Clustering

Data Mining course

Master in Information Technologies
Enginyeria Informàtica

Tomàs Aluja
Examples of Clustering Applications

- **Marketing**: discover customer groups and use them for targeted marketing and re-organization.
- **Astronomy**: find groups of similar stars and galaxies.
- **Earthquake studies**: Observed earthquake epicenters should be clustered along continent faults.
- **Genomics**: finding groups of gene with similar expressions.
- ...
Clustering

• Purpose:
  – Grouping a set of $n$ objects in $k$ classes homogeneous and distinct among them.

• Proximity measure between objects:
  – Euclidean distance:
    \[
    d^2(i, i') = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2
    \]
  – Logical (for binary data):
    \[
    d(i, i') = f(n_{11}, n_{00}, n_{10}, n_{01})
    \]
    \[
    d(i, i') = \frac{n_{11} + n_{00}}{n}
    \]
The proximity measure need to reflect the actual proximity between objects according the final aim of the clustering. Weighting the attributes might be necessary.

**Conditions**

- We will take as objects to cluster the individuals (rows) of the data matrix.
- We will suppose to have our individuals embedded in an euclidean space defined by the factorial coordinates (output of a PCA, CA, MCA, …)
Clustering in the factorial space

- Why we perform the clustering in the factorial space?
  - To take into account the structural component of data and discard the noise
  - To reduce the curse of dimensionality
  - To work with orthogonal factors (avoiding multicollinearity)
  - To embed the points in an euclidean space

- Which disadvantages do we have?
  - Interpretation is more difficult since we work with factors. Need to come back to the original variables
  - Need to specify the significative directions of the factorial space. Working with all directions is equivalent to perform the clustering with the raw data

Course DM: Clustering. T. Aluja
Coming back to the original variables

- Substitute every factor by its decomposition in the original variables

\[ \Psi_\alpha = u_{1\alpha} x_1 + u_{2\alpha} x_2 + \cdots + u_{p\alpha} x_p \]

\[ \Psi_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} \frac{1}{p} \left( \phi_{1\alpha} z_1 + \cdots + \phi_{J\alpha} z_J \right) \]

Choosing the right dimensionality

Rule of thumb
- Screeplot (in PCA)
- Select the factors with eigenvalue greater than their average (in PCA, MCA)
- Select 80% of the total inertia (in PCA)
Clustering individuals in an euclidean space
Clustering of BCN Quarters

Course DM: Clustering. T. Aluja
Clustering algorithms

• Direct Partitioning
  – K-means
  – K medoids, Balls, ...

• Bottom-up
  – Single linkage, Average linkage, ...
  – Ward, ...

• EM

• ...

• **Which algorithm to choose?**
• **Clustering is a 100% exploratory task**
Data structures

• Several populations

• One population

• Hierarchical structure
Direct Partitioning

- Exponential cost
  \[ N_{n,k} = N_{n-1,k-1} + kN_{n-1,k} \]

- Heuristics:
  - K-means, K-medoids, ...
  - Kohonen maps, Genetic algorithms, ...

- Number of classes \( k \) known a priori

- Criterion:
  \[ \text{Max} \frac{I_{\text{between clusters}}}{I_{\text{total}}} \]

<table>
<thead>
<tr>
<th>Tamaño</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total particiones</td>
<td>52</td>
<td>115975</td>
<td>1382958545</td>
<td>5,17242E+13</td>
<td>4,63859E+18</td>
<td>8,46749E+23</td>
</tr>
<tr>
<td>Particiones en 5 clases</td>
<td>1</td>
<td>42525</td>
<td>210766920</td>
<td>7,49206E+11</td>
<td>2,43668E+15</td>
<td>7,713E+18</td>
</tr>
</tbody>
</table>
K-means

