Measures of Similarity and Dissimilarity

Lecture-4
22 Jul 2020

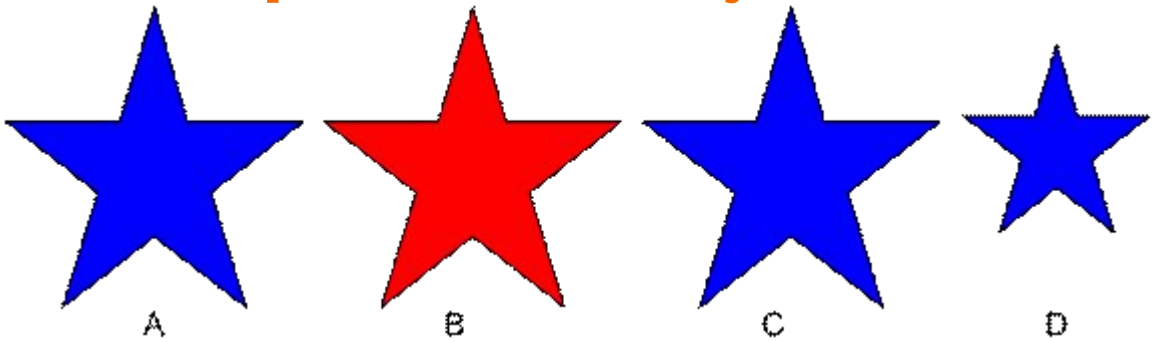
Overview

- Concept of similarity and dissimilarity
- Data matrix and Dissimilarity matrix
- Concept of proximity:
  - Nominal attribute
  - Binary attribute
  - Numeric Data
    - Minkowski Distance
    - Euclidean distance
    - Manhattan distance
  - Ordinal attribute
  - Mixed types
- Cosine similarity

Today

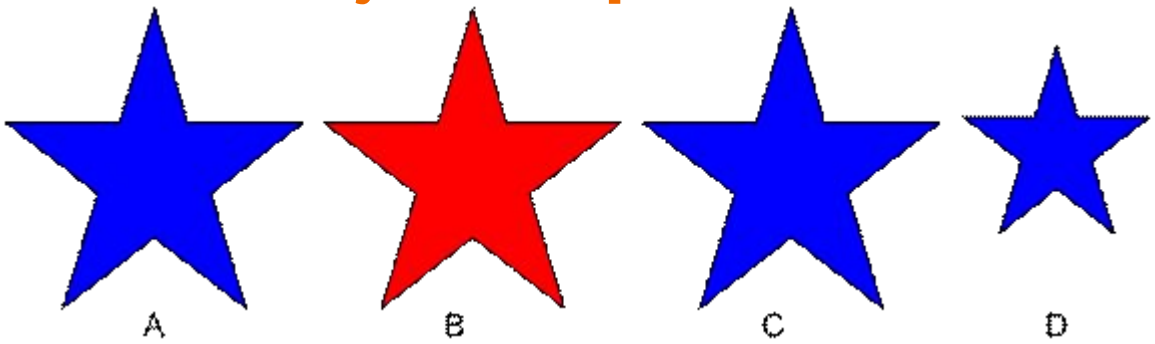
- **Concept of similarity and dissimilarity**
- **Data matrix and Dissimilarity matrix**
- **Concept of proximity:**
  - **Nominal attribute**
  - Binary attribute
  - Numeric Data
    - Minkowski Distance
    - Euclidean distance
    - Manhattan distance
  - Ordinal attribute
  - Mixed types
- Cosine similarity

Concept of similarity



- Are the stars in figure similar?
- How will you compare them?

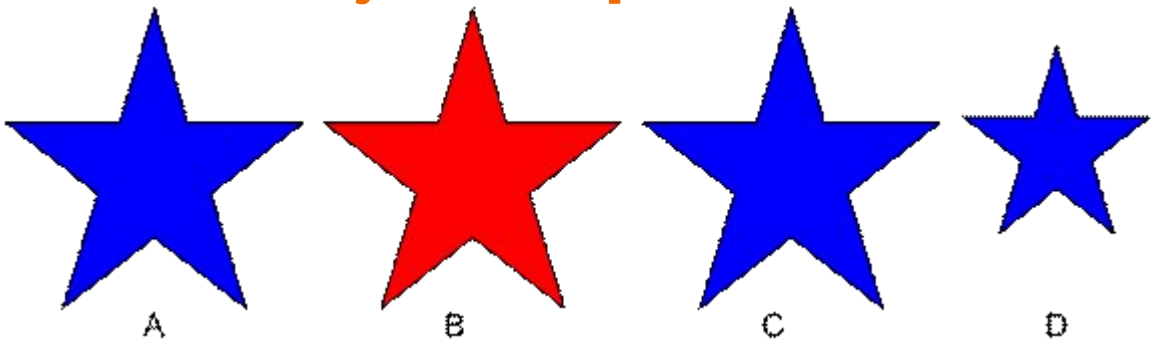
Similarity Concept



- Are the stars in figure similar?
- How will you compare them?

