Math 5364 Homework 2

Some of the problems below are quick mathematical proofs. Feel free to do these by hand on notebook paper and attach them, so you don't have to use the Microsoft equation editor. ☺

1. Label each of these variables as nominal, ordinal, or quantitative.
   (a) Weight of contents in a bag of chips.
   (b) City that a person lives in.
   (c) Answers on a survey asking people if they strongly agree, agree, disagree, or strongly disagree with statements.
2. Give examples of the following two types of problems.
   (a) Regression problems.
   (b) Classification problems.
3. Recall that $\log_2 0$ is undefined, and in fact, $\lim_{x\to 0^+} \log_2 x = -\infty$. Therefore, why is it reasonable to define $0 \cdot \log_2 0 = 0$ in the definition of entropy? (Hint: Calculate $\lim_{x\to 0^+} x \log_2 x$.)
4. Consider a classification problem with class distribution $(p_0, p_1, \ldots, p_{c-1})$ (p_i is the fraction of records in class i). The distribution is *pure* if one of the p_i's is 1 and all others are zero (if 100% of the records belong to one of the classes, and therefore, none belong to the other classes). Verify that the entropy, Gini, and classification error impurity measures are equal to 0 for a pure distribution.
5. Consider a two-class problem with class distribution (p_0, p_1). Prove mathematically that the entropy, Gini, and classification impurity measures attain their maximum values when $p_0 = p_1 = \frac{1}{2}$.
6. Write a function in R that accepts a class distribution vector $(p_0, p_1, \ldots, p_{c-1})$ and returns the Gini impurity measure. Repeat for the entropy and classification error measures, and be careful handling 0 for the entropy measure.
7. Quick example of plotting the parabola $y = x^2$, $-5 \le x \le 5$, in R:

```
x=seq(from=-5,to=5,by=0.01)    #Creates a vector from -5 to 5
                               #by increments of 0.01.

mysquare=function(a){          #Function that squares a number.
  return(a^2)}

y=sapply(x,mysquare)           #sapply applies the function mysquare to the
                               #elements of x and stores the results in the
                               #vector y.

plot(x,y,type='l')             #Plot of y versus x.  type='l' connects the
                               #points by lines to create a smooth curve
                               #instead of a scatterplot.
```

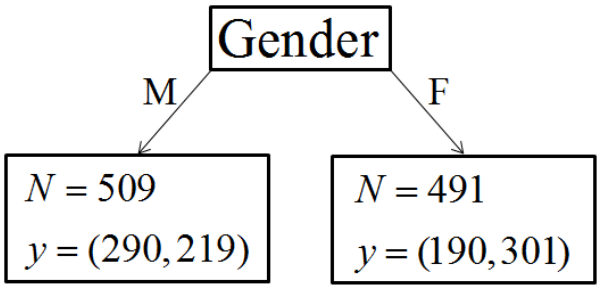
8. Reproduce the plots of the impurity measures on slide 17 of the lecture notes. It's ok to plot them on separate graphs. Note that you cannot use the functions from problem 6 to do this, because they are designed to accept a vector $(p_0, p_1, \ldots, p_{c-1})$. Instead, you could write a new function `twoclassgini` that accepts a number p_0 and returns the gini index for (p_0, p_1) (then repeat for entropy and classification error).

9. The file `Hw2.csv` contains hypothetical data for 1000 people, including the variables gender, car type, shirt size, and a class label.

   (a) Verify that the class distribution for the overall data is the one given below, and verify that the entropy of this distribution is 0.9988.

   $$N = 1000$$
   $$y = (480, 520)$$

   (b) Verify that splitting the data according to the gender variable results in the tree below, and verify that the weighted entropy is 0.9746.



   (c) Draw the tree resulting from a multiway split for car type, and calculate the weighted entropy. Again, feel free to draw this tree by hand, and we will see how to do this using R during the next class.

   (d) Do the same for the shirt size variable.

   (e) Which variable would be the best one for initially splitting the tree?