**Initialization**: Selection at random of \( k \) centers \( C_{j,o} \)

**do for** \( l=1:\text{niter} \)

– Assign every individual to the closest center \( C_{j,l-1} \)
– Definition of the new partition \( P_l \)
– Update the centers as centroid of every class \( C_{j,l} \)
– \( \sum_i C_{l-1,i} - C_l < \epsilon \text{ exit} \)

**end do**

• Convergence to a local optima
• Linear cost
K-means
Inicialization
Iteration 1
Iteration 1+1
K-means in R

\texttt{kmeans(x, centers, iter.max = 10)}

- \texttt{x}: A numeric matrix of data or a data frame with all numeric columns.
- \texttt{centers}: Either the number of clusters or a set of initial cluster centers. If the first, a random set of rows in 'x' are chosen as the initial centers.
- \texttt{iter.max}: The maximum number of iterations allowed.

\textit{Results:}

- \texttt{cluster}: A vector of integers indicating the cluster to which each point is allocated.
- \texttt{centers}: A matrix of cluster centres.
- \texttt{withinss}: The within-cluster sum of squares for each cluster.
- \texttt{size}: The number of points in each cluster.
K-means

• Choose the seeds wisely (extreme points in the cloud, by an expert, …).
• Extensions: K-medoids, Balls, Kohonen maps, ...

• Fast k-means (quick & dirty)
  – Selection at random of $C_h$ centers.
  – Assign the first individual of the data set to the closest $C_h$
  – Update the new center of affected class $C_h$
  – Assign the second individual to its closest center …
K-medoids (Nuées dynamiques)

– Choose as centers actual $R$ individuals (in k-medoids $=1$) for each class. Call the $R$ individuals of a class as its core.
– Assign every individual to the closest class core $\min_i D(x_i, C_l)$.
– Define the new classes $P_j$.
– For each class update its core $\min_{C_l} F(C_l, P_j)$.

Distance of an individual to a class core: $D(x_i, C_l) = \sum_{s=1}^{R} d(x_i, c_s^l)$

Distance of a core and a class: $F(C_l, P_j) = \sum_{i \in P_j} \sum_{s=1}^{R} d(x_i, c_s^j)$

Algorithm more stable than k-means
Hierarchical clustering

Algorithm
- \( E = \) Set of objects to cluster
- Calculate the matrix of distances of \( E \) in \( D \)
- While (\( \text{cardinal}(E) > 1 \)) do
  - Find the closest \( (a, b) \) in \( D \)
  - Set \( h = a \cup b \)
  - Update \( E = E - \{a, b\} + \{h\} \)
  - Update the matrix of distances of \( E \) in \( D \)
- End while
Dendrogram (hierarchical tree)

Important leaps in height reveal where to cut the tree to obtain a meaningful partition.
Agregation criteria

- Single linkage
  \[ d(h, h') = \min[d(i, i')] \]

- Average linkage
  \[ d(h, h') = \text{mean}[d(i, i')] \]

- Centroide
  \[ d(h, h') = d(g_1, g_2) \]

- Ward
  \[ d(h, h') = \text{Inercia}(g_1, g_2) \]
Hierarchical clustering in R

d <- dist(x, method = "euclidean")
hc <- hclust(d, method = "complete") # "ward", "single", average", ...
plot(hc)

Results:
• merge: an n-1 by 2 matrix. Row i of 'merge' describes the merging of clusters at step i of the clustering.

• height: value of the criterion associated with the clustering 'method'.

• ...

