CS 520 Final: Question 2 - Classification

Diksha Prakash - dp978

December 2019

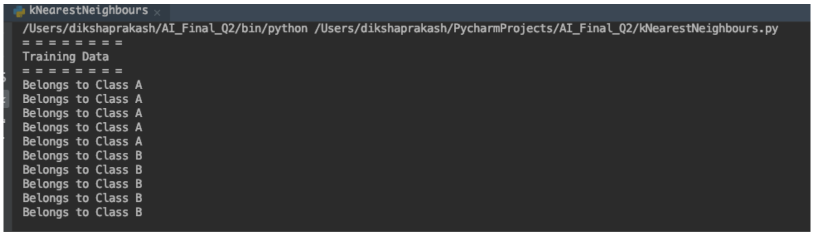
Abstract

This problem provides five instances of Class A images, five instances of Class B images, and five instances of unlabeled images. We are expected to build two distinct classification models and provide insights on the output.

1 Construct a model to classify images as Class A or B, and train it on the indicated data. Specify your trained model. What does your model predict for each of the unlabeled images? Give the details of your model, its training, and the final result. Do the predictions make sense, to you?

For the first classification algorithm, I have used ‘K-Nearest Neighbours Algorithm’ where I have taken k=3.
Images in the ‘Class A’ and ‘Class B’ were represented as a 5X5 matrix in separate ’.txt’ files. These matrices were flattened by the code to obtain a list of length 25. All the provided grids were entered into an array and a list of corresponding classes were made.
The algorithm simply evaluates k-nearest items in the dataset to predict the class of the current grid. Thus, the algorithm doesn’t need a training phase.

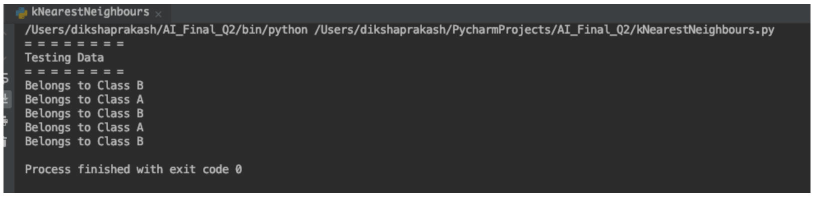
The metric used to calculate the nearest neighbours i.e. the measure of distance used in the implementation is ‘Euclidean Distance’. I have ensured that the value chosen for ‘k’ is odd to avoid conflict in the interpretation of results. After experimenting with various values of ‘k’, I chose $k = 3$.
Values greater than 3 are too big for this dataset as the dataset itself has 5 instances in each class.
The evaluation of the proposed model on the training data yields correct classification for all the grids. This can be seen in Figure 1.



```
/Users/dikshaprakash/AI_Final_Q2/bin/python /Users/dikshaprakash/PycharmProjects/AI_Final_Q2/kNearestNeighbours.py
= = = = = = = =
Training Data
= = = = = = = =
Belongs to Class A
Belongs to Class A
Belongs to Class A
Belongs to Class A
Belongs to Class A
Belongs to Class B
Belongs to Class B
Belongs to Class B
Belongs to Class B
Belongs to Class B
```

Figure 1: Evaluation of Model on Training Data

The classification for the grids in Mystery (fed as the testing data via 'Mystery.txt'), were as shown in Figure 2.



```
/Users/dikshaprakash/AI_Final_Q2/bin/python /Users/dikshaprakash/PycharmProjects/AI_Final_Q2/kNearestNeighbours.py
= = = = = = = =
Testing Data
= = = = = = = =
Belongs to Class B
Belongs to Class A
Belongs to Class B
Belongs to Class A
Belongs to Class B

Process finished with exit code 0
```

Figure 2: Output of Model for Test Data

The predictions in Figure 2 i.e. Grid 1 - Class B, Grid 2 - Class A, Grid 3 - Class B, Grid 4 - Class A and Grid 5 - Class B, make sense to me the grids in Class A have more black blocks to the left whereas the grids in Class B have more black blocks to the right.
In the grids given for Mystery, the following analysis can be made:

- Grid 1 has more black blocks to the left. Hence, it should belong to Class A, which it does.
- Grid 2 has more black blocks to the right. Hence, it should belong to Class B, which it does.
- Grid 3 has more black blocks to the left. Hence, it should belong to Class A, which it does.
- Grid 4 has more black blocks to the right. Hence, it should belong to Class B, which it does.
- Grid 5 is an interesting case as it has symmetric distribution with the black blocks distributed evenly around the corners. However, the model classified it as 'Class B' (in both the types of implemmentation).

