CLUSTERING

- **Basic Concepts**

  In clustering or unsupervised learning no training data, with class labeling, are available. The goal becomes: Group the data into a number of sensible clusters (groups). This unravels similarities and differences among the available data.

  - **Applications:**
    - Engineering
    - Bioinformatics
    - Social Sciences
    - Medicine
    - Data and Web Mining

  - To perform clustering of a data set, a clustering criterion must first be adopted. Different clustering criteria lead, in general, to different clusters.

➢ A simple example

Blue shark, sheep, cat, dog

Lizard, sparrow, viper, seagull, gold fish, frog, red mullet

1. Two clusters
2. Clustering criterion: How mammals bear their progeny

Gold fish, red mullet, blue shark

Sheep, sparrow, dog, cat, seagull, lizard, frog, viper

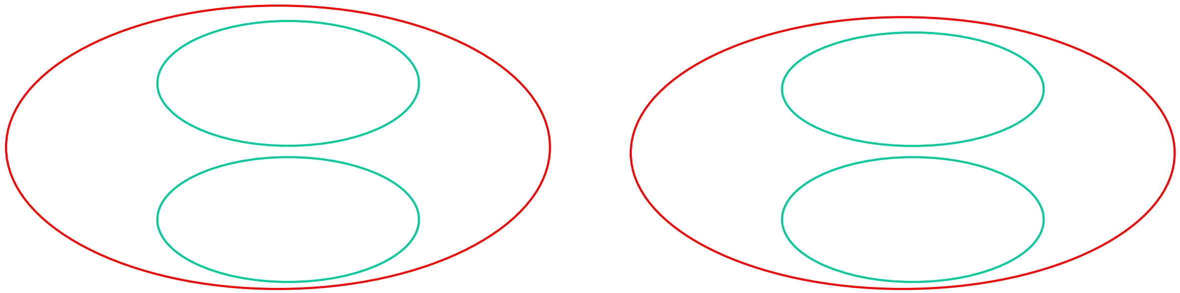
1. Two clusters
2. Clustering criterion: Existence of lungs

❖ Clustering task stages

- **Feature Selection:** Information rich features-Parsimony
- **Proximity Measure:** This quantifies the term similar or dissimilar.
- **Clustering Criterion:** This consists of a cost function or some type of rules.
- **Clustering Algorithm:** This consists of the set of **steps** followed to reveal the structure, based on the similarity measure and the adopted criterion.
- **Validation of** the results.
- **Interpretation of** the results.

- Depending on the similarity measure, the clustering criterion and the clustering algorithm different clusters may result. **Subjectivity** is a reality to live with from now on.

- A simple example: How many clusters??



2 or 4 ??

- Basic application areas for clustering
  - Data reduction. All data vectors within a cluster are substituted (represented) by the corresponding cluster representative.
  - Hypothesis generation.
  - Hypothesis testing.
  - Prediction based on groups.

❖ Clustering Definitions

- **Hard Clustering:** Each point belongs to a single cluster
  - Let $X = \{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_N\}$
  - An m-clustering R of X, is defined as the partition of X into m sets (clusters), $C_1, C_2, \ldots, C_m$, so that

$$C_i \neq \emptyset, i = 1,2,\ldots,m$$

$$\bigcup_{i=1}^{m} C_i = X$$

$$C_i \cap C_j = \emptyset, \; i \neq j, \; i,j = 1,2,\ldots,m$$

In addition, data in C_i are more similar to each other and less similar to the data in the rest of the clusters. Quantifying the terms similar-dissimilar depends on the types of clusters that are expected to underlie the structure of X.

- Fuzzy clustering: Each point belongs to all clusters up to some degree.

A fuzzy clustering of X into m clusters is characterized by m functions

$$u_j : \underline{x} \rightarrow [0,1], \quad j = 1,2,\ldots,m$$

$$\sum_{j=1}^{m} u_j(\underline{x}_i) = 1, \quad i = 1,2,\ldots,N$$

$$0 < \sum_{i=1}^{N} u_j(\underline{x}_i) < N, \quad j = 1,2,\ldots,m$$

These are known as membership functions.
Thus, each $\underline{x}_i$ belongs to any cluster "up to some degree", depending on the value of

$$u_j(\underline{x}_i), \;\; j = 1,2,\ldots,m$$

$u_j(\underline{x}_i)$ close to $1 \Rightarrow$ high grade of membership of $\underline{x}_i$ to cluster j.

$u_j(\underline{x}_i)$ close to $0 \Rightarrow$ low grade of membership.

TYPES OF FEATURES

- With respect to their <u>domain</u>
  - Continuous (the domain is a continuous subset of $\Re$).
  - Discrete (the domain is a finite discrete set).
    - *Binary* or *dichotomous* (the domain consists of two possible values).

