**Part One:**

1. Report the class labels of each instance in the test set predicted by the basic nearest neighbour method (where k=1), and the classification accuracy on the test set of the basic nearest neighbour method;

sepal length range: 3.4000000000000004

sepal width range: 2.4000000000000004

petal length range: 5.9

petal width range: 2.4

Classified as: veriscolor|| Actually is: setosa

Classified as: veriscolor|| Actually is: setosa

Classified as: veriscolor|| Actually is: setosa

Classified as: veriscolor|| Actually is: setosa

Classified as: virginica|| Actually is: veriscolor

Classified as: virginica|| Actually is: veriscolor

Classified as: setosa|| Actually is: veriscolor

Classified as: setosa|| Actually is: veriscolor

Classified as: setosa|| Actually is: veriscolor

Classified as: virginica|| Actually is: veriscolor

Classified as: virginica|| Actually is: veriscolor

Classified as: virginica|| Actually is: veriscolor

Classified as: virginica|| Actually is: veriscolor

Classified as: setosa|| Actually is: veriscolor

Classified as: setosa|| Actually is: veriscolor

Classified as: veriscolor|| Actually is: virginica

Classified as: veriscolor|| Actually is: virginica

accuracy = 77.33333333333333% 58.0/75

2. Report the classification accuracy on the test set of the k-nearest neighbour method where k=3, and compare and comment on the performance of the two classifiers (k=1 and k=3);

sepal length range: 3.4000000000000004

sepal width range: 2.4000000000000004

petal length range: 5.9

petal width range: 2.4

Classified as: veriscolor|| Actually is: setosa

Classified as: veriscolor|| Actually is: setosa

Classified as: veriscolor|| Actually is: setosa

Classified as: veriscolor|| Actually is: setosa

Classified as: virginica|| Actually is: veriscolor

Classified as: virginica|| Actually is: veriscolor

Classified as: virginica|| Actually is: veriscolor

Classified as: setosa|| Actually is: veriscolor

Classified as: setosa|| Actually is: veriscolor

Classified as: setosa|| Actually is: veriscolor

Classified as: virginica|| Actually is: veriscolor

Classified as: setosa|| Actually is: veriscolor

Classified as: setosa|| Actually is: veriscolor

Classified as: veriscolor|| Actually is: virginica

Classified as: veriscolor|| Actually is: virginica

accuracy = 80.0% 60.0/75

* Where K = 3 is more accurate than K = 1. This is because when K = 3, we take the majority class of the t3 nearest neighbours, which is more robust to ‘noisy data’.

3. Discuss the main advantages and disadvantages of k-Nearest Neighbour method.

* The main advantage of the nearest neighbour method is its simplicity to implement. It also can be effective classifier if the training data is large enough. However, the cost of computations quite high. Also with distance based learning is not clear which type of distance to use, and which attributes to use to produce the best results, when classifying more .

4. Assuming that you are asked to apply the k-fold cross validation method for the above problem with k=5, what would you do? State the major steps.

* Split the data into 5 equal subsets
* For each subset of the iterate through the subsets– combine the other 4 subsets and use as the training set . Train classifier using the training set,. Then apply it to the test set.
* The training/test process is repeated 5 times, with each of the 5 subsets used exactly once as the test set
* The 5 results from the folds can be then averaged to produce a single estimation.

5. In the above problem, assuming that the class labels are not available in the training set and the test set, and that there are three clusters, which method would you use to group the examples in the data set? State the major steps.

1. Set 3 initial “means” randomly from the data set.

2. Create 3 clusters by associating every instance with the nearest mean

based on a distance measure.

3. Replace the old means with the centroid of each of the k clusters (as

the new means).

