Unsupervised Learning

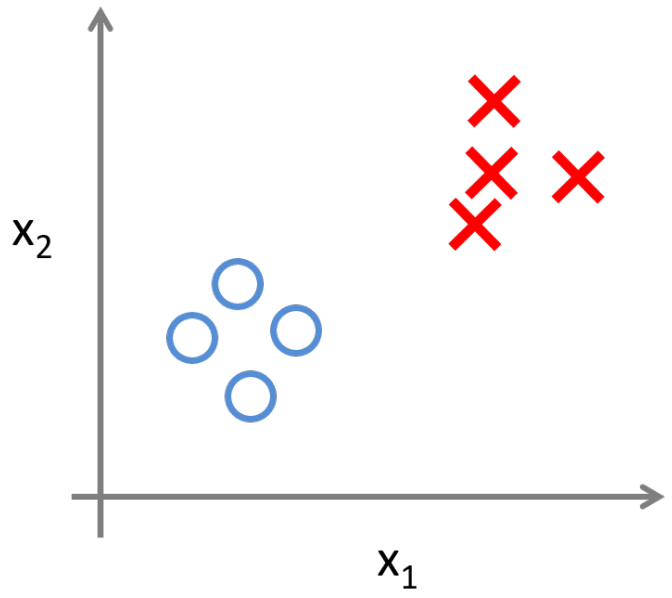
Alvaro J. Riascos Villegas
University of los Andes and Quantil

July 6 de 2018

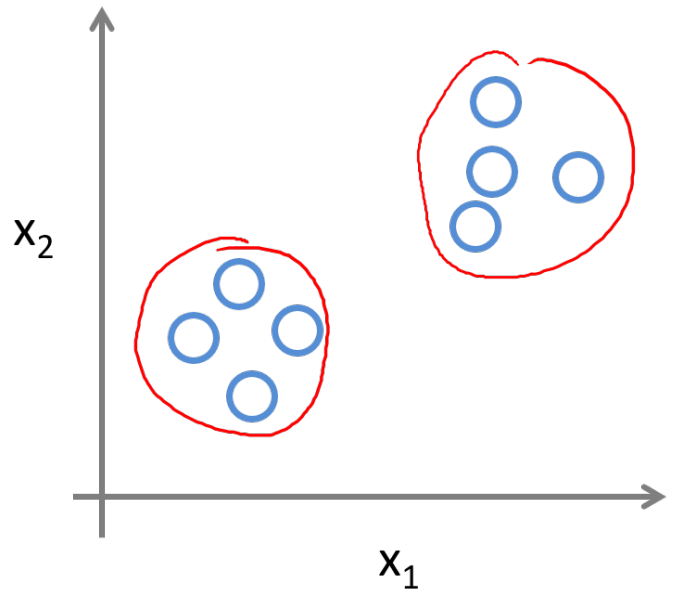
Contenido

Supervised vs. Unsupervised Learning

Supervised Learning



Unsupervised Learning



Applications

Applications

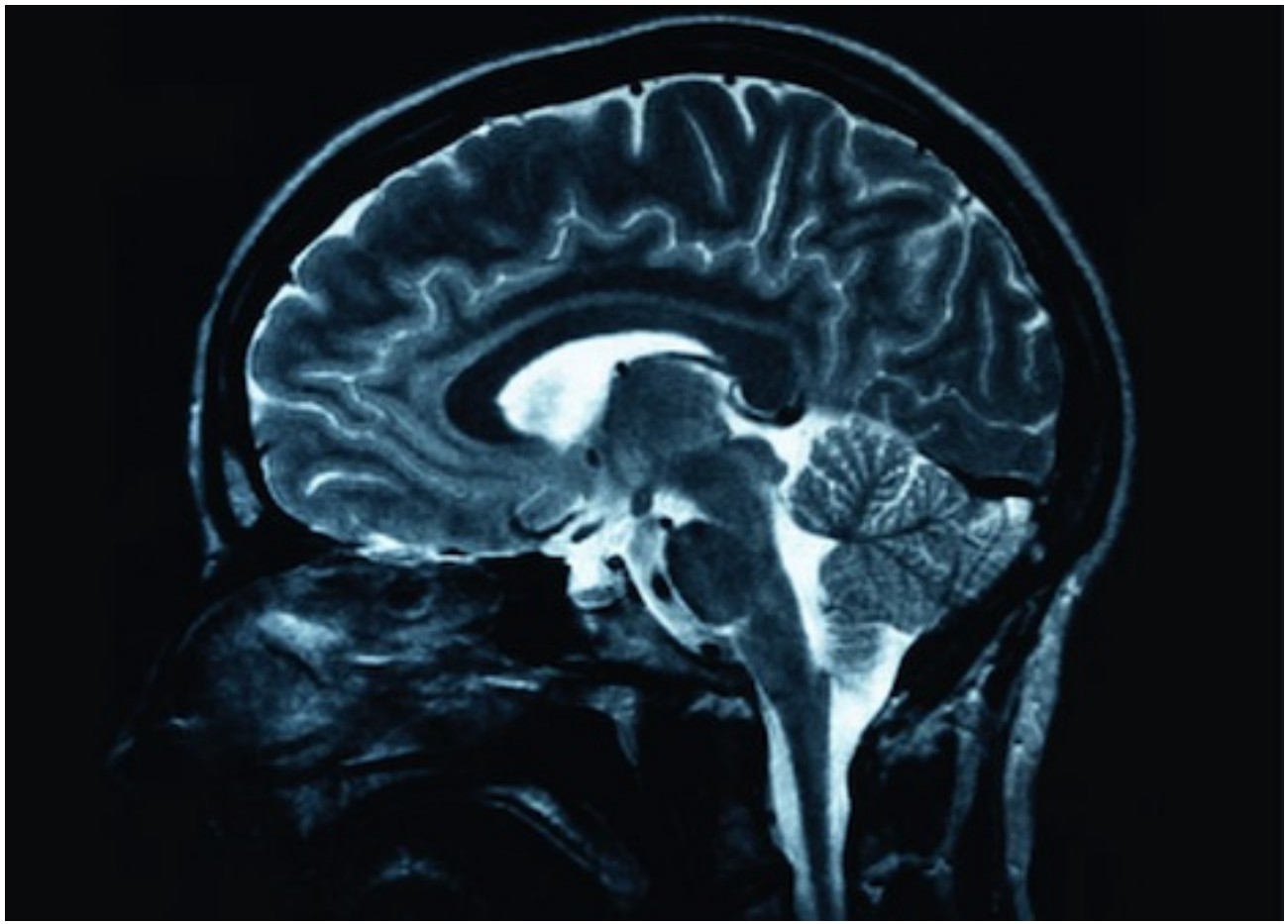
- Clients segmentation.

Applications

- Dimensionality reduction.

