Statistical Learning

Unsupervised classification

Pedro Galeano
Department of Statistics
UC3M-BS Institute on Financial Big Data
Universidad Carlos III de Madrid
pedro.galeano@uc3m.es

Academic year 2016/2017

Master in Big Data Analytics

uc3m | Universidad Carlos III de Madrid

1. Introduction

2. Clustering framework

3. Centroid-based clustering

4. Hierarchical clustering

5. Model-based clustering

1. Introduction

2. Clustering framework

3. Centroid-based clustering

4. Hierarchical clustering

5. Model-based clustering

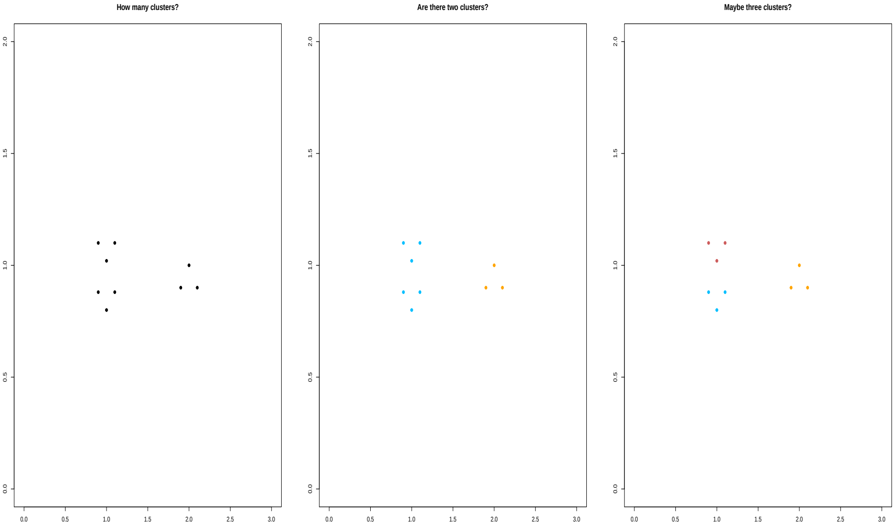
Introduction

- Unsupervised classification: Group objects in a multidimensional data set into different homogeneous groups.
- Also known as: Cluster analysis (groups are called clusters).
- How to do it?: Many, many, many, many different ways.
- Why?: Many domains of application and many different data structures.
- Consequence: Only the most important techniques can be presented here.

Introduction

- Usual approach: Group objects that are somehow similar according to some appropriate criterion that suits well with the characteristics of the data set.
- Once clusters are obtained: It is generally useful to describe each one using some descriptive tools to create a better understanding of the differences that exists among them.
- Unsupervised classification problem: Apparently, it is a simple and well defined issue.
- Nevertheless: Several questions make unsupervised classification a challenging matter:
  - What is a meaningful cluster?
  - How many clusters are appropriate?
  - How can we validate the obtained clusters?

How many clusters?


How many clusters?

Are there two clusters?

Maybe three clusters?

Introduction

- Chapter 3.R script:
  - PCA: NCI60 data set.

Introduction

- Usually: The number of clusters is unknown.

- Problem: Specify the number of clusters is not easy.

- Idea for most of methods: Explore different values and looks at potential interpretation of the clustering results.

- Few methods: Provide with the number of clusters and the clusters themselves.

Introduction

- Strength of unsupervised classification: Its exploratory nature.
- Different cluster patterns: As one varies the method, the number of clusters, tuning parameters,. . .
- Patterns might provide:
  - New insight into the structure of the data.
  - Existence of unexpected substructures, which, in turn, can lead to further or more in-depth investigations of the data.
- Subject expert: Where possible, the interpretation of a cluster analysis should involve a subject expert.

Introduction

- Unsupervised classification: There are a large vast amount of procedures.
- Focus on:
  - Centroid-based clustering: Starts from an initial random group definition and proceed by exchanging elements between groups until an appropriate cluster structure is found.
  - Hierarchical clustering: Starts with individual observations as clusters and merges clusters using cluster distances.
  - Model-based clustering: Assume that the observed variable a distribution for each cluster, fit the joint density, and assign observations based on the Bayes Theorem.
- Other methods will be covered in the machine learning courses.

