Statistical Learning

Week 4 - Unsupervised classification (I)

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

Academic Year 2017/2018

Master in Big Data Analytics

uc3m | Universidad Carlos III de Madrid

1. Introduction
2. Clustering framework
3. Partitional clustering
4. Hierarchical clustering

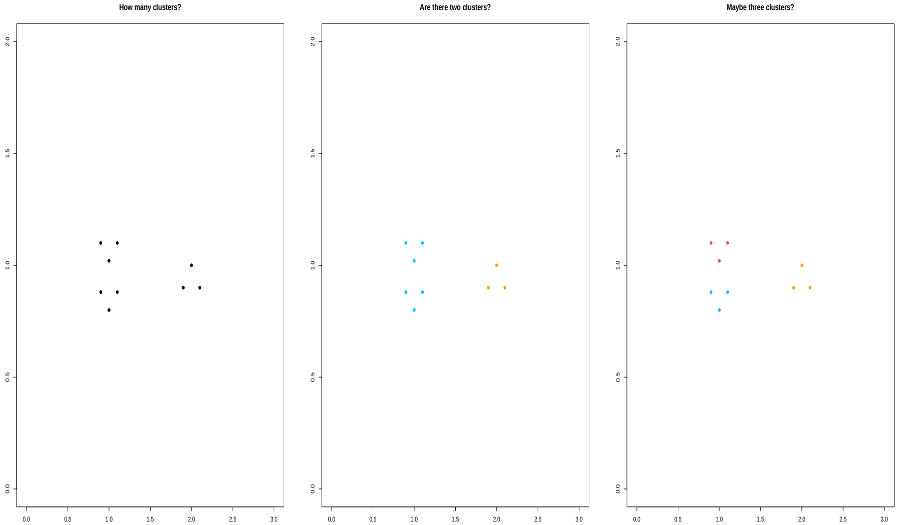
Introduction

- 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: Describe each one using descriptive tools to better understand the differences that exists between them.
- Apparently: Unsupervised classification or clustering is a simple and well defined problem.
- Nevertheless: Several questions make clustering a challenging issue:
  - What is a meaningful cluster?
  - How many clusters are appropriate?
  - How can we validate the obtained clusters?

How many clusters?



Introduction

- Week 4.R script:
  - PCA: NCI60 data set.

Introduction

- Usually: The number of clusters is unknown.
- Problem: Specify the number of clusters a priori is not easy.
- Approach for most of methods: Compare the solutions for different number of clusters.
- Nevertheless: A few methods provide with the number of clusters and the clusters themselves.

Introduction

- Categories of procedures for clustering:
  - Partitional clustering: Starts from an initial cluster definition and proceed by exchanging elements between clusters until an appropriate cluster structure is found.
  - Hierarchical clustering:
    - Agglomerative algorithms: Start with clusters containing a single observation and continues merging the clusters.
    - Divisive algorithms: Start with a single cluster containing all the observations and continues splitting clusters.
  - 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.

Introduction

- The rest of this chapter is devoted to present:
  - The general clustering framework.
  - Partitional clustering.
  - Hierarchical clustering.
  - Model-based clustering (next week).

Clustering framework

- Data matrix: X.
- Sample size: n.
- Dimension: p.
- Indexes 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 indexes of the observations in each cluster.
- $i \in C_k$: Means that $x_{i\cdot}$ belongs to cluster k, where $k = 1, \ldots, K$.
- 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 the most appropriate partition, $C_1, \ldots, C_K$, for our data matrix.
- Key interpretative point: Elements within a C_k are 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 100 observations and 3 groups, we have 5.15×10^{47} different partitions!!!

- Thus: Even if we know K, it is not possible to explore all the possible partitions.

- Then: How to get the most appropriate one?

Partitional clustering

- Main goal of partitional clustering: Find the partition of the data matrix that minimize a certain optimality criterion.
- The K-means algorithm: One of the first and most popular clustering methods.
- Other partitional algorithms: Variants of the K-means algorithm.
- Main advantage of partitional clustering: Very efficient for clustering large data matrices.
- Main disadvantage of partitional clustering: Does not find clusters of arbitrary shapes.

