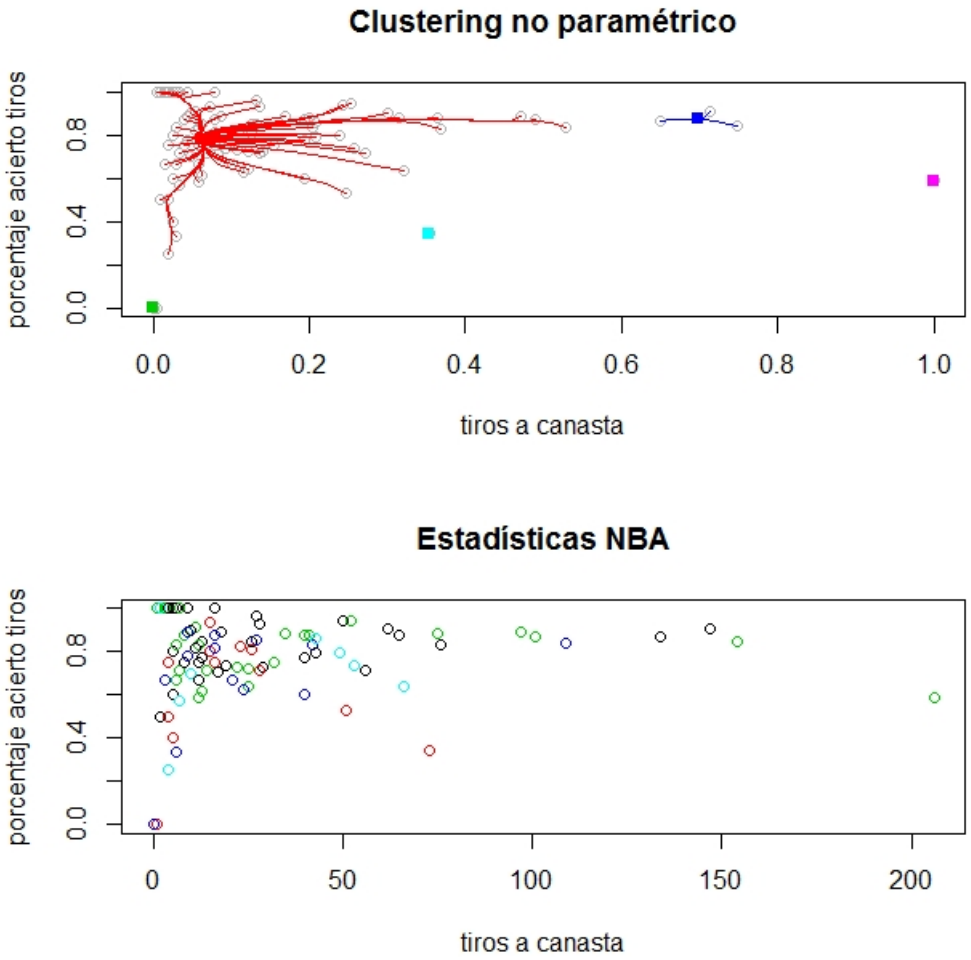

Figura 4.5: Clustering no paramétrico y clasificación real de las variables tiros realizados y porcentaje de tiros

.

un número de rebotes menor que 8) aparecen en el diagrama de dispersión observaciones de todas las posiciones de campo. Esto se refleja en un grupo rojo con muchos datos en clustering no paramétrico. Hay dos grupos en colores amarillo y azul, que se asocian a datos situados en zonas centrales con menor densidad de datos.

Como conclusión, considerando las variables rebotes por partido y porcentaje de acierto en tiros, el clustering no paramétrico determina que los pívots presentan como características de su juego más rebotes y porcentajes entre el 60 y 80. Más allá de este grupo, no hay evidencias de que el porcentaje de acierto y el número de rebotes esté relacionado con la posición del jugador en la pista de juego.

