Unsupervised Learning: Introduction

Universitat de les Illes Balears
Departament de Ciències Matemàtiques i Informàtica

11752 Aprendizaje Automático
11752 Machine Learning
Máster Universitario
en Sistemas Inteligentes

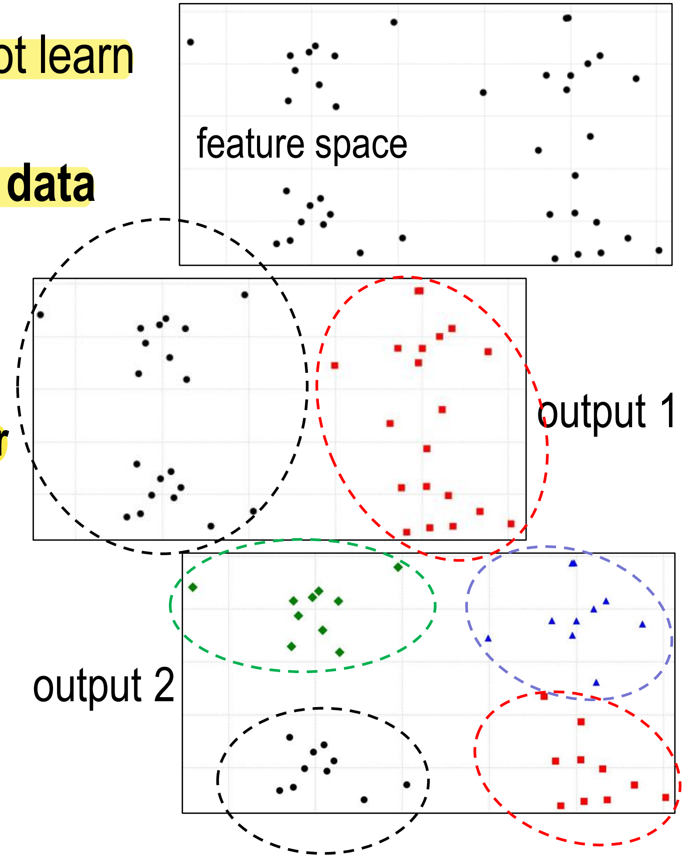
Alberto ORTIZ RODRÍGUEZ

Contents

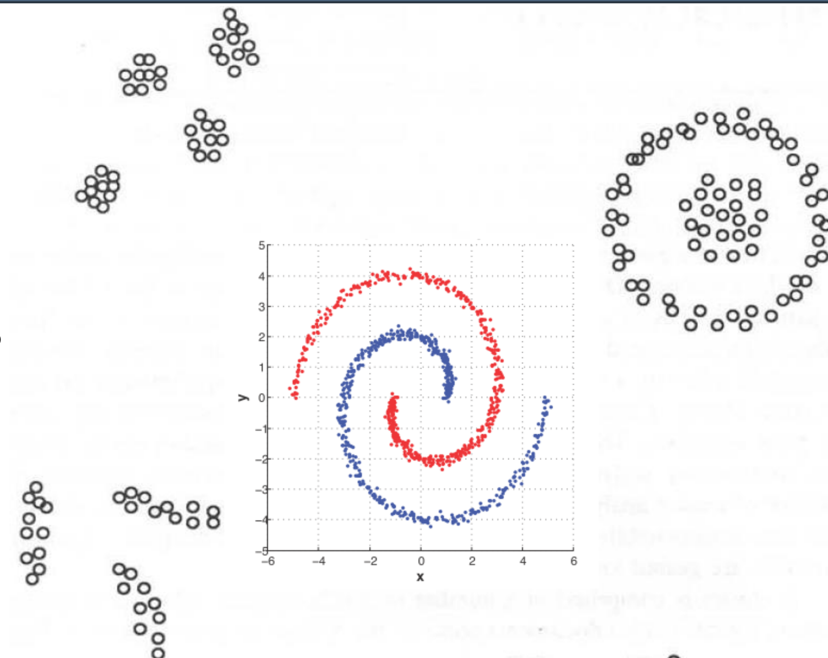
Problem description

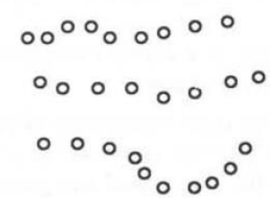
- In **unsupervised learning** (UL), a program does not learn from labelled data. Instead, it attempts to discover patterns in the data, i.e. **learn the structure of the data**
  - bring out patterns and structure within the data, maybe **informative** by itself

