

Clustering algorithms

Konstantinos Koutroumbas

Unit 2

- Proximity measures between vectors and sets
- Cluster (point) representatives

Proximity measures: Definitions

(B) Between sets

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

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

$$\wp: U \times U \rightarrow \mathfrak{R}$$

For **dissimilarity measure** the following properties should hold

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

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

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

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

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**.

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 x and a set C actually **quantifies** the **resemblance/relation** of x with the cluster C .

Let $X = \{x_1, \dots, x_N\}$ and $x \in X, C \subset X$

Definitions of $\wp(x, C)$:

(a) **All points of C contribute** to the definition of $\wp(x, C)$.

- **Max** proximity function

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

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

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

- **Min** proximity function

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

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

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

- **Average** proximity function

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

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

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

n_C is the
cardinality of C .

Proximity functions between a point and a set

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

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

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

Typical **point** representatives are:

- The **mean vector**

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

n_C is the **cardinality** of C .

- The **mean center**

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

- The **median center**

$$\mathbf{m}_{med} \in C: \text{med}(d(\mathbf{m}_{med}, \mathbf{y}) | \mathbf{y} \in C) \leq \text{med}(d(\mathbf{z}, \mathbf{y}) | \mathbf{y} \in C), \forall \mathbf{z} \in C$$

d: dissimilarity measure.

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

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 $X = \{x_1, \dots, x_N\}$ 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 $\wp(D_i, D_j)$.

- **Max** proximity function

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

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

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

- **Min** proximity function

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

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

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

- **Average** proximity function

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

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

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

Proximity functions between two sets

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

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

- **Mean** proximity function

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

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

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

- $\wp^{SS}_e(D_i, D_j) = \sqrt{\frac{n_i n_j}{n_i + n_j}} \wp(\mathbf{m}_i, \mathbf{m}_j)$

$$n_i = |D_i|$$
$$n_j = |D_j|$$

$$d^{SS}_e(D_i, D_j) = \sqrt{\frac{n_i n_j}{n_i + n_j}} d(\mathbf{m}_i, \mathbf{m}_j)$$

$$s^{SS}_e(D_i, D_j) = \sqrt{\frac{n_i n_j}{n_i + n_j}} s(\mathbf{m}_i, \mathbf{m}_j)$$

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

Proximity measures between vectors

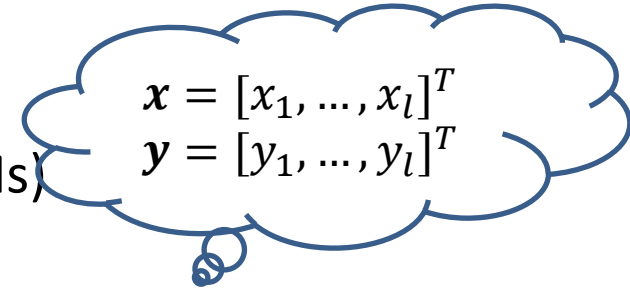
In the sequel we consider the cases:

(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


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

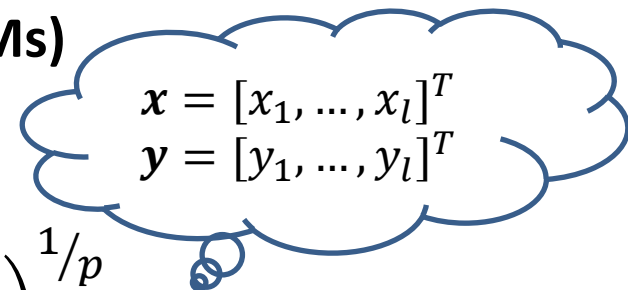
NOTE: Some of the **measures** below may seem “**weird**”. However, they have been **tailored** for certain types of applications.

Proximity measures between vectors

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

- Weighted l_p metric DMs

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


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

Interesting instances are obtained for:

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

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

$$p = \infty \rightarrow d_\infty(\mathbf{x}, \mathbf{y}) = \max_{i=1, \dots, l} w_i |x_i - y_i| \quad (l_\infty \text{ or maximum distance})$$

NOTES:

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

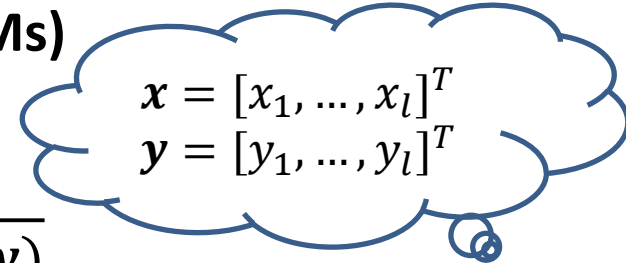
✓ It holds: $d_\infty(\mathbf{x}, \mathbf{y}) \leq d_2(\mathbf{x}, \mathbf{y}) \leq d_1(\mathbf{x}, \mathbf{y})$

