B365 Homework 4

1. Read Ch. 3 (Classification) of *Principles of Data Mining*

2. This problem works with the Chilean Voting data as in chilean_voting.r where the data matrix is x. The age variable can be simplified to retain only the decade of the person by using

   ```
   x[,5] = floor(x[,5]/10)
   ```

   (a) Using this simplification, create a 3-dimensional table on age, education and vote.

   (b) Using this table create a Bayes' classifier to predict the voting status of a person given their decade and education level. You can represent your classifier as a table where the rows account for all possible configurations of the decade and education variables, giving the vote classification for each.

   (c) How would the Bayes' classifier classify a female, post-secondary-educated person from the SA region in their 50's?

   (d) Expalin your degree of confidence in this classification and why you believe this.

   (e) Estimate the *prior* distribution on the vote (Y or N) using the data.

   (f) Separately for both the Yes and No voters, estimate the class-conditional distributions for gender, education, region, and age. For instance, for gender you would need to compute four probabilities:

   $$P(F|Y), P(M|Y), P(F|N), P(M|N)$$

   (g) How would the naive Bayes' classifier classify a female, post-secondary-educated person from the SA region, in their 50s and why?

3. A common medical condition is present in 30% of the population. We have 10 tests for the condition, which do not discriminate particularly well. To be precise, when the condition is present the tests give a positive result with probabilities: .65, .60, .57, .62, .58, .64, .67, .58, .61, .60. However when the condition is not present the test gives a *negative* result with the stated probabilities. We will assume that the tests are independent given the medical condition, thus it is reasonable to use a naive Bayes classifier. Write an R program that does the following $n = 1000$ times

   (a) simulate a boolean variable that behaves like the described medical condition.

   (b) Generate the results of the 10 tests (which depend on whether or not the condition exists)

   (c) Compute the posterior probability that the condition is present, given the test results

   (d) Classify the the instance as either "trait present" or "trait not present"

   (e) Keep a tally of the number of correctly identified individuals and compute the error rate of your classifier.

   It is interesting to see that a collection of rather weak classifiers can perform very well when used collectively.

4. Consider the same problem and suppose that $p1$ is the vector of probabilities given above, while $p2$ is the vector of complementary probabilties. Suppose the we get our test results by

   ```
   runif(1) < p1
   ```

   when the trait is present and

   ```
   runif(1) < p2
   ```

   when the trait is not present.

   (a) Is it true that, when the trait is present the tests give positive results with probabilities, $p1[1], p1[2], \ldots$ as they should?

   (b) Working as you did in the previous problem, compute the error rate for your classifier with the tests created as suggested.

   (c) Your error rate should be significantly worse than in the previous problem. Why is this so?