    or serve as a guide to **further analysis**, e.g. **learn features to build a supervised classifier**
  - known as **exploratory analysis** or **knowledge discovery**
- Using UL terminology: discover groups of related observations within the data called **clusters**
  - **clustering** or cluster analysis
    - assigns observations into groups such that samples in the same group are **most similar** to one another
  - e.g. discover segments of customers (marketing)
- **More widely applicable** than supervised learning: no need for labelled data
- Other names: Numerical taxonomy (biology), Typology (social sci.), Partitioning (graphs th.)



- Meaningful clusters can be of ***diverse shapes***:
  - This is problem dependent, i.e. problem semantics
  - Humans are very good at detecting clusters in two and three dimensions



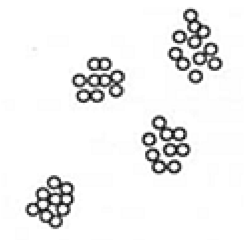


- SL and UL can be thought as occupying opposite ends of a spectrum
- A different approach called **semi-supervised learning** (SSL) makes use of both supervised and unsupervised learning
  - The dataset is **partially labelled**
  - Located somewhere in-between SL and UL

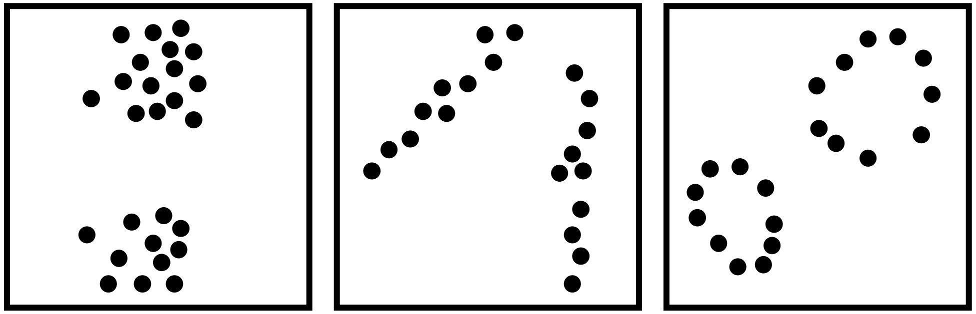
Contents

Definition of clustering

- A cluster consists of a number of ***similar*** objects, given a certain **similarity criterion**.
- Other definitions (collected from here and there):
  - A cluster is a set of entities which are **alike**, while entities from different clusters are **not alike**
  - A cluster is an aggregation of **points** in the test space such that the *distance between points in the same cluster* is less than the *distance between any point in the cluster and any point not in it*
  - Clusters may be described as connected regions of a **multi-dimensional space** containing a relatively **high density of points**, separated from other such regions by a region containing a relatively **low density of points**



- A number of definitions based on rather vaguely defined terms which lead to a **bunch of algorithms**.
- In any case, the goal is to find the **natural structure** of the data, similarly to the human skill of grouping points in 2D/3D space, but for any amount of dimensions.



Definition of clustering

- Given a set of samples $X = \{x_1, x_2, \dots, x_N\}$ defined in an L-dimensional space, a first **non-universal**, **but formal definition** of clustering would be as follows: *an M-grouping of X is a partition of X into M sets $C_1, C_2, \dots, C_M$, so that*

$$(1) \quad C_i \neq \emptyset, \quad i = 1, \dots, M$$

$$(2) \quad \bigcup_{i=1}^{M} C_i = X$$

$$(3) \quad C_i \bigcap C_j = \emptyset, \quad i \neq j, \; j = 1, \dots, M$$

$$(4a) \quad d(a,b) < d(a,c), \quad a, b \in C_i, \; c \in C_j \quad [\text{where } d(\cdot,\cdot) \text{ is a dissimilarity measure}]$$
$$(4b) \quad s(a,b) > s(a,c), \quad a, b \in C_i, \; c \in C_j \quad [\text{where } s(\cdot,\cdot) \text{ is a similarity measure}]$$

