INTRODUCTION TO DATA ANALYTICS

Class # 24

Clustering Techniques: Similarity Measures

Dr. Sreeja S R

Assistant Professor

Indian Institute of Information Technology

IIIT Sri City

Topics to be covered…

- Introduction to clustering
- Similarity and dissimilarity measures
- Clustering techniques
    - Partitioning algorithms
    - Hierarchical algorithms
    - Density-based algorithm

Introduction to Clustering

- Classification consists of assigning a class label to a set of unclassified cases.
- **Supervised Classification**
    - The set of possible classes is known in advance.
- **Unsupervised Classification**
    - Set of possible classes is not known. After classification we can try to assign a name to that class.
        - Unsupervised classification is called **clustering**.

Introduction to Clustering

- Clustering is somewhat related to classification in the sense that in both cases data are grouped.

- However, there is a major difference between these two techniques.

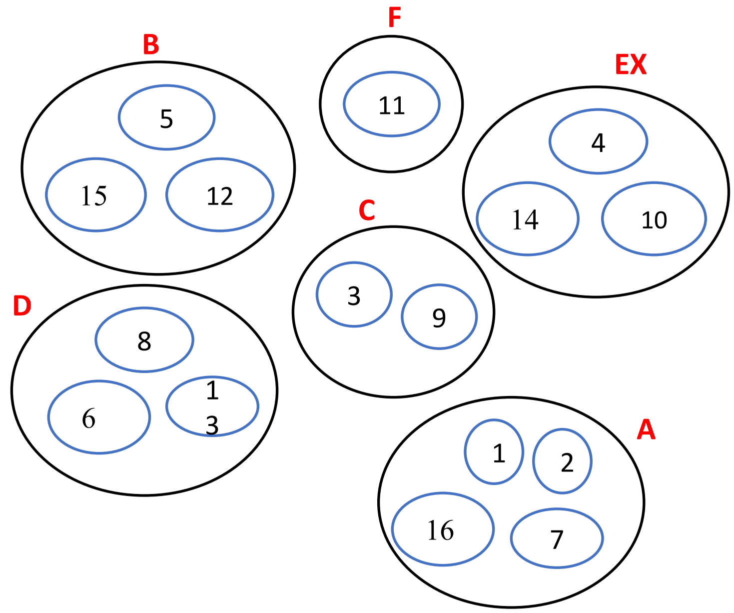
- In order to understand the difference between the two, consider a sample dataset containing marks obtained by a set of students and corresponding grades as shown in Table 24.1.

Introduction to Clustering

Table 24.1: Tabulation of Marks

Roll No	Mark	Grade
1	80	A
2	70	A
3	55	C
4	91	EX
5	65	B
6	35	D
7	76	A
8	40	D
9	50	C
10	85	EX
11	25	F
12	60	B
13	45	D
14	95	EX
15	63	B
16	88	A

