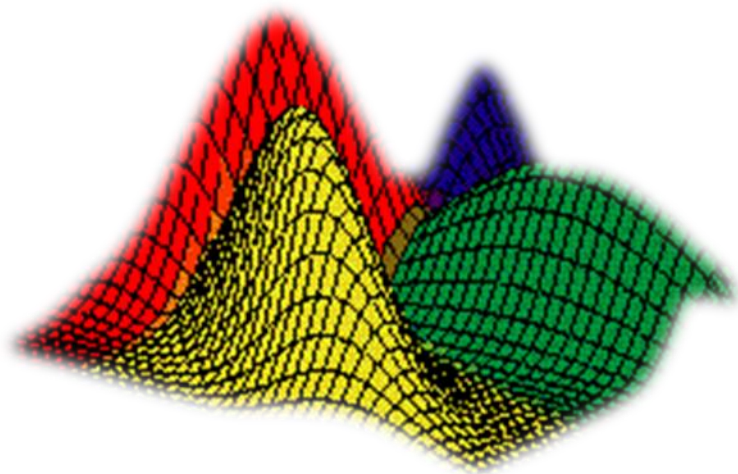
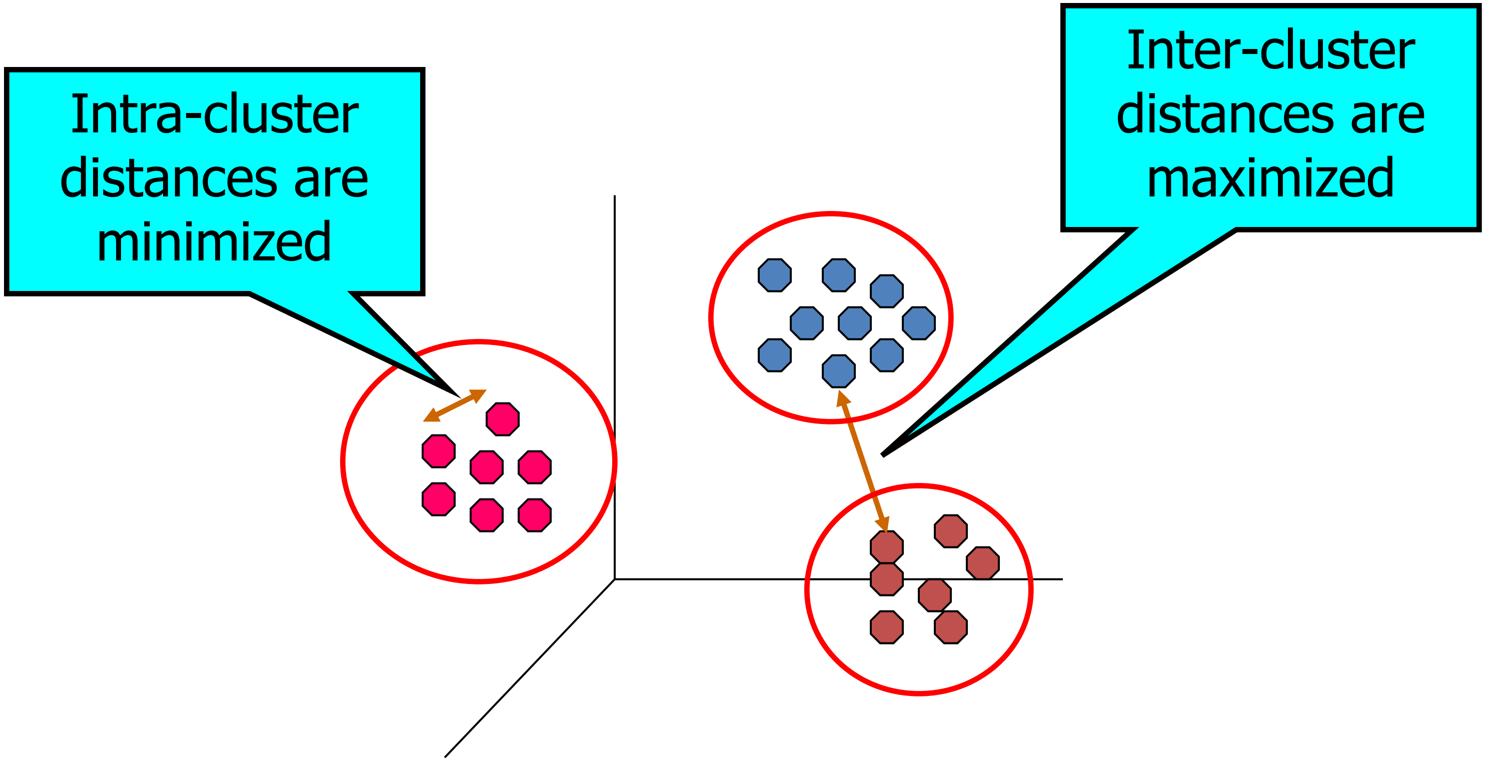