Introduction

- The rest of this chapter is devoted to present:
  - Clustering framework.
  - Centroid-based clustering.
  - Hierarchical clustering.
  - Model-based clustering.

1 Introduction

2 Clustering framework

3 Centroid-based clustering

4 Hierarchical clustering

5 Model-based clustering

Clustering framework

- Data matrix: X.
- Sample size: n.
- Dimension: p.
- Indices of the observations: $1, \ldots, n$.
- Number of clusters: K.

Clustering framework

- Partition of the observations in X into K clusters: $C_1, \ldots, C_K$, that are sets containing the indices of the observations in each cluster.
- $i \in C_k$: Means that $x_{i\cdot}$ belongs to cluster k.
- Two properties needed:
  - Each observation belongs to at least one of the K clusters, i.e., $C_1 \cup \cdots \cup C_K = \{1, \ldots, n\}$.
  - No observation belongs to more than one cluster, i.e., $C_k \cap C_{k'} = \emptyset$.
- Problem: Find an appropriate partition, $C_1, \ldots, C_K$, for our data set.
- Key interpretative point: Elements within a C_k are much more similar to each other than to any element from a different $C_{k'}$.

Clustering framework

- Note: The number of possible partitions for n observations into K clusters is given by:

$$\frac{1}{K!}\sum_{k=1}^{K}(-1)^{K-k}\left(\begin{array}{c} K \\ k \end{array}\right)k^{n}$$

- For instance: For only 100 observations and 3 groups, we have 5.15×10^{47} different partitions!!!

- Thus: How to get the best one?

1. Introduction

2. Clustering framework

3. Centroid-based clustering

4. Hierarchical clustering

5. Model-based clustering

Centroid-based clustering

- Centroid: The centroid of a cluster is a representative object of the cluster.
- Note: The centroid might not be necessarily one of the observations of the group.
- Structure of centroid-based clustering procedures:
  1. Make an initial assignment of the observations into K clusters.
  2. Compute the centroids.
  3. Re-assign points to whichever centroid is closest to them.
  4. Repeat the procedure, hoping the clustering converges.
- Consequence: The similarity between two clusters is the similarity between their respective centroids.

Centroid-based clustering

- Centroid-based clustering procedures: Essentially, differs in the centroid used.

- K-means clustering: Most popular approach that uses as a centroid the sample mean vector of the members of the group.

- Used for large-scale clustering projects: K-means is extremely efficient and fast.

- Besides the centroid, key choices to be made:
  - Initial clustering.
  - Distance to measure closeness.
  - Number of iterations of the procedure.

- Other centroid-based clustering procedures: Variants of the K-means algorithm, which is presented next in detail.

Centroid-based clustering

- K-means clustering algorithm:

  1. The analyst picks the number of clusters K.
  2. Assign randomly each observation to one of the K clusters.
  3. Compute the sample mean vectors (centroids) of the K clusters.
  4. Each centroid absorbs nearby points, based on a distance.
  5. Back to step 3, until the algorithm reaches a certain number of iterations or the algorithm converges to a solution.

Centroid-based clustering

- Characteristics:

1. The algorithm requires to know the number of clusters, K (how to make this choice?).
2. The solution depends on the initial random assignment, so it is convenient to run multiple times and select the best solution (how?).
3. The algorithm uses to provide with clusters of approximately similar size.

Centroid-based clustering

- Standard approach: Use the Euclidean distance between observations and centroids of clusters.

- Characteristics:

  1. The algorithm is only applicable to quantitative variables (do not include qualitative variables).

  2. If the variables have different units of measurement, it is better to standardize the data in advance.

  3. The algorithm seeks for the partition that minimizes the within-cluster sums of squares:

  $$WSS\left(C_1,\ldots,C_K\right)=\sum_{k=1}^{K}\sum_{i\in C_k}d_E\left(x_{i\cdot},\overline{x}_k\right)^2$$

  where:

     1. $i\in C_k$ means that $x_{i\cdot}$ is in group C_k, and

     2. $d_E\left(x_{i\cdot},\overline{x}_k\right)^2$ stands for the squared Euclidean between $x_{i\cdot}$ and the sample mean vector of the observations in group k, $\overline{x}_k$.