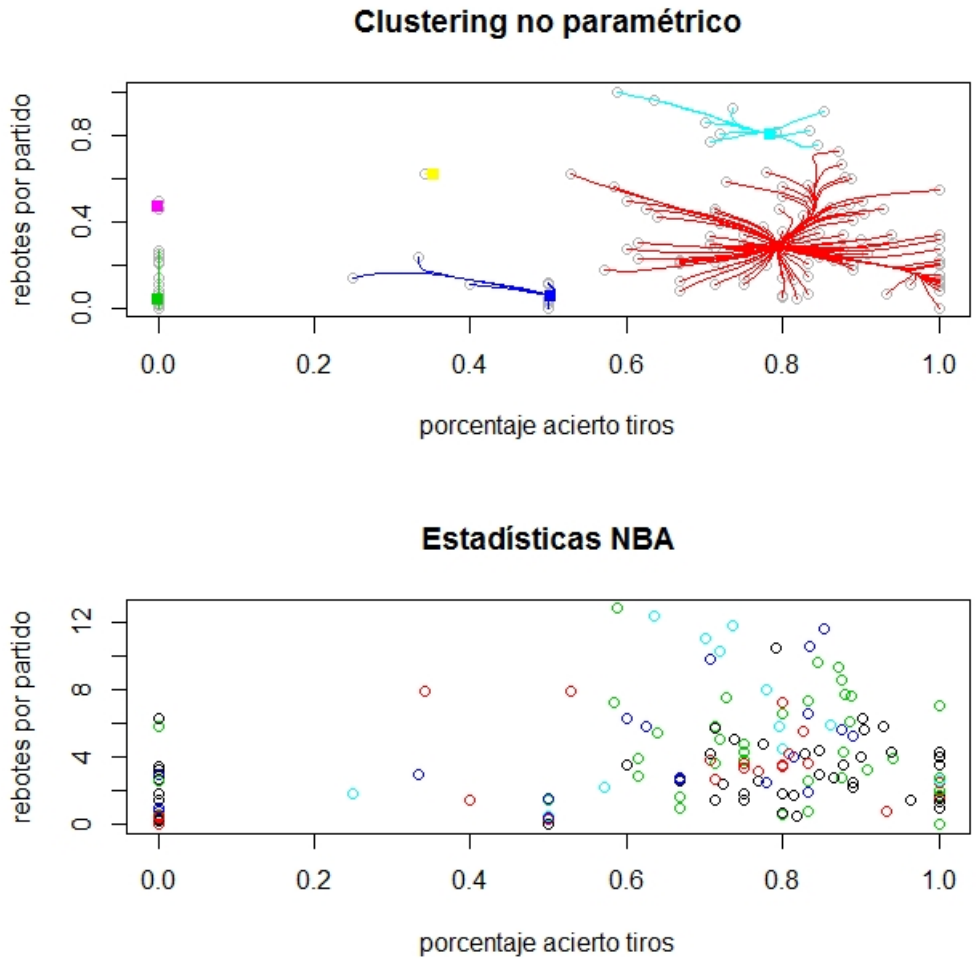

Figura 4.6: Clustering no paramétrico y clasificación real de las variables porcentaje de tiros de campo y rebotes promedio

Apéndice A

Anexo de código

Se añade el código usado en este capítulo para que se pueda reproducir. Para el conjunto de datos de crímenes, se usarán las técnicas basadas en el concepto de distancia (k-medias) y las técnicas basadas en modelos paramétricos (algoritmo EM).

```
set.seed(83)
USArrests
windows()
par(mfrow=c(2,1))
hist(datos$Murder,main="Histograma",xlab="Asesinatos",...
+ylab="frecuencia")
hist(datos$Assault,main="Histograma",xlab="Asaltos",...
+ylab="frecuencia")

#capítulo1--Clustering basado en distancias#
############################################
#k-medias
km=kmeans(USArrests,centers=3)
km$cluster
#capítulo2-clustering paramétrico##########
############################################
library(mclust)
set.seed(83)
clust_param=Mclust(datos,G=3)
summary(clust_param)
```

```
#representación gráfica
plot(datos,col=km$cluster,main="Crímenes en USA",xlab="Murder",ylab="Assault")
points(km$centers,col="purple",pch="C")
plot(clust_param,main="Crímenes en USA")
```

