Measures of Similarity and Dissimilarity

Unit - II
Datamining

Measures of Similarity and Dissimilarity

- **Similarity and dissimilarity** are important because they are used by a number of data mining techniques
  - such as
    - clustering,
    - nearest neighbor classification, and
    - anomaly detection.
- **Proximity** is used to refer to either similarity or dissimilarity.
  - proximity between objects having only one simple attribute, and
  - proximity measures for objects with multiple attributes.

Measures of Similarity and Dissimilarity

- **Similarity** between two objects is a numerical measure of the degree to which the two objects are alike.
  - Similarity - high -objects that are more alike.
  - Non-negative
  - between 0 (no similarity) and 1 (complete similarity).
- **Dissimilarity** between two objects is a numerical measure of the degree to which the two objects are different.
  - Dissimilarity - low - objects are more similar.
  - Distance - synonym for dissimilarity

Measures of Similarity and Dissimilarity

Transformations

- Transformations are often applied to
  - convert a similarity to a dissimilarity,
  - convert a dissimilarity to a similarity
  - to transform a proximity measure to fall within a particular range, such as [0,1].
- Example
  - Similarities between objects range from 1 (not at all similar) to 10 (completely similar)
  - we can make them fall within the range [0, 1] by using the transformation
    - $s' = (s-1)/9$
    - s - Original Similarity
    - s' - New similarity values

Measures of Similarity and Dissimilarity

Table 2.7. Similarity and dissimilarity for simple attributes

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$	$s = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}$
Ordinal	$d = \lvert x - y \rvert / (n - 1)$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - d$
Interval or Ratio	$d = \lvert x - y \rvert$	$s = -d$, $s = \frac{1}{1+d}$, $s = e^{-d}$, $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Measures of Similarity and Dissimilarity

Dissimilarities between Data Objects

The Euclidean distance measure given in Equation 2.1 is generalized by the **Minkowski** distance metric shown in Equation 2.2,

$$d(\mathbf{x},\mathbf{y}) = \left(\sum_{k=1}^{n} |x_k - y_k|^r\right)^{1/r}, \tag{2.2}$$

where r is a parameter. The following are the three most common examples of Minkowski distances.

- $r = 1$. City block (Manhattan, taxicab, L_1 norm) distance. A common example is the **Hamming distance**, which is the number of bits that are different between two objects that have only binary attributes, i.e., between two binary vectors.
- $r = 2$. Euclidean distance (L_2 norm).
- $r = \infty$. Supremum (L_{max} or L_∞ norm) distance. This is the maximum difference between any attribute of the objects. More formally, the L_∞ distance is defined by Equation 2.3

$$d(\mathbf{x},\mathbf{y}) = \lim_{r\to\infty}\left(\sum_{k=1}^{n} |x_k - y_k|^r\right)^{1/r}. \tag{2.3}$$

$$d(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}, \tag{2.1}$$

Measures of Similarity and Dissimilarity

Dissimilarities between Data Objects

If $d(x, y)$ is the distance between two points, x and y, then the following **properties** hold.

1. Positivity

(a) $d(x, x) \geq 0$ for all x and y,

(b) $d(x, y) = 0$ only if $x = y$.

2. Symmetry

$d(x, y) = d(y, x)$ for all x and y.

3. Triangle Inequality

$d(x, z) \leq d(x, y) + d(y, z)$ for all points x, y, and z.

Note:-Measures that satisfy all three properties are known as **metrics**.