2 The data provided is quite small, and over-fitting is a serious risk. What steps can you take to avoid it?

Since only five grid instances have been provided for each type of class, the data is quite small to cover all possible grids that belong to either class. The reason for choosing KNN model based classification is the ease to reduce the degree of over-fitting by increasing the value of k as experimentally, $k = 1$ yields maximum over-fitting.

When talking about methods to avoid over-fitting for other models, we can explore re-sampling, noise addition, data augmentation etc.
To talk in depth, I would like to expand on the idea of 'Data Augmentation' as I feel that we have less training data here. Data Augmentation will help to increase the variety of data seen during training. However, there will exist certain restrictions as to how data augmentation will be done as changing the orientation can lead to misclassifcation as we the have the notion of direction in the grids. Rotating the image can move the black blocks initially centered on the Left to right, just changing the class from A to B.

3 Construct and train a second type of model. Specify its details. How do its predictions compare to the first model? Are there any differences, and what about the two models caused the differences?

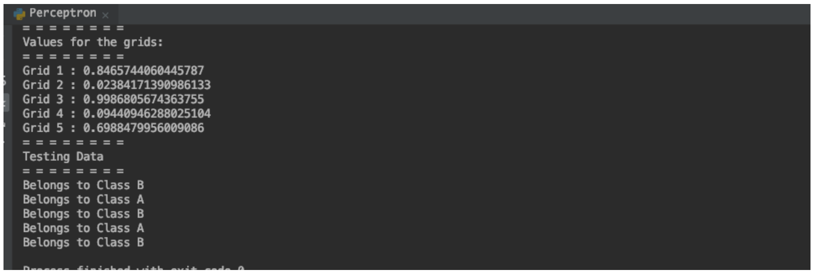
For the second classification algorithm, I have used 'Perceptron Network' where I have taken 'Sigmoid' activation function.
In this case too, images in the 'Class A' and 'Class B' were represented as a 5X5 matrix in separate '.txt' file. These matrices were flattened by the code to obtain a list of length 25. The training was done for 1000 iterations with the learning rate being 0.01
As a result of this, a regression model was built with weights and bias through the network which returned values between 0 and 1 for each grid encountered.

The values returned for the mystery grids are as under:

- Grid 1 : 0.8465744060445787
- Grid 2 : 0.02384171390986133
- Grid 3 : 0.9986805674363755

- Grid 4 : 0.09440946288025104
- Grid 5 : 0.6988479956009086

A threshold value of 0.5 has been set for a given grid to belong to Class B. As discussed in Section 1, these values make sense because the grids in Class A have more black boxes to the left side whereas the ones in Class B have more black boxes to the right side.
The classification for the grids in Mystery (fed as the testing data via 'Mystery.txt'), were as shown in Figure 3.



```
= = = = = = = =
Values for the grids:
= = = = = = = =
Grid 1 : 0.8465744060445787
Grid 2 : 0.02384171390986133
Grid 3 : 0.9986805674363755
Grid 4 : 0.09440946288025104
Grid 5 : 0.6988479956009086
= = = = = = = =
Testing Data
= = = = = = = =
Belongs to Class B
Belongs to Class A
Belongs to Class B
Belongs to Class A
Belongs to Class B

Process finished with exit code 0
```

Figure 3: Output of Model for Test Data

In the grids given for Mystery, the following analysis can be made:

- Grid 1 has more black blocks to the left. Hence, it should belong to Class A, which it does.
- Grid 2 has more black blocks to the right. Hence, it should belong to Class B, which it does.
- Grid 3 has more black blocks to the left. Hence, it should belong to Class A, which it does.
- Grid 4 has more black blocks to the right. Hence, it should belong to Class B, which it does.
- Grid 5 is an interesting case as it has symmetric distribution with the black blocks distributed evenly around the corners. However, Since its value is greater than 0.5, its classified as Class B.

It is worth noting that except for Grid 5, all other Grids belonging to Class B have values quite close to 1. However, due to the symmetry of Grid 5, its value is comparatively quite close to 0.5.
It can be observed from Figures 1 and 3 that both the models make the same predictions. However, it can be seen that the regression model was quite close to predicting Class A as well.