Figure 24.1: Group representation of dataset in Table 24.1



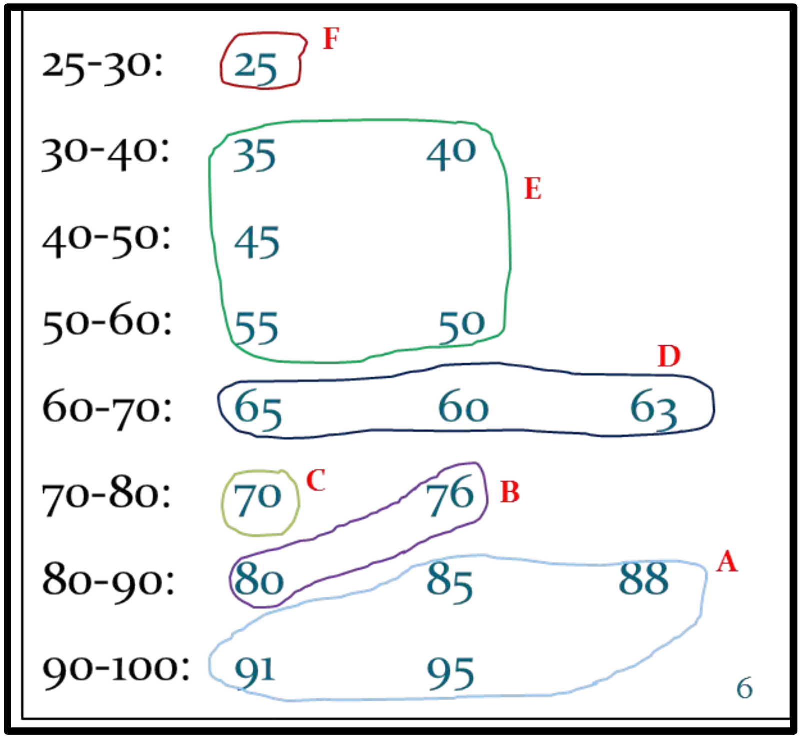
Introduction to Clustering

- It is evident that there is a simple mapping between Table 24.1 and Fig 24.1.
- The fact is that groups in Fig 24.1 are already predefined in Table 24.1. This is similar to classification, where we have given a dataset where groups of data are predefined.
- Consider another situation, where 'Grade' is not known, but we have to make a grouping.
- Put all the marks into a group if any other mark in that group does not exceed by 5 or more.
- This is similar to "Relative grading" concept and grade may range from *A* to *Z*.

Introduction to Clustering

- Figure 24.2 shows another grouping by means of another simple mapping, but the difference is this mapping does not based on predefined classes.

- In other words, this grouping is accomplished by finding similarities between data according to characteristics found in the actual data.
- Such a group making is called clustering.



Introduction to Clustering

Example 24.1 : The task of clustering

In order to elaborate the clustering task, consider the following dataset.

Table 24.2: Life Insurance database

Marital Status	Age	Income	Education	Number of children
Single	35	25000	Under Graduate	3
Married	25	15000	Graduate	1
Single	40	20000	Under Graduate	0
Divorced	20	30000	Post-Graduate	0
Divorced	25	20000	Under Graduate	3
Married	60	70000	Graduate	0
Married	30	90000	Post-Graduate	0
Married	45	60000	Graduate	5
Divorced	50	80000	Under Graduate	2

With certain similarity or likeliness defined, we can classify the records to one or group of more attributes (and thus mapping being non-trivial).

Introduction to Clustering

- Clustering has been used in many application domains:
  - Image analysis
  - Document retrieval
  - Machine learning, etc.

- When clustering is applied to real-world database, many problems may arise.

1. The (best) number of cluster is not known.
   - There is no correct answer to a clustering problem.
   - In fact, many answers may be found.
   - The exact number of cluster required is not easy to determine.

Introduction to Clustering

2. There may not be any a priori knowledge concerning the clusters.

- This is an issue that what data should be used for clustering.
- Unlike classification, in clustering, we have no supervisory learning to aid the process.
- Clustering can be viewed as similar to unsupervised learning.

3. Interpreting the semantic meaning of each cluster may be difficult.

- With classification, the labeling of classes is known ahead of time. In contrast, with clustering, this may not be the case.
- Thus, when the clustering process is finished yielding a set of clusters, the exact meaning of each cluster may not be obvious.

Definition of Clustering Problem

Definition 24.1: Clustering

Given a database $D = \{t_1, t_2, \ldots\ldots, t_n\}$ of n tuples, the clustering problem is to define a mapping $f : \mathrm{D} \rightarrow C$, where each $t_i \in D$ is assigned to one cluster $c_i \in C$. Here, $C = \{c_1, c_2, \ldots\ldots, c_k\}$ denotes a set of clusters.

- Solution to a clustering problem is devising a mapping formulation.
- The formulation behind such a mapping is to establish that a tuple within one cluster is more like tuples within that cluster and not similar to tuples outside it.