- A and B are similar
- A, B, C have same size
- A, B, and D have same color
- Any more?

Similarity Concept



- A and B are similar
- A, B, C have same size
- A, B, and D have same color
- Any more?

- Are the stars in figure similar?
- How will you compare them?

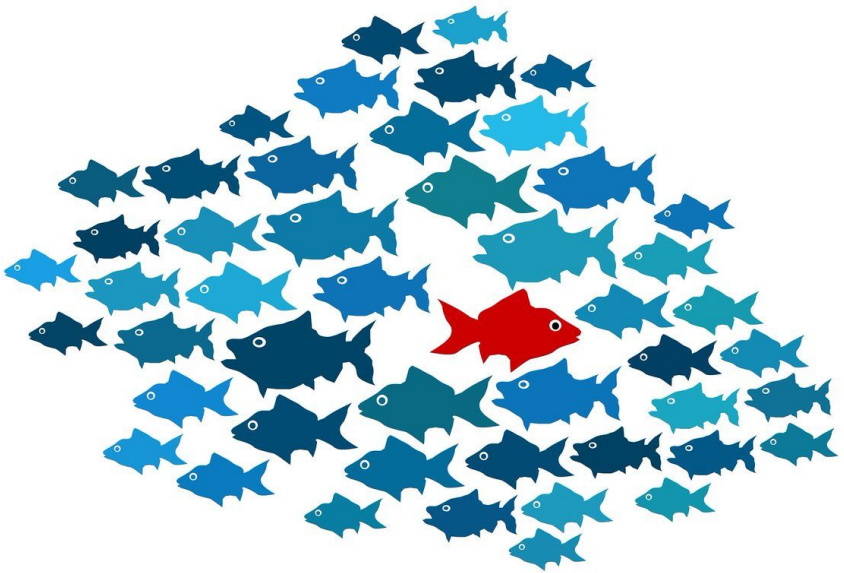
So, we are measuring similarity based some **features** like size or color.

Clusters

- Cluster groups of customers have **“similar”** interest / characteristics
- People within the same cluster are **highly similar** whereas people belongs to different clusters are highly **“dissimilar”**

Outliers

- Outliers are the data points which are highly “**dissimilar”** others
- Similarity/dissimilarity measure is important in identifying outliers



Similarity and dissimilarity

- **Distance** or similarity measures **how close two distributions are**
- Quantity that reflects the strength of relationship between two objects or two features
- Three terms:
  - **Similarity :** how alike two data objects
  - **Dissimilarity :** how different two data objects are (also called distance)
  - **Proximity:** similarity or dissimilarity

Similarity and Dissimilarity

- Both are related (inversely)
- Similarity
  - value is between 0 and 1
  - Larger the value, higher the similarity
- Dissimilarity
  - Value is between 0 and 1
  - Higher the value, more dissimilar the objects
- Similarity can be expressed as a function of dissimilarity
  - Sim(i,j) = 1- Dissim(i,j)

Two data structures

- In ML and DM, two common data structures
  - **Data matrix**
  - **Dissimilarity matrix**

Two data structures

- In ML and DM, two common data structures
  - **Data matrix : ** used to store data objects
  - **Dissimilarity matrix:** store pairwise dissimilarity

Data matrix

- Object by feature (attribute) structure
- Each row corresponds to an object, a feature vector
- With n objects, and p features, data matrix will be a real matrix of order *n by p*
- Two mode matrix (object and feature)

Example for Data Matrix

	X_1 sepal length	X_2 sepal width	X_3 petal length	X_4 petal width	X_5 class
x_1	5.9	3.0	4.2	1.5	Iris-versicolor
x_2	6.9	3.1	4.9	1.5	Iris-versicolor
x_3	6.6	2.9	4.6	1.3	Iris-versicolor
x_4	4.6	3.2	1.4	0.2	Iris-setosa
x_5	6.0	2.2	4.0	1.0	Iris-versicolor
x_6	4.7	3.2	1.3	0.2	Iris-setosa
x_7	6.5	3.0	5.8	2.2	Iris-virginica
x_8	5.8	2.7	5.1	1.9	Iris-virginica
⋮	⋮	⋮	⋮	⋮	⋮
x_{149}	7.7	3.8	6.7	2.2	Iris-virginica
x_{150}	5.1	3.4	1.5	0.2	Iris-setosa

