

UNIT - 2

UNIT - 2

(12 Hours)

Measures of Similarity and Dissimilarity

Similarity and Dissimilarity between Simple Attributes
Dissimilarities between Data Objects
Similarities between Data Objects
Examples of Proximity Measures
Issues in Proximity Calculation
Selecting the Right Proximity Measures.

Summary Statistics

Frequencies and the Mode, Percentiles,

Measures of Location: Mean and Median

Measures of Spread: Range and Variance

Multivariate Summary Statistics, Other Ways to Summarize the Data

Data Cube and OLAP

Data Cube: A multidimensional data model. *Schemas for Multidimensional data model*
: Star, Snowflakes and Fact Constellation schemas. *Dimensions*: The role of Concept Hierarchies. *Measures*: Categorization and Computation. *OLAP Operations*.

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.
- In many cases, the initial data set is not needed once these similarities or dissimilarities have been computed.
- the term **proximity** is used to refer to either similarity or dissimilarity.

Measures of Similarity and Dissimilarity

- The **similarity** between two objects is a numerical measure of the degree to which the two objects are alike.
- *similarities* are **higher** for pairs of objects that are more alike
- Similarities are usually non-negative and are often between 0 (no similarity) and 1 (complete similarity).
- The **dissimilarity** between two objects is a numerical measure of the degree to which the two objects are different.
- *Dissimilarities* are *lower* for more similar pairs of objects. Frequently, the term **distance** is used as a synonym for dissimilarity
- Dissimilarities sometimes fall in the interval $[0, 1]$, but it is also common for them to range from 0 to ∞ .

Measures of Similarity and Dissimilarity

I. Similarity Measures for Binary Data

1. Simple Matching Coefficient (SMC)

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

2. Jaccard Coefficient

$$\text{JC} = \frac{\text{Number of 1-1 matches } (f_{11})}{(\text{Number of attribute}) - (\text{Number of 0-0 matches})}$$

II. Similarity Measures for two Document Vectors

1. Cosine Similarity

$$\text{Cos}(\mathbf{x}, \mathbf{y}) = \frac{(\mathbf{x} * \mathbf{y})}{(\|\mathbf{x}\|) * (\|\mathbf{y}\|)}$$

2. Extended Jaccard Coefficient (Tanimoto Coefficient)

$$\text{EJ}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} * \mathbf{y}}{(\|\mathbf{x}\|)^2 + (\|\mathbf{y}\|)^2 - (\mathbf{x} * \mathbf{y})}$$

Measures of Similarity and Dissimilarity

I. Similarity Measures for Binary Data

1. Simple Matching Coefficient (SMC)

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

- This measure counts both presences and absences equally.
- Let **x** and **y** be two objects that consist of n binary attributes. The comparison of two such objects, i.e., two binary vectors, leads to the following four quantities (frequencies):
 - f_{00} = the number of attributes where **x** is 0 and **y** is 0
 - f_{01} = the number of attributes where **x** is 0 and **y** is 1
 - f_{10} = the number of attributes where **x** is 1 and **y** is 0
 - f_{11} = the number of attributes where **x** is 1 and **y** is 1

Measures of Similarity and Dissimilarity

I. Similarity Measures for Binary Data

2. Jaccard Coefficient

The Jaccard coefficient is frequently used to handle objects consisting of asymmetric binary attributes.

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

Measures of Similarity and Dissimilarity

II. Similarity Measures for two Document Vectors

1. Cosine Similarity

$$\text{Cos}(\mathbf{x}, \mathbf{y}) = \frac{(\mathbf{x} * \mathbf{y})}{(||\mathbf{x}|| * ||\mathbf{y}||)}$$

2. Extended Jaccard Coefficient (Tanimoto Coefficient)

$$\text{EJ}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} * \mathbf{y}}{(||\mathbf{x}||)^2 + (||\mathbf{y}||)^2 - (\mathbf{x} * \mathbf{y})}$$

I. Similarity Measures for Binary Data

Problem 1 : For the following vectors X and Y Compute Simple Matching Coefficient (SMC) and Jaccard Coefficient.