Proximity measures between vectors

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

- **Mahalanobis** distance

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

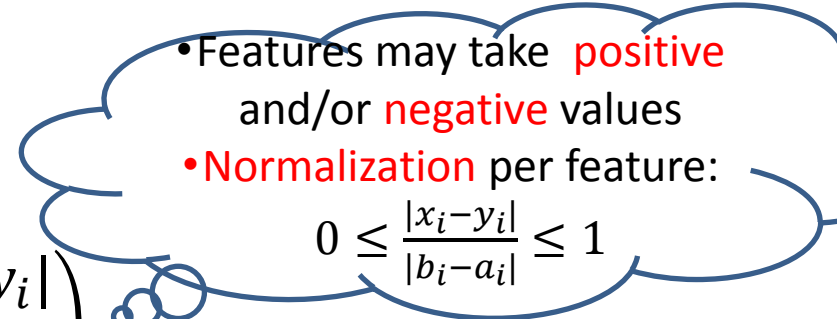


$\mathbf{x} = [x_1, \dots, x_l]^T$
 $\mathbf{y} = [y_1, \dots, y_l]^T$

B is **symmetric, positive definite** matrix

- Other measures

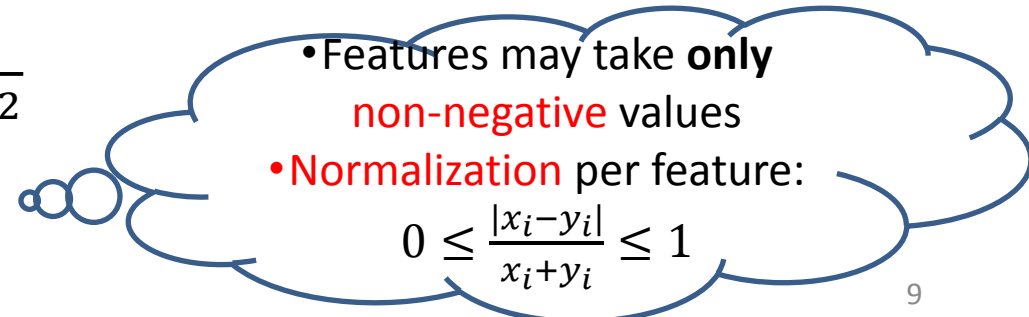
$$-d_G(\mathbf{x}, \mathbf{y}) = -\log_{10} \left(1 - \frac{1}{l} \sum_{i=1}^l \frac{|x_i - y_i|}{|b_i - a_i|} \right)$$



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

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**)

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

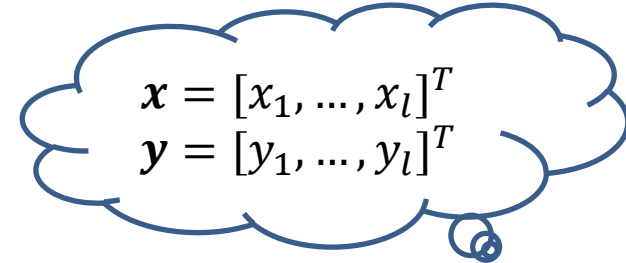


- Features may take **only non-negative** values
- **Normalization** per feature:
 $0 \leq \frac{|x_i - y_i|}{x_i + y_i} \leq 1$

Proximity measures between vectors

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

• Inner product


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

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

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

• Cosine similarity measure

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

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

Proximity measures between vectors

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

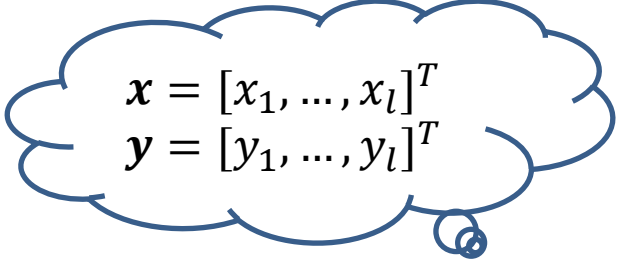
- Pearson's correlation coefficient

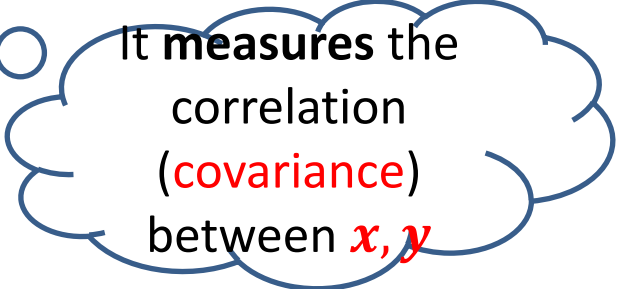
$$r_{\text{Pearson}}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}_d^T \mathbf{y}_d}{\|\mathbf{x}_d\| \cdot \|\mathbf{y}_d\|} \in [-1, 1]$$