Centroid-based clustering

- Consequences:
  - Which is the best solution for fixed K?: The one that minimizes the value of $WSS\left(C_1,\ldots,C_K\right)$.
  - Selection of K: Obtain the best solution for several values of K, and pick the one at the knee of the ratio of the within-cluster sums of squares and the between-cluster sums of squares:

$$BSS\left(C_1,\ldots,C_K\right)=\sum_{k=1}^{K} n_k\left(\overline{x}_k-\overline{x}\right)'\left(\overline{x}_k-\overline{x}\right)$$

    where:

    - ★ n_k is the number of observations assigned to cluster C_k.
    - ★ $\overline{x}$ is the sample mean vector of X.

  - Why?: $TSS=WSS\left(C_1,\ldots,C_K\right)+BSS\left(C_1,\ldots,C_K\right)$, i.e., the total sum of squares is a constant quantity independent of K and the partition $C_1,\ldots,C_K$.

Centroid-based clustering

- Is it possible to know if the cluster solution is appropriate?:
  - Silhouette: Method to validate clusters solution.
  - Let:
    - $a(x_{i\cdot})$ be the average distance of $x_{i\cdot}$ with respect all other points in its cluster.
    - $b(x_{i\cdot})$ be the lowest average distance of $x_{i\cdot}$ to any other cluster of which $x_{i\cdot}$ is not a member.
    - $s(x_{i\cdot})$ be the silhouette of $x_{i\cdot}$:

$$s(x_{i\cdot}) = \frac{a(x_{i\cdot}) - b(x_{i\cdot})}{\max\{a(x_{i\cdot}), b(x_{i\cdot})\}}$$

  - The silhouette $s(x_{i\cdot})$: Ranges from -1 to 1, such that a positive value means that the object is well matched to its own cluster and a negative value means that the object is bad matched to its own cluster.
  - The average silhouette: Gives a global measure of the assignment, such that the more positive, the better the configuration.

Centroid-based clustering

- Chapter 3.R script:
  - K-means: NCI60 data set.

Centroid-based clustering

- K-medians clustering: Replace the squared Euclidean distance with the Manhattan distance and the sample mean vector with a sample median-like vector.

- The algorithm seeks for the partition that minimizes the objective function:

$$WMedians\left(C_1,\ldots,C_K\right)=\sum_{k=1}^{K}\sum_{i\in C_k}d_M\left(x_{i\cdot},m_k\right)$$

where:

- The Manhattan distance:

$$d_M\left(x_{i\cdot},m_k\right)=\sum_{j=1}^{p}\left|x_{ij}-m_{kj}\right|$$

- Sample median-like vector: $m_k=\left(m_{k1},\ldots,m_{kp}\right)'$ is a sort of sample median for multivariate data.

Centroid-based clustering

- Characteristics:
  - The algorithm is only applicable to quantitative variables (do not include qualitative variables).
  - If the variables have different units of measurement, it is better to standardize the data in advance.
  - K-medians clustering is more resistant to outliers or strong non-Gaussianity than K-means clustering.

Centroid-based clustering

- Chapter 3.R script:
    - K-medians: NCI60 data set.

Centroid-based clustering

- K-medoids clustering: Also known as Partitioning Around Medoids (PAM).
- Medoid of a cluster: Element of the cluster whose average distance to all the observations in the cluster is minimal.
- Thus: The medoid of the cluster is the most centrally located.

Centroid-based clustering

- Algorithm:

1. Select K observations in the sample at random (initial medoids) and assign the observations to the closer medoid.

2. Compute the value of:

$$WMedoids\left(C_1,\ldots,C_K\right)=\sum_{k=1}^{K}\sum_{i\in C_k} d\left(x_{i\cdot}, med_k\right)$$

   where:

   - ⋆ med_k is the medoid of the k-th cluster.
   - ⋆ $d\left(x_{i\cdot}, med_k\right)$ is a distance (or a squared distance) between $x_{i\cdot}$ and med_k.

3. Replace one of the medoids with a non-medoid observation chosen at random:

   - ⋆ If $WMedoids\left(C_1,\ldots,C_K\right)$ is smaller than the previous one, we have a new partition.
   - ⋆ Otherwise, try with another nonmedoid point.