Definition of Clustering Problem

- Hence, mapping function f in Definition 24.1 may be explicitly stated as

$$f : \mathrm{D} \rightarrow \{c_1, c_2, \ldots\ldots, c_k\}$$

where i) each $t_i \in D$ is assigned to one cluster $c_i \in C$.

ii) for each cluster $c_i \in C$, and for all $t_{ip}, t_{iq} \in c_i$ and there exist $t_j \notin c_i$ such that

similarity $(t_{ip}, t_{iq}) >$ similarity (t_{ip}, t_j) AND similarity (t_{iq}, t_j)

- In the field of cluster analysis, this **similarity** plays an important part.
- Now, we shall learn how similarity (this is also alternatively judged as "dissimilarity") between any two data can be measured.

What is Similarity?

The quality or state of being similar; likeness; resemblance; as, a similarity of features.
Webster's Dictionary

Similarity is hard to define, but… "*We know it when we see it*"

The real meaning of similarity is a philosophical question. We will take a more pragmatic approach.

Similarity and Dissimilarity Measures

- In clustering techniques, similarity (or dissimilarity) is an important measurement.
- Informally, similarity between two objects (e.g., two images, two documents, two records, etc.) is a numerical measure of the degree to which two objects are alike.
- The dissimilarity on the other hand, is another alternative (or opposite) measure of the degree to which two objects are different.
- Both similarity and dissimilarity also termed as proximity.
- Usually, similarity and dissimilarity are non-negative numbers and may range from zero (highly dissimilar (no similar)) to some finite/infinite value (highly similar (no dissimilar)).

Note:

- Frequently, the term distance is used as a synonym for dissimilarity
- In fact, it is used to refer as a special case of dissimilarity.

Proximity Measures: Single-Attribute

- Consider an object, which is defined by a single attribute *A* (e.g., length) and the attribute *A* has *n*-distinct values $a_1, a_2, \ldots\ldots, a_n$.
- A data structure called "Dissimilarity matrix" is used to store a collection of proximities that are available for all pair of *n* attribute values.
    - In other words, the Dissimilarity matrix for an attribute *A* with *n* values is represented by an $n \times n$ matrix as shown below.

$$\begin{bmatrix} 0 & & & \\ p_{(2,1)} & 0 & & \\ p_{(3,1)} & p_{(3,2)} & 0 & \\ \vdots & \vdots & \vdots & \\ p_{(n,1)} & p_{(n,2)} & \ldots\ldots & 0 \end{bmatrix}_{n \times n}$$

    - Here, $p_{(i,j)}$ denotes the proximity measure between two objects with attribute values a_i and a_j.
- **Note:** The proximity measure is symmetric, that is, $p_{(i,j)} = p_{(j,i)}$

Proximity Calculation

- Proximity calculation to compute $p_{(i,j)}$ is different for different types of attributes according to NOIR topology.

Proximity calculation for Nominal attributes:

- For example, binary attribute, Gender = {Male, female} where Male is equivalent to binary 1 and female is equivalent to binary 0.
- Similarity value is 1 if the two objects contains the same attribute value, while similarity value is 0 implies objects are not at all similar.

Object	Gender
Ram	Male
Sita	Female
Laxman	Male

- Here, Similarity value let it be denoted by p, among different objects are as follows.

$$p(Ram, sita) = 0$$
$$p(Ram, Laxman) = 1$$

Note : In this case, if q denotes the dissimilarity between two objects i and j with single binary attributes, then $q_{(i,j)} = 1 - p_{(i,j)}$

Proximity Calculation

- Now, let us focus on how to calculate proximity measures between objects which are defined by two or more binary attributes.
- Suppose, the number of attributes be b. We can define the contingency table summarizing the different matches and mismatches between any two objects x and y, which are as follows.

Table 24.3: Contingency table with binary attributes

		1	0
Object x	1		
	0		

Here, f_{11}= the number of attributes where x=1 and y=1.

f_{10}= the number of attributes where x=1 and y=0.

f_{01}= the number of attributes where x=0 and y=1.

f_{00}= the number of attributes where x=0 and y=0.