$x = (1, 0, 0, 0, 0, 0, 0, 0, 0, 0)$ and

$y = (0, 0, 0, 0, 0, 0, 1, 0, 0, 1)$

Solution : **Simple Matching Coefficient (SMC)**

$$SMC = \frac{\text{Number of matching attribute values}(f_{00}+f_{11})}{\text{Number of attributes}(f_{00}+f_{01}+f_{10}+f_{11})} = \frac{7}{10}$$

$$SMC = 0.7$$

Jaccard Coefficient

$$JC = \frac{\text{Number of 1-1 matches } (f_{11})}{(\text{Number of attribute}) - (\text{Number of 0-0 matches})} = \frac{0}{10 - 7}$$

$$JC = 0$$

II. Similarity Measures for two Document Vectors

Problem 2 : For the following vectors X and Y Calculate Cosine Similarity and Extended Jaccard Coefficient

$x = (3, 2, 0, 5, 0, 0, 0, 2, 0, 0)$ and

$y = (1, 0, 0, 0, 0, 0, 0, 1, 0, 2)$

Solution :

Formula for Cosine Similarity $\text{Cos}(x, y) = \frac{(x * y)}{(||x||) * (||y||)}$

$$(x*y) = (3*1)+(2*0)+ (0*0)+(5*0) +(0*0)+(0*0) +(0*0)+(2*1) +(0*0)+(0*2)$$

$$(x*y) = 5$$

standard Euclidean norm on \mathbb{R}^n is defined by $||x|| = \sqrt{x \cdot x}$.

$$||x|| = \sqrt{(3*3) + (2*2) + (0*0) + (5*5) + (0*0) + (0*0) + (0*0) + (2*2) + (0*0) + (0*0)}$$

$$||x|| = 6.48$$

$$||y|| = \sqrt{(1*1) + (0*0) + (0*0) + (0*0) + (0*0) + (0*0) + (0*0) + (1*1) + (0*0) + (2*2)}$$

$$||y|| = 2.4$$

$$\text{Cosine Similarity } (x, y) = \frac{5}{(6.48) * (2.4)} = 0.31$$

$$\text{Extended Jaccard Coefficient, EJ(x, y) = } \frac{\mathbf{x * y}}{(\|\mathbf{x}\|)^2 + (\|\mathbf{y}\|)^2 - (\mathbf{x * y})}$$

$$\text{EJ(x, y) = } \frac{5}{(6.48)^2 + (2.4)^2 - (5)}$$

$$\text{EJ(x, y) = 0.1169}$$

Dissimilarity Matrix with Euclidean Distance

Euclidean distance between two points/ Vectors

$$d(A,B) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}$$

Where $A = (x_1, y_1, z_1)$

$B = (x_2, y_2, z_2)$

$$d(A, B, C) = d(A,B) + d(B,C)$$

Problem 3 : Determine Euclidean distance between

$A = (1, 0, 0)$; $B = (1, 4, 5)$ and

$C = (10, 0, 0)$

Solution :

$$\begin{aligned}\text{Formula ED, } d(A,B) &= \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2} \\ &= \sqrt{(1 - 1)^2 + (0 - 4)^2 + (0 - 5)^2} \\ &= \sqrt{0 + 16 + 25} \\ &= \sqrt{41} \\ &= 6.40\end{aligned}$$

$$\begin{aligned}d(B,C) &= \sqrt{(x_2 - x_3)^2 + (y_2 - y_3)^2 + (z_2 - z_3)^2} \\ &= \sqrt{(1 - 10)^2 + (4 - 0)^2 + (5 - 0)^2} \\ &= \sqrt{81 + 16 + 25} \\ &= \sqrt{122} \\ &= 11.04\end{aligned}$$

$$\begin{aligned}d(A,B,C) &= d(A,B) + d(B,C) \\ &= 6.40 + 11.04 \\ &= 17.44\end{aligned}$$

UNIT - 2

(12 Hours)

Measures of Similarity and Dissimilarity

Similarity and Dissimilarity between Simple Attributes
Dissimilarities between Data Objects
Similarities between Data Objects
Examples of Proximity Measures
Issues in Proximity Calculation
Selecting the Right Proximity Measures.

Summary Statistics

Frequencies and the Mode, Percentiles,

Measures of Location: Mean and Median

Measures of Spread: Range and Variance

Multivariate Summary Statistics, Other Ways to Summarize the Data

Data Cube and OLAP

Data Cube: A multidimensional data model. *Schemas for Multidimensional data model*

: Star, Snowflakes and Fact Constellation schemas. *Dimensions*: The role of Concept Hierarchies. *Measures*: Categorization and Computation. *OLAP Operations*.