Course DM: Clustering. T. Aluja
Pros and cons

- Hierarchical clustering
  - Quadratic cost
  - The tree informs about the whole process of aggregation and gives clues about the number of distinct classes in the data.
  - Suboptimal partition (overlapping classes)

- Direct partitioning
  - Linear cost
  - Number of classes must be known a priori
  - Local optimal partition
Sequential clustering

Taking profit of both approaches

Consolidation:

1. Perform a hierarchical clustering
2. Decide the number of classes present in your data and calculate the corresponding centroids
3. Perform a k-means algorithm taking as seeds the centroids previously calculated
Clustering very large data sets

1. Perform \( m (=2 \text{ or } 3) \) times a k-means algorithm (with \( k \approx 10 \))
2. Form the crosstable of \( m \) obtained partitions
3. Calculate the centroids of the (non empty) cells of the crosstable
4. Perform a Hierarchical Clustering of the centroids weighed with the number of individuals per cell
5. Decide the number of classes present in your data
6. Consolidate your clustering
DATOS A CLASIFICAR

**ETAPA 1**
PARTICION CON UN NUMERO MUY GRANDE DE CLASES

**ETAPA 2**
CLASIFICACION JERARQUICA DE LOS CENTROS DE LAS CLASES Y CORTE DEL ARBOL

DATOS CLASIFICADOS
Probabilistic clustering

Soft clustering versus crisp clustering

Assumption of Multivariate Normality, every group \( k \) follows:

\[ N_d(\mu_h, \Sigma_h) \]

where \( \mu_h \) is the centroid of the distribution and \( \Sigma_h \) is the matrix of covariances. \( \mu_h \) is the point of maximum density of the distribution and \( \Sigma_h \) defines the shape of the distribution.

Different possibilities for the groups:
1. Spherical with equal volume
2. Spherical with different volume
3. Diagonal with equal volume and shape
4. Diagonal, with different volume and shape
5. Ellipsoidal with equal volume, shape and orientation
6. Ellipsoidal with any volume, shape and orientation

We can test which assumption is more realistic given the data cloud, but the number of classes \( k \) still unknown.
EM algorithm

Choose $K$ and perform a first clustering (i.e. hierarchical)

Initialize $\bar{x}_h^0, V_h^0, \pi_h^0$ (estimates of the parameters and prob. a priori of each group, i.e. using a hierarchical clustering)

1. **Compute the posteriori probabilities per group**

$$ \pi_{ih}^l = \frac{\pi_h^l f_h(x_i)}{\sum_{h}^{K} \pi_h^l f_h(x_i)} $$

$$ \sum_{h=1}^{K} \pi_{ih}^l = 1 $$

2. **Update the parameters**

$$ \bar{x}_h^{l+1} = \frac{1}{\sum_{i=1}^{n} \pi_{ih}} \sum_{i=1}^{n} \pi_{ih} x_i $$

$$ V_h^{l+1} = \frac{1}{\sum_{i=1}^{n} \pi_{ih}} \sum_{i=1}^{n} \pi_{ih} (x_i - \bar{x}_h^{l+1})(x_i - \bar{x}_h^{l+1})' $$

$$ \pi_h^{l+1} = \frac{1}{n} \sum_{i=1}^{n} \pi_{ih} $$

Course DM: Clustering. T. Aluja
Función R: Mclust

Clustering via EM initialized by hierarchical clustering for parameterized Gaussian mixture models. The number of clusters and the clustering model is chosen to maximize the BIC.

Mclust(data, minG, maxG)

Arguments
- **data**: A numeric vector, matrix, or data frame of observations. Categorical variables are not allowed.
- **minG**: Minimum number of clusters to be considered. The default is 1 component.
- **maxG**: Maximum number of clusters to be considered. The default is 9 components.

Results
- **BIC**: A matrix giving the BIC value for each model (rows) and number of clusters (columns).
- **bic**: A scalar giving the optimal BIC value.
- **modelName**: The MCLUST name for the best model according to BIC.
- **classification**: The classification corresponding to the optimal BIC value.
- **mu**: A matrix whose columns are the means of each group in the best model.
- **sigma**: Three dimensional array in which sigma[,] gives the covariance for the kth group in the best model.
- **pro**: The mixing probabilities for each component in the best model.

