Lecture :Disimilarity between attribute of mixed Data Type

Mixed Data Types
Gower Distance
Computation with Example

Date: Monday _11/10/2023

Attributes of Data

- An **attribute** is an object's property or characteristics.
- For example. A person's hair colour, air humidity etc.
- An attribute set defines an **object**.
- The **object** is also referred to as a record of the instances or entity.

Types of attributes / Data Types

- Categorical data/Nominal.
- Categorical /Ordinal data
- Binary Data
- Quantitative/Interval Data
- Quantitative/Ratio Data

Attribute of mixed Data Type

- To compute dissimilarity between objects with mixed data types (e.g., numerical and categorical), you can use a combination of distance metrics that are appropriate for each data type.
- Example – Numeric Attributes- Euclidean ,etc.
- Example - Textual –Jaro,LCS,etc
- Example-Binary- Jaccard etc.

How to compute Mixed Types?

- One common approach is to use the Gower distance.
- What is Gower Distances?

Gower Distances

- Gower distance is a metric used to calculate the dissimilarity or similarity between objects (usually data points or rows in a dataset) with mixed data types, such as numerical and categorical attributes.
- It is a versatile approach for handling data with diverse attribute types and is often used in clustering, classification, and similarity analysis.
- The Gower distance takes into account both the magnitude and the categorical nature of attributes.

Computation of Gower Distances

- **Step 1: Data Preprocessing**:
  - **Numerical Attributes**: Ensure that numerical attributes are on a common scale, usually by normalizing or standardizing them.
  - **Categorical Attributes**: Encode categorical attributes into numerical values using techniques like one-hot encoding, label encoding, or binary encoding.

Step 2: Attribute-wise Distance Calculation

- **Attribute-wise Distance Calculation**: For each attribute (both numerical and categorical), calculate the pairwise dissimilarity or distance between data points. The specific distance metric used depends on the data type of the attribute:
  - **Numerical Attributes**:. Calculate the distance between the numerical values of two data points for each attribute. Use a distance metric such as Euclidean distance, Manhattan distance, or Mahalanobis distance etc.
  - **Categorical Attributes**: Use a categorical distance metric for categorical attributes, such as Jaccard distance, Sørensen-Dice distance, or another appropriate metric.

Step 3: Weighting

- Optionally, assign different weights to attributes based on their importance in your analysis.
- Weighting allows you to emphasize certain attributes more than others when calculating the overall dissimilarity.

Step 4: **Aggregation**

- Combine the dissimilarities calculated in step 2 into an overall dissimilarity measure for each pair of data points. To do this, you can follow these steps:

A. For each attribute, calculate the dissimilarity between the two data points, considering the attribute's type (numerical or categorical) and any weighting applied.
B. Combine the attribute-wise dissimilarities into an overall dissimilarity measure. You can use a weighted sum, weighted average, or other aggregation methods based on your preferences. Be sure to consider both numerical and categorical dissimilarities.

Step5: Normalization

- If necessary, normalize the overall dissimilarity values to ensure they are on a common scale.
- This step may be important if the attribute dissimilarities are on different scales or have different ranges.

EXAMPLE –Gower Distance

- Computing Gower distance for two objects with mixed attribute types.
- Three attributes: one numeric attribute (age)
- Two categorical attributes (gender and marital status).
- calculate the Gower distance between two individuals, Person A and Person B.

Data to Compute Gower Distance

- **Person A**:
- Age: 30
- Gender: Male
- Marital Status: Married
- **Person B**:
- Age: 25
- Gender: Female
- Marital Status: Single

Step1 -Preprocessing

- Numeric Attribute (Age): No preprocessing needed.
- Categorical Attributes (Gender and Marital Status): Encode these categorical attributes into binary values using one-hot encoding.
- For example,
  - for gender, we can encode Male as 0 and Female as 1.
  - For marital status, we can encode Married as 1 and Single as 0

Encoded Data After Preprocessing

- **Person A**:
- Age: 30
- Gender: 0 (Male)
- Marital Status: 1 (Married)
- **Person B**:
- Age: 25
- Gender: 1 (Female)
- Marital Status: 0 (Single)

Calculate Gower Distance

- Age:

  For numeric attributes like Age, you can use a standard distance metric like Euclidean distance:

  $$d_{\text{Age}}(A, B) = \sqrt{(30 - 25)^2} = 5$$

- Gender (Categorical):

  For categorical attributes like Gender, you can use a binary dissimilarity metric like the Jaccard distance, which is 1 for dissimilar values and 0 for similar values:

  $$d_{\text{Gender}}(A, B) = 1$$

- Marital Status (Categorical):

  Similarly, for the Marital Status attribute, use the Jaccard distance:

  $$d_{\text{Marital Status}}(A, B) = 1$$

Step 3-Weighting

You can assign different weights to the attributes based on their importance. Let's assume equal weights for simplicity:

- $w_{\text{Age}} = 1$
- $w_{\text{Gender}} = 1$
- $w_{\text{Marital Status}} = 1$

Compute the Gower Distance:

Now, you can calculate the Gower distance by taking the weighted average of the attribute distances:

$$d_{\text{Gower}}(A,B) = \frac{w_{\text{Age}} \cdot d_{\text{Age}}(A,B) + w_{\text{Gender}} \cdot d_{\text{Gender}}(A,B) + w_{\text{Marital Status}} \cdot d_{\text{Marital Status}}(A,B)}{w_{\text{Age}} + w_{\text{Gender}} + w_{\text{Marital Status}}}$$

Substituting the values:

$$d_{\text{Gower}}(A,B) = \frac{1 \cdot 5 + 1 \cdot 1 + 1 \cdot 1}{1+1+1} = \frac{7}{3} \approx 2.33$$

Interpretation of Result

- A Gower distance of 2.33 suggests that Person A and Person B are relatively dissimilar in terms of the attributes
- Let assume with example where user has to take decision such as, if you are using Gower Distances, for customer segmentation, a higher Gower distance might suggest that two customers are less likely to have similar preferences or needs, which could affect marketing strategies.

Thank You