Análisis No Supervisado

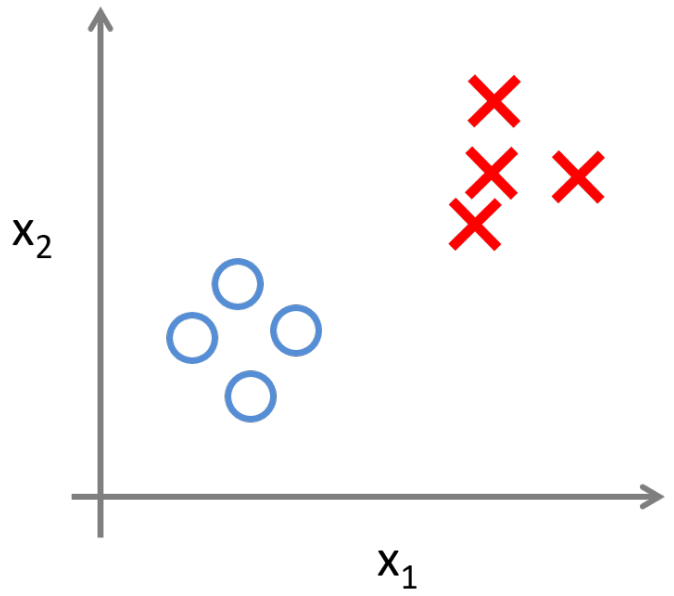
Alvaro J. Riascos Villegas
Universidad de los Andes y Quantil

Marzo de 2020

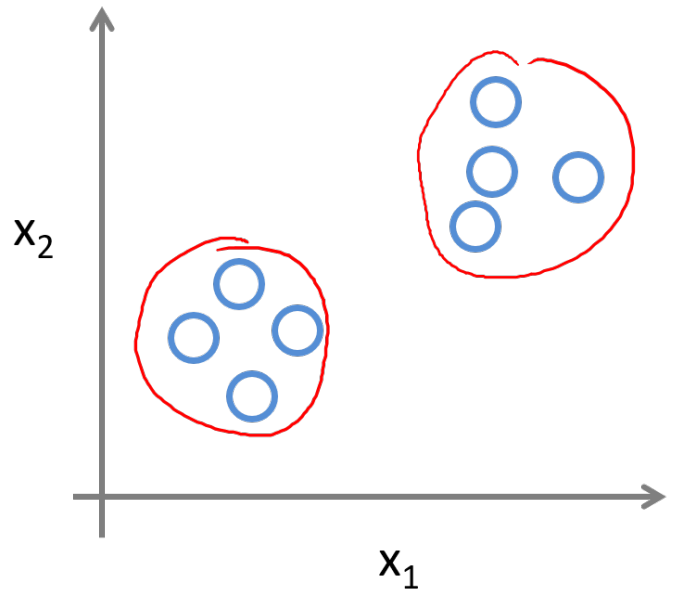
Contenido

Análisis Supervisado vs. No Supervisado

Supervised Learning



Unsupervised Learning



Análisis no supervisado

- En análisis no supervisado no tenemos "marcas".
- Tratamos de descubrir patrones o estructura entre los datos.
- Agrupar datos bajo algún criterio.
- Crear variables nuevas que "resuman" las variables.
- No tenemos como supervisar qué tan bueno es el análisis.

Análisis no supervisado

- En análisis no supervisado no tenemos "marcas".
- Tratamos de descubrir patrones o estructura entre los datos.
- Agrupar datos bajo algún criterio.
- Crear variables nuevas que "resuman" las variables.
- No tenemos como supervisar qué tan bueno es el análisis.

Análisis no supervisado

- En análisis no supervisado no tenemos "marcas".
- Tratamos de descubrir patrones o estructura entre los datos.
- Agrupar datos bajo algún criterio.
- Crear variables nuevas que "resuman" las variables.
- No tenemos como supervisar qué tan bueno es el análisis.

Análisis no supervisado

- En análisis no supervisado no tenemos "marcas".
- Tratamos de descubrir patrones o estructura entre los datos.
- Agrupar datos bajo algún criterio.
- Crear variables nuevas que "resuman" las variables.
- No tenemos como supervisar qué tan bueno es el análisis.

Análisis no supervisado

- En análisis no supervisado no tenemos "marcas".
- Tratamos de descubrir patrones o estructura entre los datos.
- Agrupar datos bajo algún criterio.
- Crear variables nuevas que "resuman" las variables.
- No tenemos como supervisar qué tan bueno es el análisis.

Aplicaciones

Aplicaciones

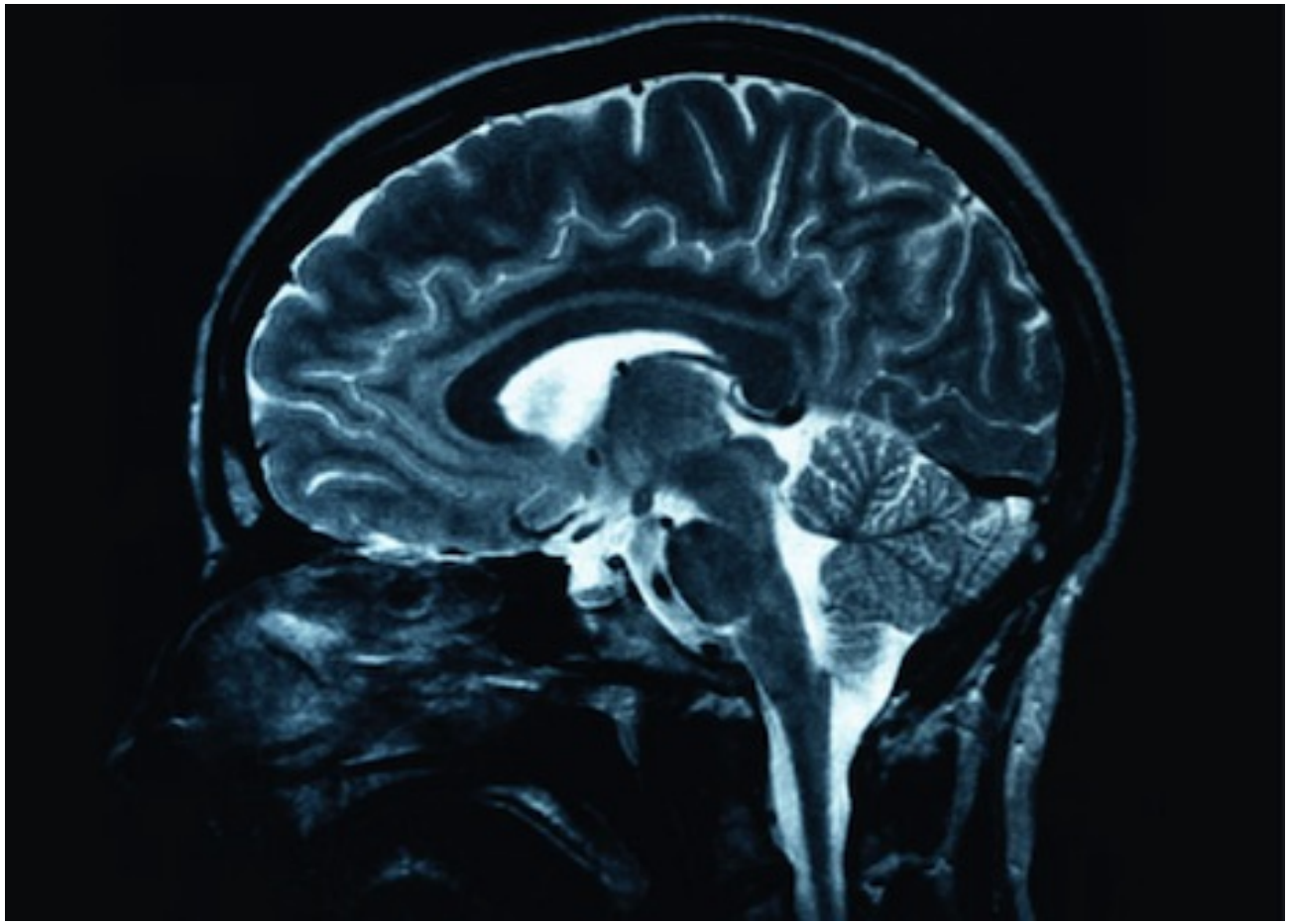
- Segmentar clientes para ofrecerles productos diferentes.

Aplicaciones

- Reducir dimensionalidad.

