Clustering algorithms

Konstantinos Koutroumbas

Unit 2

– Proximity functions between vectors
– Proximity functions between sets
– Proximity functions between a point and a set

koutroum@noa.gr

Proximity measures: Definitions

(A) Between vectors

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

$$d: X \times X \rightarrow \Re$$

with the following properties

1. $\exists d_0 \in \Re: 0 \leq d_0 \leq d(\boldsymbol{x}, \boldsymbol{y}) < +\infty, \forall \boldsymbol{x}, \boldsymbol{y} \in X$
2. $d(\boldsymbol{x}, \boldsymbol{x}) = d_0, \forall \boldsymbol{x} \in X$
3. $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x}), \forall \boldsymbol{x}, \boldsymbol{y} \in X$

Examples: Euclidean distance, Manhattan distance etc.

If in addition:

4. $d(\boldsymbol{x}, \boldsymbol{y}) = d_0 \Leftrightarrow \boldsymbol{x} = \boldsymbol{y}$
5. $d(\boldsymbol{x}, \boldsymbol{z}) \leq d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{y}, \boldsymbol{z}), \forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in X$ (triangular inequality)

d is called **metric** dissimilarity measure.

Proximity measures : Definitions

(A) Between vectors

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

$$s: X \times X \rightarrow \Re$$

with the following properties

1. $\exists s_0 \in \Re: 0 \leq s(\boldsymbol{x}, \boldsymbol{y}) \leq s_0 < +\infty, \forall \boldsymbol{x}, \boldsymbol{y} \in X$
2. $s(\boldsymbol{x}, \boldsymbol{x}) = s_0, \forall \boldsymbol{x} \in X$
3. $s(\boldsymbol{x}, \boldsymbol{y}) = s(\boldsymbol{y}, \boldsymbol{x}), \forall \boldsymbol{x}, \boldsymbol{y} \in X$

If in addition:

4. $s(\boldsymbol{x}, \boldsymbol{y}) = s_0 \Leftrightarrow \boldsymbol{x} = \boldsymbol{y}$
5. $\frac{1}{s(\boldsymbol{x},\boldsymbol{z})} \leq \frac{1}{s(\boldsymbol{x},\boldsymbol{y})} + \frac{1}{s(\boldsymbol{y},\boldsymbol{z})}, \forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in X$

s is called **metric** similarity measure.

Examples: inner product, Tanimoto distance etc.

NOTE:
Similarity measures and dissimilarity measures are also referred as **proximity measures**.

NOTATION:
- Similarity measure: *s*
  dissimilarity measure: *d*
- **proximity measures:** $\wp$

Proximity measures : Definitions

Exercise:

Consider the case where the elements of X are **scalars**.

Which of the following is

(a) a dissimilarity measure,

(b) a **metric** dissimilarity measure?

1. $d_1(x, y) = |x - y|$

2. $d_2(x, y) = |x^2 - y^2|$

3. $d_3(x, y) = \cos(x - y)$

4. $d_4(x, y) = \sin(|x - y|)$

Proximity measures: Definitions

(B) Between sets

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

A **proximity measure** (similarity or dissimilarity) $\wp$ on U is a function

$$\wp: U \times U \rightarrow \Re$$

For dissimilarity measure the following properties should hold

1. $\exists d_0 \in \Re: 0 \leq d_0 \leq d(D_i, D_j) < +\infty, \forall D_i, D_j \in X$

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

3. $d(D_i, D_j) = d(D_j, D_i), \forall D_i, D_j \in X$

If in addition:

4. $d(D_i, D_j) = d_0 \Leftrightarrow D_i = D_j$

5. $d(D_i, D_k) \leq d(D_i, D_j) + d(D_j, D_k), \forall D_i, D_j, D_k \in X$

d is called **metric** dissimilarity measure.

Question: What is the definition when $\wp$ stands for a similarity measure?

Proximity measures: Definitions

(B) Between sets

NOTE: The **definition** of the *proximity functions between sets* **passes through** the definition of *proximity functions between a point and a set*.

