CS60020: Foundations of Algorithm Design and Machine Learning
Assignment 4
Submission Deadline: 24th Feb 2018

1.a) It was observed that the train and test error of a decision tree is abnormally high after training it for a document classification task. Assuming there is no error in the tree implementation, what is a possible cause for this behavior?

b) We are learning a random forest using both instance and feature bagging on a data set containing N examples, each of which are described using D attributes. The total number of trees in the forest is T. Each tree is trained using N instances that are sampled at random with replacement from the initial set of N examples. Each tree is trained using F attributes sampled without replacement from the initial set of D attributes.

i) What is the probability of a feature not being selected for training the random forest?
ii) What is the probability of a data point not being selected for training the random forest?

c) Random forest attempts to lower the bias of the decision trees - Is the statement True or False? Explain.

d) Each tree of a random forest is built on a bootstrap sample of the training data. If n is the total number of training samples, on how many samples, on an average, is each tree built?

2. The following dataset will be used to learn a decision tree for predicting whether a mushroom is edible or not based on its shape, color and odour.

SHAPE	COLOR	ODOUR	EDIBLE
C	B	1	Yes
D	B	1	Yes
D	W	1	Yes
D	W	2	Yes
C	B	2	Yes
D	B	2	No
D	G	2	No
C	U	2	No

C	B	3	No
C	W	3	No
D	W	3	No

a) What is entropy H(Edible|Order = 1 or Odour = 3)?
b) Which attribute would the ID3 algorithm choose to use for the root of the tree (no pruning)?
c) Draw the full decision tree that would be learned for this data (no pruning).
d) Suppose we have a validation set as follows. What will be the training set error and validation set error of the tree? Express your answer as the number of examples that would be misclassified.

SHAPE	COLOR	ODOUR	EDIBLE
C	B	2	No
D	B	2	No
C	W	2	Yes

3. Consider a text classification task, such that the document X can be expressed as a binary feature vector of the words. More formally $X = [X_1, X_2, X_3, ..., X_m]$, where $X_j = 1$ if word j is present in document X, and zero otherwise. Consider using the AdaBoost algorithm with a simple weak learner, namely

$h(X;\theta) = y * X_j$
$\theta = \{j,y\}$ j is the word selector ; y is the associated class
$y \in \{-1,1\}$

More intuitively, each weak learner is a word associated with a class label. For example if we had a word football, and classes {sports,non-sports}, then we will have two weak learners from this word, namely
• Predict sports if document has word football
• Predict non-sports if document has word football.

a) How many weak learners are there?
b) This boosting algorithm can be used for feature selection. We run the algorithm and select the features in the order in which they were identified by the algorithm.
    i) Can this boosting algorithm select the same weak classifier more than once? Explain.

ii) Consider ranking the features based on their individual mutual information with the class variable y, i.e. I(y;Xj). Will this ranking be more informative than the ranking returned by AdaBoost ? Explain.

4. Suppose we are given the following dataset, where A,B,C are input binary random variables, and y is a binary output whose value we want to predict.

A	B	C	Y
0	0	1	0
0	1	0	0
1	1	0	0
0	0	1	1
1	1	1	1
1	0	0	1
1	1	0	1

a) How would a naive Bayes classifier predict y given this input: A = 0, B = 0, C = 1. Assume that in case of a tie the classifier always prefers to predict 0 for y.

b) Suppose you know for fact that A,B,C are independent random variables. In this case is it possible for any decision tree based classifier to do better than a naive Bayes classifier? (The dataset is irrelevant for this question)

5. a) Show that the cross-entropy loss function for a two-class logistic regression problem is convex by proving that the Hessian of the function is positive definite.

b) Show that the logistic sigmoid function σ satisfies the property σ(-a) = 1-σ(a) and that its inverse is given by σ^{-1}(y) = ln{y/(1-y)}.

6. a) Suppose you have regression data generated by a polynomial of degree 3. Characterize the bias-variance of the estimates of the following models on the data with respect to the true model by indicating as high or low for both bias and variance and giving a 1 line reason for each.

i) Linear regression.

ii) Polynomial regression with degree 3.

iii) Polynomial regression with degree 10.

b) Let Y = f(X) + ϵ, where ϵ has zero mean and variance σ_ϵ^2. In k-nearest neighbor (kNN) regression, the prediction of Y at point $x_0$is given by the average of the values Y at the k nearest points closest to x0.

i) Denote the l-nearest neighbor to x_0 by x_1 and its corresponding Y value by y(l). Write the prediction f(x0) of the kNN regression for x0 in terms of y(l). 1<=l<=k.

ii) What is the behaviour of the bias as k increases?

iii) What is the behaviour of the variance as k increases?

7. Prove that the decision surface is the perpendicular bisector of the line joining the means of the two classes in Fisher's linear discriminant analysis.