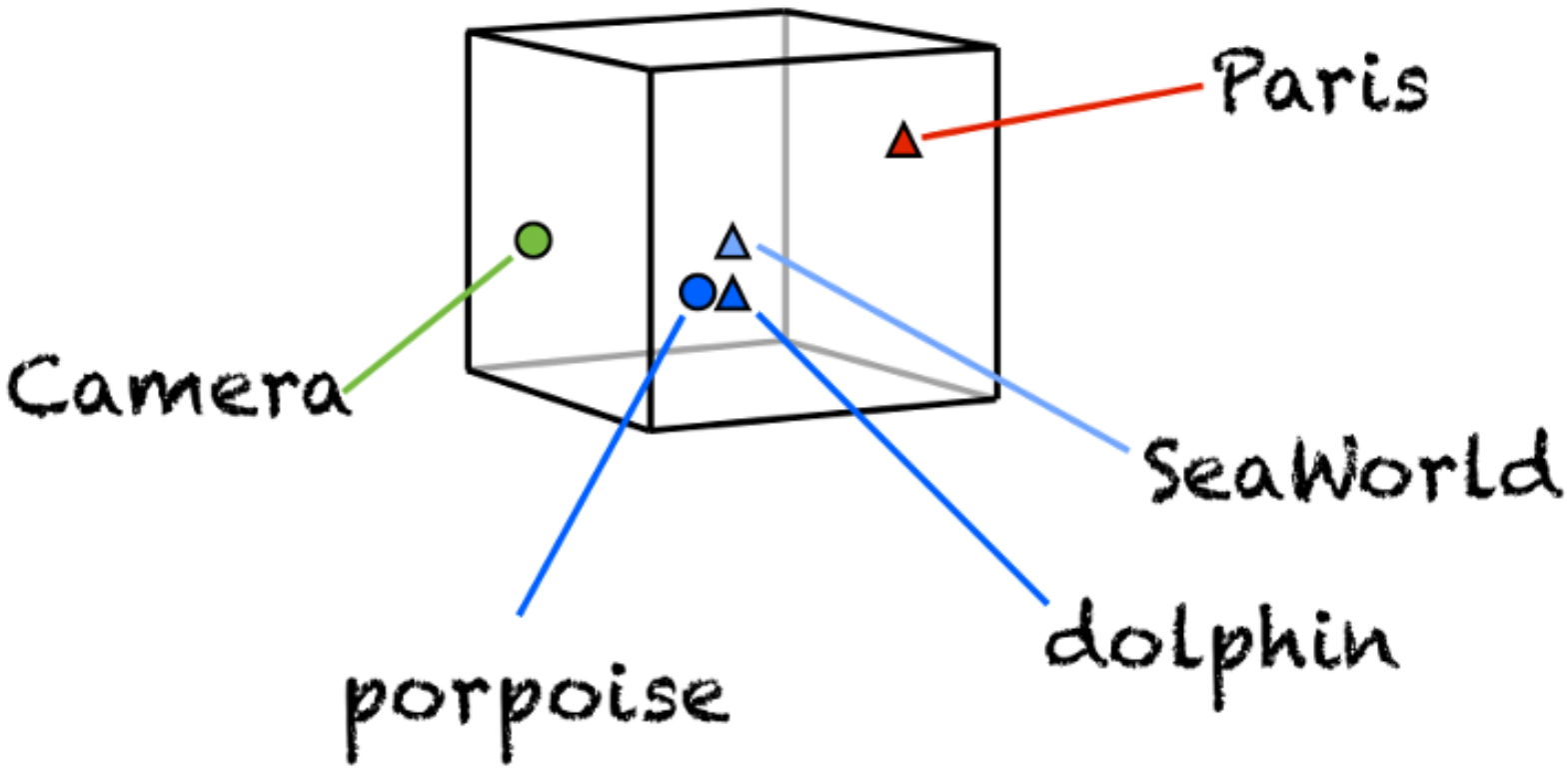
Applications

- Diagnosis.



Applications

- Meaning based on context.



Algorithms

- Association rules.
- K−medias.
- Principal components analysis.
- Mixture models.
- Local outlier factor (LOF).
- Word Embeddings (Text).
- Latent Dirichlet Allocation (Text).

Contenido

Association Rules

- The idea is to learn the distribution of X.
- We look for regions in which X has a high density and in which association between variables may be locally high.
- Let X un vector de p variables (continuos or categoric) and S_j the support of X_j. Let $s_j \subseteq S_j$. We first focus on finding $(s_1, s_2, ..., s_p)$ such that

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

where s is a minimum support level for variable j.

Association Rules

- The idea is to learn the distribution of X.
- We look for regions in which X has a high density and in which association between variables may be locally high.
- Let X un vector de p variables (continuos or categoric) and S_j the support of X_j. Let $s_j \subseteq S_j$. We first focus on finding $(s_1, s_2, ..., s_p)$ such that

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

where s is a minimum support level for variable j.

Association Rules

- The idea is to learn the distribution of X.
- We look for regions in which X has a high density and in which association between variables may be locally high.
- Let X un vector de p variables (continuos or categoric) and S_j the support of X_j. Let $s_j \subseteq S_j$. We first focus on finding $(s_1, s_2, ..., s_p)$ such that

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

where s is a minimum support level for variable j.

Association Rules: Simplification

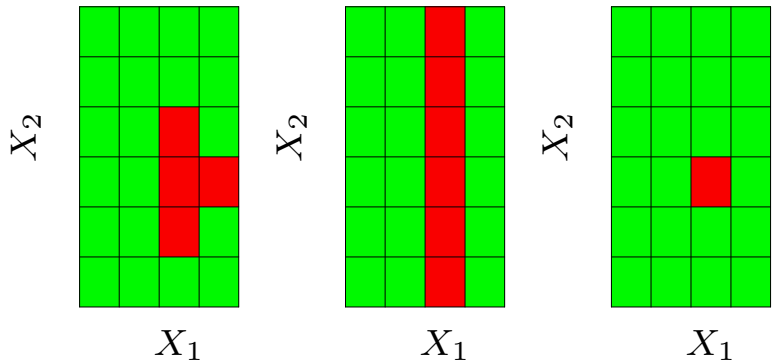


FIGURE 14.1. *Simplifications for association rules. Here there are two inputs X_1 and X_2, taking four and six distinct values, respectively. The red squares indicate areas of high density. To simplify the computations, we assume that the derived subset corresponds to either a single value of an input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

Association Rules: Dummy variable representation

- Suppose every possible value of X is represented with a dummy variable (categorize continuous variables if necessary). Assume we have K dummies.
- Each $\Upsilon \subset \{1, ..., K\}$ such that $P[\bigcap_{j \in \Upsilon}(Z_j = 1)] > s$ represents a group (basket) that has a minimum support).
- The natural estimator of this probability is the frequency of baskets that have this products.

Association Rules: Dummy variable representation

- Suppose every possible value of X is represented with a dummy variable (categorize continuous variables if necessary). Assume we have K dummies.
- Each $\Upsilon \subset \{1, ..., K\}$ such that $P[\bigcap_{j \in \Upsilon}(Z_j = 1)] > s$ represents a group (basket) that has a minimum support).
- The natural estimator of this probability is the frequency of baskets that have this products.

)

- The main questions we want to answer are:
  1. Given $\Upsilon \subset \{1, ..., K\}$ partition Υ in two sets A, B. Each expression of the form $A \Rightarrow B$, is called a rule with A the antecedent and B the consequent. the support of the rules is the support of Υ. It represents the probability that items A and B are in a basket.
  2. The confidence of the rule $A \Rightarrow B$, $C(A \Rightarrow B)$ is defined as $C(A \Rightarrow B) = \frac{Supp(A \Rightarrow B)}{Supp(A)}$. It is an estimator of de $P(B \mid A)$, a local association estimator between the two item sets.
  3. The expected confidence of B is $Supp(B = P(B)$.
  4. The *Lift* of the rule $A \Rightarrow B$ is defined as: $L(A \Rightarrow B) = \frac{C(A \Rightarrow B)}{Supp(B)}$. It is a measure of the degree of association between A and B: $\frac{P(A \wedge B)}{P(A)P(B)}$.

- The main questions we want to answer are:
  1. Given $\Upsilon \subset \{1, ..., K\}$ partition Υ in two sets A, B. Each expression of the form $A \Rightarrow B$, is called a rule with A the antecedent and B the consequent. the support of the rules is the support of Υ. It represents the probability that items A and B are in a basket.
  2. The confidence of the rule $A \Rightarrow B$, $C(A \Rightarrow B)$ is defined as $C(A \Rightarrow B) = \frac{Supp(A \Rightarrow B)}{Supp(A)}$. It is an estimator of de $P(B \mid A)$, a local association estimator between the two item sets.
  3. The expected confidence of B is $Supp(B = P(B)$.
  4. The *Lift* of the rule $A \Rightarrow B$ is defined as: $L(A \Rightarrow B) = \frac{C(A \Rightarrow B)}{Supp(B)}$. It is a measure of the degree of association between A and B: $\frac{P(A \wedge B)}{P(A)P(B)}$.

