ANLY-512
Final Exam (May 2, 2016)
Take Home Portion

200 points in five problems plus 20 Bonus points. This is the take-home portion of the exam. You may use your notes, your books, all material on the course website, and your computer or any computer in the departmental computer lab. You may also use official documentation for R , built-in or on `https://cran.r-project.org/`, but no other material on the Internet. Provide proper attribution for all such sources. You may not use any human help, except whatever help is provided by me.

Return your solutions by Thursday, 5/12/15, 11:59PM, by e-mail **as a single .RMD file together with the resulting .pdf file**, or hand in printed copies of both files, or fax them to me at 202.687.6067.

The .Rmd file should load all data and all packages, make all plots, and contain all comments and explanation. Set the seed to the year in which you maternal grandmother was born.

Problems 1 – 5 use the `Ozone` data that are available in the `mlbench` package. Be sure to read the description.

1. (10) Load the data, change the names of the variables to `mo, day, wday, maxoz, pressh, wind, hum, temp1, temp2, inverh, pressg, invert, vis`, and also include a variable `time` that runs from 1 to 366. Then generate training data (70%). Use no more that six lines of code.

2. (30) Predict ozone levels from time, using polynomial regression. Choose the best polynomial degree with cross validation. Report the residual standard error obtained from the test data.

3. (30) Predict ozone level from time, using smoothing splines for degrees of freedom ranging from 2 to 40. Summarize the residual standard error obtained from the test data for these degrees, using a suitable plot. Comment on your findings and recommend the best choice.

4. (30) Use best subset selection for linear models to predict ozone levels. Apply the Bayes Information Criterion to choose the best subset and describe the resulting model. Report the residual standard error obtained from the test data.

5. (30) Use the Lasso to predict ozone levels. Choose the best parameter, report the variables which are in the model, and report the residual standard error obtained from the test data.

6. (30) Use bagging, boosting, and random forests to predict ozone levels. Choose suitable parameters for these methods wherever appropriate and explain your choices. Report the residual standard error obtained from the test data.

Bonus (20). It is generally expected that boosting can lead to overfitting. Produce this effect for the `Ozone` data, or explain why you think this is not possible.

The last problem uses the Hepatitis data, available at the UCI Machine Learning Repository at
`http://archive.ics.uci.edu/ml/datasets/Hepatitis`.
The goal is to predict the survival of patients.

7. (40) Propose three different methods to predict patient survival and choose one that appears most promising to you, with explanation. Then use it to build a classifier from a training set of 105 cases and estimate the misclassification rate from the remaining cases.