

INTRODUCTION

This report provides a detailed analysis on the coding work done in fulfilment of COMP307:Introduction to Artificial Intelligence assignment 1. All the coding work for this assignment was done in Python 3 in a Linux Ubuntu 14.04 operating system. The following external libraries have been imported in Python: numpy, re, random, Counter, pandas, warnings, pickle and sys. The details of the three parts of the assignment are given below.

PART 1

The first part of the assignment is to implement the k-Nearest Neighbour algorithm, and evaluate the method on the iris dataset provided.

1. The class labels of all the instances in order as predicted by the k-Nearest Neighbours algorithm.

```
[2][2][2][2][2][2][2][2][2][2][2][2][2][2][2][2][2][2][2][2][2][2][4][4][4][4][4][4][4][4][6][4][4][4][4][4][4][4][4][4][4][4][4][4][4][4][6][6][4][6][6][6][6][6][6][6][6][6][6][4][6][6][6][6][6][6][6][6][6][6][6]
```

2→Iris-setosa

4→Iris-versicolor

6→Iris-virginica

When k=1, Accuracy=0.94677

2. When k=1, Accuracy=0.94677

When k=3, Accuracy=0.92

The k=1 classifier shows a better performance than the k=3 classifier. The k=3 classifier is not suitable here because the number of classes is also 3. If the number of classes and the k value are the same there is a chance that all the three nearest neighbours could belong to different classes thereby yielding no usable result.

3. **Advantages:**

Simple to implement

Flexible to feature / distance choices

Naturally handles multi-class cases

Can do well in practice with enough representative data

Disadvantages:

Large search problem to find nearest neighbours

Storage of data

Must have a meaningful distance function

4. The k-fold Cross Validation algorithm can be implemented by the following steps.
 - a. Divide the data into k equal subsets

- b. Take each subset and treat it as the test set. The remaining subsets should be taken as the training set. The classifier can be trained using the training set and then applied to the test set
 - c. The training/test process is repeated k times(the folds) with each of the k subsets used exactly once as the test set
 - d. The k results from the folds can be then averaged to produce a single estimation
5. If the class labels are not available in the training set and the test set i.e the data are unlabelled the k-means Clustering algorithm can be used. The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the K-means clustering algorithm are:
- a. The centroids of the K clusters, which can be used to label new data
 - b. Labels for the training data (each data point is assigned to a single cluster)

Rather than defining groups before looking at the data, clustering allows you to find and analyze the groups that have formed organically. The "Choosing K" section below describes how the number of groups can be determined. The K-means clustering algorithm uses iterative refinement to produce a final result. The algorithm inputs are the number of clusters K and the data set. The data set is a collection of features for each data point. The algorithm starts with initial estimates for the K centroids, which can either be randomly generated or randomly selected from the data set. The algorithm then iterates between two steps:

a. Data Assignment step

Each centroid defines one of the clusters. In this step, each data point is assigned to its nearest centroid, based on the squared Euclidean distance. More formally, if c_i is the collection of centroids in set C, then each data point x is assigned to a cluster based on

$$\operatorname{argmin}_{c_i \in C} \operatorname{dist}(c_i, x)^2$$

where $\operatorname{dist}(\cdot)$ is the standard (L2) Euclidean distance. Let the set of data point assignments for each i th cluster centroid be S_i .

b. Centroid Update step

In this step, the centroids are recomputed. This is done by taking the mean of all data points assigned to that centroid's cluster.

$$c_i = \frac{1}{|S_i|} \sum_{x_i \in S_i} x_i$$

The algorithm iterates between steps one and two until a stopping criteria is met (i.e., no data points change clusters, the sum of the distances is minimized, or some maximum number of iterations is reached).

PART 2

This part involves writing a program that implements a simple version of the Decision Tree (DT) learning algorithm on the hepatitis dataset provided.

1. The classification accuracy of the baseline classifier is 51.29
Accuracy of the decision tree is 58.84
administrator@admin:~/Desktop/COMP307/Part2\$ python3 decisiontree.py
Enter the training set filename
hepatitis-training.dat
[ASCITES < 1.000]
[BIGLIVER < 1.000]
[1]
[STEROID < 1.000]
[ANOREXIA < 1.000]
[AGE < 0.000]
[1]
[1]
[AGE < 0.000]
[0]
[0]
[AGE < 1.000]
[AGE < 0.000]
[0]
[0]
[AGE < 1.000]
[0]
[0]
[SPIDERS < 1.000]
[FIRMLIVER < 1.000]
[SGOT < 1.000]
[AGE < 1.000]
[AGE < 0.000]
[1]
[1]
[AGE < 1.000]
[1]
[1]
[BIGLIVER < 1.000]
[AGE < 1.000]
[0]
[AGE < 1.000]
[0]
[0]