- The main questions we want to answer are:
  1. Given $\Upsilon \subset \{1, ..., K\}$ partition Υ in two sets A, B. Each expression of the form $A \Rightarrow B$, is called a rule with A the antecedent and B the consequent. the support of the rules is the support of Υ. It represents the probability that items A and B are in a basket.
  2. The confidence of the rule $A \Rightarrow B$, $C(A \Rightarrow B)$ is defined as $C(A \Rightarrow B) = \frac{Supp(A \Rightarrow B)}{Supp(A)}$. It is an estimator of de $P(B \mid A)$, a local association estimator between the two item sets.
  3. The expected confidence of B is $Supp(B = P(B)$.
  4. The *Lift* of the rule $A \Rightarrow B$ is defined as: $L(A \Rightarrow B) = \frac{C(A \Rightarrow B)}{Supp(B)}$. It is a measure of the degree of association between A and B: $\frac{P(A \wedge B)}{P(A)P(B)}$.

- The main questions we want to answer are:
  1. Given $\Upsilon \subset \{1, ..., K\}$ partition Υ in two sets A, B. Each expression of the form $A \Rightarrow B$, is called a rule with A the antecedent and B the consequent. the support of the rules is the support of Υ. It represents the probability that items A and B are in a basket.
  2. The confidence of the rule $A \Rightarrow B$, $C(A \Rightarrow B)$ is defined as $C(A \Rightarrow B) = \frac{Supp(A \Rightarrow B)}{Supp(A)}$. It is an estimator of de $P(B \mid A)$, a local association estimator between the two item sets.
  3. The expected confidence of B is $Supp(B = P(B)$.
  4. The *Lift* of the rule $A \Rightarrow B$ is defined as: $L(A \Rightarrow B) = \frac{C(A \Rightarrow B)}{Supp(B)}$. It is a measure of the degree of association between A and B: $\frac{P(A \wedge B)}{P(A)P(B)}$.

- The main questions we want to answer are:
  1. Given $\Upsilon \subset \{1, ..., K\}$ partition Υ in two sets A, B. Each expression of the form $A \Rightarrow B$, is called a rule with A the antecedent and B the consequent. the support of the rules is the support of Υ. It represents the probability that items A and B are in a basket.
  2. The confidence of the rule $A \Rightarrow B$, $C(A \Rightarrow B)$ is defined as $C(A \Rightarrow B) = \frac{Supp(A \Rightarrow B)}{Supp(A)}$. It is an estimator of de $P(B \mid A)$, a local association estimator between the two item sets.
  3. The expected confidence of B is $Supp(B = P(B)$.
  4. The *Lift* of the rule $A \Rightarrow B$ is defined as: $L(A \Rightarrow B) = \frac{C(A \Rightarrow B)}{Supp(B)}$. It is a measure of the degree of association between A and B: $\frac{P(A \wedge B)}{P(A)P(B)}$.

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.

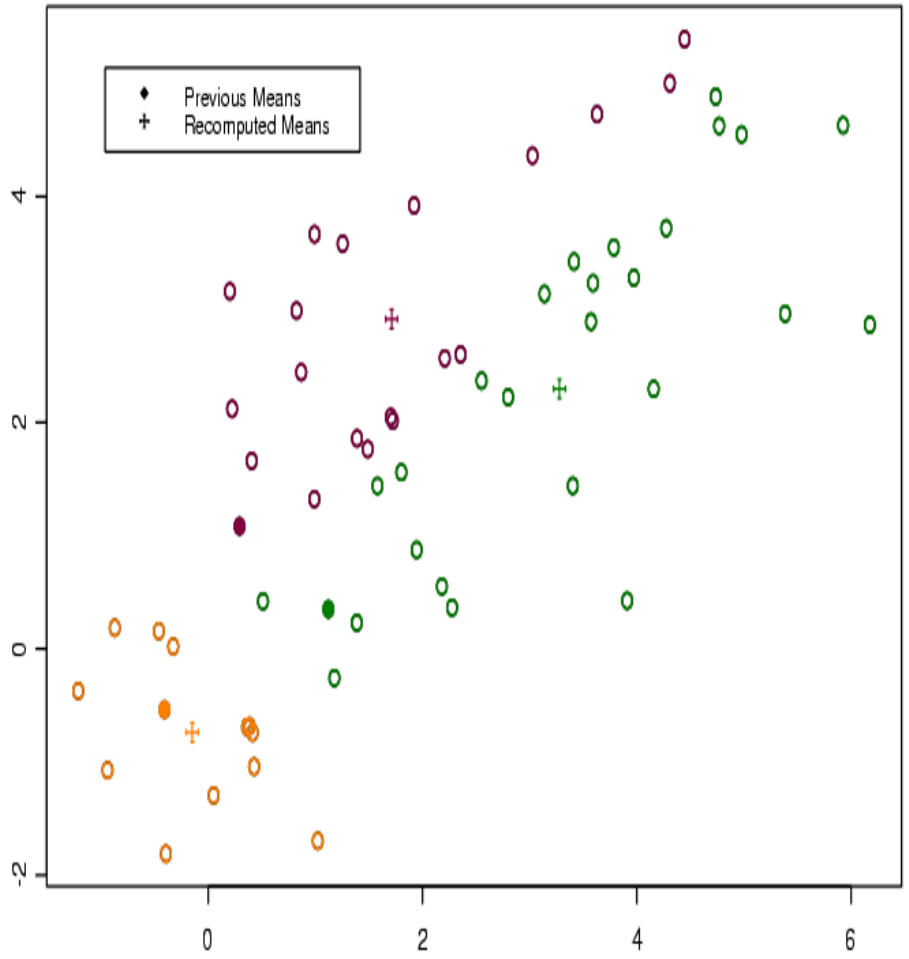
$$\left[\begin{array}{rcl} \text{language in home} & = & \textit{English} \\ \text{income} & < & \textit{\$40,000} \\ \text{marital status} & = & \textit{not married} \\ \text{number of children} & = & 0 \end{array}\right] \Downarrow \text{education} \notin \{\textit{college graduate, graduate study}\}$$

Contenido

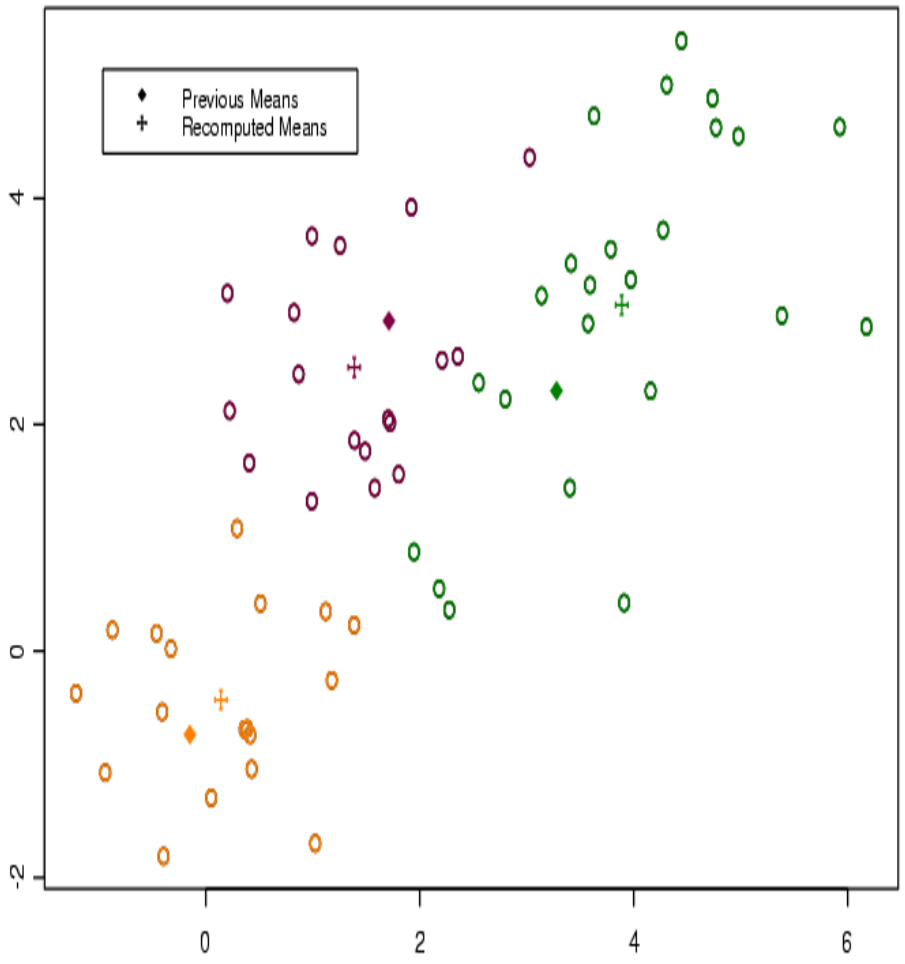
Introduction

- It is an algorithm to group data based on a dissimilarity measure (a distance).
- It is based on a distance measure in the feature space.