- Summing up, a clustering problem involves:
  - A set of **unlabelled samples**
  - A **proximity measure** $\wp$ (either a similarity *s* or a dissimilarity *d*)
  - A **clustering algorithm**, from the many alternatives readily available:

1. hierarchical	4. density-based	7. valley-seeking
2. optimization-based	5. graph-based	8. sequential
3. model-based	6. competitive	9. others …

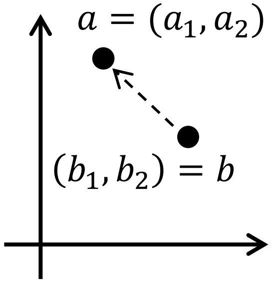
Contents

- Problem description
- Definition of clustering
- Proximity measures

- The **proximity measure** chosen plays a central role in cluster analysis
- Formal definition of **similarity measure** (SM) **/ dissimilarity measure** (DM)

$d(a,b) = \|a-b\|$

$= \sqrt{(a_1-b_1)^2 + (a_2-b_2)^2}$



$s(a,b) = e^{-\|a-b\|}$

proximity measure over X

$\wp : X \times X \to \mathbb{R}$

- A **dissimilarity** measure **d** among elements of a set X is a function such that:

$$d : X \times X \to \mathbb{R}$$

and

(1) $\exists d_0 \in \mathbb{R} : -\infty < d_0 \le d(a,b) < +\infty, \forall a,b \in X$ e.g. $d_0 = 0$

(2) $d(a,a) = d_0, \forall a \in X$

(3) $d(a,b) = d(b,a), \forall a,b \in X$

(the more dissimilar, the greater)

- A **similarity** measure **s** among elements of a set X is a function such that:

$$s : X \times X \to \mathbb{R}$$

and

(1) $\exists s_0 \in \mathbb{R} : -\infty < s(a,b) \le s_0 < +\infty, \forall a,b \in X$ e.g. $s_0 = 1$

(2) $s(a,a) = s_0, \forall a \in X$

(3) $s(a,b) = s(b,a), \forall a,b \in X$

(the more similar, the greater)

- The concept can be extended to measure **proximity between sets** (clusters)
- Formal definition of **similarity measure** (SM) **/ dissimilarity measure** (DM)

$\mathcal{P}(\{0,1,2\}) = \{\emptyset, \{0\}, \{1\}, \{2\}, \{0,1\}, \{0,2\}, \{1,2\}, \{0,1,2\}\}$

proximity measure over *P(X)*

- A **dissimilarity** measure **d** among <u>subsets</u> U ⊆ X (U ∈ *P*(X)) is a function such that:

$$d : \mathcal{P}(X) \times \mathcal{P}(X) \to \mathbb{R}$$

and
(1) $\exists d_0 \in \mathbb{R} : -\infty < d_0 \leq d(U,V) < +\infty, \forall U,V \in \mathcal{P}(X)$
(2) $d(U,U) = d_0, \forall U \in \mathcal{P}(X)$
(3) $d(U,V) = d(V,U), \forall U,V \in \mathcal{P}(X)$

(the more disimilar, the greater)

- A **similarity** measure **s** among <u>subsets</u> U ⊆ X (U ∈ *P*(X)) is a function such that:

$$s : \mathcal{P}(X) \times \mathcal{P}(X) \to \mathbb{R}$$

and
(1) $\exists s_0 \in \mathbb{R} : -\infty < s(U,V) \leq s_0 < +\infty, \forall U,V \in \mathcal{P}(X)$
(2) $s(U,U) = s_0, \forall U \in \mathcal{P}(X)$
(3) $s(U,V) = s(V,U), \forall U,V \in \mathcal{P}(X)$

(the more similar, the greater)

$$\wp : \mathcal{P}(X) \times \mathcal{P}(X) \to \mathbb{R}$$

- A more restrictive concept is that of **metric**
  - A **metric** m among elements of a set *X* is a function such that

$$m : X \times X \to \mathbb{R}$$

and

$$\begin{aligned}
&(1) \quad \exists m_0 \in \mathbb{R} : -\infty < m_0 \leq m(a,b) < +\infty, \forall a,b \in X \quad (\text{e.g. } m_0 = 0) \\
&(2) \quad m(a,a) = m_0, \forall a \in X \\
&(3) \quad m(a,b) = m(b,a), \forall a,b \in X \\
&(4) \quad m(a,b) \leq m(a,c) + m(c,b), \forall a,b,c \in X \quad (\text{triangle inequality})
\end{aligned}$$