Note : $f_{00} + f_{01} + f_{10} + f_{11} = b$, the total number of binary attributes.
Now, two cases may arise: symmetric and asymmetric binary attributes.

Similarity Measure with Symmetric Binary

- To measure the similarity between two objects defined by symmetric binary attributes using a measure called symmetric binary coefficient and denoted as $\mathcal{S}$ and defined below

$$\mathcal{S} = \frac{Number\ of\ matching\ attribute\ values}{Total\ number\ of\ attributes}$$

or

$$\mathcal{S} = \frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}}$$

The dissimilarity measure, likewise can be denoted as $\mathcal{D}$ and defined as

$$\mathcal{D} = \frac{Number\ of\ mismatched\ attribute\ values}{Total\ number\ of\ attributes}$$

or

$$\mathcal{D} = \frac{f_{01} + f_{10}}{f_{00} + f_{01} + f_{10} + f_{11}}$$

Note that, $\mathcal{D} = 1 - \mathcal{S}$

Similarity Measure with Symmetric Binary

Example 24.2: Proximity measures with symmetric binary attributes

Consider the following two dataset, where objects are defined with symmetric binary attributes.

Gender = {M, F}, Food = {V, N}, Caste = {H, M}, Education = {L, I}, Hobby = {T, C}, Job = {Y, N}

Object	Gender	Food	Caste	Education	Hobby	Job
Hari	M	V	M	L	C	N
Ram	M	N	M	I	T	N
Tomi	F	N	H	L	C	Y

$$\mathcal{S}(\text{Hari, Ram}) = \frac{1+2}{1+2+1+2} = 0.5$$

Proximity Measure with Asymmetric Binary

- Such a similarity measure between two objects defined by asymmetric binary attributes is done by Jaccard Coefficient and which is often symbolized by $\mathcal{J}$ is given by the following equation

$$\mathcal{J} = \frac{Number\ of\ matching\ presence}{Number\ of\ attributes\ not\ involved\ in\ 00\ matching}$$

or

$$\mathcal{J} = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

Proximity Measure with Asymmetric Binary

Example 24.3: Jaccard Coefficient

Consider the following two dataset.

Gender = {M, F}, Food = {V, N}, Caste = {H, M},Education = {L, I},
Hobby = {T, C}, Job = {Y, N}

Calculate the Jaccard coefficient between Ram and Hari assuming that all binary attributes are asymmetric and for each pair values for an attribute, first one is more frequent than the second.

Object	Gender	Food	Caste	Education	Hobby	Job
Hari	M	V	M	L	C	N
Ram	M	N	M	I	T	N
Tomi	F	N	H	L	C	Y

$$\mathcal{J}(\text{Hari, Ram}) = \frac{1}{2+1+1} = 0.25$$

Note: $\mathcal{J}(\text{Ram, Tomi}) = 0$ and $\mathcal{J}(\text{Hari, Ram}) = \mathcal{J}(\text{Ram, Hari})$, etc.

Example 24.4:

Consider the following two dataset.

Gender = {M, F}, Food = {V, N}, Caste = {H, M},Education = {L, I},
Hobby = {T, C}, Job = {Y, N}

Object	Gender	Food	Caste	Education	Hobby	Job
Hari	M	V	M	L	C	N
Ram	M	N	M	I	T	N
Tomi	F	N	H	L	C	Y

? How you can calculate similarity if Gender, Hobby and Job are symmetric binary attributes and Food, Caste, Education are asymmetric binary attributes?