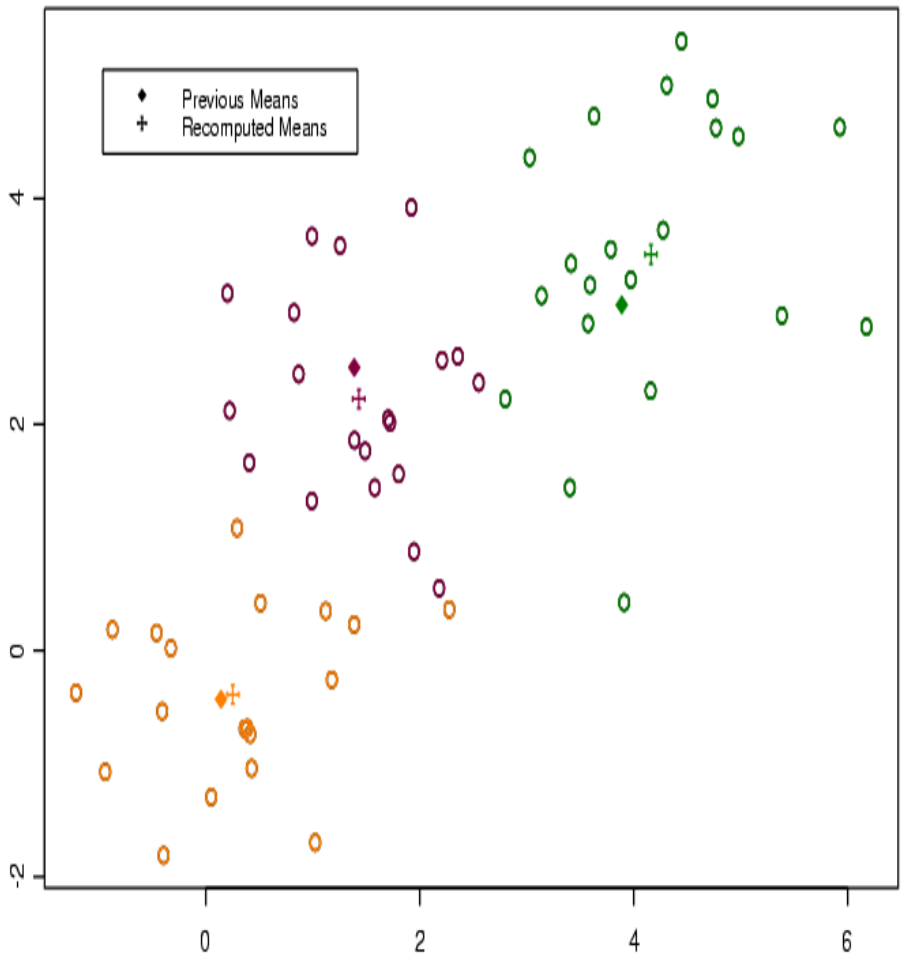
Ejemplo: Algoritmo



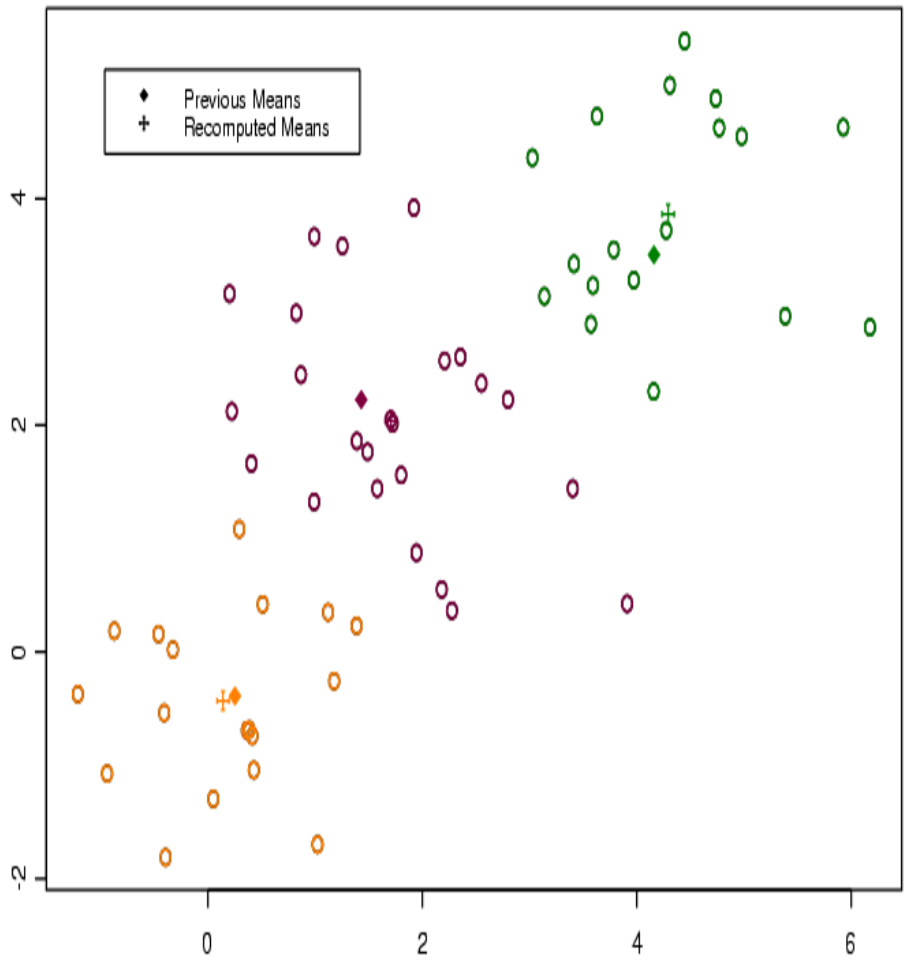
Ejemplo: Algoritmo



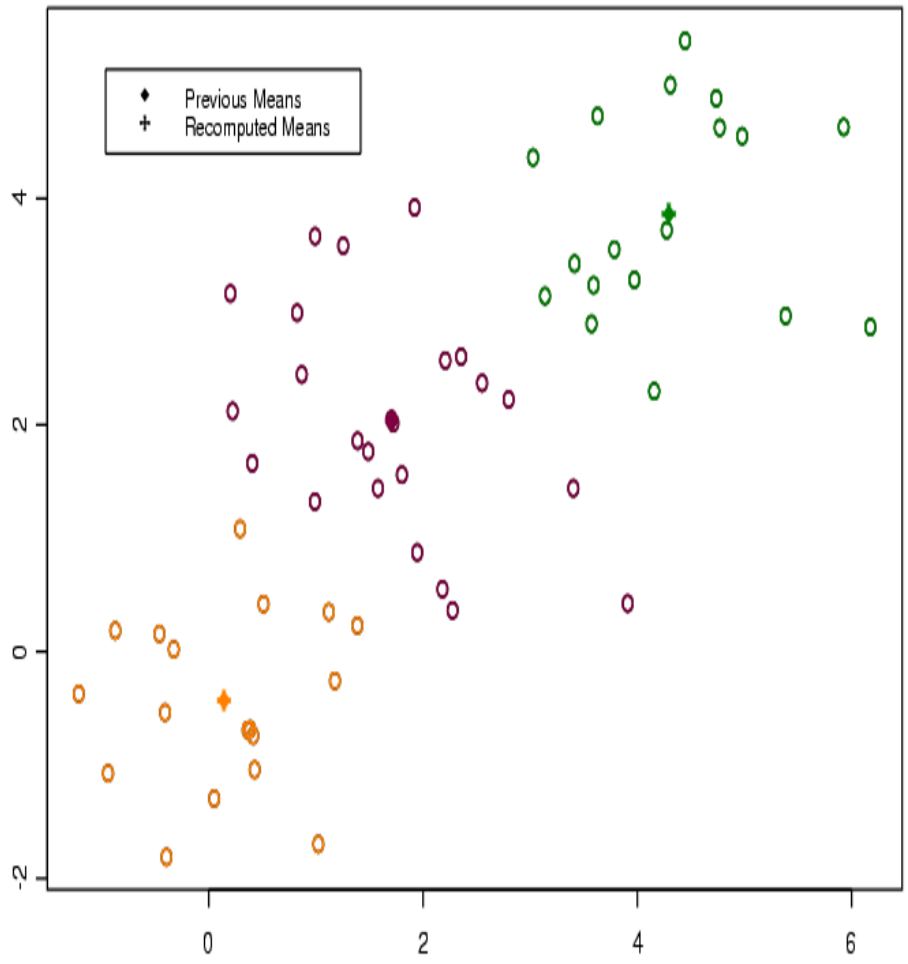
Ejemplo: Algoritmo



Ejemplo: Algoritmo