Roadmap for the next few slides:

***Proximity** functions between a point and a set*

- **Nonparametric** case

- **Parametric** case
  - **Point** representatives
    - Mean vector
    - Mean center
    - Median center
  - **Hyperplane** representatives
  - **Hypersphere** representatives
  - …

Proximity measures: Definitions

(B) Between sets

NOTE: The **definition** of the *proximity functions between sets* **passes through** the definition of *proximity functions between a point and a set*.

Roadmap for the next few slides:

***Proximity** functions between a point and a set*

- **Nonparametric** case

- **Parametric** case
  - **Point** representatives
    - Mean vector
    - Mean center
    - Median center
  - **Hyperplane** representatives
  - **Hypersphere** representatives
  - …

Proximity functions between a point and a set

Remark: Having in mind that a cluster is actually a set C, a proximity function between a point $\boldsymbol{x}$ and a set C actually **quantifies** the resemblance/relation of $\boldsymbol{x}$ with the cluster C.

Let $X = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\}$ and $\boldsymbol{x} \in X, C \subset X$

Definitions of $\wp(\boldsymbol{x}, C)$:

(a) All points of C **contribute** to the definition of $\wp(\boldsymbol{x}, C)$ (**nonparametric** repr.).

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

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

- Max proximity function

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

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

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

- Min proximity function

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

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

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

- Average proximity function

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

n_C is the cardinality of C.

Proximity measures: Definitions

(B) Between sets

NOTE: The **definition** of the *proximity functions between sets* **passes through** the definition of *proximity functions between a point and a set*.

Roadmap for the next few slides:

***Proximity** functions between a point and a set*

- **Nonparametric** case

- **Parametric** case
  - **Point** representatives
    - Mean vector
    - Mean center
    - Median center
  - **Hyperplane** representatives
  - **Hypersphere** representatives
  - …

Proximity functions between a point and a set

Definitions of $\wp(\boldsymbol{x}, C)$ (cont.):

(b) A representative of C, r_C, **contributes** to the definition of $\wp(\boldsymbol{x}, C)$ (**parametric** repr.).

In this case

$$\wp(\boldsymbol{x}, C) = \wp(\boldsymbol{x}, r_C)$$

Typical **point** representatives are:

- The mean vector

$$\boldsymbol{m}_p = \frac{1}{n_C} \sum_{\boldsymbol{y} \in C} \boldsymbol{y}$$

n_C is the cardinality of C.

- The mean center

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

d: dissimilarity measure.

- The median center

$$\boldsymbol{m}_{med} \in C\colon med(d(\boldsymbol{m}_{med}, \boldsymbol{y}) | \boldsymbol{y} \in C) \le med(d(\boldsymbol{z}, \boldsymbol{y}) | \boldsymbol{y} \in C), \forall\, \boldsymbol{z} \in C$$

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

Proximity functions between a point and a set

Definitions of $\wp(\boldsymbol{x}, C)$ (cont.):

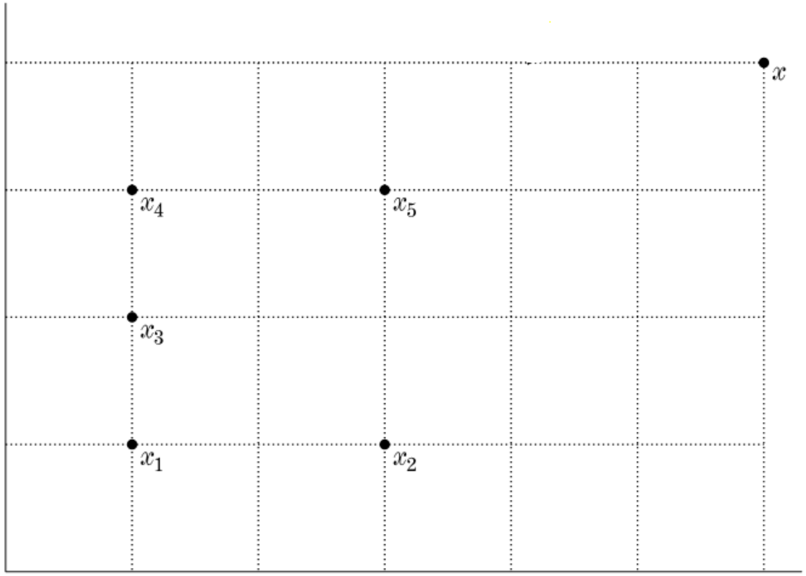
(b) A representative of C, r_C, **contributes** to the definition of $\wp(\boldsymbol{x}, C)$.