where $\mathbf{x}_d = [x_1 - \bar{x}, \dots, x_l - \bar{x}]^T$, $\mathbf{y}_d = [y_1 - \bar{y}, \dots, y_l - \bar{y}]^T$ 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.

A related dissimilarity measure:

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


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



It measures the correlation (covariance) between \mathbf{x}, \mathbf{y}

Proximity measures between vectors

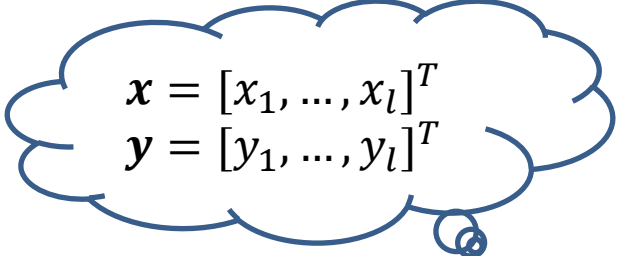
(B) Real-valued vectors –similarity measures (SMs)

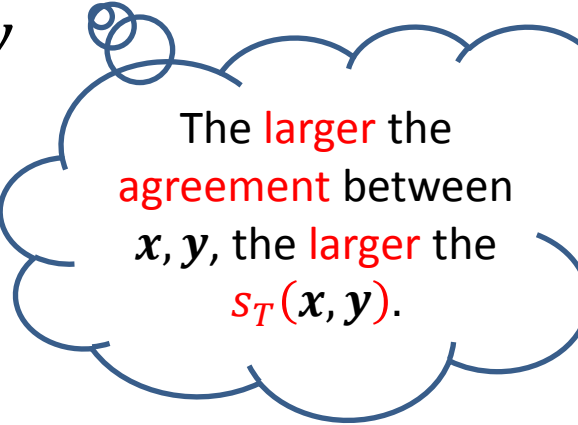
- Tanimoto distance

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

Algebraic manipulations give

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


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



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

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

- Other measure:

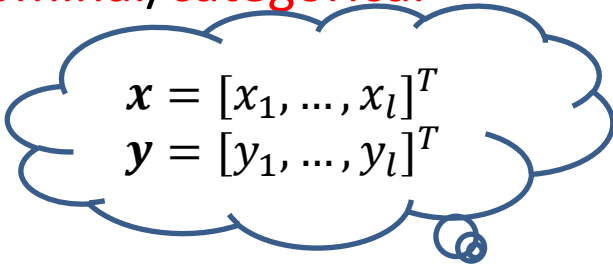
$$s_C(\mathbf{x}, \mathbf{y}) = 1 - \frac{\sqrt{(\mathbf{x} - \mathbf{y})^T (\mathbf{x} - \mathbf{y})}}{\|\mathbf{x}\| + \|\mathbf{y}\|} \in [0, 1]$$

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 feature (**nominal/categorical attribute**) can take

and n_i be its **cardinality**, $i = 1, \dots, l$.


$$\begin{aligned} \mathbf{x} &= [x_1, \dots, x_l]^T \\ \mathbf{y} &= [y_1, \dots, y_l]^T \end{aligned}$$

Consider two l -dimensional vectors

$$\mathbf{x} = [x_1, x_2, \dots, x_k, \dots, x_l]^T \in F_1 \times F_2 \times \dots \times F_k \times \dots \times F_l$$

$$\mathbf{y} = [y_1, y_2, \dots, y_k, \dots, y_l]^T \in F_1 \times F_2 \times \dots \times F_k \times \dots \times F_l$$

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

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

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 $s(\mathbf{x}, \mathbf{y})$, we need to **define** $s_k(x_k, y_k)$.