Measures of Similarity and Dissimilarity

Dissimilarities between Data Objects

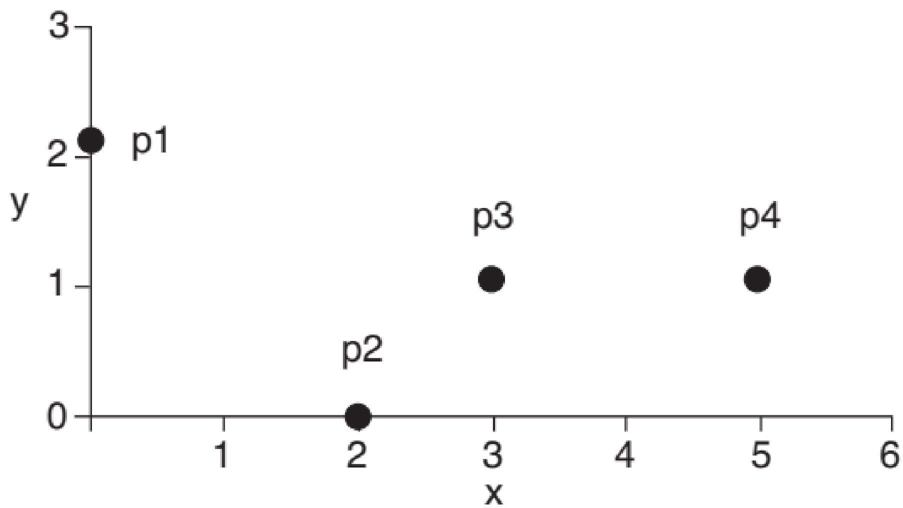


Figure 2.15. Four two-dimensional points.

Table 2.8. x and y coordinates of four points.

point	x coordinate	y coordinate
p1	0	2
p2	2	0
p3	3	1
p4	5	1

Measures of Similarity and Dissimilarity

Dissimilarities between Data Objects

Table 2.9. Euclidean distance matrix for Table 2.8.

	p1	p2	p3	p4
p1	0.0	2.8	3.2	5.1
p2	2.8	0.0	1.4	3.2
p3	3.2	1.4	0.0	2.0
p4	5.1	3.2	2.0	0.0

Measures of Similarity and Dissimilarity

Dissimilarities between Data Objects

Table 2.10. L_1 distance matrix for Table 2.8.

L_1	p1	p2	p3	p4
p1	0.0	4.0	4.0	6.0
p2	4.0	0.0	2.0	4.0
p3	4.0	2.0	0.0	2.0
p4	6.0	4.0	2.0	0.0

Measures of Similarity and Dissimilarity

Dissimilarities between Data Objects

Table 2.11. L_∞ distance matrix for Table 2.8.

L_∞	p1	p2	p3	p4
p1	0.0	2.0	3.0	5.0
p2	2.0	0.0	1.0	3.0
p3	3.0	1.0	0.0	2.0
p4	5.0	3.0	2.0	0.0

Measures of Similarity and Dissimilarity

Dissimilarities between Data Objects

Non-metric Dissimilarities: Set Differences

A = {1, 2, 3, 4} and B = {2, 3, 4},
then A − B = {1} and
B − A = ∅, the empty set.

If d(A, B) = size(A − B), then it does not satisfy the second part of the positivity property, the symmetry property, or the triangle inequality.

d(A, B) = size(A − B) + size(B − A)   (modified which follows all properties)

Measures of Similarity and Dissimilarity

Dissimilarities between Data Objects

Non-metric Dissimilarities: Time

Dissimilarity measure that is not a metric,but still useful.

$$d(t_1, t_2) = \left\{ \begin{array}{ll} t_2 - t_1 & \text{if } t_1 \leq t_2 \\ 24 + (t_2 - t_1) & \text{if } t_1 \geq t_2 \end{array} \right\}. \tag{2.4}$$

d(1PM, 2PM) = 1 hour
d(2PM, 1PM) = 23 hours

- Example:- when answering the question: “If an event occurs at 1PM every day, and it is now 2PM, how long do I have to wait for that event to occur again?”

Distance in python

```
>>> from scipy.spatial import distance
>>> p1=(0,2)
>>> p2=(2,0)
>>> print(distance.cityblock(p1,p2))
4
>>> print(distance.euclidean(p1,p2))
2.8284271247461903
>>> print(distance.minkowski(p1,p2,p=1))
4.0
>>> print(distance.minkowski(p1,p2,p=2))
2.8284271247461903
>>> print(distance.minkowski(p1,p2,p=3))
2.5198420997897464
>>> 
```

Measures of Similarity and Dissimilarity

Similarities between Data Objects

- Typical properties of similarities are the following:
  - 1. s(x, y) = 1 only if x = y. (0 ≤ s ≤ 1)
  - 2. s(x, y) = s(y, x) for all x and y. (Symmetry)
- **A Non-symmetric Similarity Measure**
  - Classify a small set of characters which is flashed on a screen.
  - **Confusion matrix** - records how often each character is classified as itself, and how often each is classified as another character.
  - “0” appeared 200 times but classified as
    - “0” 160 times,
    - “o” 40 times.
  - ‘o’ appeared 200 times and was classified as
    - “o” 170 times
    - “0” only 30 times.
- similarity measure can be made symmetric by setting
  - S`(x, y) = S`(y, x) = (s(x, y)+s(y, x))/2,
    - S` - new similarity measure.