Partitional clustering

- Some characteristics of the K-means algorithm:
  - The number of clusters, K, is assumed to be given.
  - The algorithm is only applicable to quantitative variables (do not include qualitative variables, not even binary variables).
  - If the variables have different units of measurement, it is necessary to standardize the data in advance.
  - 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 $x_{i\cdot}$ belongs to the set 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 cluster k, $\overline{x}_k$.

Partitional clustering

- The K-means algorithm:

1. Let $C_1, \ldots, C_K$ an initial partition of the data matrix leading to K initial clusters.
2. Compute the sample mean vectors of the K initial clusters, $\overline{x}_1, \ldots, \overline{x}_K$.
3. For each observation $x_{i\cdot}$:
    1. Compute the Euclidean distances between $x_{i\cdot}$ and $\overline{x}_1, \ldots, \overline{x}_K$, denoted by $d_E(x_{i\cdot}, \overline{x}_1), \ldots, d_E(x_{i\cdot}, \overline{x}_K)$.
    2. Re-assign $x_{i\cdot}$ to the cluster with closest sample mean vector, i.e., re-assign $x_{i\cdot}$ to the k-th cluster if $d_E(x_{i\cdot}, \overline{x}_k)$ is the minimum of $d_E(x_{i\cdot}, \overline{x}_1), \ldots, d_E(x_{i\cdot}, \overline{x}_K)$.
4. Back to step 3, until the algorithm reaches a certain number of iterations or the algorithm converges to a solution.

Partitional clustering

- Initial assignment:
  - The solution of the algorithm depends on the initial assignment.
  - Two alternatives:
    - Provide with a good initial solution.
    - Run the algorithm multiple times with initial random assignments and choose the solution that minimizes the value of $WSS\,(C_1, \ldots, C_K)$.

Partitional clustering

- Week 4.R script:
  - K-means: NCI60 data set.

Partitional clustering

- How to select K?:
  - Total sums of squares:

$$TSS = \sum_{i=1}^{n} (x_{i\cdot} - \overline{x})' (x_{i\cdot} - \overline{x})$$

  - Within-cluster sums of squares:

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

  - Between-cluster sums of squares:

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

    where n_k is the number of observations assigned to cluster C_k.
  - Note that: $TSS = WSS (C_1, \ldots, C_K) + BSS (C_1, \ldots, C_K)$.

Partitional clustering

- How to select K?:

1. Obtain the optimal solution for $K = 2, \ldots, K_{\max}$, for a certain threshold $K_{\max}$.
2. Obtain the ratios $WSS\left(C_1, \ldots, C_K\right) / BSS\left(C_1, \ldots, C_K\right)$, for $K = 2, \ldots, K_{\max}$.
3. Select K as the value at which the ratio decrease stabilizes at a level close to 0.

Partitional clustering

- Week 4.R script:
  - K-means: NCI60 data set.

Partitional clustering

- How to know whether or not the cluster solution is appropriate?:
  - 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.

Partitional clustering

- Week 4.R script:
  - K-means: NCI60 data set.

Partitional clustering

- Main problems of the K-means algorithm:
  - K-means only runs with quantitative variables.
  - K-means is highly affected by outliers.
  - K-means has problems when the shape of the clusters is irregular.
- Variants:
  - K-medoids.
  - CLARA.
  - K-medoids with mixed variables.

Partitional clustering

- K-means clustering: Sensitive to outliers because it uses the Euclidean distance and the sample mean vectors.
- Idea: Replace the sample mean vector as center of the cluster with an element of the cluster itself.
- Medoids: Most centrally members of the clusters.
- More specifically: The medoid of the cluster is the element of the cluster whose average distance to all the observations in the cluster is minimal.
- Moreover: Replace the Euclidean distance with another more robust distance.
- K-medoids clustering: Also known as Partitioning Around Medoids (PAM).