Ejemplo: Algoritmo



Ejemplo: Particiones

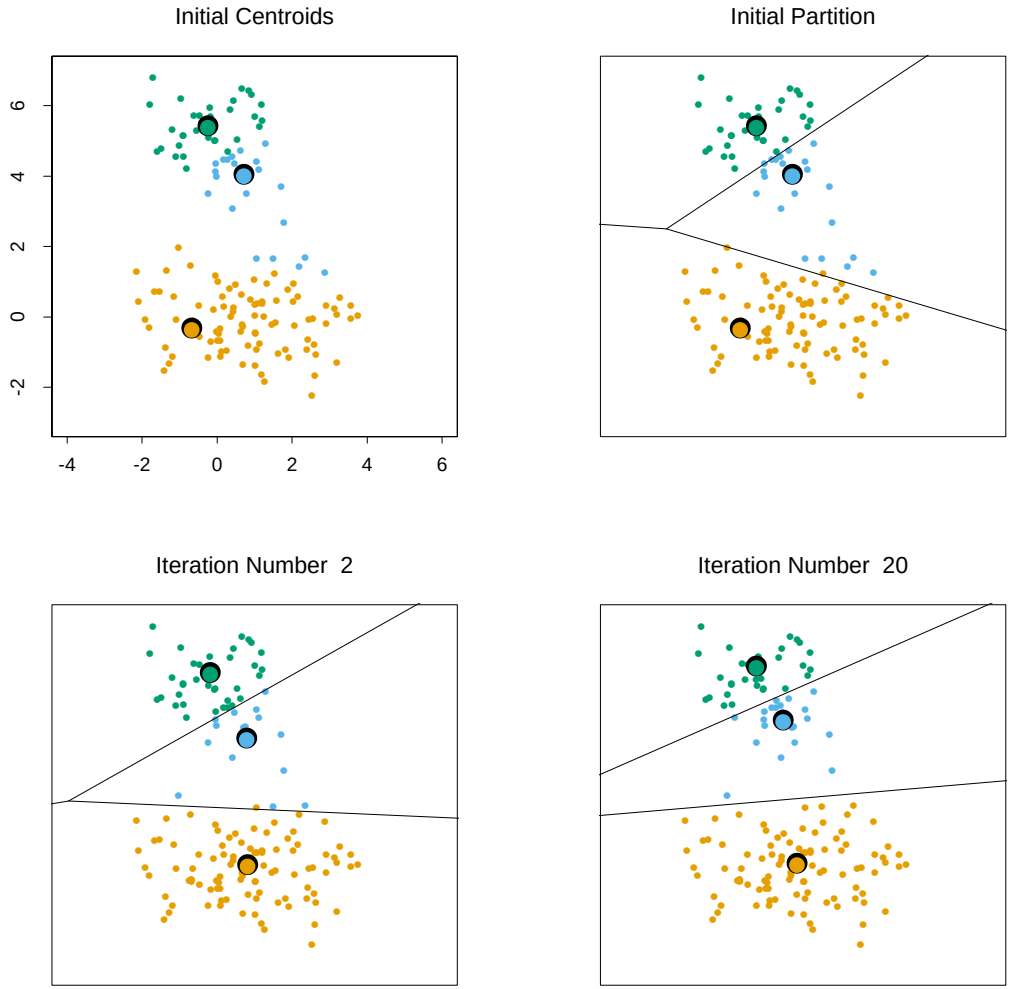


FIGURE 14.6. *Successive iterations of the K-means clustering algorithm for the simulated data of Figure 14.4.*

