Key

CSci 5523 **Homework 2** **Due: Oct 5th, 2011**

Name: **Student ID:**

Question 1

Consider the training examples shown in Table 1 for a binary classification problem

Table 1: Data set for Exercise 1.

Movie ID	Format	Movie Category	Class
1	DVD	Entertainment	C0
2	Online	Comedy	C0
3	DVD	Documentaries	C0
4	DVD	Comedy	C0
5	Online	Comedy	C0
6	DVD	Comedy	C0
7	Online	Comedy	C0
8	Online	Entertainment	C0
9	Online	Comedy	C0
10	Online	Documentaries	C0
11	DVD	Comedy	C1
12	DVD	Entertainment	C1
13	Online	Entertainment	C1
14	Online	Documentaries	C1
15	Online	Documentaries	C1
16	Online	Comedy	C1
17	Online	Comedy	C1
18	Online	Entertainment	C1
19	Online	Documentaries	C1
20	Online	Documentaries	C1

(a) Compute the Gini index for the overall collection of training examples.

Gini = 0.5

(b) Compute the Gini index for the Movie ID attribute.

Gini (Movie ID) = 0

(c) Compute the Gini index for the Format attribute.

Gini (Format) = 0.47

(d) Compute the Gini index for the Movie Category attribute using multiway split.

Gini (Movie Categorie) = 0.45

(e) Which of the three attributes has the lowest Gini index?

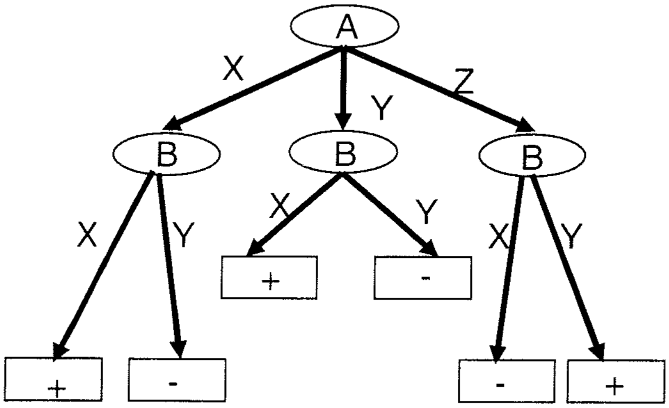
Movie ID

(f) Which of the three attributes will you use for splitting at the root node? Briefly explain your choice

We would use Movie ID since it has the lowest Gini index, excluding than Movie ID. Movie ID is unique for every sample, and would have very poor generalization properties.

Question 2

1) Given decision tree shown below, answer the following questions.



Training		Class	
A	B	Number of + instances	Number of - instances
X	X	6	3
X	Y	1	5
Y	X	18	7
Y	Y	6	29
Z	X	1	10
Z	Y	10	4

Validation		Class	
A	B	Number of + instances	Number of - instances
X	X	5	0
X	Y	4	1
Y	X	5	3
Y	Y	5	13
Z	X	5	2
Z	Y	1	6

(a) Estimate the generalization error rate of the tree using both the optimistic approach and the pessimistic approach. While computing the error with the pessimistic approach, to account for model complexity, use the simple strategy of adding a factor of 0.5 to each leaf node.

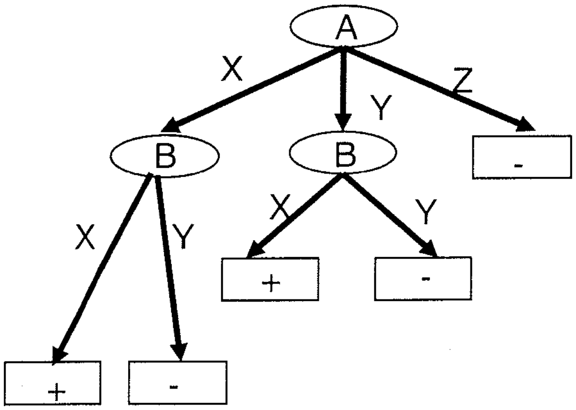
Generalization error rate (optimistic approach) = (3+1+7+6+1+4)/100 = 22/100 = 0.22
Generalization error rate (pesimistic approach) = (22+6*0.5)/100 = 0.25

(b) Estimate the generalization error rate of the tree using the reduced error pruning method with the validation set shown above.

Generalization error rate (validation) = (0+4+3+5+5+6)/50 = 23/50 = 0.46

(c) Based on the error rates obtained above, do you see the issue of model overfitting? If yes, comment briefly on how would you go about pruning the tree to increase its generalization capability? Which branches of the tree would you prune in particular? Compare the performance of the pruned tree with that of the given tree based on the generalization error rate.