In this case $\wp(\boldsymbol{x}, C) = \wp(\boldsymbol{x}, r_C)$

Exercise 5: Let $C = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5\}$, where $\boldsymbol{x}_1 = [1,1]^T$, $\boldsymbol{x}_2 = [3,1]^T$, $\boldsymbol{x}_3 = [1,2]^T$, $\boldsymbol{x}_4 = [1,3]^T$, $\boldsymbol{x}_5 = [3,3]^T$. All points lie in the discrete space $\{0,1,2, \ldots, 6\}^2$. Use the Euclidean distance to measure the dissimilarity between two vectors in C.

(a) Determine the mean vector, the mean center and the median center of C.

(b) Compute the distance of point $\boldsymbol{x} = [6,4]^T$ from C using the above defined representatives (where it is valid).



Proximity measures: Definitions

(B) Between sets

NOTE: The **definition** of the *proximity functions between sets* **passes through** the definition of *proximity functions between a point and a set*.

Roadmap for the next few slides:

***Proximity** functions between a point and a set*

- **Nonparametric** case

- **Parametric** case
  - **Point** representatives
    - Mean vector
    - Mean center
    - Median center
  - **Hyperplane** representatives
  - **Hypersphere** representatives
  - …

Proximity functions between a point and a set

Definitions of $\wp(\boldsymbol{x}, C)$ (cont.):

(b) A representative of C, r_C, **contributes** to the definition of $\wp(\boldsymbol{x}, C)$.

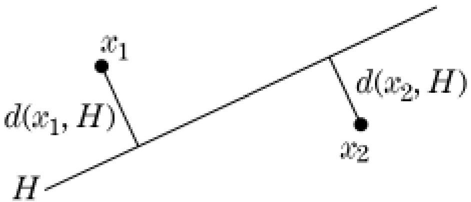
In this case $\wp(\boldsymbol{x}, C) = \wp(\boldsymbol{x}, r_C)$

Linear-shaped clusters:

- Such clusters occur e.g., in computer vision applications.
- In this case, a hyperplane is a better representative of such clusters
- Equation of a hyperplane H:

$$\sum_{j=1}^{l} a_j x_j + a_0 = \boldsymbol{a}^T \boldsymbol{x} + a_0 = 0$$

where $\boldsymbol{x} = [x_1, x_2, \dots, x_l]^T$, $\boldsymbol{a} = [a_1, a_2, \dots, a_l]^T$ is the direction vector of H and a_0 is its offset.



- **Distance** of a point $\boldsymbol{x}$ from H : $d(\boldsymbol{x}, H) = min_{\boldsymbol{z} \in H} d(\boldsymbol{x}, \boldsymbol{z})$
- If $d(\boldsymbol{x}, \boldsymbol{z})$ is the Euclidean distance, it is

$$d(\boldsymbol{x}, H) = \frac{|\boldsymbol{a}^T \boldsymbol{x} + a_0|}{||\boldsymbol{a}||} \qquad ||\boldsymbol{a}|| = \sqrt{\sum_{j=1}^{l} a_j^{\,2}}$$

Proximity functions between a point and a set

Definitions of $\wp(\boldsymbol{x}, C)$ (cont.):

(b) A representative of C, r_C, **contributes** to the definition of $\wp(\boldsymbol{x}, C)$.

