Project description: Regression or Classification

1. Pick a dataset you like to study and different from what has already been selected by others.

2. Submit a proposal (less than a page) on the Discussion Board on Blackboard in which you:

   (a) Describe the nature of the data.

   (b) Define the objective, i.e. what you are trying to accomplish. Regression or classification?

   (c) What is n and p?

   (d) If classification, then what is the proportion of 1s in the data.

   (e) If regression, what is the R^2 of linear regression.

   (f) Pick a dataset such that:

      - The number of features p is at least 20.
      - The sample size n should be at least ten times the number of features p.

3. For each $n_{learn} \in \{2p, 10p, n/2\}$, repeat the following 100 times, plot the box-plots of the errors of the different models mentioned below, and just for one random split, plot the 10-fold cross validation error. If $n/2 < 10p$, then ignore the $n_{learn} = n/2$ case.

   (a) Randomly split the dataset into two mutually exclusive datasets $D_{validation}$ and D_{learn} with size $n_{validation}$ and n_{learn} such that $n_{learn} + n_{validation} = n$.

   (b) Regression: Use D_{learn} to fit least squares, lasso, elastic-net, ridge, and random forest.

   (c) Classification: Use D_{learn} to fit logistic lasso, logistic ridge, random forest and a radial kernel svm.

   (d) In methods above which require the tuning of a hyper parameter such as λ in lasso and ridge, and C and γ in svm, find them using 10-fold cross validation and loocv (and their respective 1SE rule counterparts). For lasso, in addition to cross validation, we want to see boxplots of the predictive performance for when λ is selected using AIC.

   (e) Regression: For each estimated model calculate the R^2 for validation set as

   $$1 - \frac{\frac{1}{n_{validation}} \sum_{i \in D_{validation}} \left(y_i - \hat{y}_i\right)^2}{\frac{1}{n_{validation}} \sum_{i \in D_{validation}} \left(y_i - \bar{y}\right)^2}.$$

   (f) Classification: For each estimated model calculate the misclassification for the validation set.

4. Create a presentation with less than 15 slides. Your objective is to be as concise as possible. Hence I recommend the following:

(a) a brief description of the nature of the data, shape, etc as discussed above. (1 slide)

(b) Boxplots for each $n_{learn} \in \{2p, 10p, n/2\}$. (3 slides)

(c) For one on the 100 samples, create 10-fold CV curves for lasso and ridge, AIC curve for lasso. For radial svm there are two parameters to be optimized, one is C and the other γ: create a 10-fold CV heat map for svm. See this `http://scikit-learn.org/stable/auto_examples/svm/plot_rbf_parameters.html` for an illustration of the heat map. (3-5 slides).

(d) For lasso and ridge, plot the estimated coefficients and for random forrest, the importance of the parameters, all in the same graph, for each $n_{learn} \in \{2p, 10p, n/2\}$ (3 slides). Note that the graphs must be such that we can clearly see the difference between lasso, ridge and random forest. Something like this `http://joewheatley.net/wp-content/uploads/2014/02/variableImportance3.png` with the difference that you will have three instead of two, and the scale for random forrest will be in percentage but the scale in lasso and ridge will be not be in percentage. The idea is to see if these three statistical learning methods agree more or less on which features are important or not.

5. Bring a USB key with the pdf of your presentation. This is your only chance to present your work. After the presentation, you will receive feedback, and you will be asked to modify your analysis and or your story and upload it on blackboard.

6. Later, you will revise your presentation according to the feedback, and upload it on blackboard together with your code and data.