Para el conjunto de datos de glucosa, se trabajará con el enfoque no paramétrico y se verá cómo varía el resultado de hacer clustering cuando se cambia la ventana. Después se hará una comparación de todos los enfoques, viendo cuál es más adecuado para agrupar los datos.

```
library(heplots)
#lectura de datos
data(Diabetes)
names(Diabetes)
datos2=Diabetes[,3:5]
#boxplot por separado
library(ggplot2)
ggplot(Diabetes)+
geom_boxplot(aes(x = group, y = glutest, fill = group))+
theme_bw()
ggplot(Diabetes)+
geom_boxplot(aes(x = group, y = sspg, fill = group))+
theme_bw()
ggplot(Diabetes)+
geom_boxplot(aes(x = group, y = glufast, fill = group))+
theme_bw()
ggplot(Diabetes)+
geom_boxplot(aes(x = group, y = instest, fill = group))+
theme_bw()
#corrección de las variables
cluster=Diabetes$group
v=rep(0,145)
for (i in 1:length(cluster)){
if (cluster[i]=="Normal"){
v[i]="1"
} else if (cluster[i]=="Chemical_Diabetic") {
v[i]="2"
} else if(cluster[i]=="Overt_Diabetic"){
```

```
v[i]="3"
}
}
v
#densidad de los datos
par(mfrow=c(1,1))
library(ggplot2)
data_mode=cbind(Diabetes[,3],Diabetes[,5])
data_modas=as.data.frame(data_mode)
colnames(data_modas)=c("tolerancia a la glucosa","resistencia a la insulina")
tolerancia_glucosa=data_modas$‘tolerancia a la glucosa‘
resistencia_insulina=data_modas$‘resistencia a la insulina‘
m=ggplot(data_modas,aes(x=tolerancia_glucosa,y=resistencia_insulina))
m+
geom_point(aes(tolerancia_glucosa,resistencia_insulina)) +
stat_density2d()

windows()
par(mfrow=c(3,2))
plot(ms(datos2[,-2],h=0.08),xlab="tolerancia glucosa",...
+ylab="resistencia insulina",main="clustering no paramétrico"[1])
plot(ms(datos2[,-2],h=0.111),xlab="tolerancia glucosa",...
+ylab="resistencia insulina",main="clustering no paramétrico"[1])
plot(ms(datos2[,-2],h=0.12),xlab="tolerancia glucosa",...
+ylab="resistencia insulina",main="clustering no paramétrico"[1])

#comparación datos bien clasificados dependiendo del número de modas.
ms(datos2[,-2],h=0.111)
plot(ms(datos2[,-2],h=0.111),xlab="tolerancia glucosa",...
+ylab="resistencia insulina",main="clustering no paramétrico"[1])
clust_noparam2_h1=ms(datos2[,-2],h=0.08)
clust_noparam2_h2=ms(datos2[,-2],h=0.111)
clust_noparam2_h3=ms(datos2[,-2],h=0.12)
h1=sum(v==clust_noparam2_h1$cluster.label);h1
h2=sum(v==clust_noparam2_h2$cluster.label);h2
h3=sum(v==clust_noparam2_h3$cluster.label);h3
```

```
#representación correcta###########
###################################

windows()
par(mfrow=c(2,2))
plot(datos2[,-2],col=Diabetes$group,main="clustering real",...
+xlab="tolerancia glucosa",ylab="resistencia insulina")
legend(x="topleft",legend=c("normal","química","diabetes alta"),...
+fill=c("black","red","green"))

#métodos basados en distancias ##
#################################
#kmeans
set.seed(83)
clus_dist=kmeans(datos2[,-2],centers=3)

#reorganizo los nombres de los clusters
ckm=clus_dist$cluster
p=rep(0,length(clus_dist$cluster))
for (i in 1:length(clus_dist$cluster)){
if (ckm[i]=="1"){
p[i]="1"
} else if (ckm[i]=="2") {
p[i]="3"
} else if(ckm[i]=="3"){
p[i]="2"
}
}
p

##representación kmeans########
###############################
plot(datos2[,-2],col=clus_dist$cluster,xlab="tolerancia glucosa",...
+ylab="resistencia insulina",main="clustering con k-medias")