4. Repeat step 3 until the algorithm reaches a certain number of iterations or the algorithm converges to a solution.

Centroid-based clustering

- Characteristics:
  - K-medoids is more computationally expensive than K-means.
  - K-medoids clustering is more resistant to outliers or strong non-Gaussianity than K-means clustering.
  - If you use the Euclidean or the Manhattan distances and the variables have different units of measurement, it is better to standardize the data in advance.

Centroid-based clustering

- Chapter 3.R script:
  - K-medoids: NCI60 data set.

Centroid-based clustering

- CLARA (CLustering for lARge Applications): Extension of the k-medoids clustering method for a large number of observations.
- Idea: Apply K-medoids to a sample from the whole data set to find appropriate medoids.
- Then: Assign all observations in the data set to these medoids.
- Note: It is necessary to fix the size of the sample taken from the data set.
- Repetitions: The algorithm can be repeated several times, as K-means, to find the best solution in terms of the values of $WMedoids\left(C_1,\ldots,C_K\right)$.

Centroid-based clustering

- Chapter 3.R script:
  - CLARA: NCI60 data set.

1. Introduction
2. Clustering framework
3. Centroid-based clustering
4. Hierarchical clustering
5. Model-based clustering

Hierarchical clustering

- Hierarchical clustering methods: Unsupervised classification procedures which does not require to fix the number of groups in advance.

- Indeed: These methods produce clusters for any possible number of clusters ranging from $K = 1$ to $K = n$ by starting with n singleton clusters and merging clusters into larger groupings.

- Particularly: The K-cluster solution is obtained by merging some of the clusters from the $(K + 1)$-cluster solution.

- Distance between clusters: Hierarchical algorithms strongly depend on the distance considered between clusters.

Hierarchical clustering

- General hierarchical clustering algorithm:

1. Initially, each observation, $x_{i\cdot}$, for $i = 1, \ldots, n$, is a cluster.

2. Compute $D = \{d_{ii'}, i, i' = 1, \ldots, n\}$, the matrix that contains distances between the n observations (clusters).

3. Find the smallest distance in D, say, $d_{II'}$. Then, merge clusters I and I' to form a new cluster II'.

4. Compute distances, $d_{II',I''}$, between the new cluster II' and all other clusters $I'' \neq II'$. These distances depend upon which linkage method is used.

5. Form a new distance matrix, D, by deleting rows and columns I and I' and adding a new row and column II' with the distances computed from step 4.

6. Repeat steps 3, 4 and 5 until all observations are merged together into a single cluster.

Hierarchical clustering

- Linkage methods: Ways to compute the distance $d_{II',I''}$, between a new cluster II' and all other clusters $I'' \neq II'$:
  - Single linkage: $d_{II',I''} = \min\{d_{I,I''}, d_{I',I''}\}$.
  - Complete linkage: $d_{II',I''} = \max\{d_{I,I''}, d_{I',I''}\}$.
  - Average linkage: $d_{II',I''} = \sum_{i \in II'} \sum_{i'' \in II''} d_{i,i''} / (n_{ii'} n_{i''})$, where $n_{ii'}$ and $n_{i''}$ are the number of items in clusters II' and I'', respectively.
  - Ward linkage: $d_{II',I''}$ is the squared Euclidean distance between the sample mean vector of both clusters.

Hierarchical clustering

- Which method is better?: None of the linkage procedures is uniformly best for all clustering problems.
- Single linkage: Often leads to long clusters, joined by singleton observations near each other, a result that does not have much appeal in practice.
- Complete linkage: Tends to produce many small, compact clusters.
- Average linkage: It is dependent upon the size of the clusters, while single and complete linkage do not.
- Ward linkage: Use to provide with solutions close to the ones given by K-means.
- Thus: Compare solutions.

Hierarchical clustering

- Dendogram: Graphical representation of the procedure.
- Usefulness: Allows the user to read off the distance at which clusters are combined together to form a new cluster.
- Idea: Clusters that are similar to each other are combined at low distances, whereas clusters that are more dissimilar are combined at high distances.
- Close or far clusters?: The difference in distances defines how close (or far) clusters are of each other.