CSE 473: Pattern Recognition

Unsupervised Learning:
Clustering

What is Cluster Analysis?

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



Applications

- **Understanding**
  - Biological taxonomy
  - Group related documents for browsing (e. g., folders)
  - Group genes and proteins that have similar functionality
  - Group stocks with similar price fluctuations

- **Summarization**
  - Reduce the size of large data sets

- **Data Compression**
  - Vector quantization

- **Finding nearest neighbor**

Applications … .

- **Hypothesis generation**
  - To infer some hypothesis
  - Examples:
    - 'big companies invest overseas'
    - 'many BUET students are from Chittagong'
- **Hypothesis testing**
  - To verify an existing hypothesis
- **Prediction based on groups**
  - Predict unknown patterns

What is not Cluster Analysis?

- Supervised classification
  - Have class label information
- Simple segmentation
  - Dividing students into different registration groups alphabetically, by last name
- Results of a query
  - Groupings are a result of an external specification

Clustering Basis

- Basic Concepts

  a clustering criterion must first be adopted.

  Different criteria lead to different clusters.

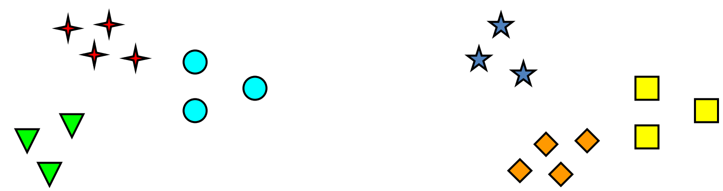
Notion of a Cluster can be Ambiguous

- 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.

Notion of a Cluster can be Ambiguous

– 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.

How many clusters?