Yes, the tree is not able to generalize well for validation data. This can be seen by small error on training set, but large error on validation set. It is important to split on B for A = Y, but may not be important to split on B for A = X and A = Z.
One version of pruned tree could be the following:



Generalization error rate (optimistic approach) = (3+1+7+6+11)/100 = 28/100 = 0.28
Generalization error rate (validation) = (0+4+3+5+6)/50 = 18/50 = 0.36

Error rate on training set increased, but the pruned tree has a better generalization capability as seen in smaller validation error rate.

2) Consider three decision trees, T1, T2 and T3, where T1 is created using classification error as the splitting criterion, T2 is obtained by pruning some of the subtrees in T1, T3 is obtained by pruning some of the subtrees in T2. Which of the following statements could be true? Briefly explain.

a) Training error of T1 > Training error of T2.

Not true. Pruning does not decrease the error rate on training set. Pruned tree T2 will most likely have smaller or unchanged training error of the original tree T1.

b) min(Test error of T2, Test error of T3) ≤ Test error of T1.

True. Pruning is usually done with the goal of decreasing test error. One or both of the pruned trees T2 and T3 will have test error smaller than that of the original tree T1.

c) Test error of T2 ≤ Test error of T3.

It depends. If T3 is pruned to much to the point of under-fitting, both the training and testing error rates will be high. In this case, the statement above will be true.
If, on the other hand, the pruning of T2 was done in order to get smaller test error rate and better generalization, the statement above will be false.

Question 3

You are given a task of predicting whether it will rain tomorrow (+) or not (-) based on a sample of 1000 consecutive days up to today. Given results of a classification algorithm given in the confusion matrix below:

a) Compute accuracy, precision, recall, and F-measure with respect to '-' class.

Accuracy = 0.87
Precision = 0.94
Recall = 0.91
F-measure = 0.93

b) Which of these metrics is a poor indicator of the overall performance of your algorithm? Which of these metrics is a good indicator of the overall performance? Give a one sentence reason why this is the case?

Accuracy is a poor indicator of overall performance of the algorithm. In the case of imbalanced class problem, accuracy does not evaluate well the performance of the algorithm on the smaller class.
F-measure is a good indicator of classifier's performance, since it takes into account both precision and recall.

	Predicted

Actual		+	-
	+	20	50
	-	80	850

c) Construct a trivial rule-based classifier that achieves better accuracy.

Rule based classifier - classify all as negative class:
{} -> -

Accuracy = 0.93

Question 4

Table 1 shows data collected on a runner's decision to go for a run or not go for a run depending on the weather conditions that day. We will use Naïve Bayes (NB) classifier to answer several questions related to this dataset.

Outlook	**Temperature**	**Humidity**	**Run**
Sunny	Hot	High	No
Overcast	Cool	Normal	No
Sunny	Mild	High	No
Overcast	Mild	High	No
Sunny	Hot	High	Yes
Rainy	Hot	High	Yes
Rainy	Mild	High	Yes
Rainy	Cool	Normal	Yes
Rainy	Cool	Normal	Yes
Sunny	Cool	Normal	Yes
Rainy	Mild	Normal	Yes
Sunny	Mild	Normal	Yes
Rainy	Mild	High	Yes
Rainy	Hot	Normal	Yes

Table 1. Running data for Question 4

a) Given the data in Table 1 is a person more likely to go for a run or not? Justify your answer.

P(Run = Yes) = 10/14 > P(Run = No) = 4/14
Person is more likely to run given no information about the weather conditions.

b) How would Naïve Bayes classify an unseen data point X = {Sunny, Mild, Normal}? Show your work. Comment on the behavior of Naïve Bayes in the case of new unseen data.

P(X | Yes) P(Yes) = 0.0514
P(X | No) P(No) = 0.0179

X would be classified as Run = Yes.
NB successfully handles unseen data.

c) Assume that the only information you have about the weather outside is that temperature is mild. What is NB's prediction whether a person will run or not? Show your work.

P(Temperature = Mild | Yes) P(Yes) = 4/14
P(Temperature = Mild | No) P(No) = 2/14.

Given Temperature = 'Mild' a person is likely to run that day.

d) In addition to knowing that temperature is mild that day, you also know that humidity is high. What is NB's prediction whether a person will go for a run or not?

P(Temperature = Mild, Humidity = High | Yes) P(Yes) = 0.114
P(Temperature = Mild, Humidity = High | No) P(No) = 0.107

Given temperature is mild and humidity is high, a person is likely to go for a run.

e) Given results in c) and d) comment on the behavior of Naïve Bayes when handling missing data.