Measures of Similarity and Dissimilarity

Examples of proximity measures

- **Similarity Measures for Binary Data**
  - Similarity measures between objects that contain only binary attributes are called **similarity coefficients**
  - Let **x** and **y** be two objects that consist of **n** binary attributes.
  - The **comparison of two objects** (or two binary vectors), leads to the following four quantities (frequencies):

    $\mathbf{f_{00}}$ = the number of attributes where x is 0 and y is 0
    $\mathbf{f_{01}}$ = the number of attributes where x is 0 and y is 1
    $\mathbf{f_{10}}$ = the number of attributes where x is 1 and y is 0
    $\mathbf{f_{11}}$ = the number of attributes where x is 1 and y is 1

Measures of Similarity and Dissimilarity

Examples of proximity measures

- **Similarity Measures for Binary Data**

  Simple Matching Coefficient(SMC)

$$SMC = \frac{\text{number of matching attribute values}}{\text{number of attributes}} = \frac{f_{11} + f_{00}}{f_{01} + f_{10} + f_{11} + f_{00}}. \qquad (2.5)$$

  Jaccard Coefficient

$$J = \frac{\text{number of matching presences}}{\text{number of attributes not involved in 00 matches}} = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}. \qquad (2.6)$$

Measures of Similarity and Dissimilarity

Examples of proximity measures

- **Similarity Measures for Binary Data**

Example 2.17 (The SMC and Jaccard Similarity Coefficients). To illustrate the difference between these two similarity measures, we calculate SMC and J for the following two binary vectors.

$\mathbf{x} = (1, 0, 0, 0, 0, 0, 0, 0, 0, 0)$
$\mathbf{y} = (0, 0, 0, 0, 0, 0, 1, 0, 0, 1)$

$f_{01} = 2$ the number of attributes where $\mathbf{x}$ was 0 and $\mathbf{y}$ was 1
$f_{10} = 1$ the number of attributes where $\mathbf{x}$ was 1 and $\mathbf{y}$ was 0
$f_{00} = 7$ the number of attributes where $\mathbf{x}$ was 0 and $\mathbf{y}$ was 0
$f_{11} = 0$ the number of attributes where $\mathbf{x}$ was 1 and $\mathbf{y}$ was 1

$$SMC = \frac{f_{11}+f_{00}}{f_{01}+f_{10}+f_{11}+f_{00}} = \frac{0+7}{2+1+0+7} = 0.7 \qquad J = \frac{f_{11}}{f_{01}+f_{10}+f_{11}} = \frac{0}{2+1+0} = 0$$

Measures of Similarity and Dissimilarity

Examples of proximity measures

Cosine similarity (Document similarity)

If x and y are two document vectors, then

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \, \|\mathbf{y}\|}, \tag{2.7}$$

$\cdot$ indicates the vector dot product, $\mathbf{x} \cdot \mathbf{y} = \sum_{k=1}^{n} x_k y_k$;

length of vector $\mathbf{x}$, $\|\mathbf{x}\| = \sqrt{\sum_{k=1}^{n} x_k^2} = \sqrt{\mathbf{x} \cdot \mathbf{x}}$.

Measures of Similarity and Dissimilarity

Examples of proximity measures

cosine similarity (Document similarity)

Example 2.18 (Cosine Similarity of Two Document Vectors). This example calculates the cosine similarity for the following two data objects, which might represent document vectors:

$$\mathbf{x} = (3, 2, 0, 5, 0, 0, 0, 2, 0, 0)$$
$$\mathbf{y} = (1, 0, 0, 0, 0, 0, 0, 1, 0, 2)$$

$$\mathbf{x} \cdot \mathbf{y} = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$$

$$\|\mathbf{x}\| = \sqrt{3*3 + 2*2 + 0*0 + 5*5 + 0*0 + 0*0 + 0*0 + 2*2 + 0*0 + 0*0} = 6.48$$