K-means: Problems with standardization

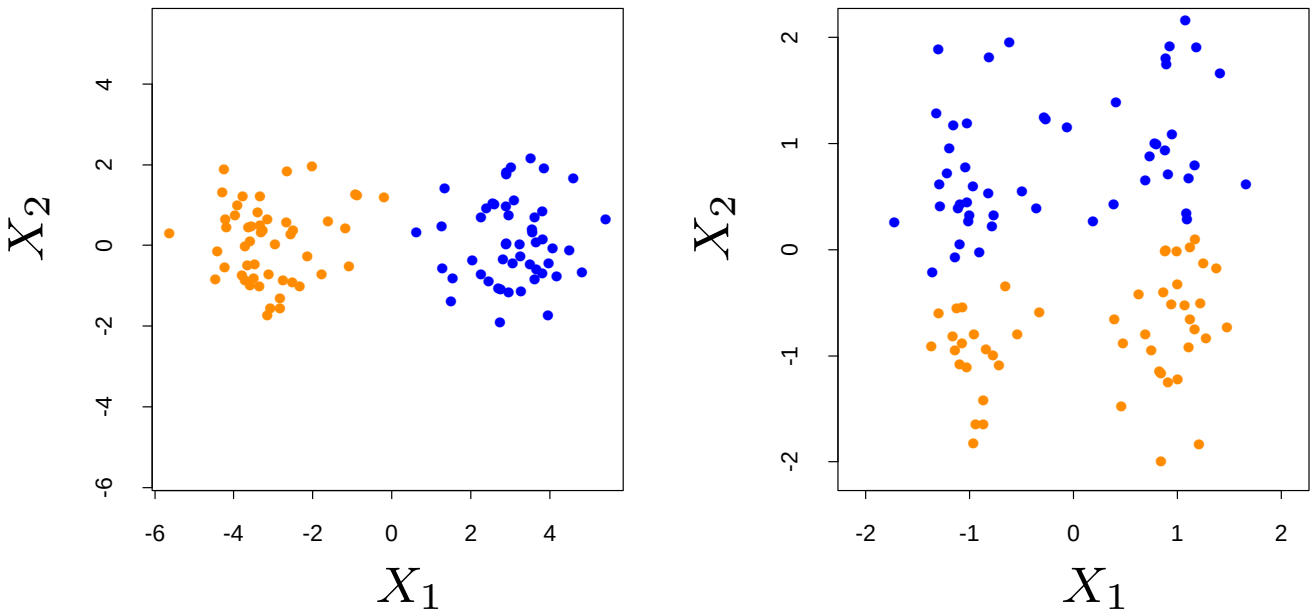
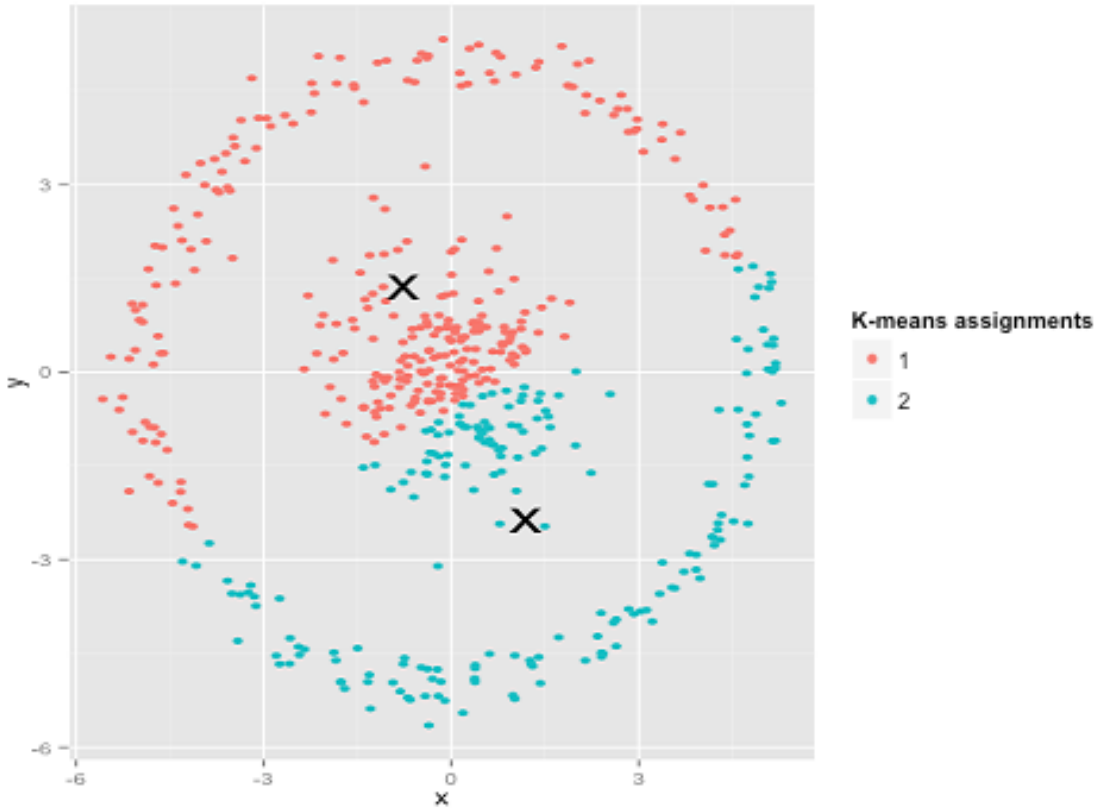
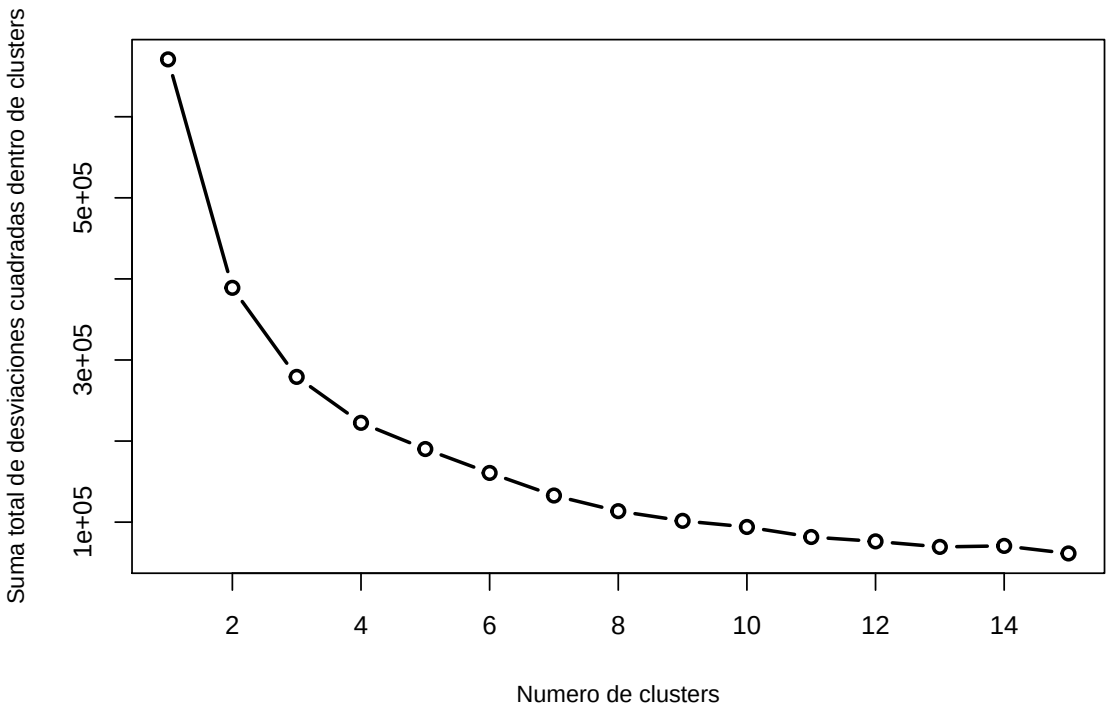


FIGURE 14.5. *Simulated data: on the left, K-means clustering (with $K=2$) has been applied to the raw data. The two colors indicate the cluster memberships. On the right, the features were first standardized before clustering. This is equivalent to using feature weights $1/[2 \cdot \mathrm{var}(X_j)]$.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

