Proximity Measures

DMPR: Lab activity for Week 05

Rahul Kumar

February 23, 2023

Q1. The given data set is raw survey data provided by your class, and in these types of survey data we can expect a lot of nonuniformities in various attributes, which need to be taken care of in pre-processing part of data analysis. **(6marks**)

For e.g., column 'Gender' contains two categories "male" and "Female" which needs to be encoded as 0's and 1's respectively to represent two distinct categories to the interpreter.

Column 'Height' contains entries in cm, ft, and inches. So, each of the heights should be represented to a common unit to perform proximity measurements. Similarly, in column "How many hours of extracurricular (non-sport) activities do you do per week (approximately on the average)?" some entries are in range form which need to be taken care of before starting any analysis.

Similar pre-processing is needed for several other features also.

List down what all the attributes that requires any pre-processing, also explain what sort of pre-processing is required for those attributes.

Q2.

Write the conditions for which the Manhattan distance and Euclidian distance will be same for two points in a given 2D feature plane. **(2 marks)**

Q3. **(9 marks)**

A) For feature "What musical instruments can you play?" , define some proximity measures in order to quantify similarity/dissimilarity between 2 instances as per this feature. **(2 marks)**

B) Define a function that calculates the Jaccard index for a pair of instance **(4 marks)**

C) Using the above function find which pairs of instances vary most as per the given features. **(3 marks)**

Q4. **(5marks)**

In the given dataset can you find what is the hair type and serial number of the person which is closest to person having

height is 178cm

nose length = 2.4inches

took 480 seconds to fill the survey form.

Note: - Distance measure used for the given problem should be the Manhattan distance.

And you should define a function which calculates the Manhattan distance.

Q5. **(8 marks)**

In the given dataset we have two features

1. "The maximum distance between your thumb and your little finger on your right hand when stretched is"
2. "The maximum distance between your thumb and your little finger on your left hand when stretched is"

For these two features we expect their values to be closely varying, but this data may contain some outliers which deviates from this trend.

A) Plot a scatterplot (feature 1 vs 2) for the given dataset and fit a linear regression model to the above dataset and see by how much percent does the slope obtained from the linear regression vary from the expected slope. **(3 marks)**

B) Using a box plot and inter quartile ranges try to find the outliers if any present and update your data by dropping the outliers. **(2 marks)**

C) Repeat step A) for this new updated dataset (without outliers) and now see what the deviation is now from the expected trend, explain your observations. **(3 marks)**