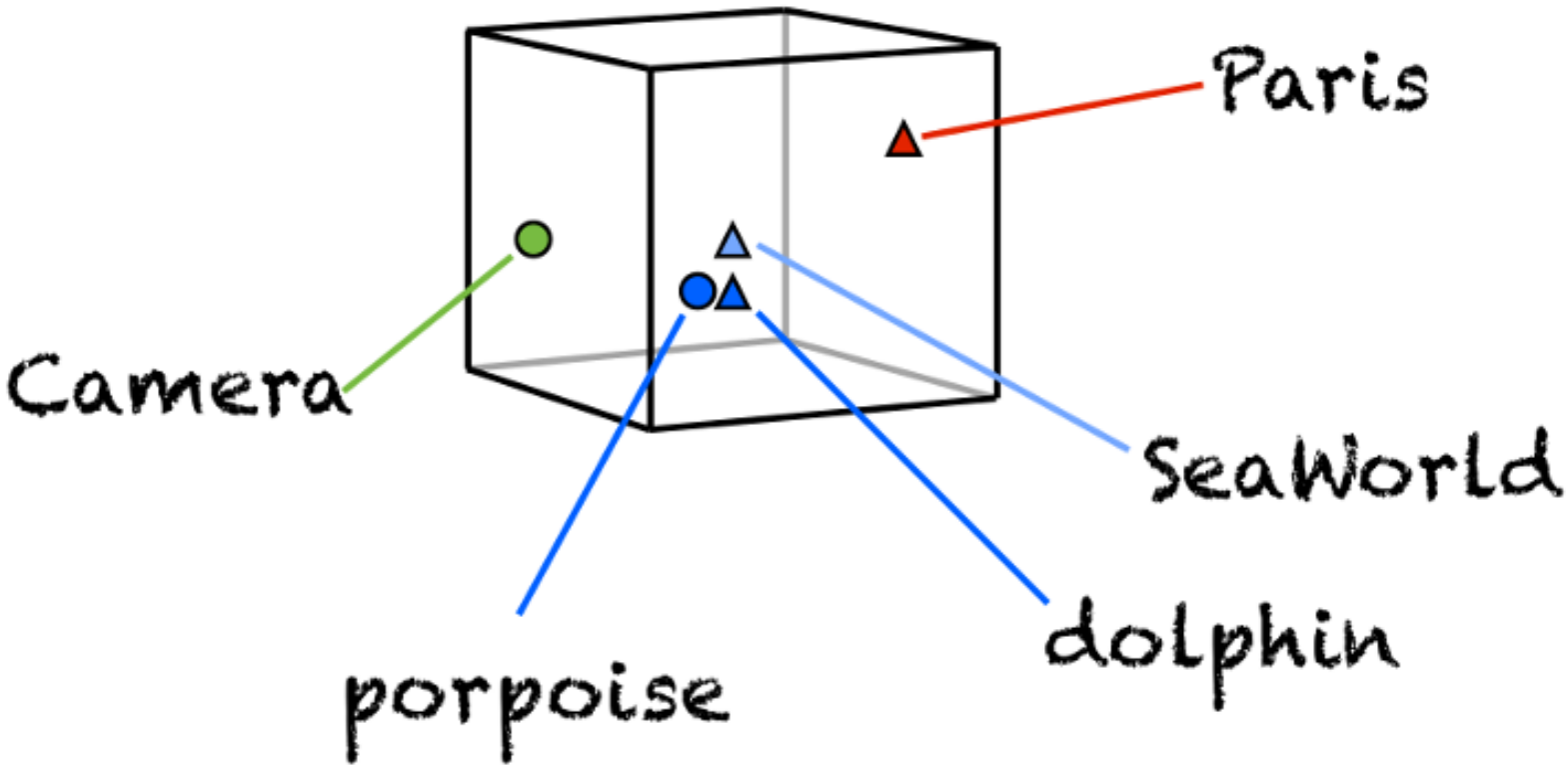
Aplicaciones

- Detectar automáticamente tumores cerebrales.



Aplicaciones

- Inferir el sentido de palabras basado en su contexto.



Supervisado o No Supervisado?

1. Predecir el clima en una ciudad dadas unas variables históricas.
2. Definir categorías entre diferentes clientes para ofrecer un servicio.
3. Detectar anomalías en flujos de datos.
4. Agrupar noticias en temas similares.
5. Detectar tumores cerebrales.

Supervisado o No Supervisado?

1. Predecir el clima en una ciudad dadas unas variables históricas.
2. Definir categorías entre diferentes clientes para ofrecer un servicio.
3. Detectar anomalías en flujos de datos.
4. Agrupar noticias en temas similares.
5. Detectar tumores cerebrales.

Supervisado o No Supervisado?

1. Predecir el clima en una ciudad dadas unas variables históricas.
2. Definir categorías entre diferentes clientes para ofrecer un servicio.
3. Detectar anomalías en flujos de datos.
4. Agrupar noticias en temas similares.
5. Detectar tumores cerebrales.

Supervisado o No Supervisado?

1. Predecir el clima en una ciudad dadas unas variables históricas.
2. Definir categorías entre diferentes clientes para ofrecer un servicio.
3. Detectar anomalías en flujos de datos.
4. Agrupar noticias en temas similares.
5. Detectar tumores cerebrales.

Supervisado o No Supervisado?

1. Predecir el clima en una ciudad dadas unas variables históricas.
2. Definir categorías entre diferentes clientes para ofrecer un servicio.
3. Detectar anomalías en flujos de datos.
4. Agrupar noticias en temas similares.
5. Detectar tumores cerebrales.

Algoritmos de Análisis no Supervisado

- Reglas de asociación.
- K−medias.
- Modelos de mixturas.
- Local outlier factor (LOF).
- Componentes Principales.
- Word Embeddings (Texto).
- Latent Dirichlet Allocation (Texto).

Contenido

Reglas de Asociación

- La idea es caracterizar la distribución de X.
- Para esto buscamos regiones del espacio en el que X toma valores donde la densidad sea alta.
- Sea X un vector de p variables (estas pueden ser continuas o categóricas) y S_j el soporte de X_j y $s_j \subseteq S_j$. La idea es encontrar valores $(s_1, s_2, ..., s_p)$ tales que

$$P[\bigcap_{j=1}^{p}(X_j \in s_j)] > s, \tag{1}$$

donde s es un nivel de soporte deseado. En otras palabras, se buscan valores de variables que tiendan a darse simultáneamente. Un estimador natural es la fracción de observaciones de X_i que asume valores en s_i.

Reglas de Asociación

- La idea es caracterizar la distribución de X.
- Para esto buscamos regiones del espacio en el que X toma valores donde la densidad sea alta.
- Sea X un vector de p variables (estas pueden ser continuas o categóricas) y S_j el soporte de X_j y $s_j \subseteq S_j$. La idea es encontrar valores $(s_1, s_2, ..., s_p)$ tales que

$$P[\bigcap_{j=1}^{p}(X_j \in s_j)] > s, \tag{1}$$

donde s es un nivel de soporte deseado. En otras palabras, se buscan valores de variables que tiendan a darse simultáneamente. Un estimador natural es la fracción de observaciones de X_i que asume valores en s_i.

Reglas de Asociación

- La idea es caracterizar la distribución de X.
- Para esto buscamos regiones del espacio en el que X toma valores donde la densidad sea alta.
- Sea X un vector de p variables (estas pueden ser continuas o categóricas) y S_j el soporte de X_j y $s_j \subseteq S_j$. La idea es encontrar valores $(s_1, s_2, ..., s_p)$ tales que

$$P[\bigcap_{j=1}^{p}(X_j \in s_j)] > s, \tag{1}$$

donde s es un nivel de soporte deseado. En otras palabras, se buscan valores de variables que tiendan a darse simultáneamente. Un estimador natural es la fracción de observaciones de X_i que asume valores en s_i.

Reglas de Asociación: Simplificación

- Para simplificar el problema vamos a permitir sólo regiones del tipo: un solo valor de la variable X_j, $v_{0,j}$ o la totalidad del soporte de esa variable S_j.
- Esto simplifica el problema a encontrar $\Gamma \subset \{1, ..., p\}$ y valores $\{v_{0,j}\}_{j \in \Gamma}$ tal que:

$$P[\bigcap_{j \in \Gamma} (X_j = v_{0,j})] > s, \tag{2}$$

donde $v_{0,j}$ es uno de los valores posibles de la variable j.

Reglas de Asociación: Simplificación

- Para simplificar el problema vamos a permitir sólo regiones del tipo: un solo valor de la variable X_j, $v_{0,j}$ o la totalidad del soporte de esa variable S_j.
- Esto simplifica el problema a encontrar $\Gamma \subset \{1, ..., p\}$ y valores $\{v_{0,j}\}_{j \in \Gamma}$ tal que:

$$P[\bigcap_{j \in \Gamma} (X_j = v_{0,j})] > s, \tag{2}$$

donde $v_{0,j}$ es uno de los valores posibles de la variable j.

Reglas de Asociación: Simplificación

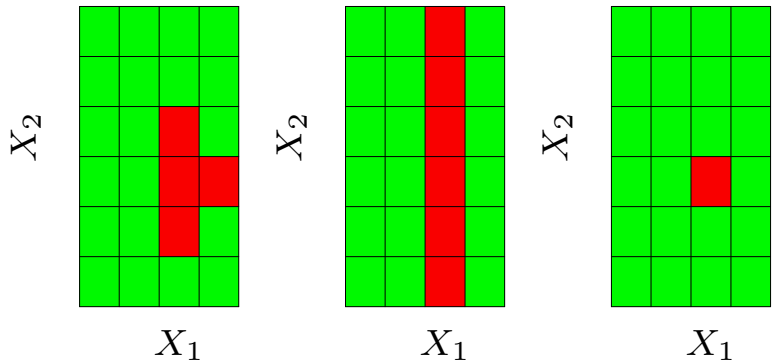


FIGURE 14.1. *Simplifications for association rules. Here there are two inputs X_1 and X_2, taking four and six distinct values, respectively. The red squares indicate areas of high density. To simplify the computations, we assume that the derived subset corresponds to either a single value of an input or all values. With this assumption we could find either the middle or right pattern, but not the left one.*

Reglas de Asociación: Representación con variables dummy

- Si además suponemos que cada variable X_j puede tomar sólo un número finito de variables, podemos definir K como el número total de dummis para describir los diferentes valores de las variables X_j:

$$K = \sum_{j=1}^{p} \mid S_j \mid$$

y creamos Z_k variables dummy por cada una de las variables K.

- El problema ahora es encontrar: $\Upsilon \subset \{1, 2, 3, ..., K\}$ tal que:

$$P[\bigcap_{j \in \Upsilon} (Z_j = 1)] > s \qquad (3)$$

Reglas de Asociación: Representación con variables dummy

- Si además suponemos que cada variable X_j puede tomar sólo un número finito de variables, podemos definir K como el número total de dummis para describir los diferentes valores de las variables X_j:

$$K = \sum_{j=1}^{p} \mid S_j \mid$$

  y creamos Z_k variables dummy por cada una de las variables K.

- El problema ahora es encontrar: $\Upsilon \subset \{1, 2, 3, ..., K\}$ tal que:

$$P[\bigcap_{j \in \Upsilon} (Z_j = 1)] > s \tag{3}$$

Reglas de Asociación

- Todo Υ tal que $P[\bigcap_{j\in\Upsilon}(Z_j = 1)] > s$ representa un grupo de items (canasta) que aparecen con una frecuencia deseada (el soporte mínimo).
- El estimador natural es la frecuencia:

$$F(\Upsilon) = Pr[\widehat{\prod_{j\in\Upsilon}(Z^j = 1)}] = \frac{|\{z : \forall j \in \Upsilon\ z_j = 1\}|}{N} = \frac{1}{N}\sum_{i=1}^{N}\prod_{j\in\Upsilon} z_{ij} \tag{4}$$

donde z_i es la observación i (canasta) y z_{ij} denota si el item j está en esa canasta.

Reglas de Asociación

- Todo Υ tal que $P[\bigcap_{j\in\Upsilon}(Z_j = 1)] > s$ representa un grupo de items (canasta) que aparecen con una frecuencia deseada (el soporte mínimo).
- El estimador natural es la frecuencia:

$$F(\Upsilon) = Pr[\widehat{\prod_{j\in\Upsilon}(Z^j = 1)}] = \frac{|\{z : \forall j \in \Upsilon\, z_j = 1\}|}{N} = \frac{1}{N}\sum_{i=1}^{N}\prod_{j\in\Upsilon} z_{ij} \tag{4}$$

donde z_i es la observación i (canasta) y z_{ij} denota si el item j está en esa canasta.

