**Homework 4: Due in class Thu April 18**

**Problem 1:**

The following link contains a data set on the popularity of new items

<https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity>

The variables of interest are # 2-59, while the variable # 60 can be turned into a class label

1. In the first part of the homework, you should explain how to threshold variable #60 (shares=number of times a news item was shared) and turn it into a categorical variable that would act as a class label.

Then, you should fit a random forest and also a boosted classification tree and compare their performance.

A good boosting package in R is the mlbench; use the boosting command.

1. In the second part of the homework, you should ignore variable #60. Using variables #2-59, you should cluster the observations using at least two methods. Since the size of the data size is large, many of the algorithms (e.g. hierarchical clustering) will not run on your laptop. You should aggressively subsample the data and apply clustering. Comment on how many clusters you selected and whether the different clustering methods give concordant results
2. In the third part of the homework, based on the subsample you used to create the clusters, you should create a new class variable that corresponds to the clusters you identified.

What is the concordance between the class label created by thresholding the shared variable in part (1) and the new class label?

Then, using this new class variable, train random forests and boosted trees classifiers. Then, predict the remaining data that were not used in clustering that led to the creation of the new class label.

Comment on the performance of the two classifiers.

Your report should not exceed 5 pages.