Table 1.1: Extract from the Iris Dataset

Dissimilarity matrix

- Pairwise dissimilarity (distance) between objects
- Its a square matrix of order n
- Each entry d_{ij} represents the dissimilarity between data points X_i and X_j
- It is a symmetric matrix :
  - Dissim(i,j) = Dissim(j,i)
- All diagonal elements are 0
  - d_{ii} = 0 (since Dissim(i,i)= 0
- One mode matrix

Group	1	2	3	4	5	6	7	8
1	0							
2	4.15	0						
3	11.02	15.01	0					
4	7.16	3.03	18.02	0				
5	43.72	47.49	32.80	50.41	0			
6	54.37	58.23	43.36	61.19	11.12	0		
7	46.34	50.20	35.34	53.16	3.78	8.03	0	
8	55.42	59.27	44.42	62.23	12.05	1.12	9.08	0

Attribute Types

- **Nominal**
  - Categorical data, numbers given as codes to certain classes
  - Zip code, EmpID, Color
  - Can be compared by = or ≠
- **Ordinal**
  - Data have meaningful order
  - But cannot know how important it is
  - Eg: Hardness of material, Toughness of exam (high/medium/low)
- **Binary**
  - True or false, Yes or No, Male or Female, 0 or 1
- **Numeric**
  - Integer/real values

Proximity Measures for Nominal Attributes

- If objects are all nominal (categorical) then proximity measures can be used to measure similarity
- Similarity/Dissimilarity can be measured by simple matching
- The dissimilarity between two objects i and j can be computed based on the ratio of mismatches
- $d(i , j) = (p-m)/p$
  - m is the number of matches
  - p is the total number of attributes
-

Proximity Measures for Nominal Attributes

- Alternatively
  - sim(i, j) = 1 - d(i,j) = m/p
- Example
-

Roll Number	Mark	Grade
1	90	A
2	82	B
3	80	B
4	90	A

Proximity Measures for Nominal Attributes

- Dissimilarity Matrix

d(1,1)	d(1,2)	d(1,3)	d(1,4)
d(2,1)	d(2,2)	d(2,3)	d(2,4)
d(3,1)	d(3,2)	d(3,3)	d(3,4)
d(4,1)	d(4,2)	d(4,3)	d(4,4)

Roll Number	Mark	Grade
1	90	A
2	82	B
3	80	B
4	90	A

Proximity Measures for Nominal Attributes

- Dissimilarity Matrix

d(1,1)= (2-2)/2 = 0	d(1,2)	d(1,3)	d(1,4)
d(2,1)	d(2,2)	d(2,3)	d(2,4)
d(3,1)	d(3,2)	d(3,3)	d(3,4)
d(4,1)	d(4,2)	d(4,3)	d(4,4)

- d(i , j) = (p-m)/p
  - m is the number of matches
  - p is the total number of attributes

Roll Number	Mark	Grade
1	90	A
2	82	B
3	80	B
4	90	A

Proximity Measures for Nominal Attributes

- Dissimilarity Matrix

d(1,1)= (2-2)/2 = 0	d(1,2)	d(1,3)	d(1,4)
d(2,1)= (2-0)/2 = 1	d(2,2)	d(2,3)	d(2,4)
d(3,1)	d(3,2)	d(3,3)	d(3,4)
d(4,1)	d(4,2)	d(4,3)	d(4,4)

- d(i , j) = (p-m)/p
  - m is the number of matches
  - p is the total number of attributes

Roll Number	Mark	Grade
1	90	A
2	82	B
3	80	B
4	90	A

Proximity Measures for Nominal Attributes

- Dissimilarity Matrix

d(1,1)= (2-2)/2 = 0	d(1,2)	d(1,3)	d(1,4)
d(2,1)= (2-0)/2 = 1	d(2,2)	d(2,3)	d(2,4)
d(3,1)	d(3,2)	d(3,3)	d(3,4)
d(4,1)	d(4,2)	d(4,3)	d(4,4)

Complete the computation!!

- d(i , j) = (p-m)/p
  - m is the number of matches
  - p is the total number of attributes

Roll Number	Mark	Grade
1	90	A
2	82	B
3	80	B
4	90	A