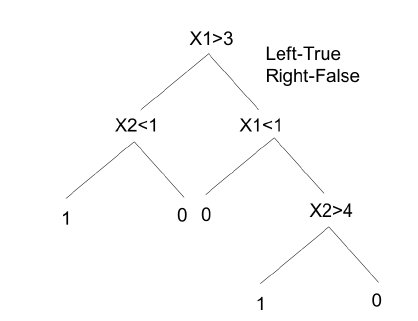
# 

# I526 Online Mid Term

### Created by:

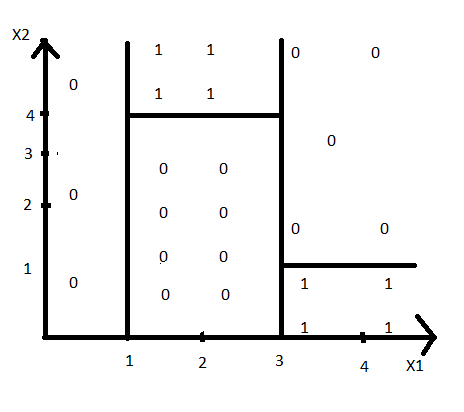
### Himanshu Goyal (hgoyal)

1. *Consider the following decision tree.*

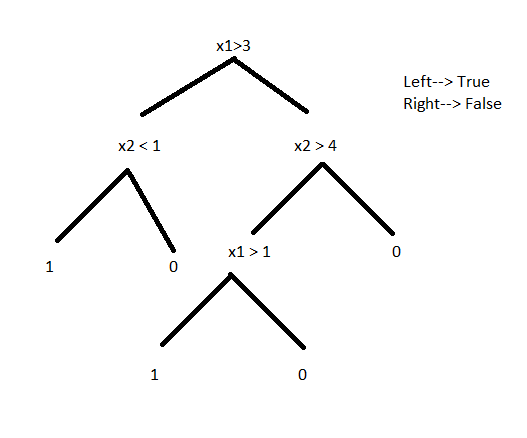


*Draw the decision boundaries defined by this tree. Each leaf is labeled with a number. Write this number in the corresponding region.*

Solution*: a)*



b) Give another tree that is syntactically different but defines the same decision boundaries.



c) When do decision-trees overfit? When do they exhibit high bias?

Decision trees overfit when they are grown such a way that it will reduce the training data error at the cost of increasing test data error or in other words if they start giving highly accurate output on training set data and gives low accurate data on test set.

High bias will happen when decision trees are shorter or in other words attributes with high information are closed to the root attribute and further does not allow the tree to grow.

1. Multiple choice questions and ROC
2. Smaller training sets lead to overfitting.

Answer: True, If the training set is smaller, our model will be biased as we will be predicting the output more accurate.

1. For K-nearest neighbors, large K leads to overfitting.

Answer: False, large K (for K=n) value will have less variance.

1. Logistic Regression learns P(X | Y ) from data and hence is discriminative.

Answer: False, this algorithm aims at learning P(Y|x) by using probabilistic approaches or by mapping classes from a set of points.

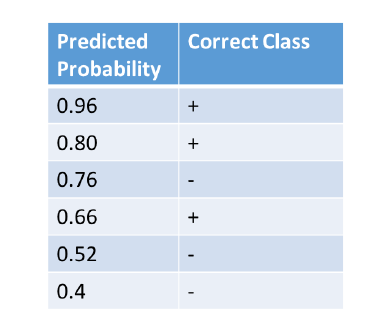
1. Naive Bayes learns P(X | Y ) and P(Y ) from data.

Answer: True, it tries to learn the classifier (P(Y)=0 or 1) given the features as it reduces the complexity from 2(2n-1) to 2n.

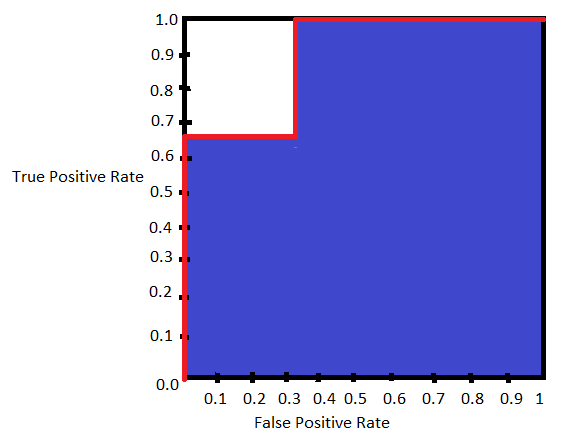
1. Accuracy is a reasonable performance measure when the data is not skewed.

Answer: True, If the data is perfectly balanced than the prediction won't be towards the frequent class.