$$\|\mathbf{y}\| = \sqrt{1*1 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 1*1 + 0*0 + 2*2} = 2.24$$

$$\mathbf{\cos(x, y) = 0.31}$$

Measures of Similarity and Dissimilarity

Examples of proximity measures

cosine similarity (Document similarity)

```
# import required libraries
import numpy as np
from numpy.linalg import norm

# define two lists or array
A = np.array([2,1,2,3,2,9])
B = np.array([3,4,2,4,5,5])

print("A:", A)
print("B:", B)

# compute cosine similarity
cosine = np.dot(A,B)/(norm(A)*norm(B))
print("Cosine Similarity:", cosine)
```

Measures of Similarity and Dissimilarity

Examples of proximity measures

cosine similarity (Document similarity)

- Cosine similarity - measure of angle between x and y.
- Cosine similarity = 1 (angle is 0°, and x & y are same (except magnitude or length))
- Cosine similarity = 0 (angle is 90°, and x & y do not share any terms (words))

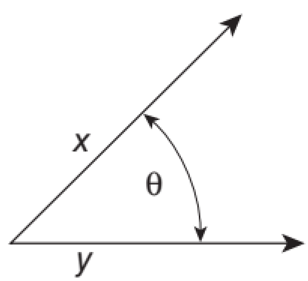


Figure 2.16. Geometric illustration of the cosine measure.

Measures of Similarity and Dissimilarity

Examples of proximity measures

cosine similarity (Document similarity)

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}}{\|\mathbf{x}\|} \cdot \frac{\mathbf{y}}{\|\mathbf{y}\|} = \mathbf{x}' \cdot \mathbf{y}', \tag{2.8}$$

$$\mathbf{x}' = \mathbf{x}/\|\mathbf{x}\|$$

$$\mathbf{y}' = \mathbf{y}/\|\mathbf{y}\|.$$

Note:-

Dividing x and y by their lengths **normalizes** them to have a length of 1 (means magnitude is not considered)

Measures of Similarity and Dissimilarity

Examples of proximity measures

Extended Jaccard Coefficient (Tanimoto Coefficient)

$$EJ(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - \mathbf{x} \cdot \mathbf{y}}. \tag{2.9}$$

Measures of Similarity and Dissimilarity

Examples of proximity measures

Pearson's correlation

$$\operatorname{corr}(\mathbf{x},\mathbf{y}) = \frac{\operatorname{covariance}(\mathbf{x},\mathbf{y})}{\operatorname{standard\_deviation}(\mathbf{x}) * \operatorname{standard\_deviation}(\mathbf{y})} = \frac{s_{xy}}{s_x\ s_y}, \quad (2.10)$$

$$\operatorname{covariance}(\mathbf{x},\mathbf{y}) = s_{xy} = \frac{1}{n-1}\sum_{k=1}^{n}(x_k-\overline{x})(y_k-\overline{y})$$

$$\operatorname{standard\_deviation}(\mathbf{x}) = s_x = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k-\overline{x})^2}$$

$$\operatorname{standard\_deviation}(\mathbf{y}) = s_y = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(y_k-\overline{y})^2}$$

$$\overline{x} = \frac{1}{n}\sum_{k=1}^{n} x_k \text{ is the mean of } \mathbf{x}$$

$$\overline{y} = \frac{1}{n}\sum_{k=1}^{n} y_k \text{ is the mean of } \mathbf{y}$$