In this case $\wp(\boldsymbol{x}, C) = \wp(\boldsymbol{x}, r_C)$

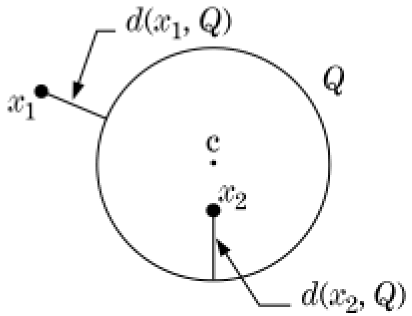
Hyperspherical clusters:

- Such clusters occur e.g., in computer vision applications.
- In this case, a hypersphere is a better representative of such clusters
- Equation of a hypersphere Q:

$$(\boldsymbol{x} - \boldsymbol{c})^T (\boldsymbol{x} - \boldsymbol{c}) = r^2$$

  where $\boldsymbol{x} = [x_1, x_2, \ldots, x_l]^T, \boldsymbol{c} = [c_1, c_2, \ldots, c_l]^T$ is the center of Q and r is its radius.

- **Distance** of a point $\boldsymbol{x}$ from Q: $d(\boldsymbol{x}, Q) = min_{\boldsymbol{z} \in Q} d(\boldsymbol{x}, \boldsymbol{z})$



- For Euclidean distance between two points, $d(\boldsymbol{x}, Q)$ has a geometric insight.

- However, other non-geometric alternatives have also been proposed.

Proximity functions between two sets

Remark: Having in mind that a cluster is actually a set C, a proximity function between two sets actually **quantifies** the resemblance/relation between two clusters.

Let $X = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\}$ and $D_i, D_j \subset X$ with $n_i = |D_i|$, $n_j = |D_j|$.

Definitions of $\wp(D_i, D_j)$:

(a) All points of each set **contribute** to the definition of $\wp(D_i, D_j)$.

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

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

- Max proximity function

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

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

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

- Min proximity function

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

$$d^{ss}{}_{avg}(D_i, D_j) = \frac{1}{n_i n_j} \sum_{x \in D_i} \sum_{y \in D_j} d(\boldsymbol{x}, \boldsymbol{y})$$

$$s^{ss}{}_{avg}(D_i, D_j) = \frac{1}{n_i n_j} \sum_{x \in D_i} \sum_{y \in D_j} s(\boldsymbol{x}, \boldsymbol{y})$$

- Average proximity function

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

Proximity functions between two sets

Definitions of $\wp(D_i, D_j)$ (cont.):

(b) Each set D_i is **represented** by a point representative $\boldsymbol{m}_i$.

- Mean proximity function

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

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

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

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

$$n_i = |D_i|$$

$$n_j = |D_j|$$

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

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

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

Proximity measures between vectors

In the sequel we consider the cases:

$$\boldsymbol{x} = [x_1, \dots, x_l]^T$$
$$\boldsymbol{y} = [y_1, \dots, y_l]^T$$

(A) Real-valued vectors – **dissimilarity** measures (DMs)

(B) Real-valued vectors – **similarity** measures (SMs)

(C) Discrete-valued vectors – **similarity-dissimilarity** measures

(D) Mixed-valued vectors – **dissimilarity** and **similarity** measures

NOTE: Some of the measures below may seem "weird". However, they have been **tailored** for certain types of applications.

Proximity measures between vectors

(A) Real-valued vectors – dissimilarity measures (DMs)

$\boldsymbol{x} = [x_1, \ldots, x_l]^T$
$\boldsymbol{y} = [y_1, \ldots, y_l]^T$

- Weighted l_p metric DMs

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

Interesting instances are obtained for:

$p = 1 \rightarrow d_1(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{l} w_i\, |x_i - y_i|$ (l_1 or Manhattan or city block dist.)

$p = 2 \rightarrow d_2(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=1}^{l} w_i (x_i - y_i)^2}$ (l_2 or Euclidean distance)

$p = \infty \rightarrow d_\infty(\boldsymbol{x}, \boldsymbol{y}) = max_{i=1,\ldots,l} w_i |x_i - y_i|$ (l_∞ or maximum distance)

NOTES:

✓ For $w_i = 1$, we obtain the unweighted versions of the l_p metrics.