SUMMARY STATISTICS

Summary Statistics are quantities such as mean, standard deviation that capture various characteristics of potentially large set of values with single numbers or a small set of numbers.

For ex: Average household income etc.

Mode: It is the data entry that occurs with the greatest frequency. Data set can have **one mode**, **more than one mode** or **no mode**.
If no entry is repeated than the data has **NO mode**

Outliers: They are not just the greatest or least values, but values that are very different from the pattern established by the rest of the data

Frequencies

Problem 4: Suppose following is the data of number of Items purchased by customers: 1, 2, 5, 4, 6, 3, 7, 1, 5, 6, 5, 3, 4, 5, 1, 4 than

Find **Absolute Frequency, Relative Frequency and Cumulative Frequency**

Solution:

X	Absolute Frequency $n(x)$	Relative Frequency $p(x)$	Cummulative Frequency $p'(x)$
1	3	$3/16 = 0.1875$	3
2	1	0.0625	$(3 + 1) = 4$
3	2	0.125	$(4 + 2) = 6$
4	3	0.1875	$(6 + 3) = 9$
5	4	0.25	13
6	2	0.125	15
7	1	0.0625	16
	SUM($n(x)$)=, 16	SUM($p(x)$)= 1	

Percentile

Step I: Write values in Ascending order & do numbering.

Step II: Find the p^{th} Percentile using formula
where n is sample size.

$$i = \frac{p}{100} * n$$

Step III: once we get the 'i' value and if 'i' is an integer than find mean of the values at i^{th} and $(i + 1)^{th}$ position. if 'i' is not an integer the value is rounded up to the next integer thus percentile is the corresponding value at that i^{th} position.

Problem 5: Using following set of stock prices(in dollars)

4, 5, 7, 8, 9, 10, 12, 12, 14, 15, 18, 20 Find 10th & 50th Percentile.

Soln: I.

DATA	4	5	7	8	9	10	12	12	14	15	18	20
Position (i)	1	2	3	4	5	6	7	8	9	10	11	12


II. Since $n=12$ and $p=10$, $i = (\frac{10}{100} * 12) = 1.2$ which isn't an integer

III. $\text{ROUNDUP}(1.2) = 2$, So the 10th Percentile is **5**

contd...

To Calculate 50th Percentile

DATA	4	5	7	8	9	10	12	12	14	15	18	20
Position (i)	1	2	3	4	5	6	7	8	9	10	11	12



II. Since $n=12$ and $p=50$, $i = \left(\frac{50}{100} * 12 \right) = 6$ (which is an integer)

So we find mean of values at 6th and 7th position in the dataset and

we get, $\frac{10+12}{2} = 11$

III. So the 50th Percentile is 11

Measures of Location: Mean & Median

Problem 6 : For the given sample data: **1, 2, 2, 3, 4, 5, 60**.

Find Median, Mode, Mean and Outlier

Solution : Median = **3**

Mode = **2**

Mean= $(1+2+2+3+4+5+60) / 7 = 77/7 =$ **11**

Outlier= **60**

Measure of Spread: Range and Variance

Range = Maximum Value – Minimum Value

Variance :

$$\text{Population Variance} = \sigma^2 = V_{\text{population}} = \frac{\sum (x - \mu)^2}{N}$$

$$\text{Sample Variance} = S^2 = V_{\text{sample}} = \frac{\sum (x - \bar{x})^2}{n-1}$$

Co-variance:

$$\text{Co-Variance}(x,y)_{\text{population}} = \frac{\sum (x_i - x_{\text{mean}}) * (y_i - y_{\text{mean}})}{N}$$

$$\text{Co-Variance}(x, y)_{\text{sample}} = \frac{\sum (x_i - x_{\text{mean}}) * (y_i - y_{\text{mean}})}{n-1}$$

Variance Problems

Problem 7: For the following values of a population of X attribute, Find **Variance** and **Range**.

$$X = 1, 2, 3, 4, 5$$

Solution: First find Mean of the values and then find variance

$$\begin{aligned}\mu &= \frac{1+2+3+4+5}{5} = 3 \\ \sigma^2 &= V_{population} = \frac{\sum(x - \mu)^2}{N} \\ &= \frac{(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2}{5} \\ \sigma^2 &= 10/5 = 2\end{aligned}$$

$$\begin{aligned}\text{Range} &= \text{Maximum Value} - \text{Minimum Value} \\ &= 5 - 1\end{aligned}$$

$$\text{Range} = 4$$

Problem 8: For the following sample data points of attributes X and Y, Find **co- variance** and also **correlation** between attributes X and Y.