Measures of Similarity and Dissimilarity

Examples of proximity measures

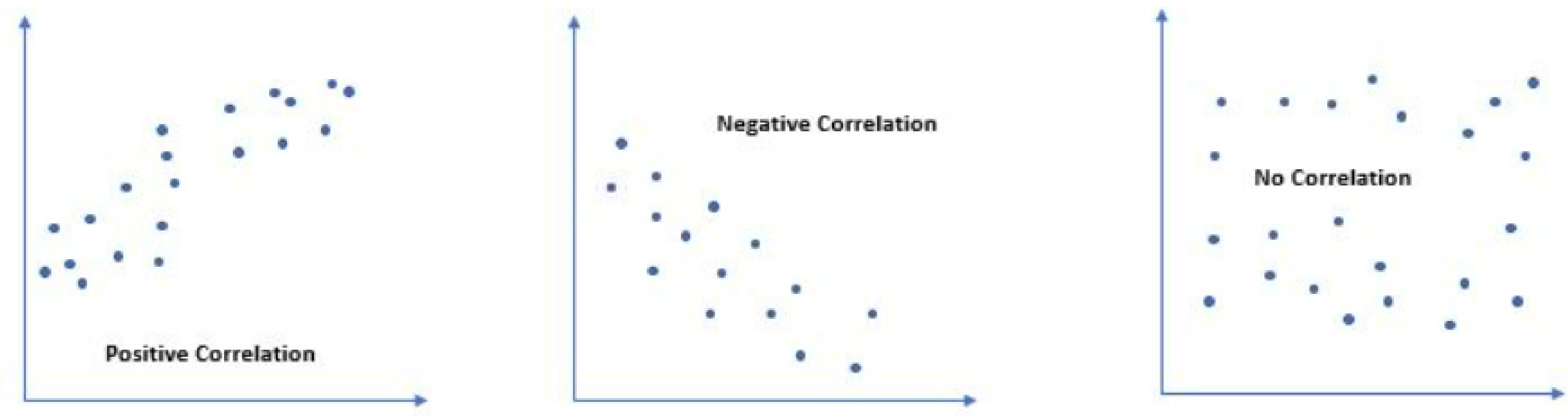
Pearson's correlation

- The more tightly linear two variables X and Y are, the closer Pearson's correlation coefficient(PCC)
  - PCC = -1, if the relationship is negative,
  - PCC=+1, if the relationship is positive.
    - an increase in the value of one variable increases the value of another variable
  - PCC = 0 Perfectly linearly uncorrelated numbers
    - an increase in the value of one decreases the value of another variable.

Measures of Similarity and Dissimilarity

Examples of proximity measures

Pearson's correlation



Measures of Similarity and Dissimilarity

Examples of proximity measures

Pearson’s correlation (scipy.stats.pearsonr() - automatic)

```
>>> df
     wheel-base  length  width  height  curb-weight  ...  horsepower  peak-rpm  city-mpg  highway-mpg  price
0          88.4   141.1   60.3    53.2         1488  ...          48      5100        47           53   5151
1          86.6   144.6   63.9    50.8         1713  ...          58      4800        49           54   6479
2          86.6   144.6   63.9    50.8         1819  ...          76      6000        31           38   6855
3          93.7   150.0   64.0    52.6         1837  ...          60      5500        38           42   5399
4          93.7   150.0   64.0    52.6         1940  ...          76      6000        30           34   6529
..          ...     ...    ...     ...          ...  ...         ...       ...       ...          ...    ...
154       110.0   190.9   70.3    58.7         3750  ...         123      4350        22           25  28248
155       105.8   192.7   71.4    55.7         2844  ...         110      5500        19           25  17710
156       105.8   192.7   71.4    55.9         3086  ...         140      5500        17           20  23875
157       113.0   199.6   69.6    52.8         4066  ...         176      4750        15           19  32250
158       115.6   202.6   71.7    56.3         3770  ...         123      4350        22           25  31600

[159 rows x 14 columns]
>>> from scipy.stats import pearsonr
>>> import pandas as pd
>>> df=pd.read_csv("Car_price_PLR.csv")
>>> x=df['price']
>>> y=df['city-mpg']
>>> pcc,r=pearsonr(x,y)
>>> print(pcc)
-0.6922730619020598
```

Measures of Similarity and Dissimilarity

Examples of proximity measures

Pearson's correlation (manual in python)

```
from scipy.stats import pearsonr
import pandas as pd
import numpy as np
def pearson(x,y):
    xbar=np.mean(x)
    ybar=np.mean(y)
    n=2
    sxy=(1/(n-1))*np.sum((x-xbar)*(y-ybar))
    sx=np.sqrt((1/(n-1))*np.sum((x-xbar)**2))
    sy=np.sqrt((1/(n-1))*np.sum((y-ybar)**2))
    pcc=sxy/(sx*sy)
    return pcc

df=pd.read_csv("Car_price_PLR.csv")
x=df['price']
y=df['city-mpg']
pcc=pearson(x,y)
print(pcc)
```

```
$ python3 pearson_m.py
-0.69227306190206
$ 
```