NB successfully handles incomplete data.

f) Now let us go back and compute prediction for a complete data point. In addition to knowing that the temperature is mild and humidity is high, assume you also know that the outlook is overcast. Is a person more likely to go for a run or not.

P(Outlook = Overcast, Temperature = Mild, Humidity = High | Yes) P(Yes) > 3/56
P(Outlook = Overcast, Temperature = Mild, Humidity = High | No) P(No) = 0

A person is not likely to go for a run that day.

g) What went wrong in f)? What approach would you use to fix it? Explain your answer.

Probability of Outlook = Overcast given Run = Yes is zero, which makes the overall posterior probability for this class zero. This can be fixed with the use of m-estimate.

Question 5

(a) You are given a data set from the automobile industry where the data objects are specific car models and the attributes describe the various features of a car. The attributes are:

Type (Small, Sports, Compact, Midsize, Large)
MPG (miles per gallon)
Drive train type (rear wheel drive, front wheel drive, all wheel drive)
Number of cylinders
Engine size (liters)
Horsepower
RPM (revs per minute at maximum horsepower)
Length (inches)
Wheelbase (inches)

Class: Low Price, Average Price, High Price

You are also given training data which has the above information for many car models. The task assigned to you is to build a classification model with the training data so that given a new car model, your model can predict whether it should have a low, average or high price.

Taking into account factors such as interpretability of the model and the presence of domain knowledge, answer the two questions below.

- What aspects of this problem would lead you to use decision trees over the knn classifier?

  Answer: There are irrelevant attributes such as wheel base

- What aspects of this problem would lead you to use a Bayes network over Naïve Bayes?

  Answer: There are correlated attributes such as cylinders and horse power.

(b) If you had to choose between the na ̈ive Bayes and k-nearest neighbor classifiers, which would you prefer for a classification problem where there are numerous missing values in the training and test data sets? Indicate your choice of classifier and briefly explain why the other one may not work so well?

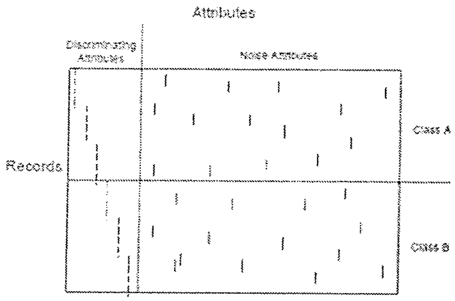
Answer: KNN can not handle missing value, Naïve Bayes can.

(c) Consider the problem of predicting whether a person is a good credit risk given the following attributes: hair color, income, weight, time in current job, marital status, height, age, and birth month. If you had to choose between Ripper and a k-nearest neighbor classifier, which would you prefer? Indicate your choice of classifier and briefly explain why the other one may not work so well?

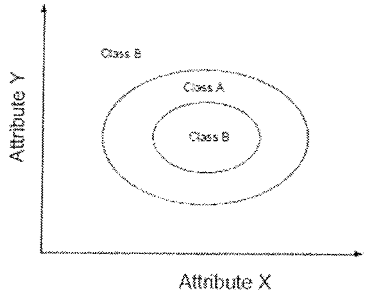
Ripper. Because KNN cannot handle irrelevant attirbutes

Question 6

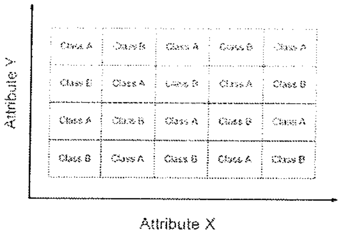
Given the data sets shown in the subfigure below, list the ones that are suitable for each of the following classifiers, naive Bayes (NB), decision tree and k-nearest neighbor (k-NN) classifiers would perform on these data sets. (a) A binary dataset represented as a binary matrix with each row being a record and each column being a attribute. The rows are grouped into two classes A and B. For each variable, the dashed or real vertical lines indicate 1's in the matrix and the rest are 0's. (b & c) Each represents a dataset with two variables (x and y axis) and two classes (A and B) of data points separated by the boundaries.



(a) Synthetic data set 1.



(b) Synthetic data set 2.



Class A	Class B	Class A	Class B	Class A
Class B	Class A	Class B	Class A	Class B
Class A	Class B	Class A	Class B	Class A
Class B	Class A	Class B	Class A	Class B

Answer:

For data 1: There are noisy attributes and thus KNN would not work well. Naïve Bayes and decision tree are both OK.

For data 2: Decision tree will have to be large in order to capture the circular decision boundaries, and thus is not the ideal solution. KNN is the best due to the proximity of the examples of the same class to each other. NB does not work well for this data set since the attributes that determine the class boundaries are not independent

For data 3: Similar as data 2, but here decision tree could work because the boundaries are either vertical or horizontal. KNN is the best.