Reglas de Asociación

- Las preguntas fundamentales que nos hacemos son:
  1. Dado un conjunto de índices $\Upsilon \subset \{1, ..., K\}$ de alto soporte se pueden particionar de muchas formas en dos subconjuntos A, B. Cada uno define una regla de la forma $A \Rightarrow B$, A se llama el antecedente y B el consecuente. El soporte de la regla es el soporte de Υ. Representa la probabilidad de observar los items A y B en una canasta (i.e., la unión de los items A y los items B).
  2. La confianza en la regla $A \Rightarrow B$, $C(A \Rightarrow B)$ se define como $C(A \Rightarrow B) = \frac{Supp(A \Rightarrow B)}{Supp(A)}$. Es un estimador de $P(B \mid A)$.
  3. La confianza esperada de B es $Supp(B) = P(B)$.
  4. El *Lift* de la regla se $A \Rightarrow B$ se define como: $L(A \Rightarrow B) = \frac{C(A \Rightarrow B)}{Supp(B)}$. Es una medida del grado de asociación: $\frac{P(A \wedge B)}{P(A)P(B)}$.

Reglas de Asociación

- Las preguntas fundamentales que nos hacemos son:
  1. Dado un conjunto de índices $\Upsilon \subset \{1, ..., K\}$ de alto soporte se pueden particionar de muchas formas en dos subconjuntos A, B. Cada uno define una regla de la forma $A \Rightarrow B$, A se llama el antecedente y B el consecuente. El soporte de la regla es el soporte de Υ. Representa la probabilidad de observar los items A y B en una canasta (i.e., la unión de los items A y los items B).
  2. La confianza en la regla $A \Rightarrow B$, $C(A \Rightarrow B)$ se define como $C(A \Rightarrow B) = \frac{Supp(A \Rightarrow B)}{Supp(A)}$. Es un estimador de $P(B \mid A)$.
  3. La confianza esperada de B es $Supp(B) = P(B)$.
  4. El *Lift* de la regla se $A \Rightarrow B$ se define como: $L(A \Rightarrow B) = \frac{C(A \Rightarrow B)}{Supp(B)}$. Es una medida del grado de asociación: $\frac{P(A \wedge B)}{P(A)P(B)}$.

Reglas de Asociación

- Las preguntas fundamentales que nos hacemos son:
  1. Dado un conjunto de índices $\Upsilon \subset \{1, ..., K\}$ de alto soporte se pueden particionar de muchas formas en dos subconjuntos A, B. Cada uno define una regla de la forma $A \Rightarrow B$, A se llama el antecedente y B el consecuente. El soporte de la regla es el soporte de Υ. Representa la probabilidad de observar los items A y B en una canasta (i.e., la unión de los items A y los items B).
  2. La confianza en la regla $A \Rightarrow B$, $C(A \Rightarrow B)$ se define como $C(A \Rightarrow B) = \frac{Supp(A \Rightarrow B)}{Supp(A)}$. Es un estimador de $P(B \mid A)$.
  3. La confianza esperada de B es $Supp(B) = P(B)$.
  4. El *Lift* de la regla se $A \Rightarrow B$ se define como: $L(A \Rightarrow B) = \frac{C(A \Rightarrow B)}{Supp(B)}$. Es una medida del grado de asociación: $\frac{P(A \wedge B)}{P(A)P(B)}$.

Reglas de Asociación

- Las preguntas fundamentales que nos hacemos son:
  1. Dado un conjunto de índices $\Upsilon \subset \{1, ..., K\}$ de alto soporte se pueden particionar de muchas formas en dos subconjuntos A, B. Cada uno define una regla de la forma $A \Rightarrow B$, A se llama el antecedente y B el consecuente. El soporte de la regla es el soporte de Υ. Representa la probabilidad de observar los items A y B en una canasta (i.e., la unión de los items A y los items B).
  2. La confianza en la regla $A \Rightarrow B$, $C(A \Rightarrow B)$ se define como $C(A \Rightarrow B) = \frac{Supp(A \Rightarrow B)}{Supp(A)}$. Es un estimador de $P(B \mid A)$.
  3. La confianza esperada de B es $Supp(B) = P(B)$.
  4. El *Lift* de la regla se $A \Rightarrow B$ se define como: $L(A \Rightarrow B) = \frac{C(A \Rightarrow B)}{Supp(B)}$. Es una medida del grado de asociación: $\frac{P(A \wedge B)}{P(A)P(B)}$.

Reglas de Asociación

- Las preguntas fundamentales que nos hacemos son:
  1. Dado un conjunto de índices $\Upsilon \subset \{1, ..., K\}$ de alto soporte se pueden particionar de muchas formas en dos subconjuntos A, B. Cada uno define una regla de la forma $A \Rightarrow B$, A se llama el antecedente y B el consecuente. El soporte de la regla es el soporte de Υ. Representa la probabilidad de observar los items A y B en una canasta (i.e., la unión de los items A y los items B).
  2. La confianza en la regla $A \Rightarrow B$, $C(A \Rightarrow B)$ se define como $C(A \Rightarrow B) = \frac{Supp(A \Rightarrow B)}{Supp(A)}$. Es un estimador de $P(B \mid A)$.
  3. La confianza esperada de B es $Supp(B) = P(B)$.
  4. El *Lift* de la regla se $A \Rightarrow B$ se define como: $L(A \Rightarrow B) = \frac{C(A \Rightarrow B)}{Supp(B)}$. Es una medida del grado de asociación: $\frac{P(A \wedge B)}{P(A)P(B)}$.

Reglas de Asociación: Ejemplo

Association rule 2: Support 13.4%, confidence 80.8%, and lift 2.13.

$$\left[\begin{array}{rcl} \text{language in home} & = & \textit{English} \\ \text{householder status} & = & \textit{own} \\ \text{occupation} & = & \{\textit{professional/managerial}\} \end{array}\right] \Downarrow \text{income} \geq \textit{\$40,000}$$

Association rule 3: Support 26.5%, confidence 82.8% and lift 2.15.

$$\left[\begin{array}{rcl} \text{language in home} & = & \textit{English} \\ \text{income} & < & \textit{\$40,000} \\ \text{marital status} & = & \textit{not married} \\ \text{number of children} & = & 0 \end{array}\right] \Downarrow \text{education} \notin \{\textit{college graduate, graduate study}\}$$

Reglas de Asociación: Algoritmo Apriori

- Obsérvese que $\Upsilon' \subset \Upsilon$ y $supp(\Upsilon) > s$ entonces $supp(\Upsilon') > s$.
- El algoritmo Apriori sigue los siguientes pasos:
  1. Sea $k = 1$. Recorra todos los items individuales que tengan soporte mayor a s. Elimine todas los items que no tengan suficiente soporte (representatividad). Recuérdese la anterior observación).
  2. Para $k + 1$ recorra todas las combinaciones de $k + 1$ items que se puedan construir con los items que han sobrevivido hasta este punto (que no fueron eliminados en la ronda anterior). Conserve solo los conjuntos de $k + 1$ items que tienen soporte mínimo s. Elimine todos los conjuntos de $k + 1$ item que no tiene suficiente representatividad.

Reglas de Asociación: Algoritmo Apriori

- Obsérvese que $\Upsilon' \subset \Upsilon$ y $supp(\Upsilon) > s$ entonces $supp(\Upsilon') > s$.
- El algoritmo Apriori sigue los siguientes pasos:
  1. Sea $k = 1$. Recorra todos los items individuales que tengan soporte mayor a s. Elimine todas los items que no tengan suficiente soporte (representatividad). Recuérdese la anterior observación).
  2. Para $k + 1$ recorra todas las combinaciones de $k + 1$ items que se puedan construir con los items que han sobrevivido hasta este punto (que no fueron eliminados en la ronda anterior). Conserve solo los conjuntos de $k + 1$ items que tienen soporte mínimo s. Elimine todos los conjuntos de $k + 1$ item que no tiene suficiente representatividad.

Reglas de Asociación: Algoritmo Apriori

- Obsérvese que $\Upsilon' \subset \Upsilon$ y $supp(\Upsilon) > s$ entonces $supp(\Upsilon') > s$.
- El algoritmo Apriori sigue los siguientes pasos:
  1. Sea $k = 1$. Recorra todos los items individuales que tengan soporte mayor a s. Elimine todas los items que no tengan suficiente soporte (representatividad). Recuérdese la anterior observación).
  2. Para $k + 1$ recorra todas las combinaciones de $k + 1$ items que se puedan construir con los items que han sobrevivido hasta este punto (que no fueron eliminados en la ronda anterior). Conserve solo los conjuntos de $k + 1$ items que tienen soporte mínimo s. Elimine todos los conjuntos de $k + 1$ item que no tiene suficiente representatividad.

Reglas de Asociación: Algoritmo Apriori

- Obsérvese que $\Upsilon' \subset \Upsilon$ y $supp(\Upsilon) > s$ entonces $supp(\Upsilon') > s$.
- El algoritmo Apriori sigue los siguientes pasos:
  1. Sea $k = 1$. Recorra todos los items individuales que tengan soporte mayor a s. Elimine todas los items que no tengan suficiente soporte (representatividad). Recuérdese la anterior observación).
  2. Para $k + 1$ recorra todas las combinaciones de $k + 1$ items que se puedan construir con los items que han sobrevivido hasta este punto (que no fueron eliminados en la ronda anterior). Conserve solo los conjuntos de $k + 1$ items que tienen soporte mínimo s. Elimine todos los conjuntos de $k + 1$ item que no tiene suficiente representatividad.

Contenido

¿Para qué sirve?

- Es un algoritmo para agrupar puntos "cercanos"
- Se crean K "prototipos" de datos que representan cada clúster.
- Se puede usar para detectar datos anómalos.
- Se puede usar para construir variables en problemas de reconocimiento de imágenes.

