CLUSTERING

- Clustering – unsupervised classification, where the class labelling of training patterns is unavailable
  - It reveals the organization of patterns into "sensible" clusters, which will allow us to discover similarities and differences among patterns and to derive useful conclusions about them
    - Ex) Image segmentation:

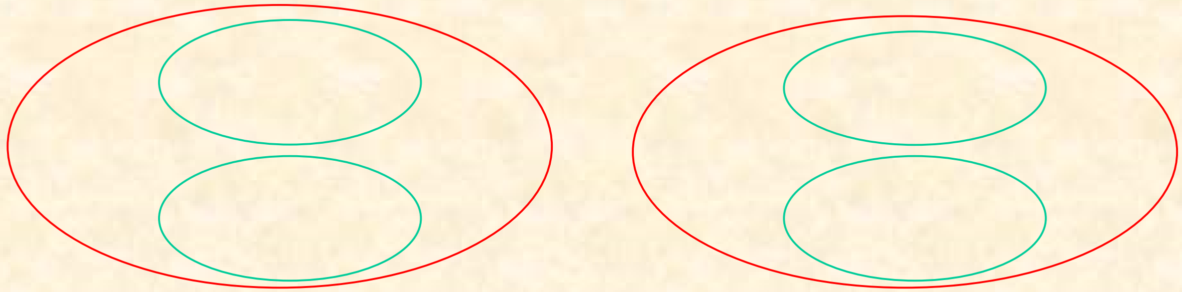
      pattern = element = color pixel
  - Clustering is one of the most primitive mental activities of humans
    - Ex) A "dog" barks

❖ Typical steps in a clustering task

1. Feature selection: Information-rich, non-redundant features
2. Proximity measure:
   - How similar or dissimilar are two feature vectors?
3. Clustering criterion
   - It depends on the type of clusters that are sensible or desirable
   - It is expressed via a cost function
4. Clustering algorithm
5. Validation of results.
6. Interpretation of 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 ??

- Natural cluster – a contiguous region of the space containing a relatively high density of points, separated from other high density regions by regions of relatively low density of points

- Application areas for clustering
  - Data reduction: vector quantization
  - Prediction based on groups: In a mart, a wine buyer usually buys cheese too

TYPES OF FEATURES

- With respect to their <u>domain</u>
  - Continuous
  - Discrete
    - *Binary* or *dichotomous* (the domain consists of two possible values).
- With respect to the <u>relative significance of values</u>
  - Nominal: a value encodes a state
    - 0 (male), 1 (female)
  - Ordinal: values are meaningfully ordered
    - 4 (excellent), 3 (very good), 2 (good), 1(bad)
  - Interval-scaled: differences of two values are meaningful, but ratios are not
    - 5^oC and $10\ ^oC$
  - Ratio-scaled: ratios are also meaning ful
    - 50 kg and 100 kg

Ratio-scaled => interval-scaled => nominal => ordinal

- 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.

- 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 (membership functions)

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

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.

PROXIMITY MEASURES

❖ *Between vectors*

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

$$d: \quad X \times X \longrightarrow \Re$$

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

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

(triangular inequality)

d is called a metric dissimilarity measure.

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

$$s: \; X \times X \longrightarrow \Re$$

with the following properties

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

If in addition

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

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

$$\wp : U \times U \longrightarrow \Re$$

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*

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

Another similarity measure

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

- Discrete-valued vectors
  - Let $F=\{0,1,\ldots,k\text{-}1\}$ be a set of symbols and $X=\{\underline{x}_1,\ldots,\underline{x}_N\} \subset F^l$
  - Let $A(\underline{x},\underline{y})=[a_{ij}]$, $i, j=0,1,\ldots,k\text{-}1$, 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 $$n_x = \sum_{i=1}^{k-1}\sum_{j=0}^{k-1} a_{ij}, \quad n_y = \sum_{i=0}^{k-1}\sum_{j=1}^{k-1} a_{ij},$$

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

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