```

```
points(clus_dist$centers,col="purple",pch="C")

#enfoque basado en modelos, algoritmo EM###
##########################################
library(mclust)
clust_param=Mclust(datos2[,-2])
summary(clust_param)
plot(clust_param,xlab="tolerancia glucosa",...
+ylab="resistencia insulina",main="clustering paramétrico")

#clustering no paramétrico, algoritmo mean-shift
library(LPCM)
set.seed(83)
clust_noparam=ms(datos2[,-2],h=0.111)
plot(ms(datos2[,-2]),xlab="tolerancia glucosa",...
+ylab="resistencia insulina",main="clustering no paramétrico"[1])

##comparación de datos bien clasificados###########
####################################################
#resultados clustering

#con kmeans
#Número de datos bien clasificados
b1=sum(v==p);b1
#Porcentaje de datos bien clasificados;
e1=1-b1/length(Diabetes$group);e1

#con el algoritmo EM
clust_param$classification

#Porcentaje de datos bien clasificados
b2=sum(v==clust_param$classification);b2
e2=1-b2/length(Diabetes$group);e2

#Porcentaje de datos bien clasificados con el algoritmo mean-shift
b3=sum(v==clust_noparam$cluster.label);b3
```

```
e3=1-b3/length(Diabetes$group);e3

library(xtable)
tabla=matrix(c(b1,b2,b3,e1,e2,e3),nrow=2,ncol=3,byrow=T)
row.names(tabla)=c("Número de datos bien clasificados","Error (%)")
colnames(tabla)=c("Enfoque distancias","Enfoque modelos de mixturas",...
+"Enfoque no paramétrico")
tabla

print(xtable(tabla))
data(Diabetes)
data=cbind(Diabetes[,1:3],Diabetes[,5])

#corrección de las variables
cluster=Diabetes$group
v=rep(0,145)
for (i in 1:length(cluster)){
if (cluster[i]=="Normal"){
v[i]="1"
} else if (cluster[i]=="Chemical_Diabetic") {
v[i]="2"
} else if(cluster[i]=="Overt_Diabetic"){
v[i]="3"
}
}
v

#ahora no represento los datos#
#métodos basados en distancias ##
###################################
#kmeans
set.seed(83)
clus_dist=kmeans(data,centers=3)

#reorganizo los nombres de los clusters
ckm=clus_dist$cluster
```

```
p=rep(0,length(clus_dist$cluster))
for (i in 1:length(clus_dist$cluster)){
if (ckm[i]=="1"){
p[i]="1"
} else if (ckm[i]=="2") {
p[i]="3"
} else if(ckm[i]=="3"){
p[i]="2"
}
}
p

#enfoque basado en modelos, algoritmo EM###
############################################
library(mclust)
set.seed(83)
clust_param=Mclust(data)
summary(clust_param)

#clustering no paramétrico, algoritmo mean-shift
library(LPCM)
set.seed(83)
clust_noparam=ms(data,h=0.111)

##comparación de datos bien clasificados###########
####################################################
#resultados clustering

#con kmeans
#Número de datos bien clasificados
b1=sum(v==p);b1
#Porcentaje de datos bien clasificados;
e1=1-b1/length(Diabetes$group);e1

#con el algoritmo EM
clust_param$classification
```

```
#Porcentaje de datos bien clasificados
b2=sum(v==clust_param$classification);b2
e2=1-b2/length(Diabetes$group);e2

#Porcentaje de datos bien clasificados con el algoritmo mean-shift
b3=sum(v==clust_noparam$cluster.label);b3
e3=1-b3/length(Diabetes$group);e3

library(xtable)
tabla=matrix(c(b1,b2,b3,e1,e2,e3),nrow=2,ncol=3,byrow=T)
row.names(tabla)=c("Número de datos bien clasificados","Error (%)")
colnames(tabla)=c("Enfoque distancias","Enfoque modelos de mixturas",...
+"Enfoque no paramétrico")
tabla

print(xtable(tabla))
```