Partitional clustering

- Which distance to use?:
  - med_k: Medoid of the k-th cluster.
  - Manhattan distance:

$$d_{Man}(x_{i\cdot}, med_k) = \sum_{j=1}^{p} |x_{ij} - med_{kj}|$$

    where $med_k = (med_{k1}, \ldots, med_{kp})'$.
  - Example: Two observations have values very close except for one or two variables.
  - Euclidean distance: Largely influenced by the discrepant variables.
  - Manhattan distance: Largely influenced by the closeness variables.

Partitional clustering

- K-medoids clustering (PAM) 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:

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

  3. Search if replacing any of the k medoids with a non-medoid observation of the corresponding cluster reduces the value of $WMedoids\left(C_1, \ldots, C_K\right)$.
     1. If we found a new medoid, re-assign the observations to the closer medoid and repeat the search.
     2. Otherwise, the algorithm stops.

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

Partitional clustering

- Week 4.R script:
  - K-medoids: NCI60 data set.

Partitional 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 random sub-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 sub-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 $WMedoids\left(C_1, \ldots, C_K\right)$.

Partitional clustering

- Week 4.R script:
  - CLARA: NCI60 data set.

Partitional clustering

- Partitional clustering with quantitative and qualitative variables:
  - Similar approach: It is possible to use the K-medoids (PAM) algorithm but replacing the Manhattan distance with a distance appropriate for mixed variables.
  - Gower distance:
    1. Express the qualitative variables as binary variables (if c is the number of classes of a variable, define $c-1$ binary variables indicating $c-1$ of the classes).
    2. Standardize all (quantitative and binary) 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 distance.

Partitional clustering

- Week 4.R script:
  - K-medoids with mixed variables: Credit data set.

Hierarchical clustering

- Hierarchical clustering methods: Unsupervised classification procedures which does not require to fix the number of groups in advance.
- Two approaches:
  - Agglomerative algorithms: Start with clusters containing a single observation and continues merging the clusters.
  - Divisive algorithms: Start with a single cluster containing all the observations and continues splitting clusters.
- Distance between clusters: Hierarchical algorithms strongly depend on the distance considered between clusters.
- Mixed variables: It is possible to cluster mixed variables if the Gower distance is used as a measure of disparity between observations.

Hierarchical clustering

- General agglomerative 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

- Week 4.R script:
  - Hierarchical clustering: NCI60 data set and Credit data set.

Hierarchical clustering

- Divisive algorithms: Proceeds the opposite way of agglomerative hierarchical algorithms.
- Idea: Initially, all the observations belongs to a single cluster, and at each step an existing cluster is divided into two clusters.
- DIvisive ANAlysis Clustering (DIANA): The most popular algorithm that performs divisive hierarchical clustering.
- Diameter of a cluster: Largest distance between two observations in the cluster.
- At each step: The cluster with largest diameter is split into two clusters.
- Repeat: This step is repeated until all observations are a single cluster.

Hierarchical clustering

- DIANA algorithm (for a single cluster):

  1. Let C the cluster to split.
  2. Find the observation that has the largest average distance from all other observations in the data set, which is set up as cluster C_1, while the rest are in cluster C_2.
  3. For all the observations in C_2, compute:
     * the average distance between that observation and all other observations in cluster C_2; and
     * the average distance between that observation and all observations in cluster C_1.
  4. Re-assign the observation to the cluster with smallest average distance.
  5. Repeat steps 3 and 4, until no more movements are found.

Hierarchical clustering

- Week 4.R script:
  - Divisive hierarchical clustering: NCI60 data set.

Hierarchical clustering

- Advantages of hierarchical clustering algorithms:
  - There is no need to fix the number of clusters in advance.
  - The dendogram is an useful descriptive tool to define the number of clusters.
- Disadvantages of hierarchical clustering algorithms:
  - Observations that have been incorrectly grouped at an early stage cannot be reallocated.
  - Computationally expensive.