Math 5364 Homework 8

1. For the following problems, consult the `HouseVotes84` data set.

   (a) What is the probability that a randomly selected representative is a Democrat?

   (b) Given that a representative is a Republican, what is the probability that he or she voted for water project cost sharing?

   (c) Given that a representative voted for adoption of the budget resolution, against the physician fee freeze, and for duty free exports, find the probability that he or she is a Democrat (Hint: If `x` is a vector, you can use `rbind(x)` to force R to treat it as a row vector. You can set missing values of `x` to `NA`, and naive Bayes will ignore them.)

2. Investigate normality of the quantitative variables in the `wdbc.data` data set, using Shapiro-Wilk tests, histograms, and qq-plots.

3. After splitting `wdbc.data` into 70% training and 30% test data, compare the test accuracies of naive Bayes, treating the quantitative variables as

   (a) normally distributed variables

   (b) categorical variables with 4 levels.

4. Use 10-fold cross validation to estimate the accuracy of naives Bayes on `wdbc.data`, treating the quantitative variables as

   (a) normally distributed variables

   (b) categorical variables with L levels, $L = 2, 3, \ldots, 10$.

   Which value of L produces the highest accuracy?

5. Investigate the effect of the Laplace smoothing argument in the `naiveBayes` function. What appears to be the optimal value for this argument?