Section A

1. a) Describe the Bagging algorithm in pseudo-code, remembering to give definitions

of any technical words you use, including any properties that you think interesting,

and explaining what models might be suited to the algorithm, and why.

BaggingTrain

for i = 1 to desired committee size  
 take a boostrap of the data  
 train a model (committee member) on the bootstrap

BaggingPredict  
 get the prediction from each committee member  
 assign the testing example to the majority vote  
  
committee member: a model trained on a bootstrap of the data whose prediction will be used to influence the overall prediction of the committee.  
Ideally each model trained should be unstable, that is small changes in the data result in large changes to the model, an example of such a model is a decision tree.

b) Describe the Boosting algorithm, and state the main difference(s) from Bagging.

BoostTrain

for i = 1 to desired committee size  
 take a bootstrap of the data  
 train a model on the bootstrap

see which examples the model got wrong

upweight the hard examples, downweight the easy ones

BoostPredict  
 get the prediction from each committee member  
 assign the testing example to the majority vote

The main difference here is that the Boosting algorithm uses weighted bootstraps to tran models ie the future models should perform better on the testing examples that the earlier ones struggled with. This is unlike in Bagging where each committee member is independent of one another.

c) I recently bought a device that claimed to predict the weather anywhere in the world. Last year in Barbados it rained for 20 of the total 365 days in the year. We will call the rainy days the ‘positive’ class. Over the rainy days, my weather-prediction device correctly predicted rain 15 times. However, several times through the year, it predicted rain when in fact it turned out nice and sunny. I counted these events – it happened 30 times! Calculate the accuracy, sensitivity, and specificity of my weather predictor.

TP = 15 TN = 315  
FN = 5 FP = 30

Sensitivity = TP / (TP + FN) = 15 / (15 + 5) = 15/20 = ⅔  
Specificity = TN / (TN + FP) = 315 / (315 + 30) = 315/345 = 0.913

Accuracy = TP + TN / total = 315 + 15 / 365 = 0.90

d) Related to the previous question, if a false negative (i.e. going out without an umbrella, and it rains) costs me a laundry bill of £20, and a false positive (i.e. buying a new umbrella to use) costs me £5, what is the cost of my classifier?

Cost of FN = 20  
Cost of FP = 5  
  
Cost of classifier = 5\*20 + 5\*30 = £250

2 a) Give pseudo-code for the ID3 algorithm. Be sure to state the base case for the

recursion and be precise when discussing the split criterion. (10 marks)

answer:  
ID3(examples)

if all examples have the same label // a base case  
 return a leaf node with that label

if there are no features left to test // a base case

return a leaf node with most common label in examples

A = feature with maximum information gain  
 create a blank tree T

for each possible value of A, v  
 sv = subset of examples such that A=v  
 tv = ID3(sv)  
 add tv as a new branch of T

return T

In the interest of precision, the ‘split criterion’ is:  
Choose the attribute A with the highest information gain, where information gain is: Entropy(examples) - (\* Entropy(examples with A=v)

2 b) State the formula for the entropy of a feature and calculate it for a binary feature

with p(X = 1) = 0.75.

answer   
The entropy of a collection of examples is equal to:

So number of classes = 2 as feature is binary  
proportion of examples in class 1 = 0.75  
proportion of examples in class 0 = 1 - 0.75 = 0.25  
- (0.75 \* log(0.75) + 0.25 \* log(0.25)) = 0.811

2 c) What are ‘filter’ and ‘wrapper’ approaches to feature selection? Give a definition

and computational complexity properties for each. (5 marks)  
**This doesn’t look familiar to me.**

2 d) What is the relationship between splitting criteria in decision trees, and filter meth-

ods for feature selection?  
**This doesn’t look familiar to me.**

Section B

2. Clustering analysis is an unsupervised learning process that groups a set of physical or abstract objects into clusters of similar or coherent objects.

a) K-means is a popular clustering algorithm. Describe this algorithm in detail and give one advantage and two disadvantages of the K-means algorithm.

K-means algorithm is used to partition a dataset into k clusters (ie groups of similar objects), various partitions of the data are constructed and then evaluated using the sum of squared distances from a cluster centre to the datapoints in that cluster.

set k seed points randomly  
until membership for each cluster no longer changes  
 for each datapoint

calculate distance to each cluster centre

assign datapoint to cluster with nearest centre

for each cluster

set cluster centre to mean position of datapoints belonging to cluster

Advantage: relatively efficient in time and space, requires O(n) space and O(tkn) time  
t = iterations k = number of cluster n = number of datapoints

Disadvantages: Relies on mean which is susceptible to outliers, hence the algorithm does not deal well with noisy data.   
Depends on knowing the number of clusters in advance, this is often difficult.

B)

it is not affected by outlier or noise.

it doesn’t calculate the mean thus other data doesn’t affect it

and it strives to always minimise the cost function which is the end goal of k-mean

c)