✓ It holds: $d_\infty(\boldsymbol{x}, \boldsymbol{y}) \leq d_2(\boldsymbol{x}, \boldsymbol{y}) \leq d_1(\boldsymbol{x}, \boldsymbol{y})$

Proximity measures between vectors

(A) Real-valued vectors – dissimilarity measures (DMs)

$\boldsymbol{x} = [x_1, \dots, x_l]^T$
$\boldsymbol{y} = [y_1, \dots, y_l]^T$

- Mahalanobis distance

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(\boldsymbol{x} - \boldsymbol{y})^T B (\boldsymbol{x} - \boldsymbol{y})}$$

B is symmetric, positive definite matrix

- *Other measures*

– $d_G(\boldsymbol{x}, \boldsymbol{y}) = -log_{10}\left(1 - \frac{1}{l}\sum_{i=1}^{l} \frac{|x_i - y_i|}{|b_i - a_i|}\right)$

•Features may take positive and/or negative values
•Normalization per feature: $0 \le \frac{|x_i - y_i|}{|b_i - a_i|} \le 1$

where b_i and a_i are the maximum and the minimum values of the i-th feature, among the vectors of X (dependence on the current data set)

– $d_Q(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\frac{1}{l}\sum_{i=1}^{l}\left(\frac{x_i - y_i}{x_i + y_i}\right)^2}$

•Features may take **only** non-negative values
•Normalization per feature: $0 \le \frac{|x_i - y_i|}{x_i + y_i} \le 1$

(B) Real-valued vectors –similarity measures (SMs)

$\boldsymbol{x} = [x_1, \dots, x_l]^T$
$\boldsymbol{y} = [y_1, \dots, y_l]^T$

- Inner product

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

- It is usually used either (i) for non-negative valued vectors or (ii) for normalized vectors, i.e., $||\boldsymbol{x}|| = \rho$.
- Concerning (ii), in order to comply with the non-negativity requirement in the definition of the similarity measure, we may consider the similarity measure $s_{inner}(\boldsymbol{x}, \boldsymbol{y}) + \rho^2$

- Cosine similarity measure

$$s_{cosine}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}^T\boldsymbol{y}}{||\boldsymbol{x}|| \cdot ||\boldsymbol{y}||}$$

where $||\boldsymbol{x}|| = \sqrt{\boldsymbol{x}^T\boldsymbol{x}} = \sqrt{\sum_{i=1}^{l} x_i^2}$ and $||\boldsymbol{y}|| = \sqrt{\boldsymbol{y}^T\boldsymbol{y}} = \sqrt{\sum_{i=1}^{l} y_i^2}$.

Proximity measures between vectors

(B) Real-valued vectors –similarity measures (SMs)

$\boldsymbol{x} = [x_1, \dots, x_l]^T$
$\boldsymbol{y} = [y_1, \dots, y_l]^T$

- Pearson's correlation coefficient

$$r_{Pearson}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x_d}^T \boldsymbol{y_d}}{||\boldsymbol{x_d}|| \cdot ||\boldsymbol{y_d}||} \in [-1,1]$$

where $\boldsymbol{x}_d = [x_1 - \bar{x}, \dots, x_l - \bar{x}]^T$, $\boldsymbol{y}_d = [y_1 - \bar{y}, \dots, y_l - \bar{y}]^T$ with $\bar{x} = \frac{1}{l}\sum_{i=1}^{l} x_i$ and $\bar{y} = \frac{1}{l}\sum_{i=1}^{l} y_i$, respectively.