Obtain the similarity matrix with Jaccard coefficient of objects for the above, e.g.

$$\mathcal{J} = \begin{array}{c c} & \begin{array}{ccc} H & R & T \end{array} \\ \begin{array}{c} H \\ R \\ T \end{array} & \begin{bmatrix} 0 & 0 & 0 \\ \mathcal{J}(R,H) & 0 & 0 \\ \mathcal{J}(T,H) & \mathcal{J}(T,R) & 0 \end{bmatrix} \end{array}$$

Proximity Measure with Categorical Attribute

- Binary attribute is a special kind of nominal attribute where the attribute has values with two states only.
- On the other hand, categorical attribute is another kind of nominal attribute where it has values with three or more states (e.g. color = {Red, Green, Blue}).
- If $s(x, y)$ denotes the similarity between two objects x and y, then

$$s(x,y) = \frac{Number\ of\ matches}{Total\ number\ of\ attributes}$$

and the dissimilarity $d(x, y)$ is

$$d(x,y) = \frac{Number\ of\ mismatches}{Total\ number\ of\ attributes}$$

- If m = number of matches and a = number of categorical attributes with which objects are defined as

$$s(x,y) = \frac{m}{a} \quad \text{and} \quad d(x,y) = \frac{a-m}{a}$$

Proximity Measure with Categorical Attribute

Example 24.4:

Object	Color	Position	Distance
1	R	L	L
2	B	C	M
3	G	R	M
4	R	L	H

The similarity matrix considering only color attribute is shown below

$$s = \begin{array}{c c} & \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \end{array} & \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

Dissimilarity matrix, $d =$?

Obtain the dissimilarity matrix considering both the categorical attributes (i.e. color and position).

Proximity Measure with Ordinal Attribute

- Ordinal attribute is a special kind of categorical attribute, where the values of attribute follows a sequence (ordering) e.g. Grade = {Ex, A, B, C} where Ex > A >B >C.
- Suppose, A is an attribute of type ordinal and the set of values of $A = \{a_1, a_2, \ldots\ldots, a_n\}$. Let n values of A are ordered in ascending order as $a_1 < a_2 < .. < a_n$. Let *i-th* attribute value a_i be ranked as *i, i=1,2,..n.*
- The normalized value of a_i can be expressed as

$$\hat{a}_i = \frac{i-1}{n-1}$$

- Thus, normalized values lie in the range $[0..1]$.
- As a_i is a numerical value, the similarity measure, then can be calculated using any similarity measurement method for numerical attribute.
- For example, the similarity measure between two objects $x \text{ and } y$ with attribute values a_i and a_j, then can be expressed as

$$s(x, y) = \sqrt{(\hat{a}_i - \hat{a}_j)^2}$$

where $\hat{a}_i$ and $\hat{a}_j$ are the normalized values of $\hat{a}_i$ and $\hat{a}_j$, respectively.

Proximity Measure with Ordinal Attribute

Example 24.5:

Consider the following set of records, where each record is defined by two ordinal attributes size={S, M, L} and Quality = {Ex, A, B, C} such that $S<M<L$ and $Ex>A>B>C$.

Object	Size	Quality
A	S (0.0)	A (0.66)
B	L (1.0)	Ex (1.0)
C	L (1.0)	C (0.0)
D	M (0.5)	B (0.33)

- Normalized values are shown in brackets.
- Their similarity measures are shown in the similarity matrix below.

$$
\begin{array}{c c}
 & \begin{array}{cccc} A & B & C & D \end{array} \\
\begin{array}{c} A \\ B \\ C \\ D \end{array} & \begin{bmatrix} 0 & 0 & 0 & 0 \\ & 0 & 0 & 0 \\ ? & & 0 & 0 \\ & & & 0 \end{bmatrix}
\end{array}
$$

Find the dissimilarity matrix, when each object is defined by only one ordinal attribute say size (or quality).

Proximity Measure with Interval Scale

- The measure called distance is usually referred to estimate the similarity between two objects defined with interval-scaled attributes.
- We first present a generic formula to express distance d between two objects $x \text{ and } y$ in n-dimensional space. Suppose, $x_i \text{ and } y_i$ denote the values of i^{th} attribute of the objects $x \text{ and } y$ respectively.

$$d(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^r \right)^{\frac{1}{r}}$$

- Here, r is any integer value.
- This distance metric most popularly known as Minkowski metric.
- This distance measure follows some well-known properties. These are mentioned in the next slide.