1. Let the following predictions be the output of a probabilistic classifier. Draw the AUC ROC curve for this prediction task. (5 points)



Answer:



1. When does k-NN have high bias? When does it have high variance? Explain in one sentence why for each case.

Answer:

a)

High Bias:

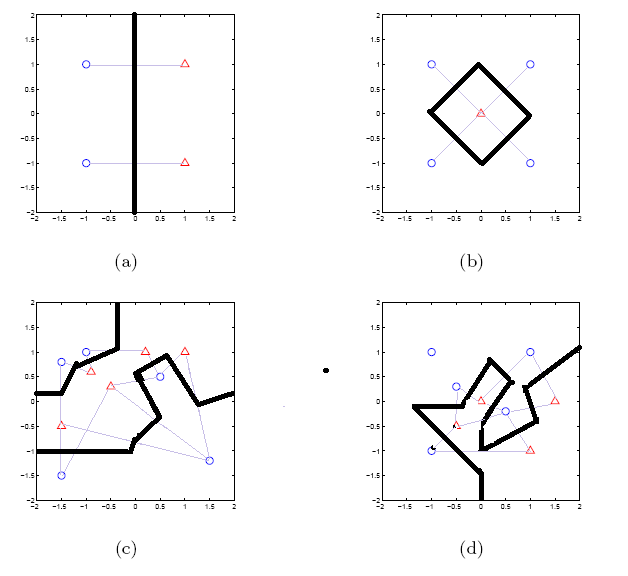
high value of K may lead to high bias; in this case we are going to predict every class.

High Variance:

Small value of K will lead to high variance. in this case everything will be

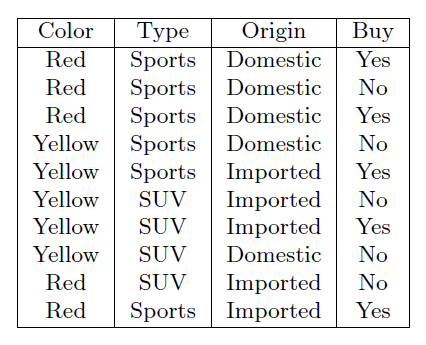
scattered, there will be no pattern.

b) For each of the subfigures below, we are given red(negative) and blue(positive) data points in 2-d space. As discussed in class, a 1-NN classifier classifies a point according to the class of its nearest neighbor. Please draw the decision boundary for a 1-NN classifier using Euclidean distance as the distance metric for each case.



4) Naive Bayes and Evaluation

Consider the following data set. Let the problem be to predict if a car is going to be bought.



1. Assume Naive Bayes condition and estimate the parameters using Maximum Likelihood estimation.

P(Buy=Yes) =5/10=0.5 P(Buy=No) =5/10=0.5

P(O=D|Buy=Y) =2/5=0.4 P(O=D|Buy=N) =3/5=0.6

P(O=I|Buy=Y) =3/5=0.6 P(O=I|Buy=N) =2/5=0.4

P(T=SP|Buy=Y) =4/5=0.8 P(T=SP|Buy=N) =2/5=0.4

P(T=SU|Buy=Y) =1/5=0.2 P(T=SU|Buy=N) =3/5=0.6

P(C=R|Buy=Y) =3/5=0.6 P(C=R|Buy=N) =2/5=0.4

P(C=YE|Buy=Y) =2/5=0.4 P(C=YE|Buy=N) =3/5=0.6

1. For the same domain, apply Laplace correction and estimate the parameters

P(Buy=Yes) =5/10=0.5 P(Buy=No) =5/10=0.5

P(O=D|Buy=Y) =3/7=0.43 P(O=D|Buy=N)=4/7=0.57

P(O=I|Buy=Y) =4/7=0.57 P(O=I|Buy=N) =3/7=0.43

P(T=SP|Buy=Y) =5/7=0.71 P(T=SP|Buy=N)=3/7=0.43

P(T=SU|Buy=Y)=2/7=0.29 P(T=SU|Buy=N)=4/7=0.57

P(C=R|Buy=Y) =4/7=0.57 P(C=R|Buy=N) =3/7=0.4

P(C=YE|Buy=Y) =3/7=0.43 P(C=YE|Buy=N)=4/7=0.6

1. Define precision and recall.

Precision:

Precision is the ratio of number of relevant records retrieved to the total number of irrelevant and relevant records retrieved.

Precision= TP/(TP+FP)

Recall:

Recall is the ratio of number of relevant records retrieved to the total number of relevant records in the datasets.

Recall = TP/(TP+FN)

1. Give an example scenario where minimizing false negatives is more important than minimizing false positives. Briefly justify your answer.

Minimizing False negative is more important in scenarios like cancer prediction, pregnancy test or any medical domain. Even if model has 90% accuracy, it is important to make sure that we are not predicting incorrectly. It’s better to be safe than sorry. For ex, If a person has cancer but if our model predicts that he is not having cancer, it will be a risk for that patient.