Six Clusters

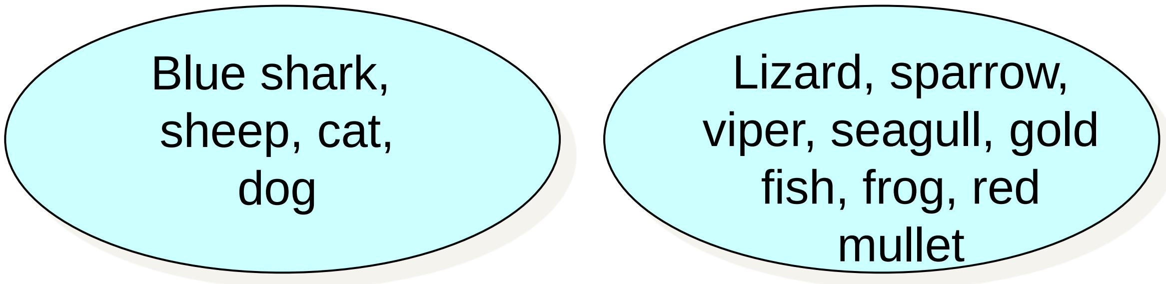
Two Clusters

Four Clusters

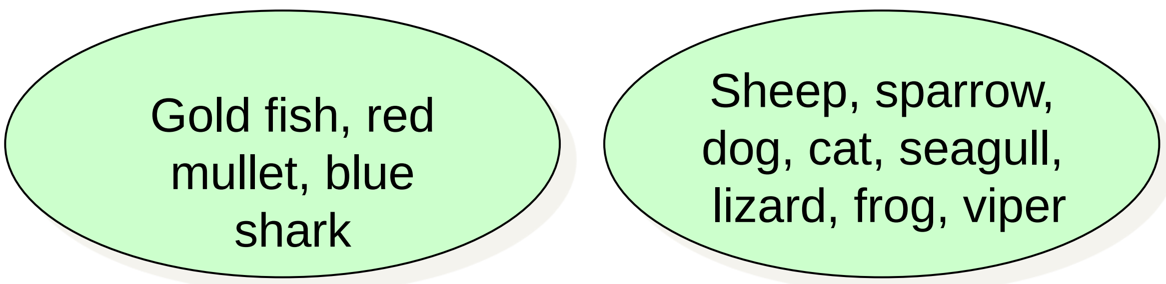
Notion of a Cluster can be Ambiguous

- Let these animals to be clustered
  - Blue shark, sheep, cat, Dog, Lizard, sparrow, viper, seagull, gold fish, frog, red mullet

– A real example

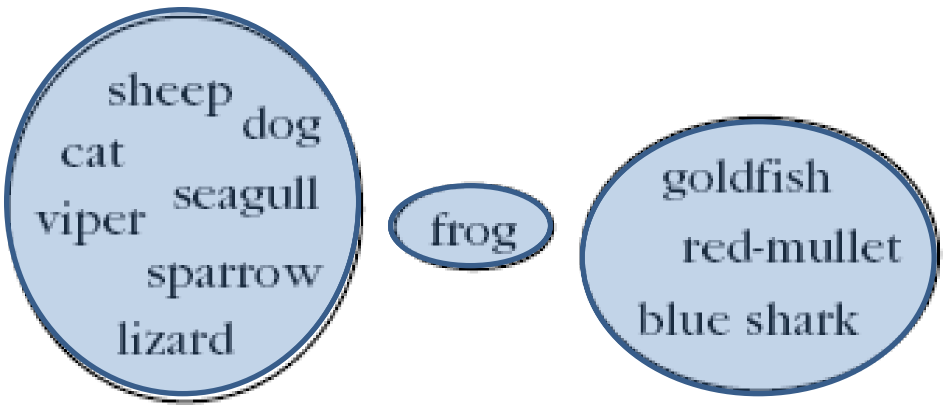


1. Two clusters
2. Clustering criterion: How mammals bear their progeny

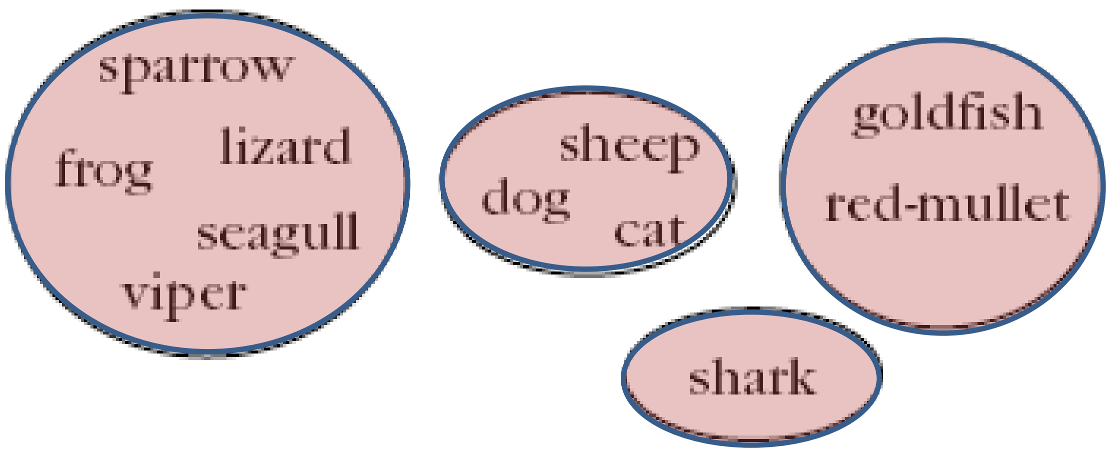


1. Two clusters
2. Clustering criterion: Existence of lungs

– A real example



1. Three clusters
2. Clustering criterion: Their living environment



1. Four clusters
2. Clustering criterion: Bear progeny and Existence of lungs

Clustering Task Stages

- Feature Selection: Information rich features-Parsimony
- Proximity Measure: This quantifies the term similar or dissimilar.
- 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.

