Part 1 Questions:

COMP307 Report

Question 1:

Prediction Actual Result

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-versicolor Iris-versicolor Match

Iris-versicolor Iris-versicolor Match

Iris-versicolor Iris-versicolor Match

Iris-versicolor Iris-versicolor Match

Iris-versicolor Iris-versicolor Match

Iris-versicolor Iris-versicolor Match

Iris-versicolor Iris-versicolor Match

Iris-versicolor Iris-versicolor Match

Iris-virginica Iris-versicolor No Match

Iris-versicolor Iris-versicolor Match

Iris-versicolor Iris-versicolor Match

Iris-versicolor Iris-versicolor Match

Iris-versicolor Iris-versicolor Match

Iris-versicolor Iris-versicolor Match

Iris-versicolor Iris-versicolor Match

Iris-versicolor Iris-versicolor Match

Iris-versicolor Iris-versicolor Match

Iris-versicolor Iris-versicolor Match

Iris-versicolor Iris-versicolor Match

Iris-versicolor Iris-versicolor Match

Iris-versicolor Iris-versicolor Match

Iris-versicolor Iris-versicolor Match

Iris-versicolor Iris-versicolor Match

Iris-versicolor Iris-versicolor Match

Iris-versicolor Iris-versicolor Match

Iris-virginica Iris-virginica Match

Iris-virginica Iris-virginica Match

Iris-versicolor Iris-virginica No Match

Iris-virginica Iris-virginica Match

Iris-virginica Iris-virginica Match

Iris-virginica Iris-virginica Match

Iris-virginica Iris-virginica Match

Iris-virginica Iris-virginica Match

Iris-versicolor Iris-virginica No Match

Iris-virginica Iris-virginica Match

Iris-virginica Iris-virginica Match

Iris-virginica Iris-virginica Match

Iris-virginica Iris-virginica Match

Iris-versicolor Iris-virginica No Match

Iris-virginica Iris-virginica Match

Iris-virginica Iris-virginica Match

Iris-virginica Iris-virginica Match

Iris-virginica Iris-virginica Match

Iris-virginica Iris-virginica Match

Iris-virginica Iris-virginica Match

Iris-virginica Iris-virginica Match

Iris-virginica Iris-virginica Match

Iris-virginica Iris-virginica Match

Iris-virginica Iris-virginica Match

71 Matches

Process finished with exit code 0

Managed to predict 71 correct using this algorithm. An algorithm of this sort is most likely not going to predict with 100%. This is probably due to a few outliers that the algorithm will struggle to predict due to the way it works.

Question 2:

Prediction Actual Result

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-setosa Iris-setosa Match

Iris-versicolor Iris-versicolor Match

Iris-versicolor Iris-versicolor Match

Iris-virginica Iris-versicolor No Match

Iris-versicolor Iris-versicolor Match

Iris-versicolor Iris-versicolor Match

Iris-versicolor Iris-versicolor Match

Iris-versicolor Iris-versicolor Match

Iris-versicolor Iris-versicolor Match

Iris-virginica Iris-versicolor No Match

Iris-versicolor Iris-versicolor Match

Iris-versicolor Iris-versicolor Match

Iris-versicolor Iris-versicolor Match

Iris-versicolor Iris-versicolor Match

Iris-versicolor Iris-versicolor Match

Iris-versicolor Iris-versicolor Match

Iris-versicolor Iris-versicolor Match

Iris-versicolor Iris-versicolor Match

Iris-versicolor Iris-versicolor Match

Iris-versicolor Iris-versicolor Match

Iris-versicolor Iris-versicolor Match

Iris-versicolor Iris-versicolor Match

Iris-versicolor Iris-versicolor Match

Iris-versicolor Iris-versicolor Match

Iris-virginica Iris-virginica Match

Iris-versicolor Iris-virginica No Match

Iris-versicolor Iris-virginica No Match

Iris-virginica Iris-virginica Match

Iris-virginica Iris-virginica Match

Iris-virginica Iris-virginica Match

Iris-virginica Iris-virginica Match

Iris-virginica Iris-virginica Match

Iris-versicolor Iris-virginica No Match

Iris-virginica Iris-virginica Match

Iris-virginica Iris-virginica Match

Iris-virginica Iris-virginica Match

Iris-virginica Iris-virginica Match

Iris-versicolor Iris-virginica No Match

Iris-virginica Iris-virginica Match

Iris-virginica Iris-virginica Match

Iris-virginica Iris-virginica Match

Iris-virginica Iris-virginica Match

Iris-virginica Iris-virginica Match

Iris-virginica Iris-virginica Match

Iris-virginica Iris-virginica Match

Iris-virginica Iris-virginica Match

Iris-virginica Iris-virginica Match

Iris-virginica Iris-virginica Match

Iris-virginica Iris-virginica Match

69 Matches

Process finished with exit code 0

As the value of k increased i found that the amount of matches decreased. This may be due to as the value of k increases, the larger the distance from test node to value node can allow for a greater risk of outliers.

Question 3:

Advantages:

- Simple algorithm with no assumptions
- Can be used both in classification and regression
- There is not training modules
- Can introduce the idea of weighting the neighbours to create a more accurate model
- Can be used in real time as it uses the given data

Disadvantages:

- There is an optimal value of K which can be hard as it depends on the data
- Imbalances in the training set will show in the testing set
- Can be sensitive to outliers
- For large datasets it can be a slow algorithm and have a hard cost iterating through training sets multiple times

Question 4:

The goal of cross validation is to to the models ability to predict new data in order to eliminate the possibility likelihood an over fitting module.