- With respect to the <u>relative significance of the values they take</u>
  - Nominal (the values code states, e.g., the sex of an individual).
  - Ordinal (the values are meaningfully ordered, e.g., the rating of the services of a hotel (poor, good, very good, excellent)).
  - Interval-scaled (the difference of two values is meaningful but their ratio is meaningless, e.g., temperature).
  - Ratio-scaled (the ratio of two values is meaningful, e.g., weight).

PROXIMITY MEASURES

❖ *Between vectors*

➢ Dissimilarity measure (between vectors of X) is a function

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

with the following properties

- $\exists d_0 \in \Re: \ -\infty < d_0 \leq d(\underline{x}, \underline{y}) < +\infty, \ \ \forall \underline{x}, \underline{y} \in X$
- $d(\underline{x}, \underline{x}) = d_0, \ \ \forall \underline{x} \in X$
- $d(\underline{x}, \underline{y}) = d(\underline{y}, \underline{x}), \ \ \forall \underline{x}, \underline{y} \in X$

If in addition

- $d(\underline{x},\underline{y}) = d_0 \text{ if and only if } \underline{x} = \underline{y}$
- $d(\underline{x},\underline{z}) \le d(\underline{x},\underline{y}) + d(\underline{y},\underline{z}), \ \forall \underline{x},\underline{y},\underline{z} \in X$

  (triangular inequality)

d is called a metric dissimilarity measure.

- **Similarity measure** (between vectors of X) is a function

$$s: \; X \times X \longrightarrow \mathbb{R}$$

with the following properties

- $\exists s_0 \in \mathbb{R}: \; -\infty < s(\underline{x}, \underline{y}) \leq s_0 < +\infty, \;\; \forall \underline{x}, \underline{y} \in X$
- $s(\underline{x}, \underline{x}) = s_0, \;\; \forall \underline{x} \in X$
- $s(\underline{x}, \underline{y}) = s(\underline{y}, \underline{x}), \;\; \forall \underline{x}, \underline{y} \in X$

If in addition

- $s(\underline{x},\underline{y}) = s_0 \text{ if and only if } \underline{x}=\underline{y}$
- $s(\underline{x},\underline{y})s(\underline{y},\underline{z}) \leq [s(\underline{x},\underline{y})+s(\underline{y},\underline{z})]s(\underline{x},\underline{z}), \ \forall \underline{x},\underline{y},\underline{z} \in X$

s is called a metric similarity measure.

❖ *Between sets*

Let $D_i \subset X$, $i = 1, \ldots, k$ and $U=\{D_1,\ldots,D_k\}$

A proximity measure $\wp$ on U is a function

$$\wp : U \times U \longrightarrow \mathbb{R}$$

A dissimilarity measure has to satisfy the relations of dissimilarity measure between vectors, where D_i's are used in place of $\underline{x}$, $\underline{y}$ (similarly for similarity measures).

PROXIMITY MEASURES BETWEEN VECTORS

❖ Real-valued vectors

➢ **Dissimilarity measures** (DMs)

- *Weighted l_p metric DMs*

$$d_p(\underline{x}, \underline{y}) = \left(\sum_{i=1}^{l} w_i \, |x_i - y_i|^p\right)^{1/p}$$

Interesting instances are obtained for

– $p = 1$ (*weighted Manhattan* norm)

– $p = 2$ (*weighted Euclidean* norm)

– $p = \infty$ ($d_\infty(\underline{x}, \underline{y}) = \max_{1 \le i \le l} w_i |x_i - y_i|$)

- *Other measures*
  - $$d_G(\underline{x}, \underline{y}) = -\log_{10}\left(1 - \frac{1}{l}\sum_{j=1}^{l}\frac{|x_j - y_j|}{b_j - a_j}\right)$$

    where b_j and a_j are the maximum and the minimum values of the j-th feature, among the vectors of X (dependence on the current data set)

  - $$d_Q(\underline{x}, \underline{y}) = \sqrt{\frac{1}{l}\sum_{j=1}^{l}\left(\frac{x_j - y_j}{x_j + y_j}\right)^2}$$

➢ Similarity measures

- *Inner product*

$$s_{inner}(\underline{x}, \underline{y}) = \underline{x}^T \underline{y} = \sum_{i=1}^{l} x_i y_i$$

- *Tanimoto measure*

$$s_T(\underline{x}, \underline{y}) = \frac{\underline{x}^T \underline{y}}{\| \underline{x} \|^2 + \| \underline{y} \|^2 - \underline{x}^T \underline{y}}$$

- $$s_T(\underline{x}, \underline{y}) = 1 - \frac{d_2(\underline{x}, \underline{y})}{\| \underline{x} \| + \| \underline{y} \|}$$

❖ Discrete-valued vectors

- Let $F=\{0,1,\ldots,k-1\}$ be a set of symbols and $X=\{\underline{x}_1,\ldots,\underline{x}_N\}\subset F^l$
- Let $A(\underline{x},\underline{y})=[a_{ij}]$, $i, j=0,1,\ldots,k-1$, where a_{ij} is the number of places where $\underline{x}$ has the i-th symbol and $\underline{y}$ has the j-th symbol.