  - Captures the notion of dissimilarity
  - Not all dissimilarity measures are metrics

- An DM can be obtained from a SM using any monotonically decreasing function, e.g.

$$d(a,b) = \max\{s(a,b)\} - s(a,b)$$

- **Examples of proximity measures**:
  - weighted **L_p metric** (or **Minkowski** measure) **-** DM

$$d_p(a,b) = \left(\sum_{i=1}^{L} w_i |a_i - b_i|^p \right)^{\frac{1}{p}}, w_i \geq 0$$

$$d_2(a,b) = \sqrt{\sum_{i=1}^{L} (a_i - b_i)^2} \quad \text{(metric } L_2 \text{ or Euclidean distance)}$$

$$d_1(a,b) = \sum_{i=1}^{L} |a_i - b_i| \quad \text{(metric } L_1 \text{ or Manhattan distance or City Block distance)}$$

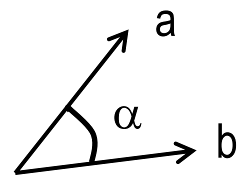
$$d_\infty(a,b) = \max_{1 \leq i \leq L} |a_i - b_i| \quad \text{(metric } L_\infty \text{ or Chebyshev distance)}$$

$$\left(= \lim_{p \to +\infty} d_p(a,b) \quad \text{if } w_i = 1\right)$$

- **Examples of proximity measures**:

Change of date of the exam

consider descriptors as nD-vectors instead of nD-points



- **dot product** – SM

$$s_\bullet(a,b) = a^T b = \sum_{i=1}^{L} a_i b_i = \|a\|\,\|b\| \cos\alpha, \;\; \|a\|, \|b\| \leq m, \forall a,b \Rightarrow s_\bullet \in [-m^2, +m^2]$$

- **cosinus** measure – SM

$$s_{\cos}(a,b) = \frac{a^T b}{\|a\|\,\|b\|}, \quad s_{\cos} \in [-1,+1], \quad s_{\cos}(a=\vec{0},b) = \frac{0}{0} \rightarrow 1$$

- **Tanimoto** measure – SM

$$s_T(a,b) = \frac{a^T b}{\|a\|^2 + \|b\|^2 - a^T b} = \frac{a^T b}{a^T a + b^T b - 2a^T b + a^T b}$$

$$= \frac{1}{1 + \frac{(a-b)^T(a-b)}{a^T b}} = \frac{1}{1 + \frac{(d_2(a,b))^2}{s_\bullet(a,b)}}$$

$$s_T(a,a) = \frac{1}{1 + \frac{0}{m^2}} = 1, \quad s_T(a,-a) = \frac{1}{1 + \frac{(2m)^2}{-m^2}} = -\frac{1}{3}$$

- The previous proximity measures are intended for **quantitative, real-valued descriptors**
- Other proximity measures for these kind of data are the following:
  - **Canberra** distance – DM

$$d_{\text{can}}(a,b) = \sum_{i=1}^{L} \frac{|a_i - b_i|}{|a_i| + |b_i|}, \quad d_{\text{can}}(\vec{0},\vec{0}) = \frac{0}{0} \to 0$$

  - **Bray-Curtis** distance (Sorensen distance) – DM

$$d_{\text{bc}}(a,b) = \frac{\sum_{i=1}^{L} |a_i - b_i|}{\sum_{i=1}^{L} (a_i + b_i)} \quad [\text{intended for } a_i, b_i \geq 0, \forall i]$$