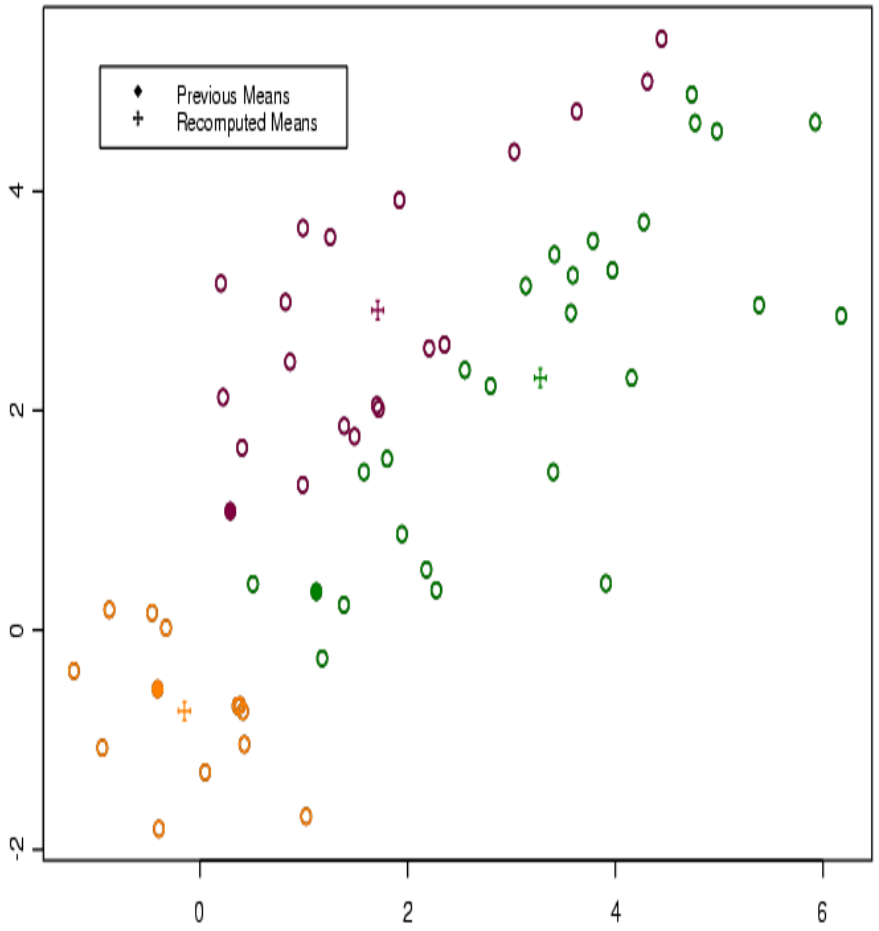
El algoritmo

1. Escoger K centroides iniciales aleatoriamente.
2. Repetir los siguientes pasos hasta convergencia:
   - Asignar cada observación al clúster cuyo centroide sea más cercano.
   - Reasignar el centroide del clúster *i* al promedio de las observaciones de ese clúster.

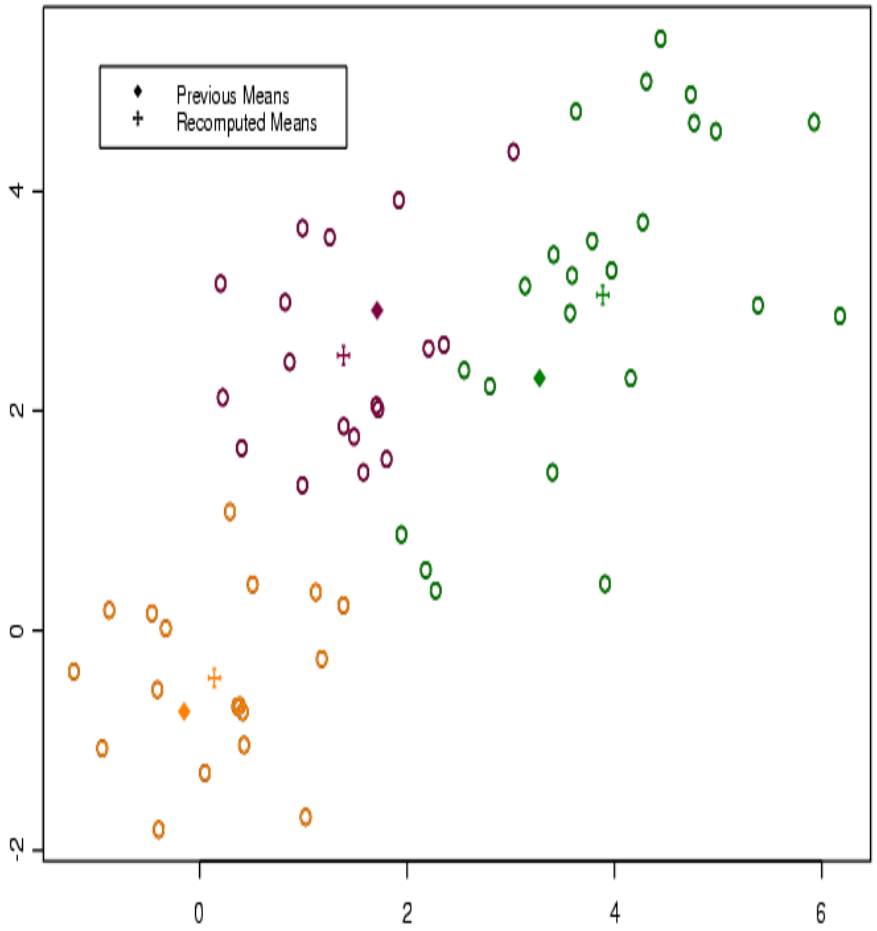
Algoritmo Alternativo

1. Asignar a cada observación aleatoriamente un número entre 1 y K.
2. Repetir los siguientes pasos hasta convergencia:
   - Definir el centroide del clúster i como el promedio de las observaciones de ese clúster.
   - Asignar cada observación al clúster cuyo centroide sea más cercano.