En la última parte del capítulo, el objetivo es que una vez se realizó un método clustering (en este caso mediante el enfoque clustering no paramétrico) en un conjunto de datos reales, poder analizar si los grupos resultantes tienen relación con valores concretos en las variables que componen los datos. Por ejemplo, si se consideran las variables cantidad de lanzamientos a canasta y porcentaje de tiros. La pregunta que se quiere responder es si tendrá relación una posición como la de base (que tiene el balón en sus manos en gran parte de la posesión) con tirar más a canasta o mejor.

```
###estudio datos NBA

#lectura de datos
library(readxl)
nba_stats=  read_excel("C:/Users/Toshiba/Desktop/TFM/2020-2021...
+ NBA Stats  Player Box Score  Advanced Metrics.xlsx")
names(nba_stats)
position=nba_stats$...4
```

```
FT=nba_stats$...12
PPG=nba_stats$...19
RPG=nba_stats$...20
NT=nba_stats$...11
set.seed(83)
indices <- sample( 1:nrow(nba_stats),170)
indices
data_nba=matrix(c(nba_stats[indices,]$...4,nba_stats[indices,]$...11,...
+nba_stats[indices,]$...12,nba_stats[indices,]$...19,...
+nba_stats[indices,]$...20),nrow=170,byrow=F)
data_nba=data_nba[-c(75),]
colnames(data_nba)=c("position","NT","FT","PPG","RPG" )
data_nba

#clustering no paramétrico####
##############################
library(LPCM)

#clustering no paramétrico para las dos variables tiros a canasta
#y porcentaje de acierto en tiros

data_1=as.data.frame(matrix(c(as.numeric(data_nba[,2]),...
+as.numeric(data_nba[,3])),nrow=169,ncol=2,byrow=F))
clust_np=ms(data_1,h=0.07)
clust_np$cluster.center
clust_np$cluster.label
plot(clust_np,main="Clustering no paramétrico",...
+xlab="tiros a canasta",ylab="porcentaje acierto tiros")

#representación gráfica
windows()
par(mfrow=c(2,1))
plot(clust_np,main="Clustering no paramétrico",...
+xlab="tiros a canasta",ylab="porcentaje acierto tiros")
plot(data_nba[,2],data_nba[,3],...
+main="Estadísticas NBA",col=clasif,xlab="tiros a canasta",...
```

```
+ylab="porcentaje acierto tiros")


#clustering no paramétrico para las dos variables porcentaje de tiros y rebotes
data_2=as.data.frame(matrix(c(as.numeric(data_nba[,3]),...
+as.numeric(data_nba[,5])),nrow=169,ncol=2,byrow=F))
clust_np2=ms(data_2,h=0.073)
clust_np2$cluster.center
clust_np2$cluster.label

#representación gráfica
windows()
par(mfrow=c(2,1))
plot(data_nba[,3],data_nba[,5],main="Estadísticas NBA",...
+col=clasif,xlab="porcentaje acierto tiros",...
+ylab="rebotes por partido")
plot(clust_np2,main="Clustering no paramétrico",...
+xlab="porcentaje acierto tiros",ylab="rebotes por partido")
```

