CS451 - Assignment 1
Data

Due: Friday September 13, at the beginning of class

1. **You know the drill:** Read through the administrative handout on the course web page. *What is the late day policy for the course?*

2. **Data, data, data:** There are now lots of really interesting data sets publicly available to play with. They range in size, quality and the type of features and have resulted in many new machine learning techniques being developed.

   Find a public, free, supervised (i.e. it must have features *and* labels), machine learning dataset. You may NOT list a data set from 1) The UCI Machine Learning Repository or 2) from Kaggle.com. Once you've found the data set, provide the following information:

   (a) The name of the data set.

   (b) Where the data can be obtained.

   (c) A brief (i.e. 1-2 sentences) description of the data set including what the features are and what is being predicted.

   (d) The number of examples in the data set.

   (e) The number of features for each example. If this isn't concrete (i.e. it's text), then a short description of the features.

   Extra credit will be given for particularly interesting data sets, e.g. the most unique, the data set with the largest number of examples and the data set with the largest number of features.

3. **Data analysis:** One of the first things to do before trying any formal machine learning technique is to dive into the data. This can include looking for funny values in the data, looking for outliers, looking at the range of feature values, what features seem important, etc. At:

   `http://www.cs.middlebury.edu/~dkauchak/classes/cs451/assignments/assign1/titanic-train.csv`

   I have provided a modified version of passenger survival data for the Titanic[1].

   This data set has six binary features: `First_class` (whether the passenger was in first class or not), `Sex` (0 = Male, 1 = Female), `Age` (0 = <25, 1 = 25+), `SibSp` (had siblings/spouses aboard?), `ParCh` (had parents/children aboard?) and `Embarked` (Left from Southhampton?). Based on these features, the Titanic task is to learn to predict the last column, whether or not the passenger survived.

   (a) For each of the features calculate (and write down) the *training error* if you used **only** that feature to classify the data. To do this you will need to do the following for each feature:

   - Split the data based on that feature. Call bin_0 all examples that have 0 for that features and bin_1 all examples that have 1 for that feature.
   - Calculate the majority count for the label in each bin, i.e. for bin_0,

   $$majority(bin_0) = max(count(bin_0 = survive), count(bin_0 = notsurvive))$$

     This value is how many examples you would get right in bin_0 if you split on that feature. Make sure you understand why!
   - Calculate the training error for that feature. The accuracy on the training set (i.e. percentage correct) can be calculate as:

   $$accuracy = \frac{majority(bin_0) + majority(bin_1)}{totalNumberOfTrainingExamples}$$

     and then

   $$error = 1 - accuracy$$

   You can either write a program to do this in any language you'd like (you don't need to submit the code) or you could also do this in a spreadsheet program like excel.

   (b) Which feature would be the best to use? Put another way, if we were building a 1-level decision tree using Algorithm 1 from the book, which feature would it pick?

   (c) Do you agree that this is the best choice to make? Just 1-2 sentences explaining yes/no is sufficient.

[1] The original data can be found at: `http://www.kaggle.com/c/titanic-gettingStarted`