Proximity measures between vectors

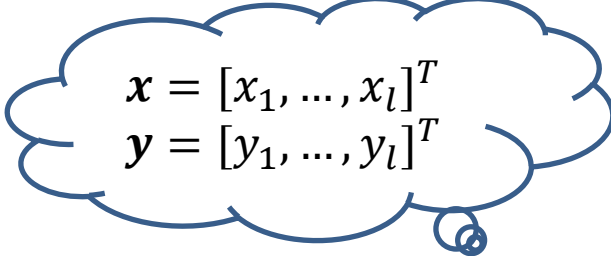
(C) Discrete-valued vectors – similarity & dissimilarity measures (SMs-DMs)

Example: Let $l=3$ and

$$F_1 = \{a, b, c\}$$

$$F_2 = \{1, 2, 3, 4\}$$

$$F_3 = \{A, B, C\}$$


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

Consider the vectors:

$$\mathbf{x} = [x_1, x_2, x_3]^T = [a, 2, A]^T$$

$$\mathbf{y} = [y_1, y_2, y_3]^T = [a, 3, B]^T$$

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

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

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**, $i=1, \dots, l$.

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

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

Recall that, in order to define $s(\mathbf{x}, \mathbf{y})$, 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	...	q
1	$s_k(1,1)$	$s_k(1,2)$. . .	$s_k(1,q)$
2	$s_k(2,1)$	$s_k(2,2)$. . .	$s_k(2,q)$
.	\ddots	. . .
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** \Leftrightarrow **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	$s_k(2,1)$	$s_k(2,2)$. . .	$s_k(2,q)$
.	\ddots	. . .
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

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

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

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

- Goodall3 measure

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

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

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

Proximity measures between vectors

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

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

• Eskin measure

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

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

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

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

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

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

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

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

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

$s_k(x_k, y_k) \in [-2 \log \frac{N}{2}, 0]$ 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
x_1	a	1	A
x_2	b	4	B
x_3	a	3	B
x_4	c	2	A
x_5	a	2	A
x_6	a	2	B
x_7	b	1	B
x_8	c	1	A
x_9	b	1	A
x_{10}	a	3	B
x_{11}	a	4	A
x_{12}	b	4	C
x_{13}	b	3	A
x_{14}	c	2	A
x_{15}	a	2	C

Exercise 1: Consider the data set X given in the adjacent table.

Determine the similarity between the vectors

$$\mathbf{x} = [a, 2, A]^T \text{ and}$$

$$\mathbf{y} = [a, 3, B]^T \text{ utilizing}$$

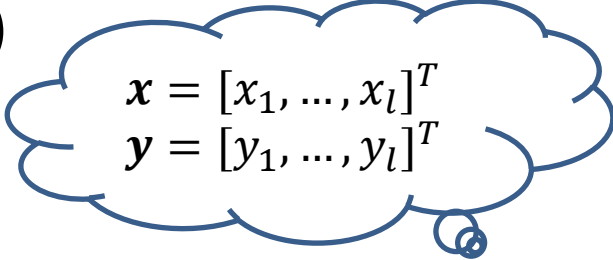
- (a) The **overlap** measure
- (b) The **Goodall3** measure
- (c) The **Eskin** measure
- (d) The **IOF** measure
- (e) The **Lin** measure.

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

Proximity measures between vectors

(D) Mixed-valued vectors –similarity measures (SMs)

Here **some coordinates** of the feature vectors are **real-valued**, while **others** are **discrete-valued**.


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

How to **measure** the **proximity** between \mathbf{x} and \mathbf{y} ?

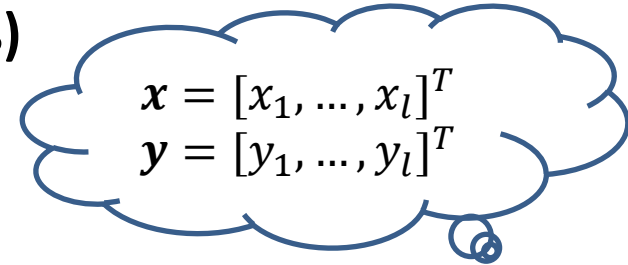
- **Adopt** a **proximity measure suitable** for **real-valued vectors** (**only** for **ordinal discrete-valued** features).
- **Convert** the **real-valued features to discrete-valued ones** (e.g., via quantization) and **employ** a **discrete proximity measure** (again, **only** for **ordinal discrete-valued** features).
- For the more general case where **nominal, ordinal, interval-scaled** and **ratio-scaled** features **co-exist**, we treat each one of them separately, as follows:

Proximity measures between vectors

(D) Mixed-valued vectors –similarity measures (SMs)

The similarity between \mathbf{x} and \mathbf{y} is defined as:

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



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

where:

- $w_k = 0$, if **at least one** of x_k and y_k is **undefined** or (optionally) both x_k and y_k are equal to 0. Otherwise $w_k = 1$.
- If x_k and y_k are **binary**, $s_k(x_k, y_k) = \begin{cases} 1, & \text{if } x_k = y_k = 1 \text{ (or } x_k = y_k) \\ 0, & \text{otherwise} \end{cases}$
- If x_k and y_k are **nominal** or **ordinal**, $s_k(x_k, y_k) = \begin{cases} 1, & x_k = y_k \\ 0, & \text{otherwise} \end{cases}$
- If x_k and y_k are **interval** or **ratio scaled**-valued

$$s_k(x_k, y_k) = 1 - \frac{|x_k - y_k|}{r_k}$$

This is the **overlap measure**. Other options can also be used.

where r_k is the width of the interval where the k -th coordinates of the vectors of \mathbf{X} lie.

Proximity measures between vectors

(D) Mixed-valued vectors –similarity measures (SMs)

Exercise 2: Consider the data set given in the following table. Each row corresponds to a vector and each column to a feature. The first three features are ratio scaled, the 4th one is nominal and the 5th one is ordinal. Utilizing the previous similarity measure, compute the similarities between any pair of feature vectors.

Company	1 st year budget	2 nd year budget	3 rd year budget	Activity abroad	Rate of services 0: not good 1: good 2: very good
1 (x_1)	1.2	1.5	1.9	0	1
2 (x_2)	0.3	0.4	0.6	0	0
3 (x_3)	10	13	15	1	2
4 (x_4)	6	6	7	1	1

Proximity measures between vectors

Fuzzy measures – an alternative perspective

- Let $\mathbf{x} \in [0,1]^l$.
- In this context, x_k is **not** the **outcome** of a **measuring device**.
- Rather, it **indicates** the **degree** to which \mathbf{x} **possesses** the k -th characteristic.
- The closer the x_k to **1** (**0**), the **more likely** is that \mathbf{x} **possesses** (**does not possess**) the k -th characteristic.
- As x_k **approaches 0.5**, the **certainty** about the **possession or not** of the i -th feature from \mathbf{x} decreases.
- Let

$$\mathbf{x} = [x_1, x_2, \dots, x_k, \dots, x_l]^T \in [0,1]^l$$

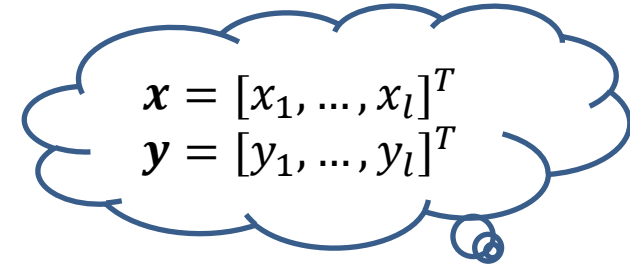
$$\mathbf{y} = [y_1, y_2, \dots, y_k, \dots, y_l]^T \in [0,1]^l$$

- A **measure of similarity between** x_k and y_k is the following

$$s(x_k, y_k) = \max(\min(1 - x_k, 1 - y_k), \min(x_k, y_k))$$

Then, as **measure of similarity between** \mathbf{x} and \mathbf{y} we can use the following

$$s^q(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^l s(x_k, y_k)^q \right)^{1/q}, \quad q \in [1, +\infty)$$


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

Proximity measures between vectors

Fuzzy measures – an alternative perspective

Exercise 3: Let $l = 3$ and $q = 1$.

(a) Consider the vectors $\mathbf{x}_1 = [1,1,1]^T$, $\mathbf{x}_2 = [0,0,1]^T$, $\mathbf{x}_3 = [\frac{1}{2}, \frac{1}{3}, \frac{1}{4}]^T$,
 $\mathbf{x}_4 = [\frac{1}{2}, \frac{1}{2}, \frac{1}{2}]^T$. Determine the similarities $s^1(\mathbf{x}_i, \mathbf{x}_j)$, $i, j = 1,2,3,4$.

(b) Consider the vectors $\mathbf{y}_1 = [\frac{3}{4}, \frac{3}{4}, \frac{3}{4}]^T$, $\mathbf{y}_2 = [1,1,1]^T$, $\mathbf{y}_3 = [\frac{1}{4}, \frac{1}{4}, \frac{1}{4}]^T$,
 $\mathbf{y}_4 = [\frac{1}{2}, \frac{1}{2}, \frac{1}{2}]^T$. Determine the similarities $s^1(\mathbf{y}_i, \mathbf{y}_j)$, $i, j = 1,2,3,4$, $i \neq j$.

(c) Draw your conclusions.