It **measures** the correlation (covariance) between $\boldsymbol{x}, \boldsymbol{y}$

A related dissimilarity measure:

$$D(\boldsymbol{x}, \boldsymbol{y}) = \frac{1 - r_{Pearson}(\boldsymbol{x}, \boldsymbol{y})}{2} \in [0,1]$$

Proximity measures between vectors

(B) Real-valued vectors –similarity measures (SMs)

$x = [x_1, \dots, x_l]^T$
$y = [y_1, \dots, y_l]^T$

- Tanimoto distance

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

Algebraic manipulations give

$$s_T(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{1 + \frac{(\boldsymbol{x} - \boldsymbol{y})^T (\boldsymbol{x} - \boldsymbol{y})}{\boldsymbol{x}^T \boldsymbol{y}}}$$

The larger the agreement between $\boldsymbol{x}, \boldsymbol{y}$, the larger the $s_T(\boldsymbol{x}, \boldsymbol{y})$.

NOTE: $s_T(\boldsymbol{x}, \boldsymbol{y})$ is inversely proportional to the Euclidean distance and proportional to the inner product.

- *Other measure:*

$$s_C(\boldsymbol{x}, \boldsymbol{y}) = 1 - \frac{\sqrt{(\boldsymbol{x} - \boldsymbol{y})^T (\boldsymbol{x} - \boldsymbol{y})}}{||\boldsymbol{x}|| + ||\boldsymbol{y}||} \in [0,1]$$

(C) Discrete-valued vectors – similarity & dissimilarity measures (SMs-DMs)

Let F_i be the **discrete** set of values the i-th feature (nominal/categorical attribute) can take
and n_i be its cardinality, $i = 1, \ldots, l$.

$$\boldsymbol{x} = [x_1, \ldots, x_l]^T$$
$$\boldsymbol{y} = [y_1, \ldots, y_l]^T$$

Consider two l-dimensional vectors

$$\boldsymbol{x} = [x_1, x_2, \ldots, x_k, \ldots, x_l]^T \in F_1 \mathrm{x} F_2 \mathrm{x} \ldots \mathrm{x} F_k \mathrm{x} \ldots \mathrm{x} F_l$$
$$\boldsymbol{y} = [y_1, y_2, \ldots, y_k, \ldots, y_l]^T \in F_1 \mathrm{x} F_2 \mathrm{x} \ldots \mathrm{x} F_k \mathrm{x} \ldots \mathrm{x} F_l$$

The similarity measure $s(\boldsymbol{x}, \boldsymbol{y})$ is defined as

$$s(\boldsymbol{x}, \boldsymbol{y}) = \sum_{k=1}^{l} w_k s_k(x_k, y_k)$$

where $s_k(x_k, y_k)$ is the **feature** similarity measure between the values x_k, y_k of the k-th feature.

Thus, in order to define $s(\boldsymbol{x}, \boldsymbol{y})$, we need to **define** $s_k(x_k, y_k)$.

Proximity measures between vectors

(C) Discrete-valued vectors – similarity & dissimilarity measures (SMs-DMs)

Example: Let $l=3$ and

$$\boldsymbol{x} = [x_1, \ldots, x_l]^T$$
$$\boldsymbol{y} = [y_1, \ldots, y_l]^T$$

$$F_1 = \{a, b, c\}$$
$$F_2 = \{1,2,3,4\}$$
$$F_3 = \{A, B, C\}$$

Consider the vectors:

$$\boldsymbol{x} = [x_1, x_2, x_3]^T = [a, 2, A]^T$$
$$\boldsymbol{y} = [y_1, y_2, y_3]^T = [a, 3, B]^T$$

That is, $x_1 = a$, $y_1 = a$,
$x_2 = 2$, $y_2 = 3$,
$x_3 = A$, $y_3 = B$.