[SPLEENPALPABLE < 1.000]
[ANOREXIA < 1.000]
[1]
[AGE < 1.000]
[0]
[0]
[AGE < 0.000]
[1]
[1]
[AGE < 1.000]
[SGOT < 1.000]
[ANTIVIRALS < 1.000]
[STEROID < 1.000]
[0]
[1]
[AGE < 0.000]
[0]
[0]
[1]
[1]
[VARICES < 1.000]
[0]
[FIRMLIVER < 1.000]
[BIGLIVER < 1.000]
[AGE < 0.000]
[1]
[1]
[STEROID < 1.000]
[ANTIVIRALS < 1.000]
[FEMALE < 1.000]
[0]
[1]
[FATIGUE < 1.000]
[AGE < 0.000]
[1]
[1]
[FEMALE < 1.000]
[0]
[1]
[AGE < 1.000]
[AGE < 0.000]
[1]
[1]
[1]

[AGE < 0.000]

[1]

[1]

2. The 10 pairs of training files were classified using the 'decisiontreetrainer.py' program. The program stores the trained classifier in a pickle file. The pickle file can be accessed by the program 'decisiontreetester.py' which is used for testing the accuracy of the 10 pairs of testing files provided.

Mean accuracy of 10 training files=62.41

3. Pruning works by eliminating nodes that are not clearly relevant. We start with a full tree, as generated by Decision-tree-learning. We then look at a test node that has only leaf nodes as descendants. If the test appears to be irrelevant—detecting only noise in the data then we eliminate the test, replacing it with a leaf node. We repeat this process, considering each test with only leaf descendants, until each one has either been pruned or accepted as is. Suppose we are at a node consisting of p positive and n negative examples. If the attribute is irrelevant, we would expect that it would split the examples into subsets that each have roughly the same proportion of positive examples as the whole set, $p/(p+n)$, and so the information gain will be close to zero. Thus, the information gain is a good clue to irrelevance.

Pruning essentially reduces overfitting in the decision tree. Overfitting is when the program finds patterns in a dataset when the data is random. Pruning cuts out the unwanted data points or nodes thereby reducing overfitting.

4. The value of the impurity measure will always be between 0 and 1 despite the number of classes. So when the number of classes is more than 3 the values get more in accurate.

PART 3

This part of the assignment involves writing a program that implement a perceptron with “random” features that learns to distinguish between two classes of black and white images (X's and O's).

1. The perceptron eventually converged on a correct set of weights. The number of epochs needed depended on the initial random initialisation of the weights and corrections. The converged correct set of weights for one such instance are given below.

[0.08666072118161361,	0.12027826122527774,	0.7305000074653596,
0.5326042531267859,	0.5206860323239939,	0.5589709547855768,
0.4812026769444545,	0.15170188171312138,	0.7467523370232728,
0.5841444561916894,	0.37033688409355325,	0.1480065545220538,
0.44927350684113354,	0.35202993908161595,	0.4109210297658963,
0.41047822723900074,	0.3413080336029777,	0.10026777879637117,
0.690953607577633,	0.552807433130851,	0.7452137913503757,
0.8407596772149913,	0.052761630865427955,	0.045966184192057225,

0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0], [1, 0, 1, 0, 0, 0, 1, 1,
1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0], [1, 1, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 1, 0, 1, 1, 0, 0, 0, 0, 1, 1, 1,
1, 0, 1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1, 1, 1, 1, 0, 1], [1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0], [1, 1, 0,
1, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0,
1, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0], [1, 1, 1, 1, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0,
1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0], [1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,
1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 1, 0,
0, 1], [1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 1, 0, 0, 1, 1, 0, 0, 1, 0, 1, 0,
0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 1, 0], [1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0,
1, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1], [1, 1, 0, 0, 1, 0, 0,
0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0,
0, 0, 1, 0, 1, 1, 1], [1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0,
0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 1, 1, 0, 0, 1, 0, 0, 0, 1, 1, 0], [1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1,
1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0], [1, 0,
0, 1, 0, 0, 1, 0, 1, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0,
1, 0, 1, 1, 0, 0, 0, 1, 0, 1, 0, 1], [1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 1, 0, 1, 1, 0, 0,
0, 1, 0, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1], [1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0,
0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 1, 0, 0, 0, 1, 0, 0, 1,
0, 0, 0], [1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0,
1, 1, 0, 0, 1, 0, 1, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0], [1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1,
0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0], [1, 1, 0, 0, 0, 0,
0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,
0, 0, 0, 1, 0, 0, 1, 1], [1, 1, 0, 1, 0, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0,
0, 0, 1, 1, 0, 0, 1, 0, 1, 1, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0], [1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 1,
0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0], [1,
1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1,
0, 0, 0, 1, 1, 1, 1, 1, 0, 1, 1, 0], [1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0,
1, 0, 1, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1], [1, 0, 1, 0, 0, 0, 1, 0, 0, 1,
0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 1, 0], [1, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 1, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,
0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1], [1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0,
0, 0, 1, 0, 0, 1, 0, 1, 0, 1, 0, 1, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0],
[1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0

[illegible]

