EE5180

End term exam C

VARASALA SANJAY
EE19B126

Problem:

You are a student in IITM and you want to build up your extra-curricular part of the resume. However, there are multiple options such as student secretary, youth festival coordinator, sport etc. Assume that there are 4 possible options and you have to now decide which one suits you best, which has maximum impact on your resume, which you have a high probability of getting in etc.
Devise a learning algorithm which takes as input a set of features such as 1) maximum impact on resume, high chance of getting in etc. and outputs that option which is the best for you.
Describe your input space, the feature vectors, the logic behind this input space/feature vectors, the loss function etc.
Clearly define input domain, labels, loss function, hypothesis class etc.
Propose a learning algorithm best suited for this task and evaluate its performance using a toy data set or a data set you can generate using ppl you know.

Intro:

Classification is a classic machine learning application. Classification basically categories your output in two classes i.e., your output can be one of two things. For example, a bank wants to know whether a customer will be able pay his/her monthly investments or not? We can use machine learning algorithms to determine the output of this problem, which will be either Yes or No (Two classes). But what if you want to classify something that has more than 2 categories and isn't as simple as a yes/no problem?

This is where multi-class classification comes in. Multiclass classification can be defined as the classifying instances into one of three or more classes

This is where multi-class classification comes in. Multiclass classification can be defined as the classifying instances into one of three or more classes. In this article we are going to do multi-class classification using K Nearest Neighbours.
KNN is a super simple algorithm, which assumes that similar things are in close proximity of each other. So if a datapoint is near to another datapoint, it assumes that they both belong to similar classes.

Outlook:

There are four viable options that a student can take and the gaining knowledge of algorithm

should return the pleasant viable choice. The given problem is an instance of multi-class type.

The four positions I have considered for this venture are:

- Student Secretary

- Hostel secretary
- E-cell core
- Sports secretary.

We can now label the 4 feasible options as y= {0,1,2,3} respectively. The mastering algorithm proposed for this hassle is the K-Nearest Neighbour (KNN) algorithm.

Feature vector:

Feature vectors are used to represent numeric or symbolic characteristics, called features, of an object in a mathematical, easily analyzable way. The model of the kNN classifier is based on feature vectors and class labels from the training data set

The Features of the position of responsibility can be listed as follows:

- ### Maximum impact on resume:

The impact of PORs is one of the primary reasons students choose them. Coming to the placements Many managerial positions, like consulting etc necessitate POR.
As a result, this functionality was taken into consideration.

- ### Possibility of gaining the job:

A single student will not be able to apply for all of the POR that are available. Hence
He or she may wish to know what their chances are of getting the POR before applying.
This ensures that the student's efforts are not wasted. As a result, this functionality was created.

- ### The amount of work involved in the position:

POR is performed by students in addition to their academic obligations. As a result, students are hesitant to pursue jobs that require too much effort from them. As a result, this functionality was taken into consideration.

- ### Power of the position:

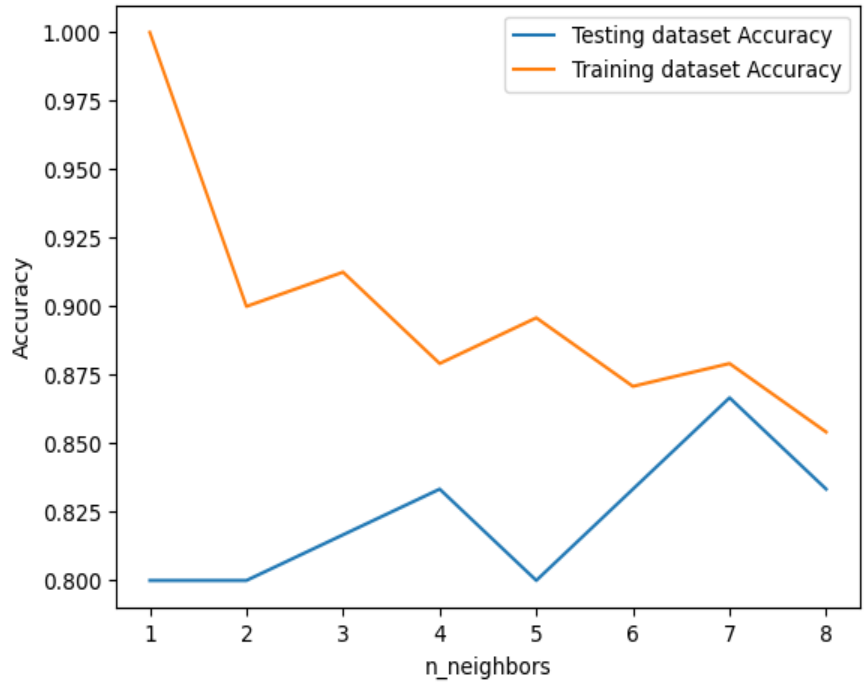
Many students participate in POR in order to prepare for their future endeavors. POR tied to academics may be beneficial in the future, however some POR may contain social engagement, which is not beneficial to a student academically. As a result, this functionality was taken into consideration.

- ### Impact of PORs on connections:

In the professional world, having good contacts/network is a huge plus, and PORs are one of the ways a student can get connections. As a result, this functionality was taken into consideration.

Plots:

When the model is trained for the training data and tested for real data, the predictions made were as follows. The axes of the plots are features of the position.



For the optimal results we choose the value of k =7(Accuracy on both test and training sets being around 87% for 7, high variance for other values of k). **Accuracy = 87%**

CODE: code in python

```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import train_test_split
from sklearn.datasets import load_iris
import numpy as np
import matplotlib.pyplot as plt
import random as rand
f = open("data.txt", "r")
```

```
x = []
y = []
dt = f.read()
arr = dt.split("\n")
for i in range (300):
    temp = arr[i].split(" ")
    for j in range(6):
        temp[j] = int(temp[j])
    x.append(temp[:5])
    y.append(temp[5])
f.close()
X_train, X_test, y_train, y_test = train_test_split(
            x, y, test_size = 0.2, random_state=42)
neighbors = np.arange(1, 9)
train_accuracy = np.empty(len(neighbors))
test_accuracy = np.empty(len(neighbors))
for i, k in enumerate(neighbors):
    knn = KNeighborsClassifier(n_neighbors=k)
    knn.fit(X_train, y_train)
    train_accuracy[i] = knn.score(X_train, y_train)
    test_accuracy[i] = knn.score(X_test, y_test)


plt.plot(neighbors, test_accuracy, label = 'Testing dataset Accuracy')
plt.plot(neighbors, train_accuracy, label = 'Training dataset Accuracy')
plt.legend()
plt.xlabel('n_neighbors')
plt.ylabel('Accuracy')
plt.savefig("plot.png")
```

The code is written for KNN classification. The data is stored in a text file as data.txt

The dataset that we are going to be using is data.txt and it basically has 5 features for its 300 data points and is categorized into 4 different classes which are mentioned.

The dataset can be downloaded from the following link: data

We use the built-in KNN algorithm from sci-kit learn. We split our input and output data into training and testing data, as to train the model on training data and testing model's accuracy on the testing model.

Here, we see that the classifier chose 7 as the optimum number of nearest Neighbors to classify the data best.

TWIST:

All of the characteristics appear to fill only a portion of the input domain space, as demonstrated in the sample of test data. The Algorithm fails to forecast if the input vector is X = [1, 1, 1, 1, 1, 3] because it has never seen such test data.

In general, if the input features take on unusual values (values that appear unlikely), then the algorithm is unable to forecast the outcome.

References

- Multiclass classification using scikit-learn - geeks for geeks
- Multiclass Classification Using K-Nearest Neighbors - towards data science