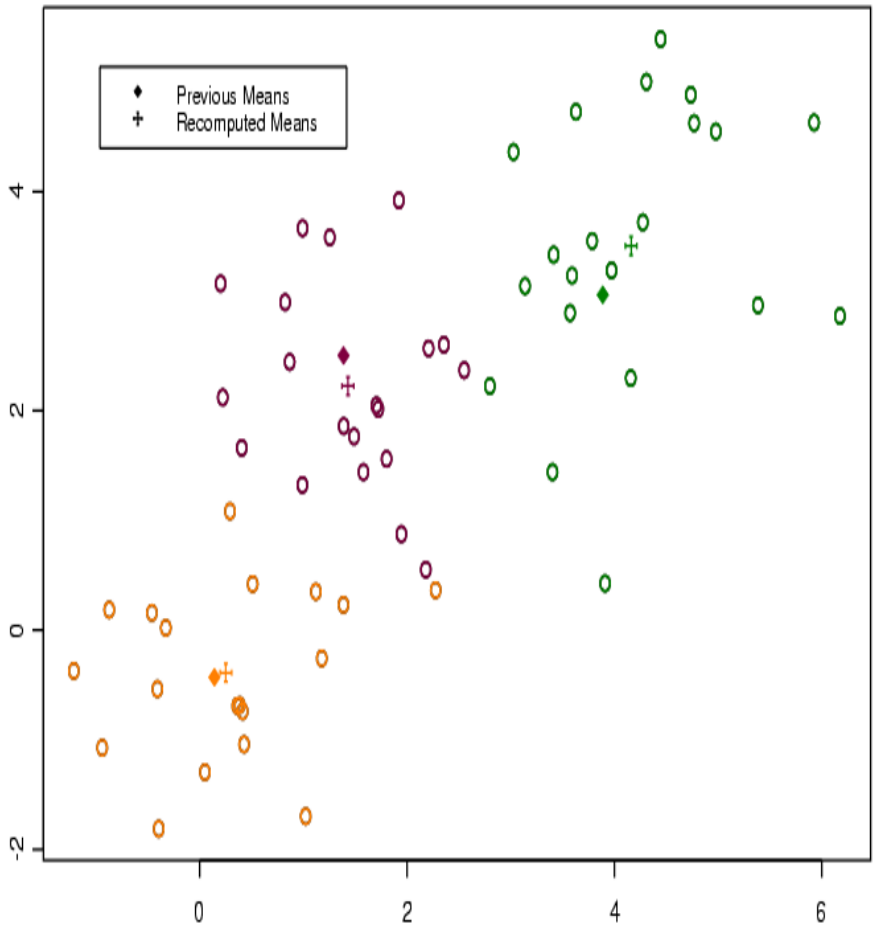
Ejemplo: Algoritmo



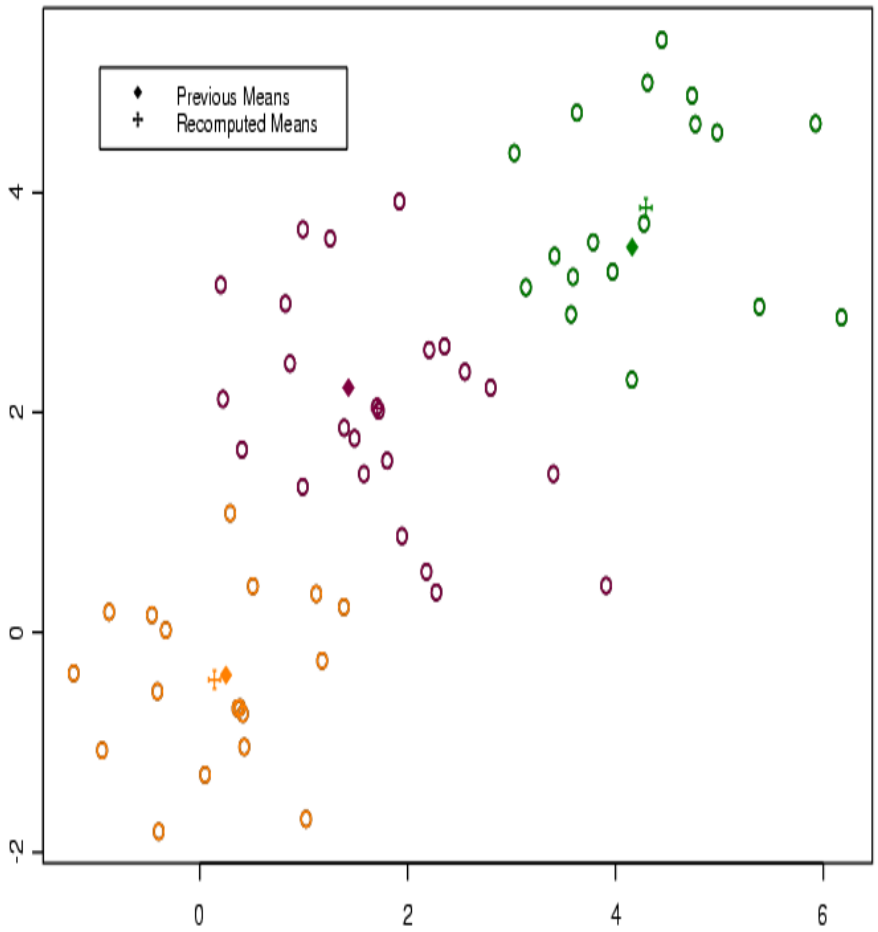
Ejemplo: Algoritmo



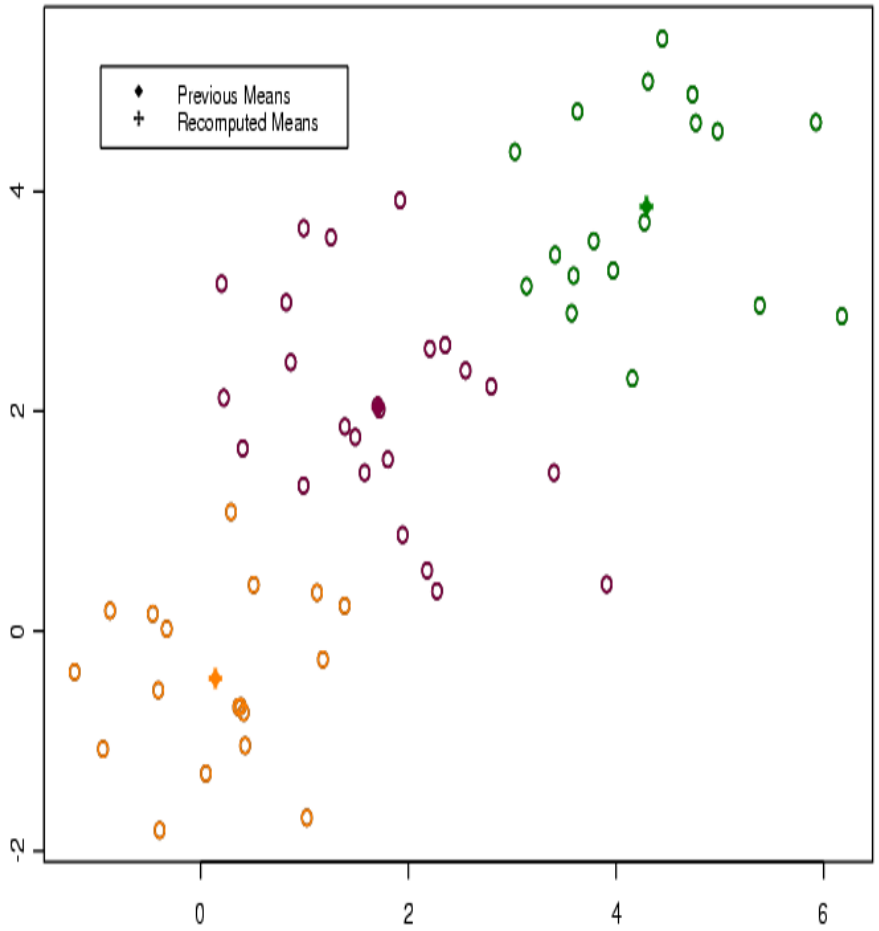
Ejemplo: Algoritmo



Ejemplo: Algoritmo



Ejemplo: Algoritmo



Ejemplo: Particiones

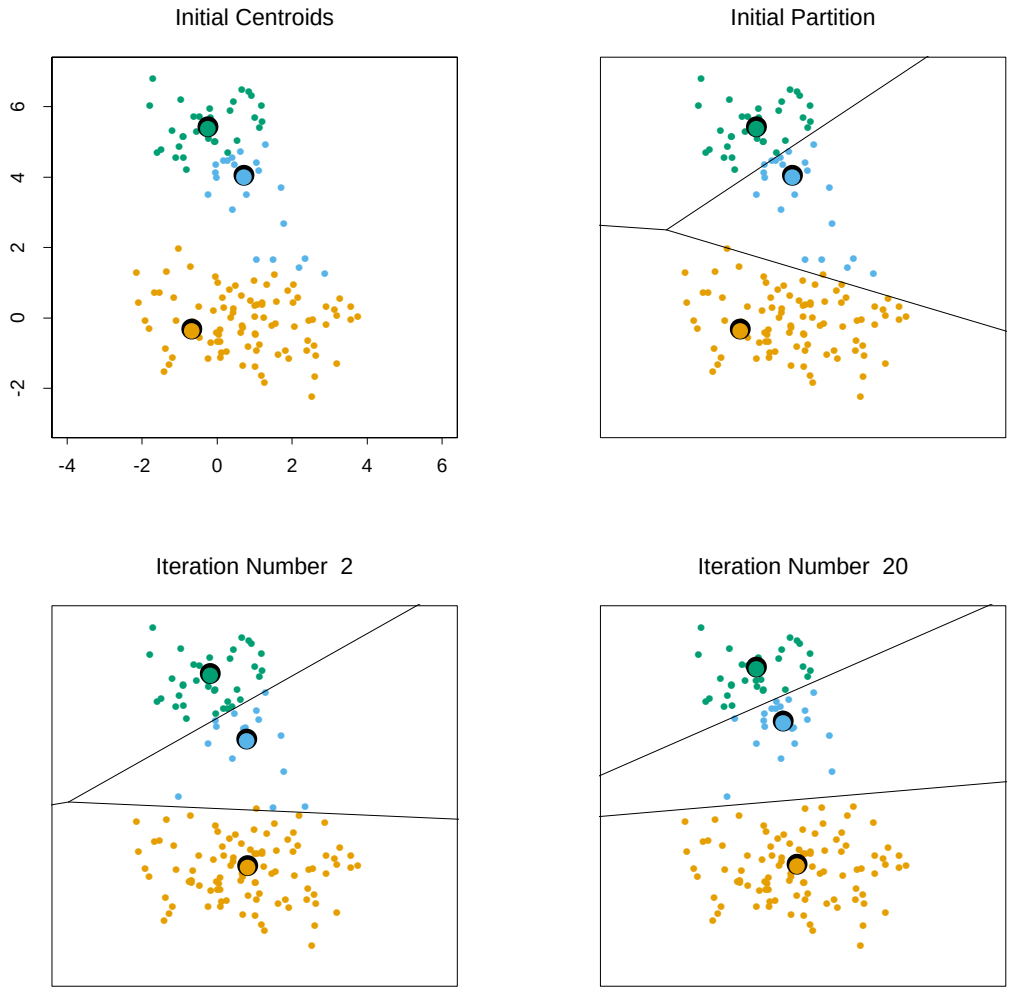
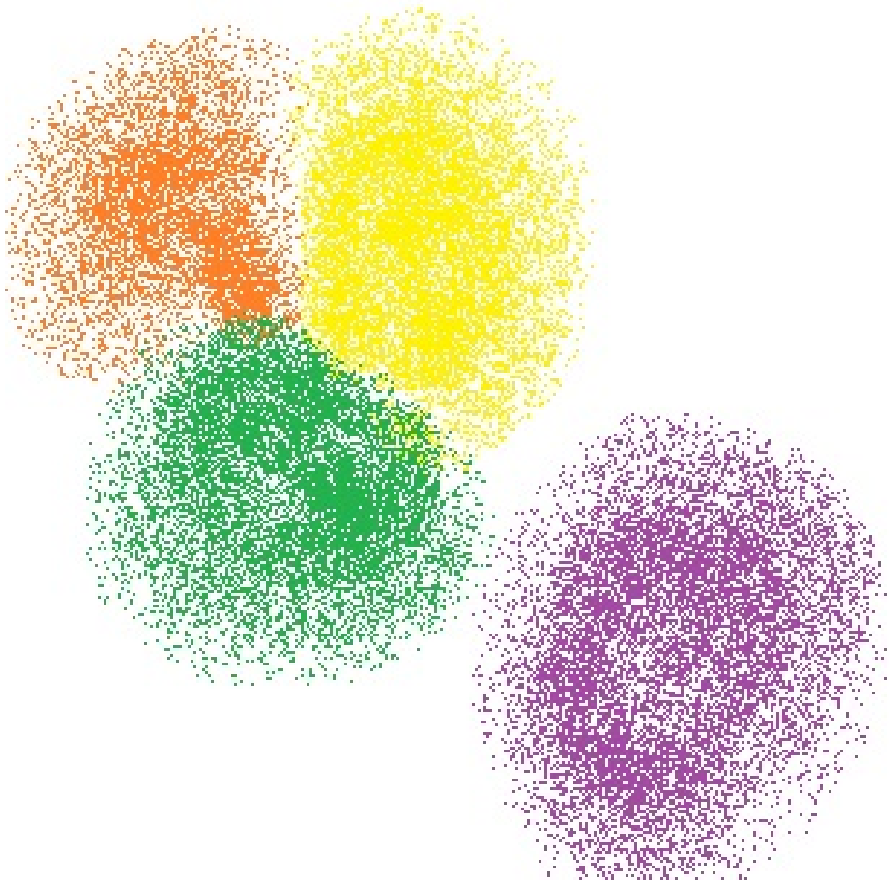


FIGURE 14.6. *Successive iterations of the K-means clustering algorithm for the simulated data of Figure 14.4.*