Bibliografía

[1] Consultado el 31/08/2021. https://towardsdatascience.com/grouping-soccer-players-with-similar-skillsets-in-fifa-20-part-1-k-means-clustering-c4a845db78bc.

[2] Anil K. Jain, 2009. Data clustering: 50 years beyond K-means. Pattern recognition letters, 31(8), 651-666.

[3] McNeil, D. R. (1977) Interactive Data Analysis. New York: Wiley.

[4] Reaven, G. M. and Miller, R. G. (1979), An attempt to define the nature of chemical diabetes using a multidimensional analysis. Diabetologia, 16, 17–24.

[5] NBAstuffer. Consultado el 27 de Agosto de 2021, URL: https://www.nbastuffer.com/2020-2021-nba-player-stats/

[6] Fraley, C. Raftery, A. (1998). How many cluster? Which clustering method? Answers via model-based cluster analysis. Comput. J., 41, 580-582.

[7] Hartigan, J. A. (1975). Clustering algorithms. John Wiley y Sons, Inc, Nueva York,113-125.

[8] Pearson K. (1894). Contributions to the mathematical theory of evolution. Phil. Trans. R. Soc. Land. A. 185: 71-110.

[9] Furman W. D., Lindsey B. G. (1994). Measuring the effectiveness of moment estimators as starting values in maximizing mixture likelihoods. Comput. Stat. Data. Anal., 17:493-507.

[10] McLachlan G. J., Peel D., (1998). Robust cluster analysis via mixtures of multivariate t-distributions. In: Advances in Pattern Recognition, ed: A. Amin., D. Dori., P. Pudil, H. Freeman, Berlin: Springer.

[11] Geoffrey J. McLachlan, Sharon X. Lee, and Suren I. Rathnayake. (2018). Finite Mixture Models. Annu. Rev. Stat. Appl. 2019. 6:355–78

[12] Fraley, C. Raftery, A.E. (2002). Model-based clustering, discriminant analysis and density estimation. J. Am. Stat. Assoc., 97, 611–627.

[13] Bradley. P. S.,Fayyad. U., Reina. C. (1998), Scaling EM(Expectation-Maximization) Clustering to Large Datasets, Technical Report MSR-TR-95-35, Microsoft Research.

[14] Chacón, J. E. (2020). The modal age of statistics. International Statistical Review, 88(1), 122-125.

[15] Chernoff H. (1964). Estimation of the mode. Ann. Inst. Statist. Math., 16, 31-41.

[16] Romano J. P. (1988a). On weak convergence and optimality of kernel density estimates of the mode, Ann. Statist., 16, 629-647.

[17] Devroye L. P. and Wagner T. J. (1977). The strong uniform consistency of nearest neighbor density estimates, Ann. Statist., 5, 536-540.

[18] Menardi, G. (2016). A review on modal clustering., International Statistical Review, 84(3), 413- 433.

[19] Ameijeiras-Alonso J., Crujeiras R. M. and Rodríguez-Casal A. (2018). Multimode: an R package for mode assessment. arXiv:1803.00472.

[20] Li, J., Ray, S. y Lindsay, B.G. (2007). A nonparametric statistical approach to clustering via mode identification J. Mach. Learn. Res., 8, 1687–1723.

[21] Cheng (1995). Mean Shift, Mode Seeking, and Clustering, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 17, NO. 8, 790-793.

[22] Alboukadel Kassambara, 2017. Practical Guide To Cluster Analysis in R. Ed STHDA, 36-47.

[23] MeanShift: Clustering via the Mean Shift Algorithm. Consultado el 25 de Julio, URL: https://mran.microsoft.com/snapshot/2017-02-04/web/packages/MeanShift/index.html

[24] Daniel Peña, 2002 . Análisis de datos multivariantes. Madrid, ed. McGraw-hill, 227-228.

[25] Fraley, C. (1999) Algorithms for model-based Gaussian hierarchical clustering. SIAM J. Sci. Comput., 20, 270–281.

[26] Gaussian Mixture Models Clustering Algorithm Explained. (2019). Consultado el 21 Abril 2021, URL: https://towardsdatascience.com/gaussian-mixture-models-d13a5e915c8e.

[27] Mukherjee. S., Feigelson. E. D., Babu. G. J., Fraley. C., and Raftery. A. E. (1998), Three Types of Gamma Ray Bursts, The Astrophysical Journal, 508, 314-327.

[28] McLachlan GJ, Lee SX. 2016. Comment on On nomenclature for, and the relativemerits of, two formulations of skew distributions by A Azzalini, R Browne,MGenton, and P McNicholas. Stat. Probab. Lett. 116:1–5.