B565: Homework 4

1. A disease is present in a population with probability $p(D)$. Our test for the disease gives a continuous score, x, according to $p(x|D)$ when the disease is present and $p(x|\bar{D})$ when the disease is not present. That is, $p(x|D)$ and $p(x|\bar{D})$ are the class-conditional densities. You should assume these quantities are *known*.

   Generally speaking, the downside of diagnosing a disease when it is *not* present is much different from failing to diagnose when it *is* present. Suppose that $L(D, \bar{D})$ is the *loss* we incur when a person has the disease but is classified as *not* having the disease, with similar interpretations of $L(\bar{D}, D), L(\bar{D}, \bar{D}), L(D, D)$. We will assume that $L(D, D) = L(\bar{D}, \bar{D}) = 0$ — there is no loss when we classify correctly.

   (a) Suppose we observe a fixed value of x and consider the two possible classiciation decisions: $\hat{c}(x) = D$ and $\hat{c}(x) = \bar{D}$. For each of these two decisions write expressions for the expected loss that will be incurred.

   (b) Find the classification strategy, that leads to the minimum expected loss.

2. Download the "naive_bayes_binary.csv" data from the course web site. These data are for a 3-class classification problem with 10 binary variables. The true class is the 11th column of the data file.

   (a) Using the first half of the data set, train a naive Bayes classifier.

   (b) Using the 2nd half of the data set, classify each vector and construct the confusion matrix. We we have C different classes, then the confusion matrix is the $C \times C$ matrix where the ij entry counts the number of times an observation from class i was classified as from class j.

3. Download the *Student Performance Data Set* at the UCI Machine Learning Repository,

   https://archive.ics.uci.edu/ml/datasets/Student+Performance

   We will use the math data.

   (a) Create a class variable for each student by testing if the final score "G3" satisfies $G3 > 10$ or $G3 \leq 10$. Create a decision tree predicting this class using all other variables *except* "G1" and "G2." Prune the tree to avoid overfitting and submit a plot of your tree by choosing the tree with smallest estimated generalization error.

   (b) For your pruned tree, what is the error rate on the training data and what is the estimated generalization error.

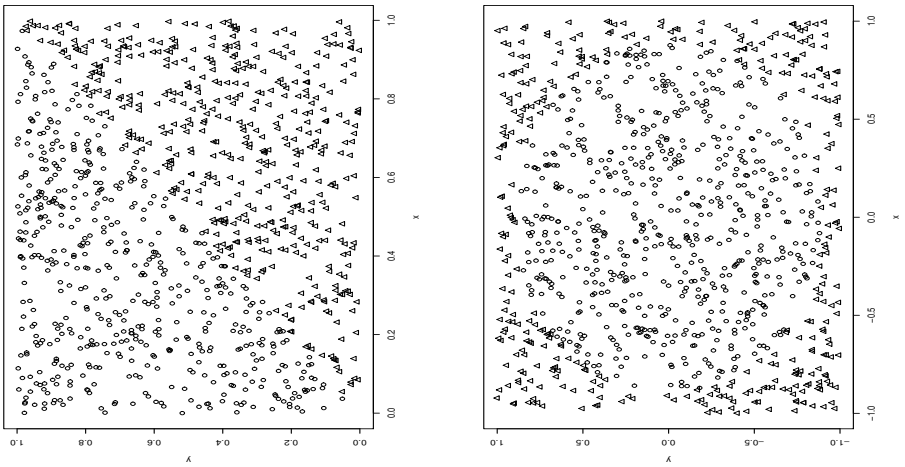
   (c) What appears to be the most useful variable for prediction, and why do you say so?

   (d) You can also use *rpart* to predict the score of a *continuous value*. That is, we can treat the problem as *regression* rather than classification. To do this with *rpart* just change the method to "anova" and use the original continuous "G3" variable. For regression the notion of "error rate" isn't meaningful since we are trying to predict a continuous value, so we use sum of squared errors (SSE) of the predition. That is, if y_i is the true value for the i the observation and $y(\hat{x}_i)$ is our estimate of y, which depends on the features, then

   $$\mathrm{S}SE = \sum_i (y_i - \hat{y}(x_i))^2$$

   Plot the tree and answer the same questions.

4. For the 2 cases below, find new features (functions of x and y) that would increase the efficiency of a classification tree.



5. Using the data in "strange_binary.csv," build a classification tree that distinguishes the "good" examples from the "bad" ones using no more than 3 splits.

   (a) Report the classification error rate on this training set. Is it reasonable to assume that your classification accuracy would be similar on test data from the same model? Explain your answer.

   (b) Introduce an additional feature — a function of the current features — that allows you to significantly decrease the error rate, still using only 3 splits. Report the training error rate for this new classifier. It should be possible to get about 80% correct on the training.

6. A convex function, $k(x)$, of a single variable x satisfies $pk(x_1) + qk(x_2) \geq k(px_1 + qx_2)$ where $p, q \geq 0$ and $p + q = 1$. Jensen's inequality says that for a convex function, k, $E(k(X)) \geq k(E(X))$. Using the fact that $-\log$ is convex, it follows that

$$E(\log(X)) \leq \log(E(X))$$

   (a) Use this inequality to show that the average entropy caused by a tree split is no greater than the original entropy at the node. That is, if q_l and q_r are the proportions of examples going to the left and right nodes and p, p_l, p_r are vectors giving the class distributions at the original, left, and right nodes, then

$$q_l H(p_l) + q_r H(p_r) \leq H(p)$$

   (b) Let C be the class of an example and T be the leaf node of the tree for that example, regarded both as random variables. That is, if we sample radomly from our data set then both the class and the leaf node of the tree associated with the sample are random. Define the conditional entropy of the class given the tree, $H(C|T)$, to be

$$H(C|T) = \sum_t p_t H(C|T = t)$$

   where p_t is the probability of reaching leaf node t and $H(C|T = t)$ is the entropy of the class distribution at leaf node t. Show that each split reduces $H(C|T)$. It is fine to think of all probabilities in this case as proportions.

   (c) The joint entropy of the pair of random variables, (C, T), is defined to be $-\sum_{t,c} p_{t,c} \log p_{t,c}$. Show that

$$H(T, C) = H(T) + H(C|T)$$

   This is a general fact about entropy or "information," not depending on the particular example of classification trees.

7. **Extra Credit** Using the list of 5-letter words, words5.txt on the Canvas page in the data directory, find the word that gives the maximal entropy on the Wordle response. In doing so, you should assume that the hidden Wordle word is chosen so that all 8492 words have equal probability. Submit your code, the word, and the entropy it produces.

A minor modification of this strategy can be used to play the entire game of Wordle, choosing at each stage the word giving the maximal entropy response on the currently-active words. But you may need to modify the strategy so there is a slight bias for choosing a word that is consistent with the currently observed responses.