K-means: Difficult cases

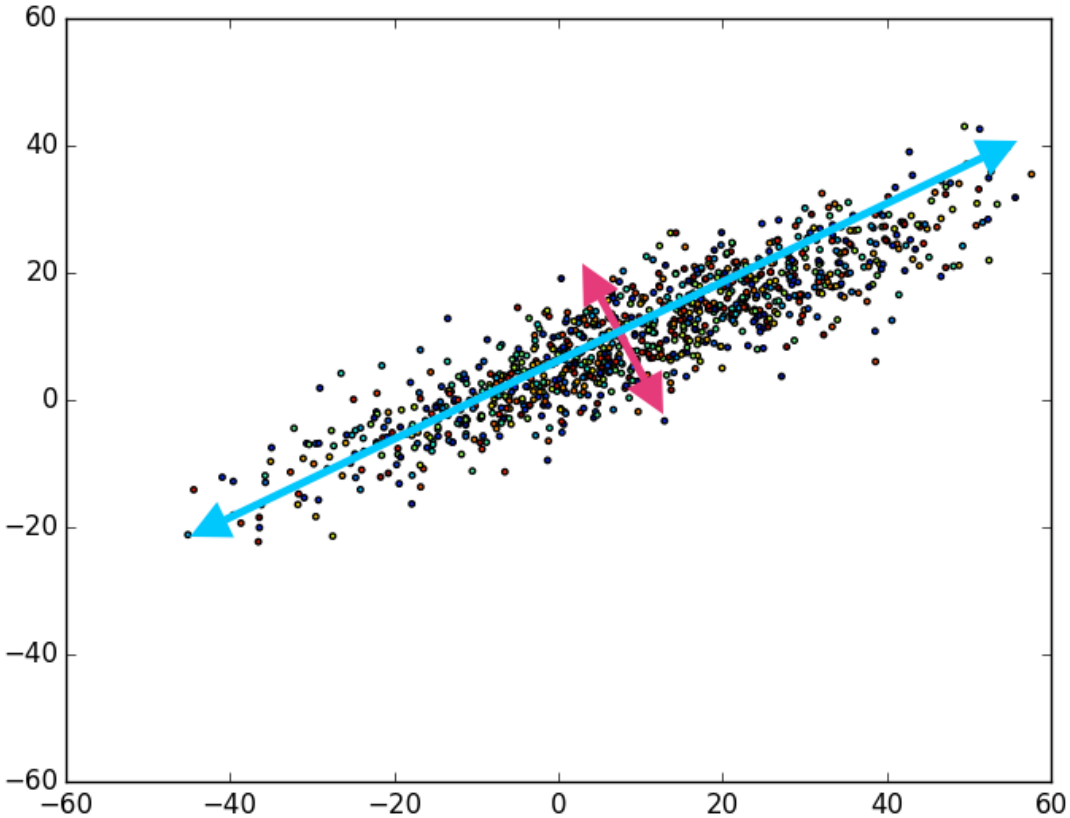


Choosing the number of clusters

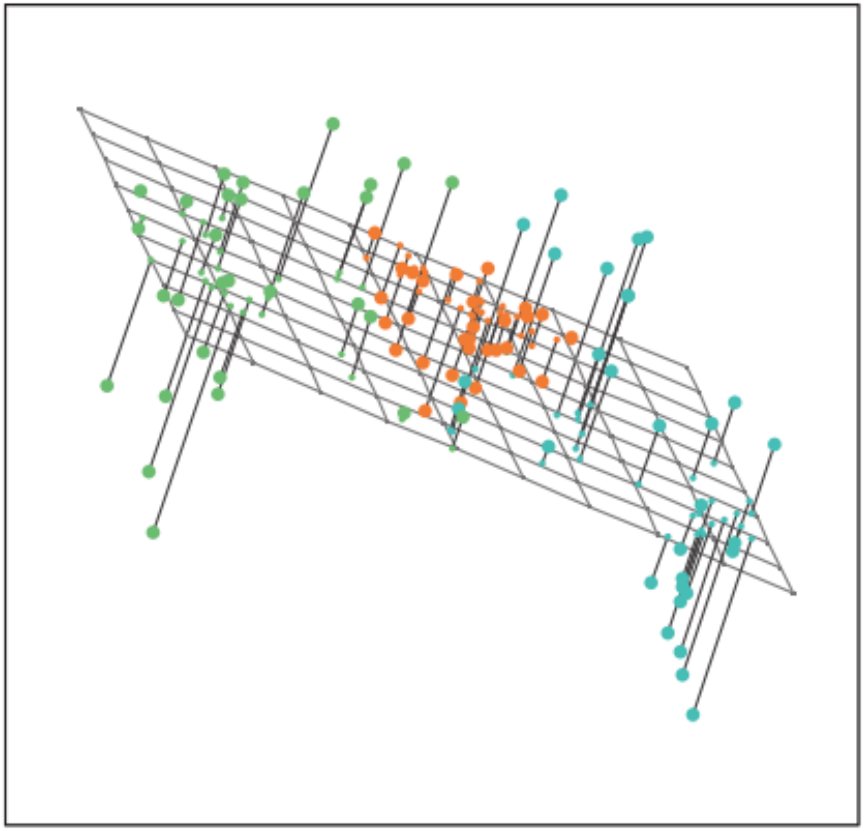


Contenido

Principal Components: Example



Principal Components: Example II



Application: Handwritten numbers

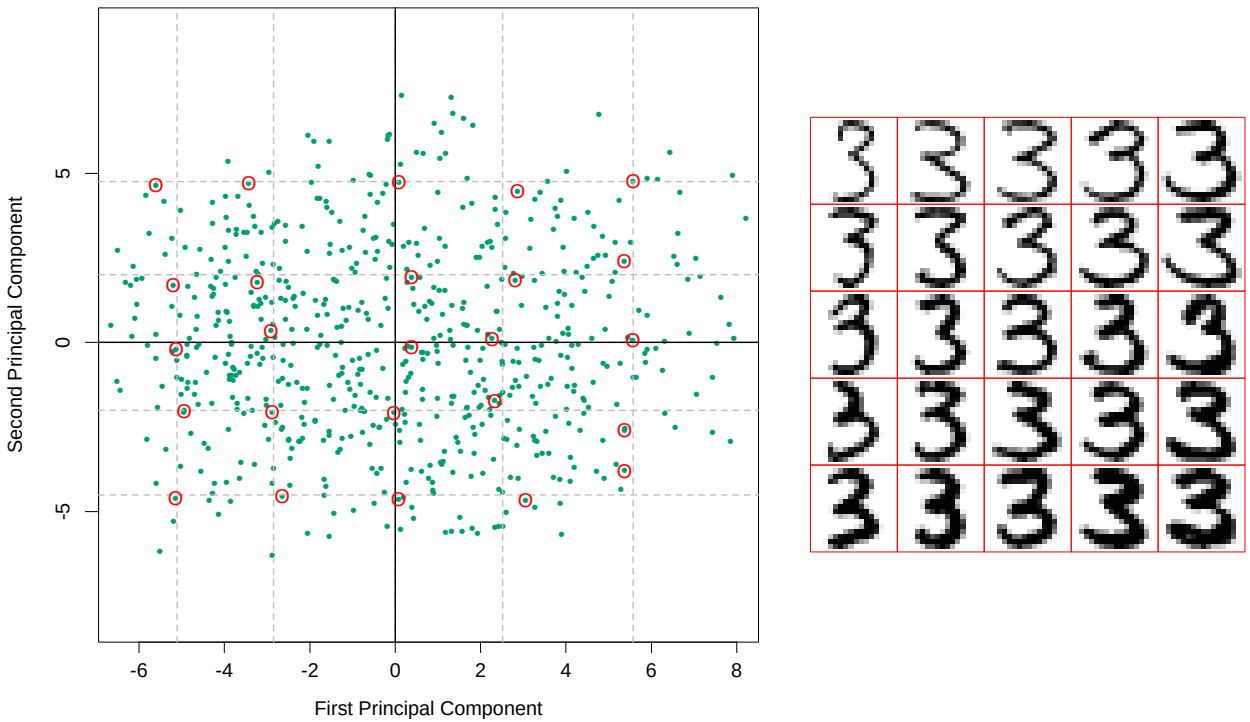


FIGURE 14.23. *(Left panel:) the first two principal components of the handwritten threes. The circled points are the closest projected images to the vertices of a grid, defined by the marginal quantiles of the principal components. (Right panel:) The images corresponding to the circled points. These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

Contenido

Estimation of densities

- Suppose $x_1, ..., x_N$ is a sample of data taken from a distribution with density $f_X(x)$.
- A first local estimator is:

$$f_X(x_0) = \frac{num\{N(x_0)\}}{N\lambda} \tag{2}$$

where $N(x_0)$ is a neighbourhood of size λ.

- A smoothed version is:

$$f_X(x_0) = \frac{1}{N\lambda}\sum_{i=1}^{N} K_\lambda(x_0, x_i) \tag{3}$$

where for example $K_\lambda(x_0, x_i) = \psi(\frac{|x-x_0|}{\lambda})$ and ψ is the standard Gaussian density $(0, 1)$.

Estimation of densities

- Suppose $x_1, ..., x_N$ is a sample of data taken from a distribution with density $f_X(x)$.
- A first local estimator is:

$$f_X(x_0) = \frac{num\{N(x_0)\}}{N\lambda} \tag{2}$$

where $N(x_0)$ is a neighbourhood of size λ.

- A smoothed version is:

$$f_X(x_0) = \frac{1}{N\lambda}\sum_{i=1}^{N} K_\lambda(x_0, x_i) \tag{3}$$

where for example $K_\lambda(x_0, x_i) = \psi(\frac{|x-x_0|}{\lambda})$ and ψ is the standard Gaussian density $(0, 1)$.

Estimation of densities

- Suppose $x_1, ..., x_N$ is a sample of data taken from a distribution with density $f_X(x)$.
- A first local estimator is:

$$f_X(x_0) = \frac{num\{N(x_0)\}}{N\lambda} \tag{2}$$

where $N(x_0)$ is a neighbourhood of size λ.