NOTE: $$\sum_{i=0}^{k-1}\sum_{j=0}^{k-1} a_{ij}=l$$

Several proximity measures can be expressed as combinations of the elements of $A(\underline{x},\underline{y})$.

- Dissimilarity measures:
  - The Hamming distance (number of places where $\underline{x}$ and $\underline{y}$ differ)

$$d_H(\underline{x},\underline{y})=\sum_{i=0}^{k-1}\sum_{\substack{j=0\\ j\neq i}}^{k-1} a_{ij}$$

  - The l_1 distance

$$d_1(\underline{x},\underline{y})=\sum_{i=1}^{l}|x_i-y_i|$$

- Similarity measures:

  - Tanimoto measure : $$s_T(\underline{x}, \underline{y}) = \frac{\sum_{i=1}^{k-1} a_{ii}}{n_x + n_y - \sum_{i=1}^{k-1}\sum_{j=1}^{k-1} a_{ij}}$$

    where $n_x = \sum_{i=1}^{k-1}\sum_{j=0}^{k-1} a_{ij}$, $n_y = \sum_{i=0}^{k-1}\sum_{j=1}^{k-1} a_{ij}$,

  - Measures that exclude a_{00}: $\sum_{i=1}^{k-1} a_{ii} / l$ $\quad \sum_{i=1}^{k-1} a_{ii} /(l - a_{00})$

  - Measures that include a_{00}: $\sum_{i=0}^{k-1} a_{ii} / l$

❖ Mixed-valued vectors

Some of the coordinates of the vectors $\underline{x}$ are real and the rest are discrete.

Methods for measuring the proximity between two such $\underline{x}_i$ *and* $\underline{x}_j$:

- Adopt a proximity measure (PM) suitable for real-valued vectors.
- Convert the real-valued features to discrete ones and employ a discrete PM.

The more general case of mixed-valued vectors:

- Here nominal, ordinal, interval-scaled, ratio-scaled features are treated separately.

The similarity function between $\underline{x}_i$ and $\underline{x}_j$ is:

$$s(\underline{x}_i, \underline{x}_j) = \sum_{q=1}^{l} s_q(\underline{x}_i, \underline{x}_j) / \sum_{q=1}^{l} w_q$$

In the above definition:

- $w_q = 0$, if at least one of the q-th coordinates of $\underline{x}_i$ and $\underline{x}_j$ are undefined or both the q-th coordinates are equal to 0. Otherwise $w_q=1$.
- If the q-th coordinates are binary, $s_q(\underline{x}_i, \underline{x}_j) = 1$ if $x_{iq}=x_{jq}=1$ and 0 otherwise.
- If the q-th coordinates are nominal or ordinal, $s_q(\underline{x}_i, \underline{x}_j) = 1$ if $x_{iq}=x_{jq}$ and 0 otherwise.
- If the q-th coordinates are interval or ratio scaled-valued

$$s_q(\underline{x}_i, \underline{x}_j) = 1 - | x_{iq} - x_{jq} | / r_q,$$

where r_q is the interval where the q-th coordinates of the vectors of the data set X lie.

❖ Fuzzy measures

Let $\underline{x}$, $\underline{y} \in [0,1]^l$. Here the value of the i-th coordinate, x_i, of $\underline{x}$, is **not** the outcome of a measuring device.

- The closer the coordinate x_i is to 1 (0), the more likely the vector $\underline{x}$ possesses (does not possess) the i-th characteristic.
- As x_i approaches 0.5, the certainty about the possession or not of the i-th feature from $\underline{x}$ decreases.

$$s(x_i, y_i) = \max(\min(1 - x_i, 1 - y_i), \min(x_i, y_i))$$

A possible similarity measure that can quantify the above is:

Then $$s_F^q(\underline{x}, \underline{y}) = \left(\sum_{i=1}^{l} s(x_i, y_i)^q \right)^{1/q}$$

❖ Missing data

For some vectors of the data set X, some features values are unknown

Ways to face the problem:

- Discard all vectors with missing values (not recommended for small data sets)
- Find the mean value m_i of the available i-th feature values over that data set and substitute the missing i-th feature values with m_i.
- Define b_i=0, if both the i-th features x_i, y_i are available and 1 otherwise. Then

$$\wp(\underline{x}, \underline{y}) = \frac{l}{l - \sum_{i=1}^{l} b_i} \sum_{all\ i:\ b_i=0} \phi(x_i, y_i)$$

where $\phi(x_i, y_i)$ denotes the PM between two scalars x_i, y_i.

