CMSC-478 Machine Learning - Spring 2018
Homework Assignment 4
Due at the start of class on May 1st

1. Consider a problem in which the class label $y \in \{T, F\}$ and each training example X has 2 binary attributes $X_1, X_2 \in \{T, F\}$.

   Let the class prior be $p(Y = T) = 0.5$ and $p(X_1 = T | Y = T) = 0.8$ and $p(X_2 = T | Y = T) = 0.5$. Likewise, $p(X_1 = F | Y = F) = 0.7$ and $p(X_2 = F | Y = F) = 0.9$. Attribute X_1 provides slightly stronger evidence about the class label than X_2.

   - Assume X_1 and X_2 are truly independent given Y. Write down the naive Bayes decision rule.

   - What is the expected error rate of naive Bayes if it uses only attribute X_1? What if it uses only X_2?

     The expected error rate is the probability that the naive Bayes classifier is wrong summed over all possible combinations of X values. You compute the probability of the two classes for each possible value of X, find the most probable class in each case, and sum the probabilities of the other class. Those probabilities correspond to probabilities that the classifier is wrong, the probability that the lesss likely class is the true class.

   - Show that if naive Bayes uses both attributes, X_1 and X_2, the error rate is 0.235, which is better than if using only a single attribute (X_1 or X_2).

   - Now suppose that we create new attribute X_3 which is an exact copy of X_2. So for every training example, attributes X_2 and X_3 have the same value. What is the expected error of naive Bayes now?

   - Explain what is happening with naive Bayes? Does logistic regression suffer from the same problem? Explain why.

2. There are two ways to adapt naive Bayes to real-valued attributes. The first is to discretize the real values by, for example, dividing the range of observed values into k non-overlapping, adjacent ranges of the same size and replacing each real value by the discrete index of the range into which it falls. If you take this approach:

   - What is one advantage and one disadvantage of large values of k?

   - How might you choose a "good" value of k?

Another way to approach the problem of real values is to assume that each attribute value is sampled from a normal distribution of mean μ_i and standard deviation σ_i for class i. Describe what quantities are needed for this form of naive Bayes classifier and how you would estimate them from training data. Finally, write down the naive Bayes decision rule for this classifier.