- A smoothed version is:

$$f_X(x_0) = \frac{1}{N\lambda} \sum_{i=1}^{N} K_\lambda(x_0, x_i) \tag{3}$$

where for example $K_\lambda(x_0, x_i) = \psi(\frac{|x-x_0|}{\lambda})$ and ψ is the standard Gaussian density $(0, 1)$.

Estimation of densities

- In reduced form:

$$f_X(x) = \frac{1}{N} \sum_{i=1}^{N} \psi_\lambda(x - x_i) \tag{4}$$

  where ψ_λ is the Gaussian density with standard deviation λ.
- This easily generalizes to several dimensions.

Estimation of densities

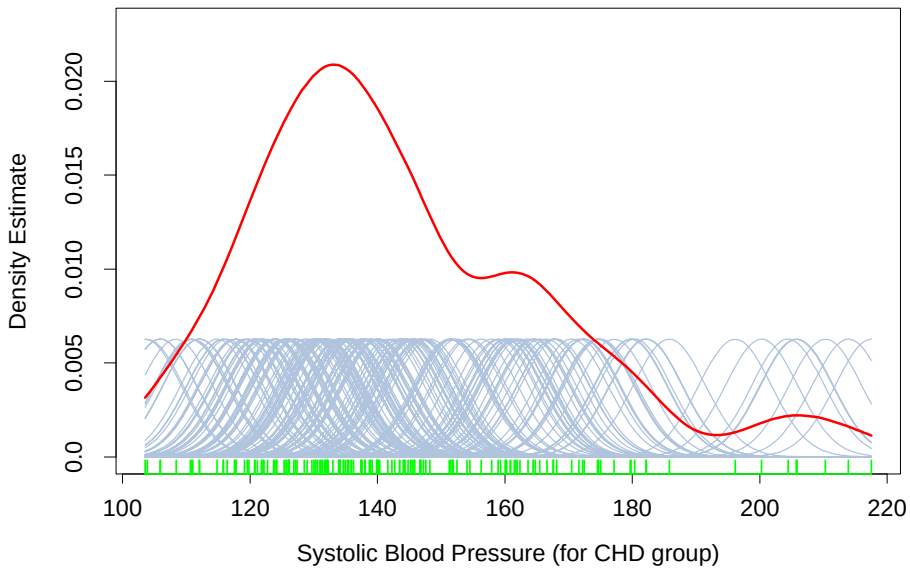


FIGURE 6.13. *A kernel density estimate for systolic blood pressure (for the CHD group). The density estimate at each point is the average contribution from each of the kernels at that point. We have scaled the kernels down by a factor of 10 to make the graph readable.*

Classification

- Using the estimations of Kernel densities by class f_j, the optimal classifier can be written as (using Bayes rule):

$$P(G = j \mid X = x_0) = \frac{\pi_j f_j(x_0)}{\sum_{k=1}^{J} \pi_k f_k(x_0)} \tag{5}$$

where π_j are the relative frequencies of each class.

Estimation of densities

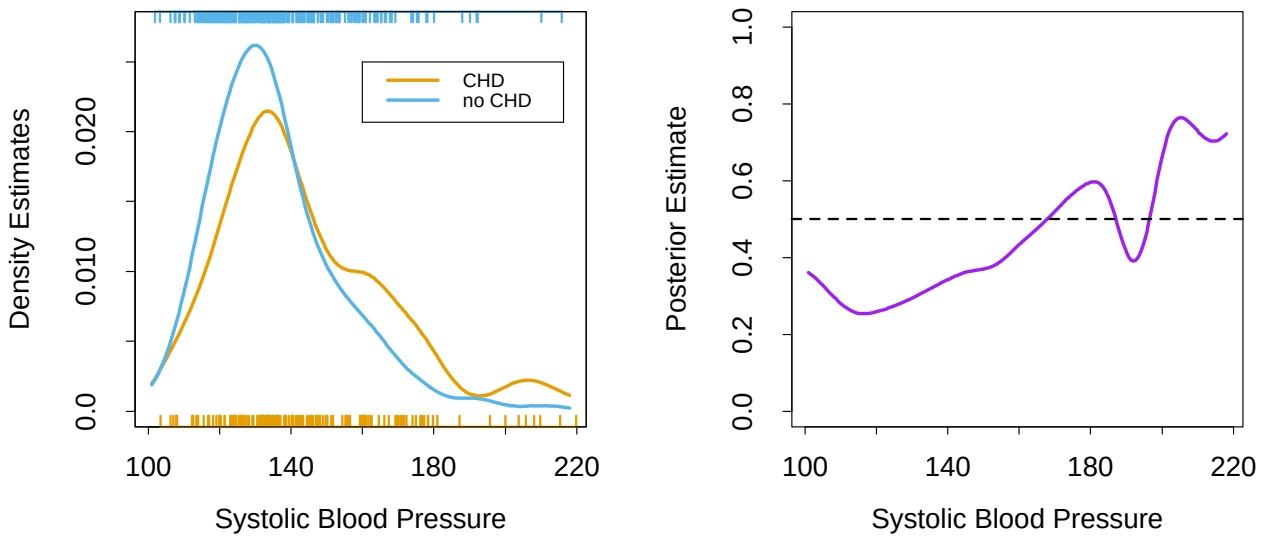


FIGURE 6.14. *The left panel shows the two separate density estimates for systolic blood pressure in the CHD versus no-CHD groups, using a Gaussian kernel density estimate in each. The right panel shows the estimated posterior probabilities for CHD, using (6.25).*

Classification: Naive Bayes

- When the feature space is to big (big p) the estimation by kernels has large variance
- Bayes classifier assumes independence among features and in that case the Bayes classifier is reduced to:

$$P(G = j \mid X = x_0) = \frac{\pi_j f_j(x_0)}{\sum_{k=1}^{J} \pi_k f_k(x_0)} \tag{6}$$

where:

$$f_j(x) = \Pi_{l=1}^{p} f_{j,l}(x_l) \tag{7}$$

- Each $f_{j,l}$ is a unidimensional kernel. The complexity of the problem is reduced immensely.
- If any feature is discrete, it allows to mix easily continuous and categorical features.

Classification: Naive Bayes

- When the feature space is to big (big p) the estimation by kernels has large variance
- Bayes classifier assumes independence among features and in that case the Bayes classifier is reduced to:

$$P(G = j \mid X = x_0) = \frac{\pi_j f_j(x_0)}{\sum_{k=1}^{J} \pi_k f_k(x_0)} \tag{6}$$

where:

$$f_j(x) = \Pi_{l=1}^{p} f_{j,l}(x_l) \tag{7}$$

- Each $f_{j,l}$ is a unidimensional kernel. The complexity of the problem is reduced immensely.
- If any feature is discrete, it allows to mix easily continuous and categorical features.

Classification: Naive Bayes

- When the feature space is to big (big p) the estimation by kernels has large variance
- Bayes classifier assumes independence among features and in that case the Bayes classifier is reduced to:

$$P(G = j \mid X = x_0) = \frac{\pi_j f_j(x_0)}{\sum_{k=1}^{J} \pi_k f_k(x_0)} \tag{6}$$

where:

$$f_j(x) = \Pi_{l=1}^{p} f_{j,l}(x_l) \tag{7}$$

- Each $f_{j,l}$ is a unidimensional kernel. The complexity of the problem is reduced immensely.
- If any feature is discrete, it allows to mix easily continuous and categorical features.

Classification: Naive Bayes

- When the feature space is to big (big p) the estimation by kernels has large variance
- Bayes classifier assumes independence among features and in that case the Bayes classifier is reduced to:

$$P(G = j \mid X = x_0) = \frac{\pi_j f_j(x_0)}{\sum_{k=1}^{J} \pi_k f_k(x_0)} \tag{6}$$

where:

$$f_j(x) = \Pi_{l=1}^{p} f_{j,l}(x_l) \tag{7}$$

- Each $f_{j,l}$ is a unidimensional kernel. The complexity of the problem is reduced immensely.
- If any feature is discrete, it allows to mix easily continuous and categorical features.