This is done by firstly splitting the data into training and test segments. Each segment is trained and tested with the results of each individual execution being recorded. For a K-fold of 5, we would split the data into 5 segments. The k nearest neighbour algorithm will be completed a total of 5 times, each time using a different segments of the data for training and testing. The results of each segments execution will be aggregated which is used to test the fit of the algorithm being used.

Question 5:

k-Mean clustering. K-means clustering is used to partition a dataset into k classifications. This can be done without initially knowing the class labels as it uses *n* observations in order to group the data. Below are the steps required to implement k-means clustering:

- Find a value of k where k is the amount of clusters

- Find k random values in the dataset
- Measure the distance from the k values to the data points
- Each unknown data point will be assigned the value of the nearest k value
- Calculate the mean of each cluster
- Re-evaluate the clusters as did in steps 3 and 4 using the mean values
- Find the variance and repeat n times with new values of k to find the best fit

Part 2 Questions:

Question 1:

```
2Categories
16attributes
Read 110 instances
2Categories
16attributes
Read 27 instances
 ASCITES = True:
  SPIDERS = True:
   BILIRUBIN = True:
    FIRMLIVER = True:
  Class alive, prob=1.00
    FIRMLIVER = False:
     BIGLIVER = True:
      STEROID = True:
  Class alive, prob=1.00
      STEROID = False:
       FEMALE = True:
  Class alive, prob=1.00
       FEMALE = False:
        ANTIVIRALS = True:
         FATIGUE = True:
  Class dead, prob=1.00
          FATIGUE = False:
  Class alive, prob=1.00
        ANTIVIRALS = False:
  Class dead, prob=1.00
     BIGLIVER = False:
  Class alive, prob=1.00
   BILIRUBIN = False:
  Class dead, prob=1.00
  SPIDERS = False:
   FIRMLIVER = True:
    AGE = True:
  Class alive, prob=1.00
    AGE = False:
     HISTOLOGY = True:
  Class alive, prob=1.00
     HISTOLOGY = False:
      ANTIVIRALS = True:
  Class dead, prob=1.00
      ANTIVIRALS = False:
       STEROID = True:
  Class alive, prob=1.00
```

```
            STEROID = False:
    Class dead, prob=1.00
      FIRMLIVER = False:
        HISTOLOGY = True:
          BIGLIVER = True:
            SPLEENPALPABLE = True:
    Class alive, prob=1.00
            SPLEENPALPABLE = False:
              ANOREXIA = True:
    Class dead, prob=1.00
              ANOREXIA = False:
    Class alive, prob=1.00
          BIGLIVER = False:
    Class dead, prob=1.00
        HISTOLOGY = False:
    Class alive, prob=1.00
  ASCITES = False:
    BIGLIVER = True:
      STEROID = True:
    Class dead, prob=1.00
      STEROID = False:
        ANOREXIA = True:
    Class dead, prob=1.00
        ANOREXIA = False:
    Class alive, prob=1.00
    BIGLIVER = False:
    Class alive, prob=1.00
```

11/27 Correct

Question 2:

```
Training 1 ------- Test 1: 23/37
Training 2 ------- Test 2: 22/37
Training 3 ------- Training 3: 22/37
Training 4 ------- Training 4: 18/37
Training 5 ------- Training 5: 20/37
Training 6 ------- Training 6: 19/37
Training 7 ------- Training 7: 21/37
Training 8 ------- Training 8: 23/37
Training 9 ------- Training 9: 18/37
Training 10 ------- Training 10: 20/37
Total accuracy = 306/370 -> 63%
```

Question 3:

Pruning is the act of not splitting on a decision as the impurity is low enough for you to determine to viable outcome. In order to prune we can set a threshold. If the gained accuracy of splitting the node is greater than the threshold, then we will decide to split. If the gained accuracy of splitting is less than the threshold then we will not continue splitting, therefore pruning the decision tree. The gained accuracy will be determined through the calculation of the impurity (previous branched weighted average impurity – current branched weighted average impurity). This reduces accuracy of the training set as there is less purity in leaf nodes, therefore greater uncertainty. Pruning decreases complexity as there is less steps in the model however, pruning reduces the amount of over fitting to a training set. Because pruning reduces the amount of over fitting it creates a more general tree making it a better model to test data.

Question 4:
Binary decision algorithms use greedy decisions in the splitting process. As the subsets that you are splitting get smaller and smaller therefore when there is a large number of splits that are required the computational expense gets higher and higher. As we increase the amount of categories the max purity decrease. However, it should increase. As the number of categories increase the level of max impurity will plateau therefore making the accuracy of the impurity decrease.

Part 3 Questions:

Question 1:
my perceptron was able to correctly measure 52/100 against my image data. Through debugging my code i found that majority of the time my Y values lower than the the expected d value. Upon training the weight with the calculation - `double learnWeight = ff.getWeight() + (learn * (DValue - y) * ff.getValue());` should have trained the weights increase them to give me a better accuracy on its next run through. However, this was not the case as my accuracy stays the same. The reason as to why my perceptron is not training the weights to increase them is something that i am unable to find. I will see my lecturers for feedback so that i can learn from this in the future.

Question 4:
Evaluating performance on the training set is not a good measure of the perceptrons effectiveness as it will likely cause over training. The model will specify to the training set, rather than learning the optimal algorithm will learn the data of the training set and will compute to fit this data each time.