Hierarchical clustering

- How many groups?: A partition of the data into a specified number of groups can be obtained by cutting the dendogram at an appropriate distance.
- Draw a horizontal line: The number, K, of vertical lines cut by that horizontal line identifies a K-cluster solution.
- Members of the clusters: The intersection of the horizontal line and one of those K vertical lines then represents a cluster, and the items located at the end of all branches below that intersection constitute the members of the cluster.
- However: If the number of observations is high, the dendogram might be not very useful.

Hierarchical clustering

- Chapter 3.R script:
    - Hierarchical clustering: NCI60 data set.

Hierarchical clustering

- Distance between observations:
  - Quantitative variables: The Euclidean or Manhattan distances are used, after standardize the variables if they have different units of measurements.
  - Quantitative and qualitative variables: The Gower distance is used.
- Gower distance:
  1. Express the qualitative variables as indicator variables (as seen in Chapter 1).
  2. Standardize all variables individually such that the sample mean of each variable is 0 and the sample variance is 1.
  3. Compute the distance between observations using the Manhattan (or the Euclidean) distance.

Hierarchical clustering

- Chapter 3.R script:
  - Hierarchical clustering with Gower distance: Flower data set.

1 Introduction

2 Clustering framework

3 Centroid-based clustering

4 Hierarchical clustering

5 Model-based clustering

Model-based clustering

- Model-based clustering: It is assumed that the data set has been generated by a mixture of K unknown distributions.
- Mixture distribution: Observations are generated by different distributions with certain probabilities.
- Approximate model: Any continuous density can be approximated to arbitrary accuracy with a mixture density with enough mixture components.
- Gaussian mixtures: By far, the most popular approach, thus we focus on it.
- Maximum-likelihood estimation: Method to estimate the parameters associated to the Gaussian mixture.
- Then: One model parameters have been estimated, each observation is assigned to the mixture (cluster) with larger probability of having generated the observation.

Model-based clustering

- PDF of a mixture distribution:

$$f_x\left(x|\theta\right) = \sum_{k=1}^{K} \pi_k f_{x,k}\left(x|\theta_k\right)$$

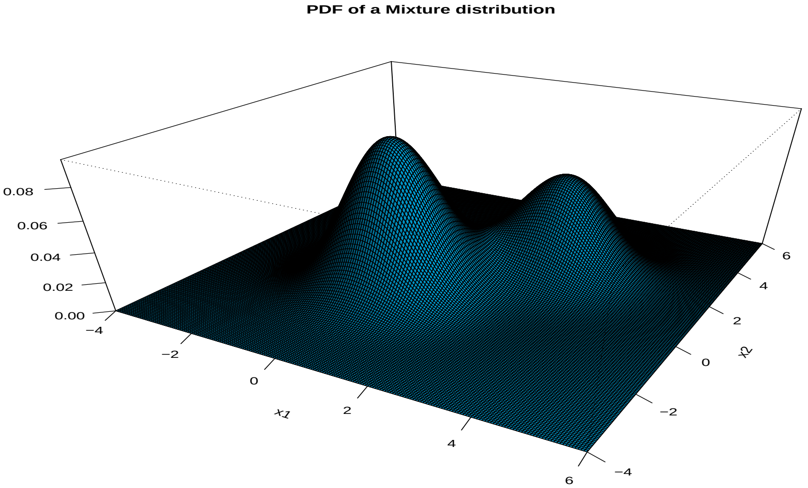
where:

  - π_k, for $k = 1, \ldots, K$, are the weights such that $\pi_1 + \cdots + \pi_K = 1$.
  - θ_k, for $k = 1, \ldots, K$, are the parameters of the distributions $f_{x,k}\left(\cdot|\theta_k\right)$.
  - θ is a vector with all the parameters of the model.