K medias: Ejemplo



K medias: Efecto estándarización variables

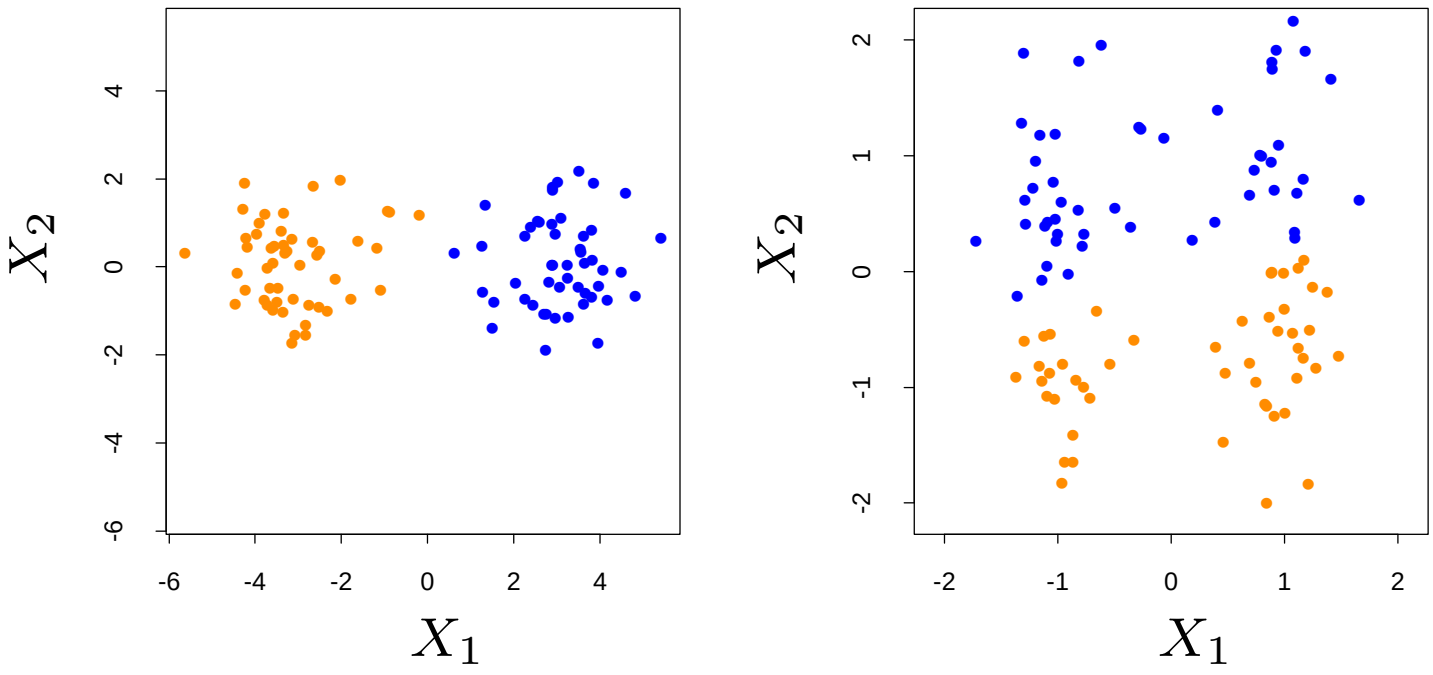
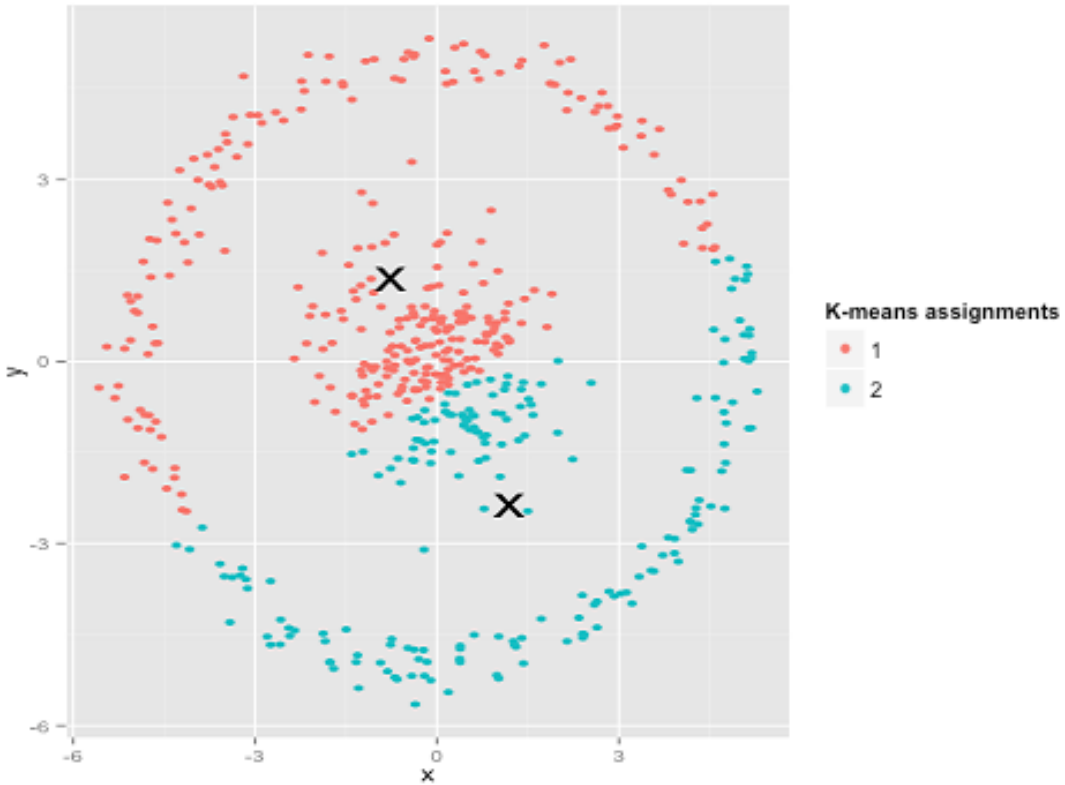


FIGURE 14.5. *Simulated data: on the left, K-means clustering (with K=2) has been applied to the raw data. The two colors indicate the cluster memberships. On the right, the features were first standardized before clustering. This is equivalent to using feature weights* $1/[2 \cdot \mathrm{var}(X_j)]$. *The standardization has obscured the two well-separated groups. Note that each plot uses the*

Casos Problemáticos



Optimización

- K−medias también soluciona el siguiente problema de optimización

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|x_j - \mu_i\|^2$$

  donde μ_i es la media de puntos en S_i.

- También puede ser:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \frac{1}{|S_i|} \sum_{x_j, x_{j'} \in S_i} \left\|\mathbf{x}_j - \mathbf{x}_{j'}\right\|^2$$

Problemas con K−medias

- Es un problema NP.
- Hay varios óptimos locales (pero no infinitos).

Ejemplo: Tumores y expresiones de genes

TABLE 14.2. *Human tumor data: number of cancer cases of each type, in each of the three clusters from K-means clustering.*

Cluster	Breast	CNS	Colon	K562	Leukemia	MCF7
1	3	5	0	0	0	0
2	2	0	0	2	6	2
3	2	0	7	0	0	0
Cluster	Melanoma	NSCLC	Ovarian	Prostate	Renal	Unknown
1	1	7	6	2	9	1
2	7	2	0	0	0	0
3	0	0	0	0	0	0

The data are a 6830×64 matrix of real numbers, each representing an expression measurement for a gene (row) and sample (column). Here we cluster the samples, each of which is a vector of length 6830, corresponding to expression values for the 6830 genes. Each sample has a label such as `breast` (for breast cancer), `melanoma`, and so on; we don't use these labels in the clustering, but will examine *posthoc* which labels fall into which clusters.

Ejemplo: Tumores y expresiones de genes

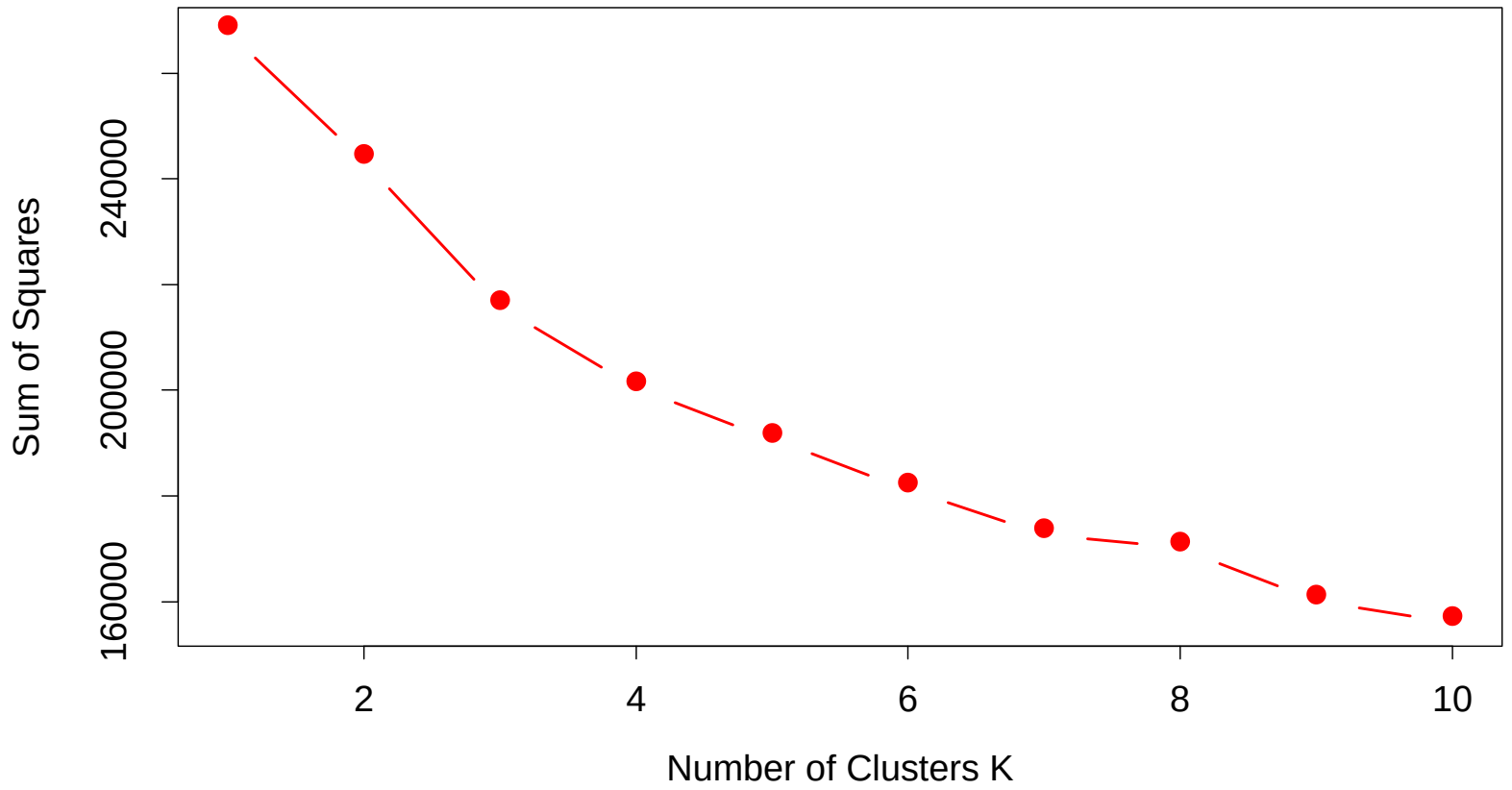


FIGURE 14.8. *Total within-cluster sum of squares for K-means clustering applied to the human tumor microarray data.*

Contenido

K – Medoids

- El algoritmo K-Medias depende de:
  1. Una noción de disimilitud.
  2. Calcula los centros de los grupos en cada iteración como la media de los puntos más cercanos.
- Lo primero puede sustituirse por cualquier medidad de disimilitud $D(x_i, x_{i'})$.
- Lo segundo puede generalizarse a expensas de tiempo computacional.

K – Medoids

- El algoritmo K-Medias depende de:
  1. Una noción de disimilitud.
  2. Calcula los centros de los grupos en cada iteración como la media de los puntos más cercanos.
- Lo primero puede sustituirse por cualquier medidad de disimilitud $D(x_i, x_{i'})$.
- Lo segundo puede generalizarse a expensas de tiempo computacional.