Proximity Measure with Interval Scale

Properties of Minkowski metrics:

1. Non-negativity:
   a. $d(x, y) \geq 0\ for\ all\ x\ and\ y$
   b. $d(x, y) = 0$ only if $x = y$. This is also called identity condition.

2. Symmetry:
   $d(x, y) = d(y, x)\ \ for\ all\ x\ and\ y$
   This condition ensures that the order in which objects are considered is not important.

3. Transitivity:
   $d(x, z) \leq d(x, y) + d(y, z)\ for\ all\ x\ , y\ and\ z.$
   - This condition has the interpretation that the least distance $d(x, z)$ between objects $x\ and\ z$ is always less than or equal to the sum of the distance between the objects $x\ and\ y$, and between $y\ and\ z$.
   - This property is also termed as Triangle Inequality.

Proximity Measure with Interval Scale

Depending on the value of r, the distance measure is renamed accordingly.

1. **Manhattan distance (L_1 Norm: $r = 1$)**

The Manhattan distance is expressed as

$$d = \sum_{i=1}^{n} |x_i - y_i|$$

where $|\ldots|$ denotes the absolute value. This metric is also alternatively termed as **Taxicals metric, city-block metric**.

Example: x = [7, 3, 5] and y = [3, 2, 6].

The Manhattan distance is $|7 - 3| + |3 - 2| + |5 - 6| = 6$.

- As a special instance of Manhattan distance, when attribute values $\in [0, 1]$ is called Hamming distance.
- Alternatively, Hamming distance is the number of bits that are different between two objects that have only binary values (i.e. between two binary vectors).

Proximity Measure with Interval Scale

2. Euclidean Distance (L_2 Norm: $r = 2$)

This metric is same as Euclidean distance between any two points $x\ and\ y\ in\ \mathcal{R}^n$.

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Example: x = [7, 3, 5] and y = [3, 2, 6].

The Euclidean distance between $x\ and\ y$ is

$$d(x, y) = \sqrt{(7-3)^2 + (3-2)^2 + (5-6)^2} = \sqrt{18} \approx 2.426$$

Proximity Measure with Interval Scale

3. Chebychev Distance ($L_\propto$ Norm: $r \in \mathcal{R}$)

This metric is defined as

$$d(x, y) = \max_{\forall i}\{|x_i - y_i|\}$$

- We may clearly note the difference between Chebychev metric and Manhattan distance. That is, instead of summing up the absolute difference (in Manhattan distance), we simply take the maximum of the absolute differences (in Chebychev distance). Hence, $\mathbf{L_\propto < L_1}$

Example: x = [7, 3, 5] and y = [3, 2, 6].

The Manhattan distance = $|7 - 3| + |3 - 2| + |5 - 6| = 6.$

The chebychev distance = $\text{Max}\ \{|7 - 3|, |3 - 2|, |5 - 6|\} = 4.$

Proximity Measure with Interval Scale

4. Other metrics:

a. Canberra metric:

$$d(x, y) = \sum_{i=1}^{n} \frac{|x_i - y_i|}{(|x_i| + |y_i|)^q}$$

- where q is a real number. Usually $q = 1$, because numerator of the ratio is always $\leq$ denominator, the ratio ≤ 1, that is, the sum is always bounded and small.
- If $q \neq 1$, it is called Fractional Canberra metric.
- If $q > 1$, the oppositive relationship holds.

b. Hellinger metric:

$$d(x, y) = \sum_{i=1}^{n} (\sqrt{x_i} - \sqrt{y_i})^2$$

This metric is then used as either squared or transformed into an acceptable range [-1, +1] using the following transformations.

$$i. \quad d(x, y) = (1 - r(x, y))/2$$

$$ii. \quad d(x, y) = 1 - r(x, y)$$

Where $r(x, y)$ is correlation coefficient between x and y.