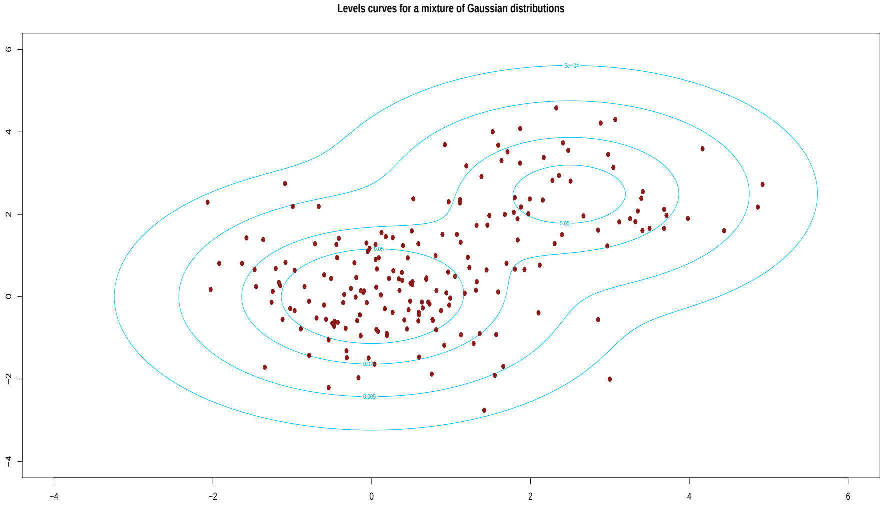
- Gaussian mixture: $f_{x,k}\left(\cdot|\theta_k\right)$ are PDFs of Gaussian distributions $N\left(\mu_k, \Sigma_k\right)$, with mean vectors $\mu_1, \ldots, \mu_K$ and covariance matrices $\Sigma_1, \ldots, \Sigma_K$, respectively.

- Thus: x can be $N\left(\mu_1, \Sigma_1\right)$ with probability π_1, $N\left(\mu_2, \Sigma_2\right)$ with probability π_2, and so on.

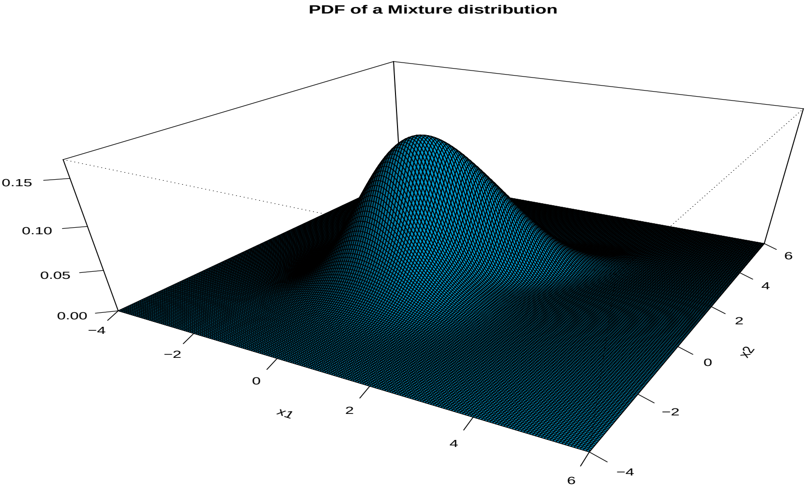
How many clusters?



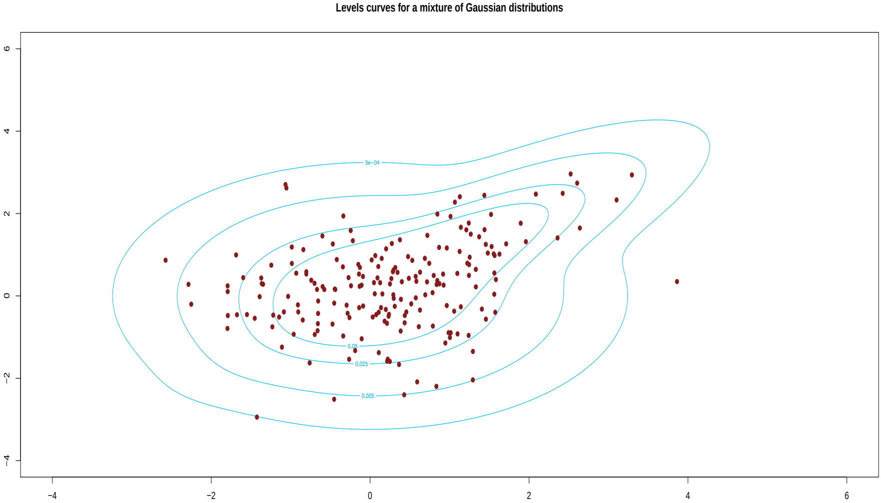
How many clusters?