The following models are compared in Mclust:
- "EI" : spherical, equal volume
- "VII" : spherical, unequal volume
- "EEI" : diagonal, equal volume, equal shape
- "VVI" : diagonal, varying volume, varying shape
- "EEE" : ellipsoidal, equal volume, shape, and orientation
- "VVV" : ellipsoidal, varying volume, shape, and orientation
Assignment of new individuals

- Define rules to assign new individuals
- Compute the distance of the new individual to each class centroid

\[
\text{Input: } x_{i1}, x_{i2}, \ldots, x_{ip}
\]

\[
\text{Centers} \rightarrow \text{Output: Assigned class}
\]

Crisp

\[
x_i \rightarrow C_h \ \text{argmin} \ d(x_i, C_h)
\]

Soft

\[
x_i \rightarrow C_h \ \text{random draw} \ f(d(x_i, C_h), h=1 \ldots K)
\]

Course DM: Clustering. T. Aluja
Interpreting the classes

• Differential characterisation among classes

Original Variables
(continuous & categorical)

Partition

Class 1
Class 2
Class 3

• Statistical characterization
Typology of employees according psychological factors
Interpreting the classes

<table>
<thead>
<tr>
<th>V.TEST</th>
<th>PROBA</th>
<th>MOYENNES</th>
<th>ECARTS TYPES</th>
<th>VARIABLES CARACTERISTIQUES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>CLASSE GENERALE</td>
<td>CLASSE GENERAL</td>
<td>NUM.LIBELLE</td>
</tr>
<tr>
<td>--------</td>
<td>-------</td>
<td>-------------------</td>
<td>------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>7.94</td>
<td>0.000</td>
<td>13.90 10.73</td>
<td>1.83 5.24</td>
<td>7'autoridad natural</td>
</tr>
<tr>
<td>6.09</td>
<td>0.000</td>
<td>13.46 10.58</td>
<td>1.95 6.21</td>
<td>27'capacidad hablar en público</td>
</tr>
<tr>
<td>5.91</td>
<td>0.000</td>
<td>13.01 10.72</td>
<td>1.50 5.09</td>
<td>35'confianza en si mismo</td>
</tr>
<tr>
<td>5.32</td>
<td>0.000</td>
<td>13.22 11.35</td>
<td>1.53 4.60</td>
<td>5'sentido de la competitividad</td>
</tr>
<tr>
<td>5.24</td>
<td>0.000</td>
<td>13.14 10.88</td>
<td>1.74 5.65</td>
<td>13'capacidad para innovar</td>
</tr>
<tr>
<td>-7.07</td>
<td>0.000</td>
<td>10.10 11.50</td>
<td>1.71 3.33</td>
<td>17'presentación</td>
</tr>
<tr>
<td>-7.24</td>
<td>0.000</td>
<td>10.10 12.31</td>
<td>1.79 5.12</td>
<td>20'facilidad de contacto</td>
</tr>
<tr>
<td>-7.33</td>
<td>0.000</td>
<td>10.20 12.49</td>
<td>2.26 5.22</td>
<td>2'capacidad por el trabajo</td>
</tr>
<tr>
<td>-7.42</td>
<td>0.000</td>
<td>10.07 12.34</td>
<td>2.15 5.10</td>
<td>1'punto por el trabajo</td>
</tr>
<tr>
<td>-7.56</td>
<td>0.000</td>
<td>10.17 11.92</td>
<td>1.81 3.87</td>
<td>14'sentido de la organización y del método</td>
</tr>
</tbody>
</table>
## Exploring a Data Matrix

| 1. Multivariate Description | Visualisation
|                            | Outlier detection
|                            | Reduction of dimensionality
|                            | Extraction of latent factors
|                            | Interpreting the information of the data
|                            | Generating new hypothesis
| 2. Clustering               | Synthesis to simplify the complex reality
|                            | Classes must be operative regard to the goal of the clustering
|                            | Define the rules to assign new individuals
| 3. Interpret the classes    | Differential characteristics per each class. Give them a name. |