KNN Classifier Implementation Project

Jacob Diba
CSE 5160
Dr Yan Zhang
California State University San Bernardino
November 25th 2020

What is K Nearest Neighbor(KNN)

- K nearest Neighbor is a supervised machine learning algorithm that classifies inputs based on already labeled data.
- How it works is that the algorithm will take in an input from testing data and measure the distance between the input and all training data values to find the distance between them.
- Then after finding the distance, sort from closest to farthest of the distance and pick out a K given. After looking at the amount of K of the distance based on their classification decided the test data classification.

Programing Language Used

- The Programing Language that I used for the creation of this algorithm was Python as it was the easiest for me to use due to its dynamic data types and many useful libraries to help me with data management .

DATA

Data

- The Data that I chose for this project was the **Somerville Happiness Survey Data Set**
- This Data Set is a measurement of the Happiness of people in Somerville based on certain attributes that lead to their mood.

- This Data Set was chosen from the UCI Machine Learning Repository
  - URL:http://archive.ics.uci.edu/ml/datasets/Somerville+Happiness+Survey

Data Attributes

- In this Data Set there are a total of 6 attributes that ranging from 1 to 5

- D = decision attribute (D) with values 0 (unhappy) and 1 (happy)
- X1 = the availability of information about the city services
- X2 = the cost of housing
- X3 = the overall quality of public schools
- X4 = your trust in the local police
- X5 = the maintenance of streets and sidewalks
- X6 = the availability of social community events
- Total Training Instance are 143 where 31 is taken as test data

Code Process

First Step

- The first step that is gonna happen is we are gonna import our data from our .csv file and import it into the program and store it.
- We then setup or dictionary file structure and import all our data into it and label it.

Second Step-Feature Scaling/Min Max Normalization

- The next step that we due is to pass our data through our min max function which is gonna min max normalize our data to not have it skew a certain way later one

$$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Third Step Euclidean Distance

- For the third step we will then find the distance of one instance of the testing data from all of the training data. Then from there we put our result into another column in our dictionary and put it under the distance label.

$$d(\mathbf{p}, \mathbf{q}) = d(\mathbf{q}, \mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

Fourth Step Sorting and K

- After we get our distances and add them to our data we then want to sort our data from the shortest distance to the longest distance from.

- After given a k or the program will automatically pick 15 as your k go in and count how many Happy or Unhappy points, and from there predict what your test instance will be and output it. This will happen for all of our test instances.

Demonstration

Evaluation

Prediction Rate of Different K Values

K=5 8/31 Wrong [74% Prediction Rate]

k=15 12/31 Wrong [60% Prediction Rate]

k=20 12/31 Wrong [60% Prediction Rate]

k=30 11/31 Wrong[64% Prediction Rate]

Limit of Program

- When looking at the limitations of the program something can come up. One thing that came up is that this program is designed specifically for this data set so that it will be very difficult to use another data set
- Also another limitation that came up would be that the attributes have to be premade and can not be inputed which holds back.

Limit of KNN

- Then when we look at KNN there are some limitations that can be seen.
- One of the limitations that can be seen is that KNN is a Lazy Learner and or does not make a model thus when running it you have to start the entire process over again and cannot rely on a model.
- While it is a Lazy Learner it is also Inefficient due to those reasons.
- Then lastly it is also sensitive to whatever that K value that is chosen which will determine the output greatly.

Thank you