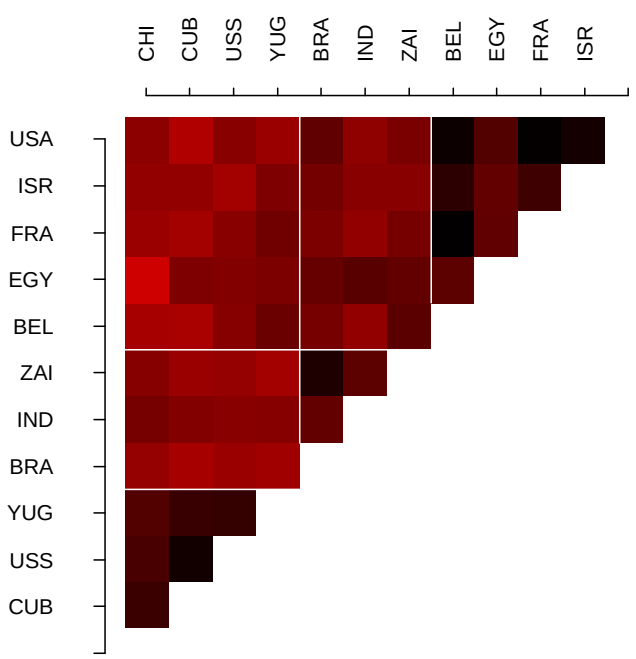
K – Medoids

- El algoritmo K-Medias depende de:
  1. Una noción de disimilitud.
  2. Calcula los centros de los grupos en cada iteración como la media de los puntos más cercanos.
- Lo primero puede sustituirse por cualquier medidad de disimilitud $D(x_i, x_{i'})$.
- Lo segundo puede generalizarse a expensas de tiempo computacional.

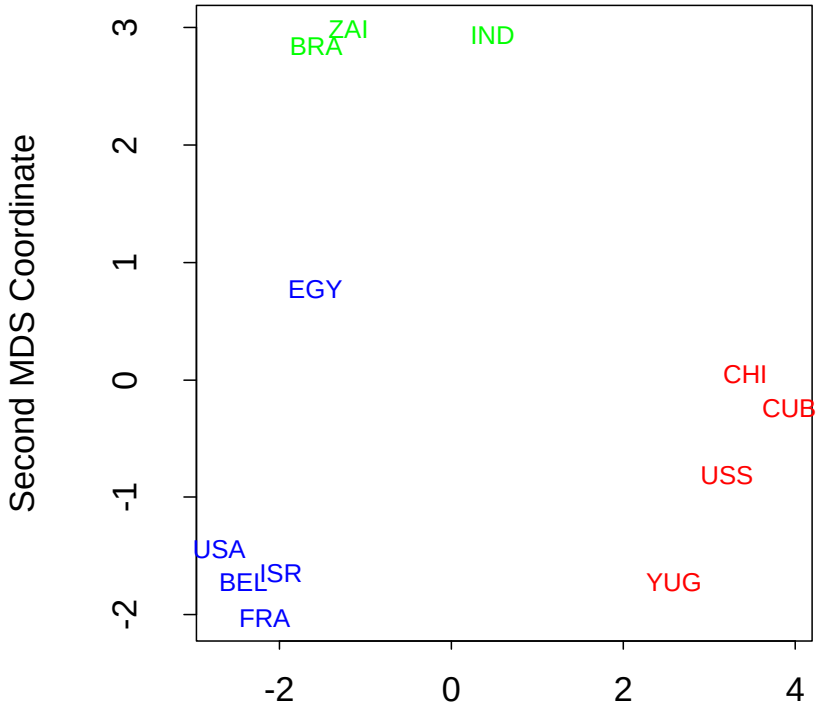
K−Medoids

- Dada una asignación de clusters C:
  1. Para cada $C_i \in C$ encontrar el punto x_i^* que minimiza la suma de la disimilitud a los otros puntos del cluster C_i. Defina los centros $m_i = x_i^*, i = 1, ..., K$.
  2. Reasignar los puntos a los centros más cercanos de acuerdo a disimilitud.
  3. Iterar los pasos anteriores hasta que no haya cambios en los clusters.

K−Medoids: Ejemplo



Reordered Dissimilarity Matrix



Contenido

Selección del número de clusters

Dos enfoques:

- Medidas intrísecas.
- Medias extrínsecas.

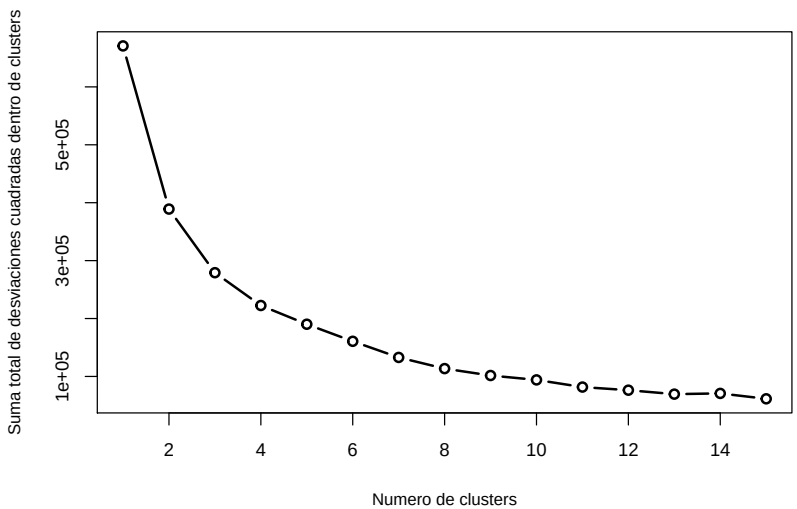
Selección del número de clusters

Medidas intrínsecas:

- El gráfico del "codo".
- El índice de Calinski-Harabasz (CH).

Gráfico del "codo"

- Relaciona la suma de las desviaciones totales dentro de los clusters con el número de clusters. Se elige el número de clusters que hacen que la suma de desviaciones sea mínima y que agregar un cluster adicional a ese número no genere una reducción significativa en la suma de los errores.

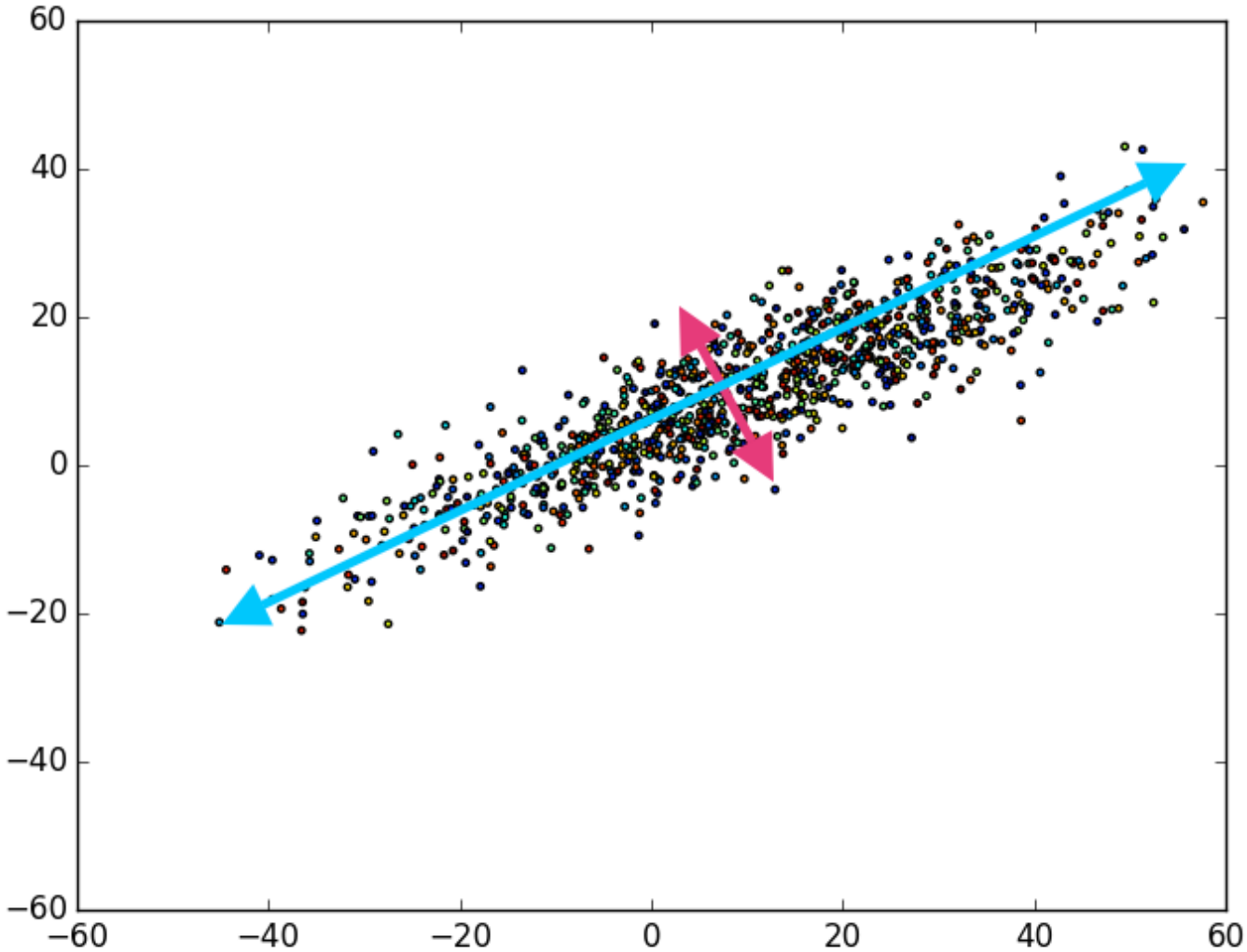


Contenido

Componentes Principales

- A veces tenemos muchas variables correlacionadas.
- La información que contienen es **redundante**
- ¿Podemos construir pocas variables que expliquen mucha de la variación?
- ¿Si quisiéramos reemplazar todas las variables por una sola que "aproxime" bien todas las otras, como lo hacemos?

Componentes Principales



Componentes Principales

- Las componentes principales son variables nuevas.
- Son combinaciones lineales de las variables originales.
- La primera componente es la combinación lineal que maximiza la varianza.
- La segunda componente maximiza la varianza condicional a ser ortogonal a la primera.
- Así sucesivamente podemos construir p componentes.
- Comenzamos discutiendo primero el caso en el que queremos reducir la dimensionalidad de un vector aleatorio X.
- Sea Σ la matriz de varianza covarianza de X.

Componentes Principales: Las CP maximizan varianza

- La idea es encontrar una combinación lineal que disperse al máximo la muestra en el espacio de características.
- Para forzar que el resultado se concentre en la dirección y no en la magnitud de la transformación se impone que el vector de transformación tenga norma uno.
- El problema que se quiere resolver es:

$$\max_{\{a:a^T a=1\}} a^T \Sigma a$$

No es difícil demostrar que el vector a^1 que resuelve este problema es el vector propio de X con el mayor valor propio asociado.