  - **correlation coefficient** – SM

$$s_{\text{corr}}(a,b) = \frac{\sum_{i=1}^{L} (a_i - \overline{a})(b_i - \overline{b})}{\sqrt{\sum_{i=1}^{L} (a_i - \overline{a})^2 \sum_{i=1}^{L} (b_i - \overline{b})^2}}, \quad \overline{a} = \frac{1}{L} \sum_{i=1}^{L} a_i \,, \; \overline{b} = \frac{1}{L} \sum_{i=1}^{L} b_i$$

$$a = (p,q), b = (-p,-q), s_{\text{corr}}(a,a) = 1, s_{\text{corr}}(a,b) = -1$$

- For other kinds of data we need other proximity functions
- Proximity functions for **binary-valued descriptors**, i.e. a_i = yes/no, e.g. $a = (0,0,1,1,0)$
  $b = (0,1,1,0,1)$

$$m_{00}(a,b) = \text{number of features at value 0 both in } a \text{ and } b$$
$$m_{01}(a,b) = \text{number of features at value 0 in } a \text{ and 1 in } b$$
$$m_{10}(a,b) = \text{number of features at value 1 in } a \text{ and 0 in } b$$
$$m_{11}(a,b) = \text{number of features at value 1 both in } a \text{ and } b$$

  - **Hamming/matching distances** – DM

$$d_{\text{ham}} = 3 \qquad d_{\text{mat}} = \frac{3}{5} = 0.6$$

$$d_{\text{ham}}(a,b) = m_{01}(a,b) + m_{10}(a,b) \qquad d_{\text{mat}}(a,b) = \left.\frac{m_{01}+m_{10}}{m_{00}+m_{01}+m_{10}+m_{11}}\right|_{(a,b)}$$

  - **Jaccard's coefficient** – SM (frequency of occurrence of 0s and 1s is not the same)

$$s_{\text{jac}} = \frac{1}{4} = 0.25 \qquad s_{\text{jac}}(a,b) = \left.\frac{m_{11}}{m_{11}+m_{01}+m_{10}}\right|_{(a,b)}$$

  - **Jaccard's distance** – DM (frequency of occurrence of 0s and 1s is not the same)

$$d_{\text{jac}} = \frac{3}{4} = 0.75 \qquad d_{\text{jac}}(a,b) = 1 - s_{\text{jac}}(a,b) = \left.\frac{m_{01}+m_{10}}{m_{11}+m_{01}+m_{10}}\right|_{(a,b)}$$

- Proximity functions for **nominal/categorical variables**, e.g. a_i = red | green | blue
  - **Encode** categorical values as **binary values** and make use of corresp. proximity functions
  - Example: a = (gender, colour) = (male | female, red | white | blue)

1. $\text{gender} \to \{0,1\}, \text{colour} \to \{(1,0,0),(0,1,0),(0,0,1)\}$

$$a = [\text{male}, \text{red}], b = [\text{female}, \text{white}], c = [\text{female}, \text{blue}]$$

$$a = [0,(1,0,0)], b = [1,(0,1,0)], c = [1,(0,0,1)]$$

$$d_{\text{ham}}(a,b) = (1,2) \to 1+2 = 3 \text{ or } d'_{\text{mat}}(a,b) = (1/1 + 2/3)/2 = 5/6$$

$$d_{\text{ham}}(a,c) = (1,2) \to 1+2 = 3 \text{ or } d'_{\text{mat}}(a,c) = (1/1 + 2/3)/2 = 5/6$$

$$d_{\text{ham}}(b,c) = (0,2) \to 0+2 = 2 \text{ or } d'_{\text{mat}}(b,c) = (0/1 + 2/3)/2 = 2/6$$

2. $\text{gender} \to \{0,1\}, \text{colour} \to \{(0,0),(0,1),(1,0)\}$

$$a = [\text{male}, \text{red}], b = [\text{female}, \text{white}], c = [\text{female}, \text{blue}]$$

$$a = [0,(0,0)], b = [1,(0,1)], c = [1,(1,0)]$$

$$d_{\text{ham}}(a,b) = (1,1) \to 1+1 = 2 \text{ or } d'_{\text{mat}}(a,b) = (1/1 + 1/2)/2 = 3/4$$

$$d_{\text{ham}}(a,c) = (1,1) \to 1+1 = 2 \text{ or } d'_{\text{mat}}(a,c) = (1/1 + 1/2)/2 = 3/4$$

$$d_{\text{ham}}(b,c) = (0,2) \to 0+2 = 2 \text{ or } d'_{\text{mat}}(b,c) = (0/1 + 2/2)/2 = 1/2$$

be careful with coding !! d(red,white) = d(red,blue) < d(white,blue)