How many clusters?



How many clusters?



Model-based clustering

- Maximum likelihood estimation: Method to estimate the parameters of the Gaussian mixture, i.e., $\pi_1, \ldots, \pi_K$, $\mu_1, \ldots, \mu_K$, and $\Sigma_1, \ldots, \Sigma_K$.
- Parameter estimates: $\widehat{\pi}_1, \ldots, \widehat{\pi}_K$, $\widehat{\mu}_1, \ldots, \widehat{\mu}_K$, and $\widehat{\Sigma}_1, \ldots, \widehat{\Sigma}_K$.
- Bayes Theorem: The estimated posterior probabilities that observation $x_{i\cdot}$ belongs to population k are obtained by applying the Bayes Theorem:

$$\widehat{\Pr}\left(k|x_{i\cdot}\right) = \frac{\widehat{\pi}_k f_{x,k}\left(x_{i\cdot}|\widehat{\mu}_k, \widehat{\Sigma}_k\right)}{\sum_{g=1}^{K} \widehat{\pi}_g f_{x,g}\left(x_{i\cdot}|\widehat{\mu}_g, \widehat{\Sigma}_g\right)}$$

- Cluster assignment: The observation $x_{i\cdot}$ is assigned to the cluster with maximum value of $\widehat{\Pr}\left(k|x_{i\cdot}\right)$.

Model-based clustering

- Number of groups: In model-based clustering, it is possible to select the number of groups, K, from the data set.

- Idea: Compare solutions with different values of $K = 1, 2, \ldots$ and choose the best result.

- Bayesian Information Criterion (BIC): Method to select the optimal K.

- Select the value of K that minimizes:

$$BIC\left(k\right) = -2 \times L_k\left(\widehat{\mu}_k, \widehat{\Sigma}_k | X\right) + \log\left(n\right) \times q_k$$

  where:

  - $L_k\left(\widehat{\mu}_k, \widehat{\Sigma}_k | X\right)$ denotes the maximized log-likelihood assuming k groups and q_k is the number of parameters of the model assuming the k groups.

Model-based clustering

- Dimensionality problem: When p is large, the number of parameters needed to perform model-based clustering is quite large.
- Dimension reduction: Once more, the idea is to apply a dimension reduction technique (PCA) before performing clustering, if needed.
- M-clust: The most popular method to perform model-based clustering with Gaussian mixtures.
- Reduce the number of parameters to fit: M-clust works with the spectral decomposition of the covariance matrices Σ_k, given by:

$$\Sigma_k = \lambda_{1,k} V_k \widetilde{\Lambda}_k V_k',$$

  where $\lambda_{1,k}$ is the largest eigenvalue, V_k is the matrix that contains the eigenvectors of Σ_k and $\widetilde{\Lambda}_k$ is the diagonal matrix of eigenvalues divided by $\lambda_{1,k}$.

Model-based clustering

- The decompostion allows for different configurations:

    1. spherical and equal volume.
    2. spherical and unequal volume.
    3. diagonal and equal volume and shape.
    4. diagonal, varying volume and equal shape.
    5. diagonal, equal volume and varying shape.
    6. diagonal, varying volume and shape.
    7. ellipsoidal, equal volume, shape, and orientation.
    8. ellipsoidal, equal volume and equal shape.
    9. ellipsoidal and equal shape.
    10. ellipsoidal, varying volume, shape, and orientation.

- Here: (i) spherical, diagonal and ellipsoidal are relative to the covariance matrices; (ii) similar volume means that $\lambda_{1,1} = \cdots = \lambda_{1,K}$; (iii) equal shape means $\widetilde{\Lambda}_1 = \cdots = \widetilde{\Lambda}_K$; and (iv) equal orientation means $V_1 = \cdots = V_K$.

Model-based clustering

- Chapter 3.R script:
  - Model-based clustering: NCI60 data set.

1. Introduction

2. Clustering framework

3. Centroid-based clustering

4. Hierarchical clustering

5. Model-based clustering