Note: Dissimilarity measurement is not relevant with distance measurement.

Proximity Measure for Ratio-Scale

The proximity between the objects with ratio-scaled variable can be carried with the following steps:

1. Apply the appropriate transformation to the data to bring it into a linear scale. (e.g. logarithmic transformation to data of the form $X = Ae^B$.

2. The transformed values can be treated as interval-scaled values. Any distance measure discussed for interval-scaled variable can be applied to measure the similarity.

Note:

There are two concerns on proximity measures:

- Normalization of the measured values.
- Intra-transformation from similarity to dissimilarity measure and vice-versa.

Proximity Measure for Ratio-Scale

Normalization:

- A major problem when using the similarity (or dissimilarity) measures (such as Euclidean distance) is that the large values frequently swamp the small ones.
- For example, consider the following data.

Make	Cost 1	Cost 2	Cost 3
X	2,00,000	70	10
Y	2,50,000	100	5

Here, the contribution of Cost 2 and Cost 3 is insignificant compared to Cost 1 so far the Euclidean distance is concerned.

- This problem can be avoided if we consider the normalized values of all numerical attributes.

- Another normalization may be to take the estimated values in a normalized range say [0, 1]. Note that, if a measure varies in the range, then it can be normalized as

$$s' = \frac{1}{1+s} \text{ where } s \in [0..\infty]$$

Proximity Measure for Ratio-Scale

Intra-transformation:

- Transforming similarities to dissimilarities and vice-versa is also relatively straightforward.
- If the similarity (or dissimilarity) falls in the interval [0..1], the dissimilarity (or similarity) can be obtained as

$$d = 1 - s$$

or

$$s = 1 - d$$

- Another approach is to define similarity as the negative of dissimilarity (or vice-versa).

Proximity Measure with Mixed Attributes

- The previous metrics on similarity measures assume that all the attributes were of the same type. Thus, 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 attribute using a method that results in a similarity between 0 and 1.
- Typically, the overall similarity is defined as the average of all the individual attribute similarities.
- See the algorithm in the next slide for doing this.

Similarity Measure with Mixed Attributes

Example 24.6:

Consider the following set of objects. Obtain the similarity matrix.

Object	A (Binary)	B (Categorical)	C (Ordinal)	D (Numeric)	E (Numeric)
1	Y	R	X	475	10^{8}
2	N	R	A	10	10^{-2}
3	N	B	C	1000	10^{5}
4	Y	G	B	500	10^{3}
5	Y	B	A	80	1

[For C: X>A>B>C]

$$
\begin{array}{c c}
 & \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\
\begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} & \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ ? & 0 & 0 & 0 & 0 \\ ? & ? & 0 & 0 & 0 \\ ? & ? & ? & 0 & 0 \\ ? & ? & ? & ? & 0 \end{bmatrix}
\end{array}
$$

?

How cosine similarity can be applied to this?

Non-Metric similarity

- In many applications (such as information retrieval) objects are complex and contains a large number of symbolic entities (such as keywords, phrases, etc.).
- To measure the distance between complex objects, it is often desirable to introduce a non-metric similarity function.
- Here, we discuss few such non-metric similarity measurements.

Cosine similarity

Suppose, x and y denote two vectors representing two complex objects. The cosine similarity denoted as $\cos(x, y)$ and defined as

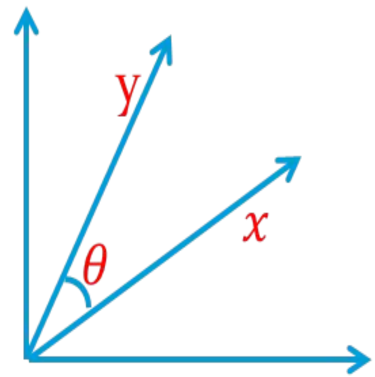
$$\cos(x, y) = \frac{x \cdot y}{\|x\| \cdot \|y\|}$$