- Find the average proximities $\phi_{avg}(i)$ between all feature vectors in X along all components. Then

$$\wp(\underline{x}, \underline{y}) = \sum_{i=1}^{l} \psi(x_i, y_i)$$

where $\psi(x_i, y_i) = \phi(x_i, y_i)$, if both x_i and y_i are available and $\phi_{avg}(i)$ otherwise.

PROXIMITY FUNCTIONS BETWEEN A VECTOR AND A SET

❖ Let $X=\{\underline{x}_1,\underline{x}_2,\ldots,\underline{x}_N\}$ and $C\subset X$, $\underline{x}\in X$

❖ *All points of C contribute to the definition of* $\wp(\underline{x}, C)$

- Max proximity function

$$\wp^{ps}_{\max}(\underline{x},C)=\max_{\underline{y}\in C}\wp(\underline{x},\underline{y})$$

- Min proximity function

$$\wp^{ps}_{\min}(\underline{x},C)=\min_{\underline{y}\in C}\wp(\underline{x},\underline{y})$$

- Average proximity function

$$\wp^{ps}_{avg}(\underline{x},C)=\frac{1}{n_C}\sum_{\underline{y}\in C}\wp(\underline{x},\underline{y})$$

n_C is the cardinality of C

❖ *A representative(s) of C, r_C, contributes to the definition of $\wp(\underline{x},C)$*

In this case: $\wp(\underline{x},C)=\wp(\underline{x},\underline{r}_C)$

Typical representatives are:

- The mean vector:

$$\underline{m}_p = \left(\frac{1}{n_C}\right)\sum_{\underline{y}\in C}\underline{y}$$

where n_C is the cardinality of C

- The mean center:

$$\underline{m}_C \in C: \ \sum_{\underline{y}\in C} d(\underline{m}_C,\underline{y}) \le \sum_{\underline{y}\in C} d(\underline{z},\underline{y}), \ \forall \underline{z}\in C$$

d: a dissimilarity measure

- The median center:

$$\underline{m}_{med} \in C: \ med(d(\underline{m}_{med},\underline{y}) \mid \underline{y}\in C) \le med(d(\underline{z},\underline{y}) \mid \underline{y}\in C), \ \forall \underline{z}\in C$$

NOTE: Other representatives (e.g., hyperplanes, hyperspheres) are useful in certain applications (e.g., object identification using clustering techniques).

PROXIMITY FUNCTIONS BETWEEN SETS

- Let $X=\{\underline{x}_1,\ldots,\underline{x}_N\}$, D_i, $D_j \subset X$ and $n_i=|D_i|$, $n_j=|D_j|$
- All points of each set contribute to $\wp(D_i,D_j)$
  - Max proximity function (measure but not metric, only if $\wp$ is a similarity measure)

$$\wp^{ss}_{\max}(D_i,D_j)=\max_{\underline{x}\in D_i,\underline{y}\in D_j}\wp(\underline{x},\underline{y})$$

  - Min proximity function (measure but not metric, only if $\wp$ is a dissimilarity measure)

$$\wp^{ss}_{\min}(D_i,D_j)=\min_{\underline{x}\in D_i,\underline{y}\in D_j}\wp(\underline{x},\underline{y})$$

  - Average proximity function (not a measure, even if $\wp$ is a measure)

$$\wp^{ss}_{avg}(D_i,D_j)=\left(\frac{1}{n_i n_j}\right)\sum_{x\in D_i}\sum_{x\in D_j}\wp(\underline{x},\underline{y})$$

- Each set D_i is represented by its representative vector $\underline{m}_i$
  - Mean proximity function (it is a measure provided that $\wp$ is a measure):

$$\wp^{ss}_{mean}(D_i, D_j) = \wp(\underline{m}_i, \underline{m}_j)$$

  - $$\wp^{ss}_{e}(D_i, D_j) = \sqrt{\frac{n_i n_j}{n_i + n_j}}\wp(\underline{m}_i, \underline{m}_j)$$

NOTE: Proximity functions between a vector $\underline{x}$ and a set C may be derived from the above functions if we set $D_i = \{\underline{x}\}$.

- Remarks:
  - Different choices of proximity functions between sets may lead to totally different clustering results.
  - Different proximity measures between vectors in the same proximity function between sets may lead to totally different clustering results.
  - The only way to achieve a proper clustering is
    - by trial and error and,
    - taking into account the opinion of an expert in the field of application.