4. Repeat the above two steps until convergence.

**Part two:**

ass1-data/part2/hepatitis-training.dat

ass1-data/part2/hepatitis-test.dat

Read 110 instances

ASCITES = True:

SPIDERS = True:

VARICES = True:

FIRMLIVER = True:

Class live

FIRMLIVER = False:

BIGLIVER = True:

STEROID = True:

Class live

STEROID = False:

FEMALE = True:

Class live

FEMALE = False:

ANTIVIRALS = True:

FATIGUE = True:

Class die

FATIGUE = False:

Class live

ANTIVIRALS = False:

Class die

BIGLIVER = False:

Class live

VARICES = False:

Class die

SPIDERS = False:

SPLEENPALPABLE = True:

ANOREXIA = True:

AGE = True:

Class live

AGE = False:

MALAISE = True:

SGOT = True:

Class live

SGOT = False:

HISTOLOGY = True:

Class die

HISTOLOGY = False:

BILIRUBIN = True:

Class live

BILIRUBIN = False:

Class die

MALAISE = False:

Class die

ANOREXIA = False:

Class live

SPLEENPALPABLE = False:

Class die

ASCITES = False:

Class die

classified as: live|| expected class: live

classified as: die|| expected class: live

classified as: live|| expected class: live

classified as: live|| expected class: live

classified as: live|| expected class: live

classified as: live|| expected class: die

classified as: live|| expected class: live

classified as: live|| expected class: live

classified as: live|| expected class: live

classified as: live|| expected class: live

classified as: die|| expected class: die

classified as: live|| expected class: live

classified as: die|| expected class: die

classified as: live|| expected class: live

classified as: die|| expected class: die

classified as: live|| expected class: live

classified as: die|| expected class: live

classified as: die|| expected class: live

classified as: die|| expected class: live

classified as: live|| expected class: live

classified as: live|| expected class: live

classified as: live|| expected class: live

classified as: live|| expected class: live

classified as: die|| expected class: live

classified as: live|| expected class: live

classified as: live|| expected class: live

classified as: die|| expected class: live

number of wrong instances:7

Accuracy = 74.07%

2. You should then construct 10 other pairs of training/test files, train and test your classifiers on each pair, and calculate the average accuracy of the classifiers over the 10 trials. There is a script split-datafile that takes the name of the full data set (eg, hepatitis), the number of training instances, and a suffix for the filenames, and will construct pairs of training and test files. For example ./split-datafile hepatitis.dat 100 run1 will construct the files hepatitis-training-run1.dat and hepatitis-test-run1.dat with 100 and 37 instances respectively. This process need to run 10 times.

Run 01 - number of wrong instances:7

Accuracy = 81.08%

Run02- number of wrong instances:6

Accuracy = 83.78%

Run03-number of wrong instances:3

Accuracy = 91.89%

Run04-number of wrong instances:6

Accuracy = 83.78%

Run05- number of wrong instances:6

Accuracy = 83.78%

Run06-number of wrong instances:8

Accuracy = 78.38%

Run07-number of wrong instances:8

Accuracy = 78.38%

Run08-number of wrong instances:14

Accuracy = 62.16%

Run09-number of wrong instances:8

Accuracy = 78.38%

Run10-number of wrong instances:8

Accuracy = 78.38%

Average = 79.99 %

3. “Pruning” (removing) some of leaves of the decision tree will always make the decision tree less accurate on the training set. Explain (a) How you could prune leaves from the decision tree; (b) Why it would reduce accuracy on the training set, and (c)Why it might improve accuracy on the test set.

1. Traverse the tree, and check if the elimination of any pair of features increase impurity by a desired, small, desired amount. If it this amount is within the threshold. The pair of features are eliminated from the tree and the common parent node becomes the leaf node.
2. It would decrease the accuracy of the training set, because it deletes nodes that create more impurity. The more pure feature tests are, the higher the accuracy.
3. However this can increase accuracy on the test set, as pruning reduces over fitting – a situation where the set is training data ‘too well’ occurs. This happens because noise/random fluctuations are learned by the tree, and thus used as part of classification. Thus minimising the effect that these random fluctuations in the data can have. Will improve performance on the test set.

**Part 3**

2. Explain why evaluating the perceptron’s performance on the training data is not a good measure of its effectiveness. You may wish to create additional data to get a better measure. If you do, report on the perceptron’s performance on this additional data.

You can not use the training set to measure performance of the algorithm because the accuracy will be distorted by over fitting. If the data which you train & test your algorithm is the same, obviously the accuracy will always be higher than expected.