- Proximity functions for **ordinal variables** (i.e. there exists an order), e.g. a_i = excellent | good | average | poor
  - **Encode** ordinal values as **real values** and make use of corresp. proximity functions
  - Example:

$$a_i = \text{excellent}|\text{good}|\text{average}|\text{poor} \rightarrow 3 \mid 2 \mid 1 \mid 0 \rightarrow 1 \mid 2/3 \mid 1/3 \mid 0$$

$$a = [\text{excellent}, \text{average}, \text{average}], b = [\text{poor}, \text{good}, \text{average}]$$

$$a = [1, 1/3, 1/3], b = [0, 2/3, 1/3]$$

$$d_2(a,b)^2 = 1^2 + (1/3)^2 + 0^2 = 10/9 \quad [\text{Spearman distance}]$$

$$d_1(a,b) = 1 + 1/3 + 0 = 4/3 \quad [(\text{Spearman) footrule distance}]$$

Proximity measures

- **Examples of proximity measures between points and clusters**

$\wp$ = DM or SM

cluster has no representive

$$\wp_{\max}(a, C) = \max_{b \in C} \; \wp(a, b)$$

$$\wp_{\min}(a, C) = \min_{b \in C} \; \wp(a, b)$$

$$\wp_{\text{avg}}(a, C) = \frac{1}{n_C} \sum_{b \in C} \wp(a, b)$$

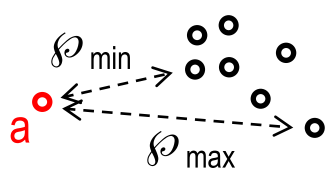
cluster has a representative μ

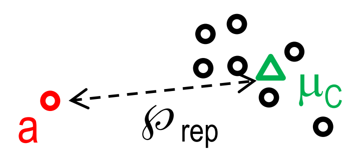
$$\wp_{\text{rep}}(a, C) = \wp(a, \mu_C)$$

where typically

$$\mu_C = \frac{1}{n_C} \sum_{b \in C} b$$

e.g. $\wp$ is a DM

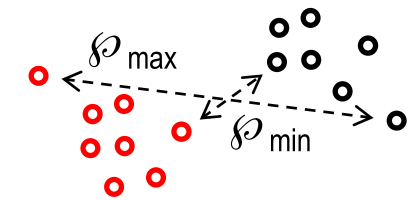




... at a formal level, they are relevant,
but, at a practical level, proximity measures
are more useful when defined between clusters

- **Examples of proximity measures between clusters**

e.g. $\wp$ is a DM

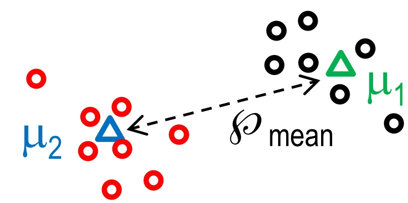


clusters have no representatives

$$\begin{cases} \wp_{\max}(C_1, C_2) = \max\limits_{a \in C_1, b \in C_2} \wp(a, b) \\ \wp_{\min}(C_1, C_2) = \min\limits_{a \in C_1, b \in C_2} \wp(a, b) \\ \wp_{\text{avg}}(C_1, C_2) = \dfrac{1}{n_1 n_2} \sum\limits_{a \in C_1} \sum\limits_{b \in C_2} \wp(a, b) \end{cases}$$

each cluster has a representative μ

$$\begin{cases} \wp_{\text{mean}}(C_1, C_2) = \wp(\mu_1, \mu_2) \\ \wp_{\text{ward}}(C_1, C_2) = \sqrt{\dfrac{n_1 n_2}{n_1 + n_2}} \wp(\mu_1, \mu_2) \end{cases}$$



  - Some of these functions may not be metric functions: e.g. $\wp_{max}$ when $\wp$ is a DM, $\wp_{min}$ when $\wp$ is an SM, or $\wp_{avg}$ either if $\wp$ is a DM or a SM

Unsupervised Learning: Introduction

Universitat de les Illes Balears
Departament de Ciències Matemàtiques i Informàtica

11752 Aprendizaje Automático
11752 Machine Learning
Máster Universitario
en Sistemas Inteligentes

Alberto ORTIZ RODRÍGUEZ