Thus

$$s_1(x_1, y_1) = s_1(a, a)$$
$$s_2(x_2, y_2) = s_2(2, 3)$$
$$s_3(x_3, y_3) = s_3(A, B)$$

and

$$s(\boldsymbol{x}, \boldsymbol{y}) = w_1 \cdot s_1(a, a) + w_2 \cdot s_2(2, 3) + w_3 \cdot s_3(A, B)$$

Proximity measures between vectors

(C) Discrete-valued vectors – similarity & dissimilarity measures (SMs-DMs)

Let F_i be the **discrete** set of values the i-th (nominal/categorical) feature can take
and n_i be its cardinality, $i=1,\dots,l$.

$$\boldsymbol{x} = [x_1, \dots, x_l]^T$$
$$\boldsymbol{y} = [y_1, \dots, y_l]^T$$

$$s(\boldsymbol{x}, \boldsymbol{y}) = \sum_{k=1}^{l} w_k s_k(x_k, y_k)$$

Recall that, in order to define $s(\boldsymbol{x}, \boldsymbol{y})$, we need to **define** $s_k(x_k, y_k)$.

Each $s_k(\cdot,\cdot)$ is completely **defined** by the associated similarity matrix.

If $F_k = \{1,2, \dots, q\}$, the similarity matrix associated with the k-th feature is

	1	**2**	**. . .**	$\boldsymbol{q}$
1	$s_k(1,1)$	$s_k(1,2)$	. . .	$s_k(1,q)$
2	$s_k(2,1)$	$s_k(2,2)$	. . .	$s_k(2,q)$
. . .	. . .	. . .	⋱	. . .
q	$s_k(q,1)$	$s_k(q,2)$	. . .	$s_k(q,q)$

NOTE: (a) The similarity matrix is **completely defined** if all of its entries are defined.
(b) Such a similarity matrix is **associated** with a similarity measure for a **single** discrete-valued feature.

Proximity measures between vectors

(C) Discrete-valued vectors – similarity & dissimilarity measures (SMs-DMs)

There are plenty of similarity measures for single discrete-valued features.

Defining such a similarity measure ⇔ **filling** the entries of the similarity matrix.

The entries filling may be carried out by utilizing:

- Simply 0 and 1 entries
- The size of the data set N
- The number of attributes n involved in the current problem
- The cardinality of F_q, n_q.
- The number of times, $f_k(j)$, the j-th symbol is encountered as k-th feature in the data set
- The frequency of occurrence of the j-th symbol as k-th feature in the data set, defined as $\hat{p}_k(j) = f_k(j)/N$, or, in some cases, $p_k{}^2(j) = \frac{f_k(j)(f_k(j)-1)}{N(N-1)}$

	1	2	. . .	q
1	$s_k(1,1)$	$s_k(1,2)$	. . .	$s_k(1,q)$
2	$s_k(2,1)$	$s_k(2,2)$	. . .	$s_k(2,q)$
. . .	. . .	. . .	⋱	. . .
q	$s_k(q,1)$	$s_k(q,2)$	. . .	$s_k(q,q)$

Proximity measures between vectors

(C) Discrete-valued vectors – similarity & dissimilarity measures (SMs-DMs)

These similarity measures can be categorized in terms of:

- ✓ The *way they fill the entries of the similarity matrix*
  - **I.** **Fill** the diagonal entries only
  - **II.** **Fill** the non-diagonal entries only
  - **III.** **Fill** both diagonal and non-diagonal entries

- ✓ The *arguments they use to define the measure* (information theoretic, probabilistic etc).

Proximity measures between vectors

(C) Discrete-valued vectors – similarity & dissimilarity measures (SMs-DMs)

Indicative measures from category **I**: **Fill** the diagonal entries only.

- Overlap measure

$$s_k(x_k, y_k) = \begin{cases} 1, & if\ x_k = y_k \\ 0, & otherwise \end{cases}, \quad w_k = \frac{1}{l}$$

$$s(\boldsymbol{x}, \boldsymbol{y}) = \sum_{k=1}^{l} w_k s_k(x_k, y_k)$$

$$s_k(x_k, y_k) \in \{0,1\}$$

- Goodall3 measure

$$s_k(x_k, y_k) = \begin{cases} 1 - {p_k}^2(x_k), & if\ x_k = y_k \\ 0, & otherwise \end{cases}, \quad w_k = \frac{1}{l}$$

$$s_k(x_k, y_k) \in [0, 1 - \frac{2}{N(N-1)}]$$

Comment: It **assigns** a high similarity **if** the matching values are **infrequent** regardless of the frequencies of the other values.