Types of Features

- With respect to their domain
  - 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 relative significance of the values they take
  - Nominal (the values code states, e.g., the home district 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).

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.

Clustering Definitions

– 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 is the representative of jth cluster
- $u_j : \underline{x} \rightarrow [0,1], \quad j = 1,2,\ldots,m$

Clustering Definitions

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.

Clustering Definitions

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

Proximity Measures

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

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

with the following properties

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

Proximity Measures

If, in addition

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

  (triangular inequality)

d is called a metric dissimilarity measure.

Proximity Measures

– Similarity measure (between vectors of X) is a function

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

with the following properties

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

Proximity Measures

If, in addition

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

s is called a metric similarity measure.

Proximity Measures

- ***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

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

Proximity Measures Between Points/Vectors

- Real-valued vectors
  - Dissimilarity measures (DMs)

    - *Weighted l_p metric DMs*

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

Interesting instances are obtained for

– p=1 (*weighted Manhattan* norm)

– p=2 (*weighted Euclidean* norm)

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

Proximity Measures Between Vectors

– 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*

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

Proximity Measures Between 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.

Example: l=6, k =3

$$\boldsymbol{x} = [0, 1, 2, 1, 2, 1]^T$$
$$\boldsymbol{y} = [1, 0, 2, 1, 0, 1]^T$$

$$A(\boldsymbol{x}, \boldsymbol{y}) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

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

Proximity Measures Between Discrete-Valued Vectors

- 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)

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

Proximity Measures Between Discrete-Valued Vectors

- 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)

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

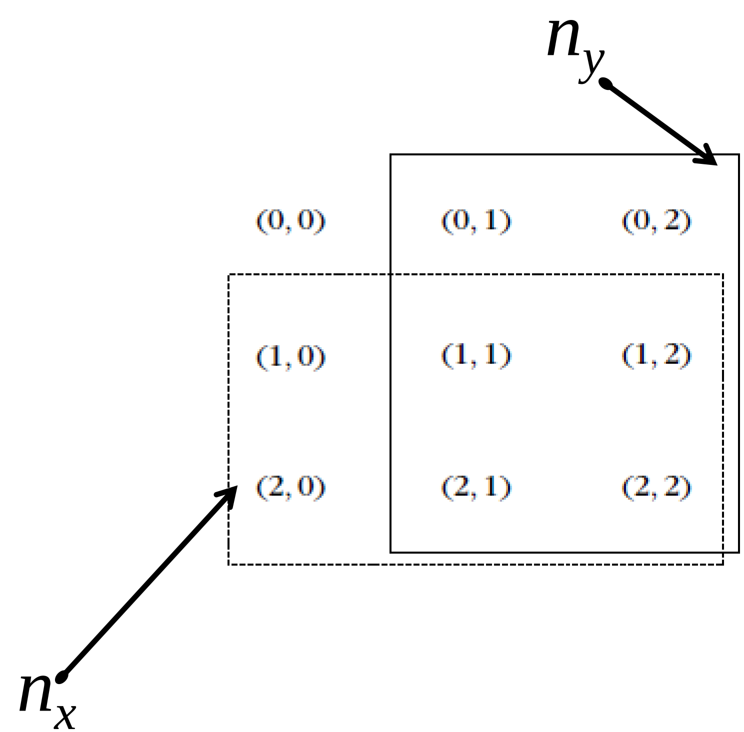
$$\boldsymbol{x} = [0, 1, 2, 1, 2, 1]^T$$
$$\boldsymbol{y} = [1, 0, 2, 1, 0, 1]^T$$

$$A(\boldsymbol{x}, \boldsymbol{y}) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

Proximity Measures Between Discrete-Valued Vectors

– Similarity measures:

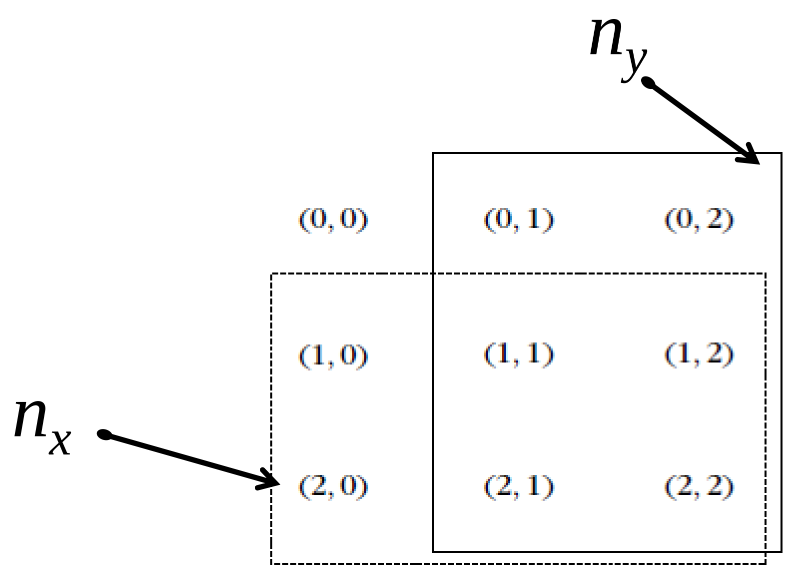
$$A(\boldsymbol{x},\boldsymbol{y}) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$



Proximity Measures Between Discrete-Valued Vectors

– Similarity measures:

$$A(x, y) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$



- Tanimoto measure :

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

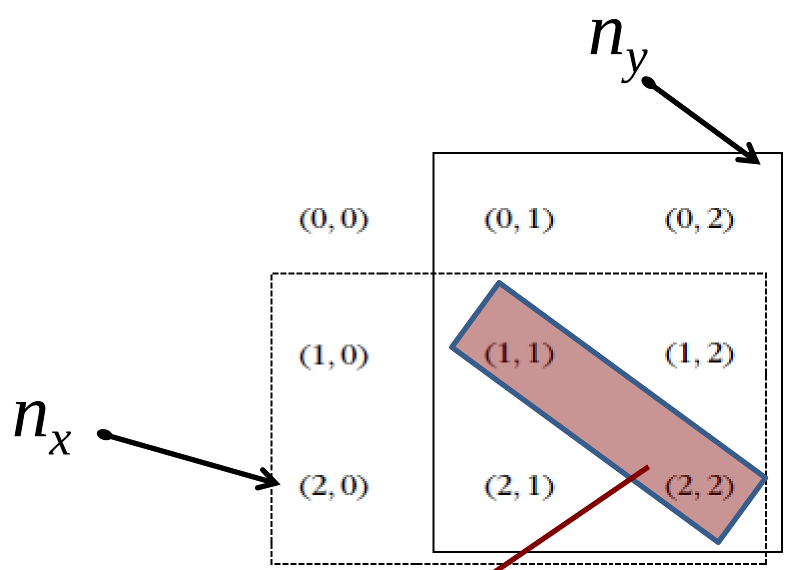
where

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

Proximity Measures Between Discrete-Valued Vectors

– Similarity measures:

$$A(\boldsymbol{x}, \boldsymbol{y}) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$



- 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

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

Proximity Measures Between Mixed-Valued Vectors

- Some of the coordinates of the vectors $\underline{\boldsymbol{x}}$ 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.

Proximity Measures Between Mixed-Valued Vectors

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 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 Proximity 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.

Fuzzy Proximity Measures

A possible similarity measure that can quantify the above is:

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

Then

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

Proximity Measures For 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.

Proximity Measures For Missing Data

– Define $b_i=0$, if both the *i-th* features x_i, y_i are available and $b_i=1$ 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.

Proximity Measures For Missing Data

– 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 $C \subset X$, $\underline{x} \in X$

- *<u>All points of C contribute to the definition of</u>* $\wp(\underline{x}, C)$
  - Max proximity function

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

  - Min proximity function

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

  - Average proximity function

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

(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:

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

  where n_C is the cardinality of C

  – The mean center:

  d: a dissimilarity measure

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

  – 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

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

  - Min proximity function

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

  - Average proximity function

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

Proximity Functions Between Sets

- 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):

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

➢ 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.