X	1	3	2	5	8	7	12	2	4
Y	8	6	9	4	3	3	2	7	7

Solution

	x	y	x - Mean(x)	y - Mean(y)	(x - Mean(x)) * (y-Mean(y))
	1	8	-3.89	2.56	-9.9584
	3	6	-1.89	0.56	-1.0584
	2	9	-2.89	3.56	-10.2884
	5	4	0.11	-1.44	-0.1584
	8	3	3.11	-2.44	-7.5884
	7	3	2.11	-2.44	-5.1484
	12	2	7.11	-3.44	-24.4584
	2	7	-2.89	1.56	-4.5084
	4	7	-0.89	1.56	-1.3884
Mean =>	4.89	5.44		SUM =>	-64.5556

Numbers of data points for attributes X and Y are, **n = 9**

Using formula, **Co-variance(x, y)** =
$$\frac{\sum (x_i - x_{mean}) * (y_i - y_{mean})}{(n-1)}$$

$$= \frac{-64.5556}{(9-1)}$$

$$= -8.069$$

Find Standard deviation, σ_x and σ_y

$$\sigma_x = \sqrt{\frac{(-3.89)^2 + (-1.89)^2 + (-2.89)^2 + (0.11)^2 + (3.11)^2 + (2.11)^2 + (7.11)^2 + (-2.89)^2 + (-0.89)^2}{(9-1)}}$$

$$= \sqrt{(100.89 / 8)} = \underline{3.551}$$

$$\sigma_y = \sqrt{\frac{(2.56)^2 + (0.56)^2 + (3.56)^2 + (-1.44)^2 + (-2.44)^2 + (-2.44)^2 + (-3.44)^2 + (1.56)^2 + (1.56)^2}{(9-1)}}$$

$$= \underline{2.506}$$

Calculate Correlation Coefficient, $r_{(x,y)}$ =
$$\frac{\text{Co-Variance}(x,y)}{\sigma_x * \sigma_y} = \frac{-8.069}{3.551 * 2.506}$$

$$= -0.907$$

Since $r_{(x,y)}$ is negative, Attributes X and Y are negatively correlated.

Issues in Proximity Calculation

Important issues related to proximity measures:

(1) how to handle the case in which attributes have *different scales and/or are Correlated*

(2) how to calculate proximity between objects that are composed of *different types of attributes*, e.g., quantitative and qualitative

(3) and how to handle proximity calculation when attributes have *different weights*; i.e., when not all attributes contribute equally to the proximity of objects.

Issues in Proximity Calculation

(1) how to handle the case in which attributes have *different scales and/or are Correlated*

- how to handle the situation when attributes do not have the same range of values.
- how to compute distance when there is correlation between some of the attributes
- A generalization of Euclidean distance, the **Mahalanobis distance**, is useful when attributes are correlated, have different ranges of values (different variances), and the distribution of the data is approximately Gaussian (normal).

Issues in Proximity Calculation

(2) how to calculate proximity between objects that are composed of *different types of attributes*, e.g., quantitative and qualitative

- A general approach is needed when the attributes are of different types.
- One straightforward approach is to compute the similarity between each attribute separately, and then combine these similarities using a method that results in a similarity between 0 and 1.
- The overall similarity is defined as the average of all the individual attribute similarities.
- this approach does not work well if some of the attributes are asymmetric attributes.
- The easiest way to fix this problem is to omit asymmetric attributes from the similarity calculation when their values are 0 for both of the objects whose similarity is being computed.

Issues in Proximity Calculation

(3) and how to handle proximity calculation when attributes have *different weights*; i.e., when not all attributes contribute equally to the proximity of objects.

- All attributes were treated equally when computing proximity.
- This is not desirable when some attributes are more important to the definition of proximity than others.
- To address these situations the formulas for proximity can be modified by weighting the contribution of each attribute.