(C) Discrete-valued vectors – similarity & dissimilarity measures (SMs-DMs)

Indicative measures from category **II: Fill** the non-diagonal entries only.

- Eskin measure

$$s_k(x_k, y_k) = \begin{cases} 1, & if\ x_k = y_k \\ \frac{n_k^2}{n_k^2+2}, & otherwise \end{cases}, \quad w_k = \frac{1}{l}$$

$$s_k(x_k, y_k) \in [\frac{2}{3}, 1]$$

Comments:

- It **gives** more weight to mismatches for attributes that take **many** values.
- It has been **used** for record-based network intrusion detection data.

- Inverse Occurrence Frequency (**IOF**) measure

$$s_k(x_k, y_k) \in [\frac{1}{1+(log\frac{N}{2})^2}, 1]$$

$$s_k(x_k, y_k) = \begin{cases} 1, & if\ x_k = y_k \\ \frac{1}{1+log f_k(x_k) \cdot log f_k(y_k)}, & otherwise \end{cases}, \quad w_k = \frac{1}{l}$$

Comments:

- It **assigns** lower similarity to mismatches on **more frequent** values..
- It is related to the concept of inverse document frequency which comes from information retrieval, where it is used to signify the relative number of documents that contain a specific word.

Proximity measures between vectors

(C) Discrete-valued vectors – similarity & dissimilarity measures (SMs-DMs)

Indicative measures from category **III**: **Fill** both diagonal & non-diagonal entries

- Lin measure

$$s_k(x_k, y_k) = \begin{cases} 2 \cdot log\hat{p}_k(x_k), & if\ x_k = y_k \\ 2 \cdot \log(\hat{p}_k(x_k) + \hat{p}_k(y_k)), & otherwise \end{cases},$$

$$w_k = \frac{1}{\sum_{i=1}^{l}(\log \hat{p}_i(x_i) + log\hat{p}_i(y_i))}$$

$s_k(x_k, y_k) \in [-2logN, 0]$ for **match**

$s_k(x_k, y_k) \in [-2log\frac{N}{2}, 0]$ for **mismatch**

Comments:

It **gives**

- higher weight to matches on frequent values, and
- lower weight to mismatches on infrequent values.

It has been **used** in word similarity procedure.

(*) S. Boriah, V. Chandola, and V. Kumar, "Similarity measures for categorical data: A Comparative Evaluation," in *Proc. SDM*, pp. 243-254, 2008.

Proximity measures between vectors

(C) Discrete-valued vectors – similarity & dissimilarity measures (SMs-DMs)

	Feat. 1	Feat. 2	Feat. 3
$\boldsymbol{x}_1$	a	1	A
$\boldsymbol{x}_2$	b	4	B
$\boldsymbol{x}_3$	a	3	B
$\boldsymbol{x}_4$	c	2	A
$\boldsymbol{x}_5$	a	2	A
$\boldsymbol{x}_6$	a	2	B
$\boldsymbol{x}_7$	b	1	B
$\boldsymbol{x}_8$	c	1	A
$\boldsymbol{x}_9$	b	1	A
$\boldsymbol{x}_{10}$	a	3	B
$\boldsymbol{x}_{11}$	a	4	A
$\boldsymbol{x}_{12}$	b	4	C
$\boldsymbol{x}_{13}$	b	3	A
$\boldsymbol{x}_{14}$	c	2	A
$\boldsymbol{x}_{15}$	a	2	C

Exercise 1: Consider the data set X given in the adjacent table.
Determine the similarity between the vectors $\boldsymbol{x} = [a, 2, A]^T$ and $\boldsymbol{y} = [a, 3, B]^T$ utilizing

(a) The overlap measure
(b) The Goodall3 measure
(c) The Eskin measure
(d) The IOF measure
(e) The Lin measure.

Exercise 2: Define corresponding dissimilarity measures for the above defined similarity measures.