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 30.
   - The sample size n should be at least ten times the number of features p.

3. For each $n_{learn} \in \{2p, 10p\}$, 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.

   (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 in svm, find them using 10-fold cross validation and loocv (and their respective 1SE rule counterparts).

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

$$1 - \frac{\frac{1}{|D_{validation}|}\sum_{i\in D_{validation}} (y_i - \hat{y}_i)^2}{\frac{1}{n}\sum_{i=1}^{n} (y_i - \bar{y})^2}.$$

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

4. Write a final report and create a presentation. Your objective, when writing the final report, is to be as concise as possible. As you know reports, if long, are not read. The same is true with presentations. Hence I recommend the following:

(a) A 1-3 page (pdf) final report covering the points above. The report should mainly include a brief description of the nature of the data, shape, etc as discussed above.

(b) At most 7 slides (pdf) slide presentation. The presentation must be less than 9 minutes.

(c) R code.

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.