B565: Homework 3

1. For a mosaic plot of a 2-way table of categorical variables X and Y we know that

   - The column widths are proportional to the marginal counts on the possible outcomes of the 1st variable.
   - The cell areas are proportional to the number of counts for each cell.

   (a) Show that $P(Y = y|X = x)$ is proportional to the *height* of the (x, y) cell.

   (b) Express the independence of X and Y as a statement about $P(Y = y|X = x)$. Phrase this in terms of the appearance of the mosaic plot for the (x, y) counts.

   (c) Construct a probability distribution on 3 *binary* variables, X, Y, Z, so that X and Y are independent when nothing is known about Z, but are not independent when we know $Z = z$.

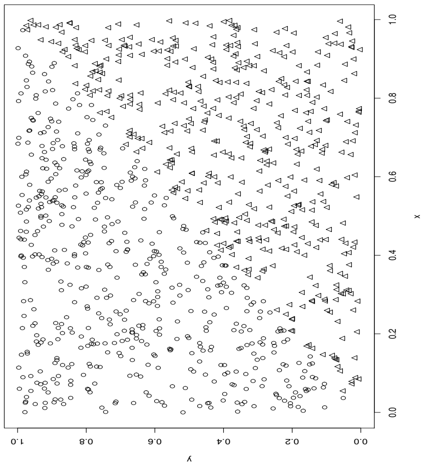
2. Download the *Student Performance Data Set* at the UCI Machine Learning Repository,

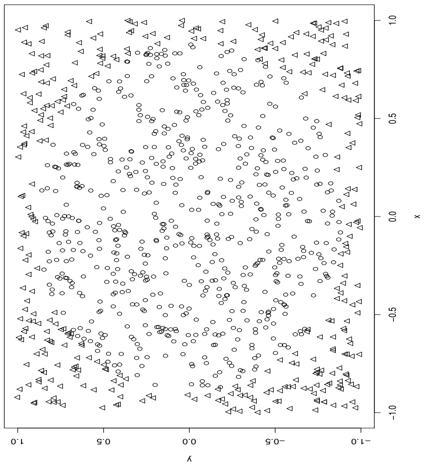
   https://archive.ics.uci.edu/ml/datasets/Student+Performance

   (a) Create a class variable for each student by testing if the final score "G3" is greater than 10. Create a decision tree predicting this class using all other variables (except "G1" and "G2" which are other scores, thus strongly correlated with "G3"). Prune the tree to remove variables that do not show statistically significant improvement. Submit a plot of your tree. What is the estimated generalization error? What is the most important variable? How many splits does your tree use?

   (b) 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 "G3" variable. Plot the tree and answer the same questions as in the first part.

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





4. 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?

(b) Introduce an additional feature 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.

5. Consider the following code fragment which produces a sequence of numbers:

```
> n = 1000;
> x = rep(0,n);
> s = c(1,5,4,2,3);
> k = length(s);
> for (i in (k+1):n) {
>  j = s[(i %% k) + 1];  # i %% k is i mod k
>  x[i] = 1 - x[i-j];
> }
```

Create a classification tree that predicts x_i as a function of the past m values, $x_i, \ldots, x_{i-m+1}$, giving nearly perfect predictions on these data. What is the minimal value for m and the minimal depth of the tree needed to get nearly perfect accuracy?

6. Jensen's inequality says that for a convex function, c, $E(c(X)) \geq c(E(X))$. 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 split is no greater than the original entropy. That is, if q_l and q_r are the proportions going to the left and right nodes and p, p_l, p_r are 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 Y be the class of an example and T be the leaf node of the tree for that example, regarded both as random variables. Define the conditional entropy of the class given the tree, $H(Y|T)$, to be $\sum_t p_t H(Y|T=t)$ where p_t is the probability of reaching leaf node t and $H(Y|T=t)$ is the entropy of the class distribution at leaf node t. Show that each split reduces $H(Y|T)$. It is fine to think of all probabilities in this case as proportions.

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

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

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

7. This problem is taken from the text on page 204 in my edition. The "classification_accuracy" table on canvas gives classification accuracy of 3 different classification techniques: decision trees, naive Bayes, and support vector machines. Compare each pair of techniques on each data set, deciding the comparision as a win, loss, or draw for the first technique of the pair. Produce a 3x3 table with rows labeled by the techniques and columns labeled by win/loss/draw, counting the number of data sets that fall into each cell.