Componentes Principales: Las CP maximizan varianza

- Ahora la idea es econtrar una segunda combinación lineal, tal que el vector que la define sea ortogonal al anterior (a^1), tenga norma uno y se maximice la dispersión.
- El problema que se quiere resolver ahora es:

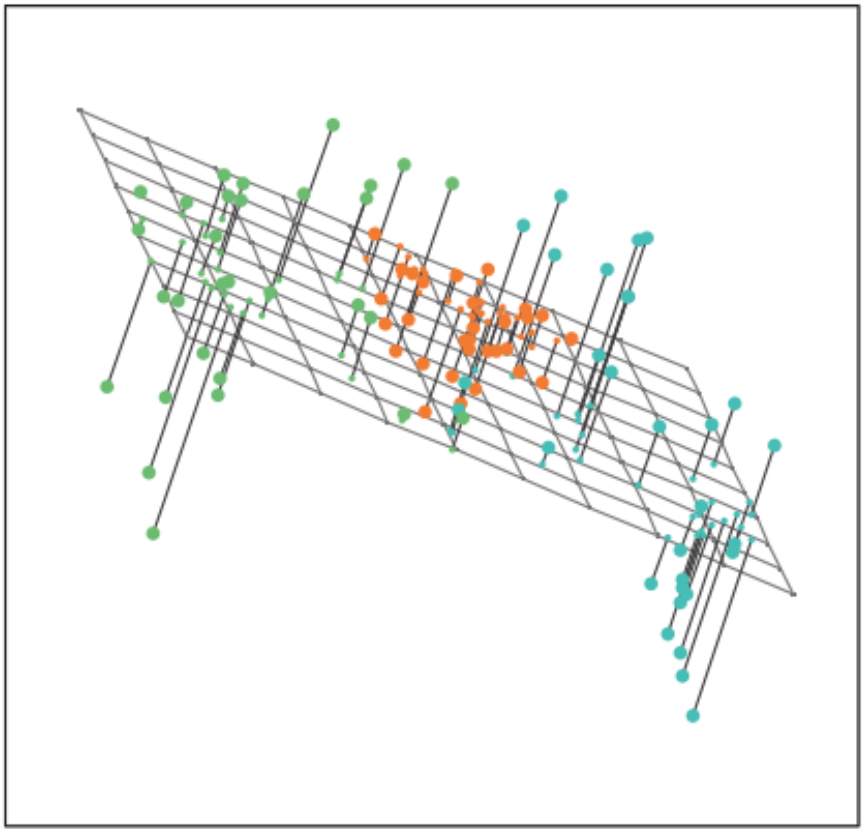
$$\max_{\{a: a^T a=1, a^T a^1=0\}} a^T \Sigma a$$

Se puede demostrar que el vector a^1 que resuelve este problema es un vector propio de X con el segundo mayor valor propio asociado.

- Para medir qué tanto las componentes explican la varianza del vector aleatorio X:

$$\sum_{j=1}^{p} var(X_j) = traza(\Sigma) = \sum_{j=1}^{p} \lambda_j = \sum_{j=1}^{p} var(Y_j)$$

Sobre las componentes principales



Aplicación: Números escritos a mano

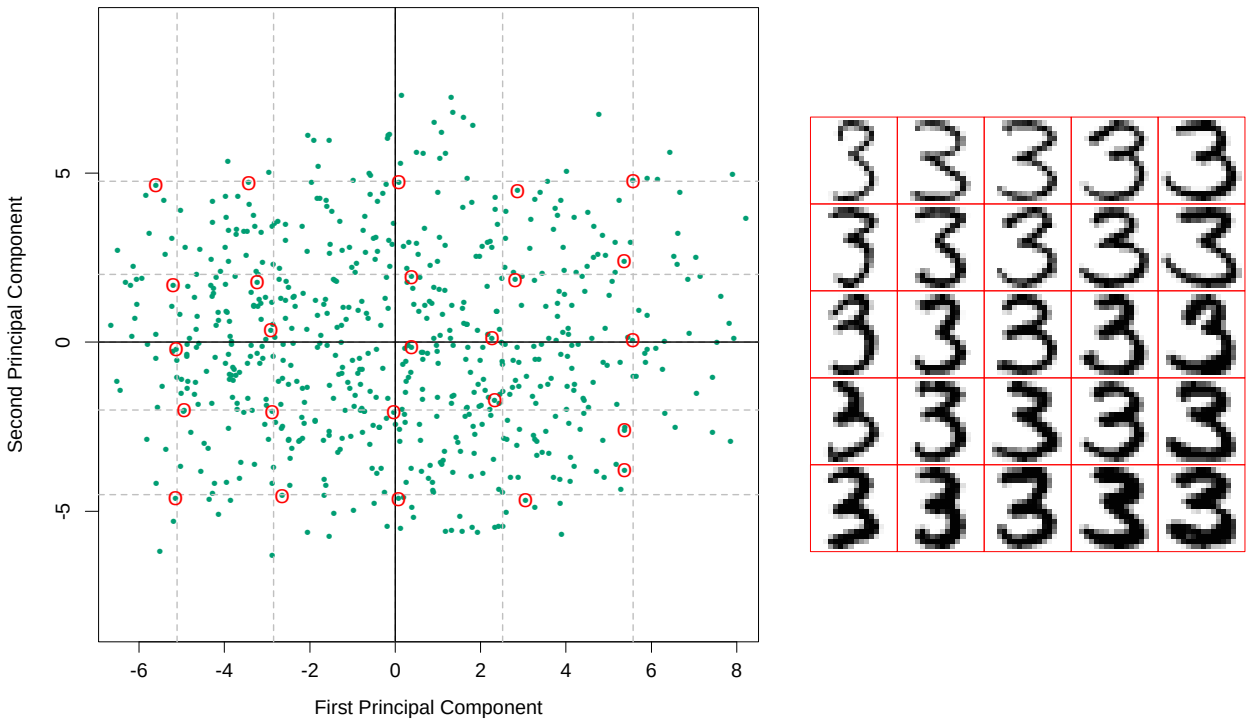


FIGURE 14.23. *(Left panel:) the first two principal components of the handwritten threes. The circled points are the closest projected images to the vertices of a grid, defined by the marginal quantiles of the principal components. (Right panel:) The images corresponding to the circled points. These show the nature of the first two principal components.*

PCA en la práctica

1. Aplicar PCA.
2. Escoger primeras componentes que expliquen x % de la varianza.
3. Hacer un modelo de clasificación basado en esas componentes.
4. Esto puede transformar un modelo de 200 variables en uno igual de bueno de 5 variables.

Componentes Principales: Basados en Muestras

- En la práctica observamos una muestra de realizaciones i.i.d del vector aleatorio X, $\{x_1, ..., x_n\}$.
- Sea Σ el estimador muestral de la varianza covarianza de X.
- Los datos transformados se definen como:

$$y_i = \Gamma^T (x_i - \mu)$$

  donde interpetamos y_i como una realización de las CP (rotaciones de los datos originales).
- La matrix de varianza covarianza muestral de $\{y_1, ..., y_n\}$ es Δ la matriz diagonal de valores propios de Σ.
- Las demás propiedades son análogas interpretando correctamente los parámetros como estimadores muestrales.

Componentes Principales: Teorema espectral

- Necesitamos el teorema espectral: Toda matriz simétrica A se puede descomponer como:

$$A = \Gamma \Delta \Gamma^T$$

  donde Δ es una matriz diagonal de valores propios (ordenados de mayor a menor) y Γ es una matriz de vectores propios (cada vector propio es una columna) normalizados.
- Si la matriz A es simidefinida positiva se puede garantizar que los valores propios son no negativos.

Componentes Principales: Transformación de CP

- La transformación de componentes principales de X se define como:

$$Y = \Gamma^T (X - \mu)$$

que puede interpretarse como centrar y rotar los datos.

- Y_j se conoce como la j-esima CP. Γ se conoce como la matriz de carga. Por simplicidad supongamos $\mu = 0$.
- Es fácil demostrar que $E[Y] = 0$ y $cov(Y) = \Delta$.

Modelo de Factores

- Sea Σ la matriz de varianza - covarianza de X_t.
- X_t tiene un modelo de $p-$ factores (lineal) si se puede escribir como:

$$X = a + BF + \varepsilon$$

Donde:

1. $F = (F_1, ..., F_p)'$, $p < d$. Se llaman factores.
2. La matriz de varianza covarianza Ω de F es positiva definida.
3. $\varepsilon = (\varepsilon_1, ...\varepsilon_d)'$ es un vector de choques idiosyncráticos, no correlacionados, con media cero y matriz de varianza covarianza (diagonal) Υ.
4. $B \in R^{d \times p}$ (matriz de loading) y $a \in R^{d \times p}$ son matrices y $cov(F, \varepsilon) = 0$.

Modelo de Factores

- En general es posible transformar los factores comúnes para que tengan media cero y sean ortogonales (componentes principales).
- Como Σ es una matriz simétrica, por el teorema de descomposición espectral ésta se puede escribir como:

$$\Sigma = \Gamma \Lambda \Gamma'$$

donde Λ es diagonal con valores propios ordenados de mayor a menor (reales) y Γ es ortogonal: $\Gamma\Gamma' = \Gamma'\Gamma = I$.

Modelo de Factores

- La transformación de componentes principales de X se define como:

$$F = \Gamma'(X - \mu),$$

luego $E[F] = 0$ y $cov(F) = \Lambda$ y X se puede expresar como un modelo de factores con $d = p$:

$$X = \mu + \Gamma F$$

Modelo de Factores

- La reducción de la dimension se logra eligiendo $p < d$.
- Se puede demostrar que la varianza de cada componente es igual al valor propio asociado a esa componente.
- Una vez hecho esto se descompone F de la siguiente forma $F = (F_1', F_2')$ donde F_1 son los primeros p factores y F_2' los siguientes $d - p$ factores. De forma similar particionamos $\Gamma = (\Gamma_1', \Gamma_2')$ y reescribimos X como:

$$\begin{aligned} X &= \mu + \Gamma_1 F_1 + \Gamma_2 F_2 \\ &= \mu + \Gamma_1 F_1 + \varepsilon \end{aligned}$$

donde $\varepsilon = \Gamma_2 F_2$ lo intepretamos como un error.

Modelo de Factores

- Hay una interpretación más intuitiva.
- La primera componente principal es la recta que mejor aproxima los datos.
- Va en la dirección de mayor varianza.
- La primera y segunda componente forman el plano que mejor aproxima los datos.