COMP 307 Assignment 1

Part 1 Report

Predicted class labels in the test set using basic KNN method where k=1

```
Accuracy for K = 1 : 0.9438202247191011

Predicted Class: 3 Actual class: 3
Predicted Class: 3 Actual class: 3
Predicted Class: 3 Actual class: 3
Predicted Class: 1 Actual class: 1
Predicted Class: 1 Actual class: 1
Predicted Class: 1 Actual class: 1
Predicted Class: 1 Actual class: 2
Predicted Class: 2 Actual class: 2
Predicted Class: 1 Actual class: 1
Predicted Class: 2 Actual class: 2
Predicted Class: 2 Actual class: 2
Predicted Class: 3 Actual class: 2
Predicted Class: 3 Actual class: 3
Predicted Class: 3 Actual class: 3
Predicted Class: 1 Actual class: 1
Predicted Class: 2 Actual class: 2
Predicted Class: 3 Actual class: 3
Predicted Class: 3 Actual class: 3
Predicted Class: 1 Actual class: 1
Predicted Class: 1 Actual class: 1
Predicted Class: 3 Actual class: 3
Predicted Class: 2 Actual class: 2
Predicted Class: 2 Actual class: 2
Predicted Class: 3 Actual class: 3
Predicted Class: 2 Actual class: 2
Predicted Class: 3 Actual class: 2
Predicted Class: 2 Actual class: 2
Predicted Class: 3 Actual class: 3
Predicted Class: 2 Actual class: 2
Predicted Class: 1 Actual class: 1
Predicted Class: 2 Actual class: 2
Predicted Class: 1 Actual class: 1
Predicted Class: 2 Actual class: 2
Predicted Class: 1 Actual class: 1
Predicted Class: 2 Actual class: 2
Predicted Class: 2 Actual class: 2
Predicted Class: 2 Actual class: 2
Predicted Class: 2 Actual class: 2
Predicted Class: 2 Actual class: 2
Predicted Class: 1 Actual class: 1
Predicted Class: 2 Actual class: 2
Predicted Class: 2 Actual class: 2
Predicted Class: 3 Actual class: 3
Predicted Class: 1 Actual class: 1
Predicted Class: 2 Actual class: 2
Predicted Class: 1 Actual class: 1
Predicted Class: 3 Actual class: 3
Predicted Class: 2 Actual class: 2
Predicted Class: 2 Actual class: 2
Predicted Class: 1 Actual class: 1
Predicted Class: 3 Actual class: 3
Predicted Class: 1 Actual class: 1
Predicted Class: 1 Actual class: 1
Predicted Class: 3 Actual class: 3
Predicted Class: 3 Actual class: 3
Predicted Class: 1 Actual class: 1
Predicted Class: 1 Actual class: 1
Predicted Class: 3 Actual class: 3
Predicted Class: 1 Actual class: 1
Predicted Class: 3 Actual class: 3
Predicted Class: 3 Actual class: 3
Predicted Class: 1 Actual class: 2
Predicted Class: 2 Actual class: 2
Predicted Class: 3 Actual class: 3
Predicted Class: 2 Actual class: 2
Predicted Class: 3 Actual class: 3
Predicted Class: 3 Actual class: 3
Predicted Class: 1 Actual class: 1
Predicted Class: 1 Actual class: 1
Predicted Class: 2 Actual class: 2
Predicted Class: 1 Actual class: 2
Predicted Class: 3 Actual class: 3
Predicted Class: 2 Actual class: 2
Predicted Class: 2 Actual class: 2
Predicted Class: 1 Actual class: 1
Predicted Class: 1 Actual class: 1
Predicted Class: 1 Actual class: 1
Predicted Class: 3 Actual class: 3
Predicted Class: 1 Actual class: 1
Predicted Class: 1 Actual class: 1
Predicted Class: 2 Actual class: 2
Predicted Class: 2 Actual class: 2
Predicted Class: 3 Actual class: 3
Predicted Class: 1 Actual class: 1
Predicted Class: 2 Actual class: 2
Predicted Class: 1 Actual class: 1
Predicted Class: 1 Actual class: 1
Predicted Class: 2 Actual class: 2
Predicted Class: 1 Actual class: 1
```

Classification accuracy for k=3 vs k=1

```
Accuracy for K = 1 : 0.9438202247191011
Accuracy for K = 3 : 0.9550561797752809
```

As you can see there was a slight improvement in accuracy where k=3 this is because in some cases for k=1 the nearest neighbor for an instance is affected by noise, when k=3 the noise has a lower influence on the result.

Advantages and disadvantages of KNN

KNN is very easy to implement and use while at the same time achieving good results/accuracy. Though for larger data sets it is very computationally expensive testing each instance. It is also quite sensitive to noise/outliers in datasets.

Applying k-fold cross validation with k=5

If I were to apply K-Fold Cross Validation to the above problem with k=5 I would:

1. First split the whole data set into 5 subsets
2. Then for each subset
    1. Take the selected subset as the test data set
    2. Take the remaining 4 subsets as the training data set
    3. Train a classifier using the training set and apply it to the test set
    4. Store the results
3. Find the average of the 5 different results for the final result

Approaching the same problem where there are no class labels in the data sets and 3 known clusters

In a situation where the same data was given but no class labels were given in the data sets though we knew there were 3 obvious clusters in the data, we could use K-Means Clustering to group the unlabeled data.

1. Set 3 initial means randomly in the data set
2. Create 3 clusters by clustering every instance with the nearest mean using a distance measure (e.g. Euclidean distance)
3. Replace the old means with the centroid of each of the 3 clusters
4. Repeat steps 2 & 3 until there is no change in each of the clusters centers.