- where $x \cdot y$ denotes the vector dot product, namely $x \cdot y = \sum_{i=1}^{n} x_i \cdot y_i$ such that $x = [x_1, x_2, .., x_n]$ and $y = [y_1, y_2, .., y_n]$.
- $\|x\|$ and $\|y\|$ denote the Euclidean norms of vector x and y, respectively (essentially the length of vectors x and y), that is

$$\|x\| = \sqrt{{x_1}^2 + {x_2}^2 + .. + {x_n}^2} \text{ and } \|y\| = \sqrt{{y_1}^2 + {y_2}^2 + .. + {y_n}^2}$$

Cosine Similarity

- In fact, cosine similarity essentially is a measure of the (cosine of the) angle between x and y.
- Thus if the cosine similarity is 1, then the angle between x and y is 0° and in this case, x and y are the same except for magnitude.
- On the other hand, if cosine similarity is 0, then the angle between x and y is 90° and they do not share any terms.
- Considering, this cosine similarity can be written equivalently



$$\cos(x, y) = \frac{x \cdot y}{\|x\| \cdot \|y\|} = \frac{x}{\|x\|} \cdot \frac{y}{\|y\|} = \hat{x} \cdot \hat{y}$$

  where $\hat{x} = \frac{x}{\|x\|}$ and $\hat{y} = \frac{y}{\|y\|}$. This means that cosine similarity does not take the magnitude of the two vectors into account, when computing similarity.

- It is thus, one way normalized measurement.

Non-Metric Similarity

Example 24.7: Cosine Similarity

Suppose, we are given two documents with count of 10 words in each are shown in the form of vectors x and y as below.

$$x = [3, 2, 0, 5, 0, 0, 0, 2, 0, 0] \text{ and } y = [1, 0, 0, 0, 0, 0, 0, 1, 0, 2]$$

Thus, $x \cdot y = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2$
$= 5$

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

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

$$\therefore \quad \cos(x, y) = 0.31$$

Extended Jaccard Coefficient

The extended Jaccard coefficient is denoted as EJ and defined as

$$EJ = \frac{x \cdot y}{\|\mathrm{x}\|^2 \cdot \|y\|^2 - x \cdot y}$$

- This is also alternatively termed as Tanimoto coefficient and can be used to measure like document similarity.

Compute Extended Jaccard coefficient (EJ) for the above example 24.7.

Pearson's Correlation

- The correlation between two objects x and y gives a measure of the linear relationship between the attributes of the objects.
- More precisely, Pearson's correlation coefficient between two objects x and y is defined in the following.

$$P(x, y) = \frac{S_{xy}}{S_x \cdot S_y}$$

where $S_{xy} = covariance\ (x, y) = \frac{1}{n-1} \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

$S_x = Standard\ deviation\ (x) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

$S_y = Standard\ deviation\ (y) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$

$\bar{x} = mean\ (x) = \frac{1}{n}\sum_{i=1}^{n} x_i$

$\bar{y} = mean\ (y) = \frac{1}{n}\sum_{i=1}^{n} y_i$

and n is the number of attributes in x and y.

Note 1: Correlation is always in the range of -1 to 1. A correlation of 1(-1) means that x and y have a perfect positive (negative) linear relationship, that is, $x_i = a \cdot y_i + b$ for some a and b.

Example 24.8: Pearson's correlation

Calculate the Pearson's correlation of the two vectors x and y as given below.

$$x = [3, 6, 0, 3, 6]$$
$$y = [1, 2, 0, 1, 2]$$

Note: Vector components can be negative values as well.

Note:

If the correlation is 0, then there is no linear relationship between the attribute of the object.

Example 24.9: Non-linear correlation

Verify that there is no linear relationship among attributes in the objects x and y given below.

$$x = [-3, -2, -1, 0, 1, 2, 3]$$
$$y = [9, 4, 1, 0, 1, 4, 9]$$

$P(x, y) = 0$, and also note $x_i = {y_i}^2$ for all attributes here.

Any question?