CSE574 Introduction to Machine Learning
Programming Assignment 1
Modeling slump flow of concrete using MLE, Ridge and LASSO regression
Due Date: **March 18, 2018 before midnight**

For programming assignment 1, we will be working with high-performance concrete (HPC) data. The data is available on the UCI Machine Learning Repository, here. The paper describing the data set is:

> Yeh, I-Cheng, *Modeling slump flow of concrete using second-order regressions and artificial neural networks*, Cement and Concrete Composites, Vol.29, No. 6, 474-480, 2007. PDF

The required tasks are:

**TASK 0 Preparation 10 points (state in your report the infrastructure you use)**
Set up your machine learning infrastructure. You may use any language you prefer. The following implementations of GLMNET may be helpful for this assignment.

| | |
|---|---|
| Matlab | http://web.stanford.edu/~hastie/glmnet_matlab/index.html |
| Python | http://web.stanford.edu/~hastie/glmnet_python/index.html |
| R | https://cran.r-project.org/web/packages/glmnet/index.html |
| | https://cran.r-project.org/web/packages/glmnet/vignettes/glmnet_beta.pdf |

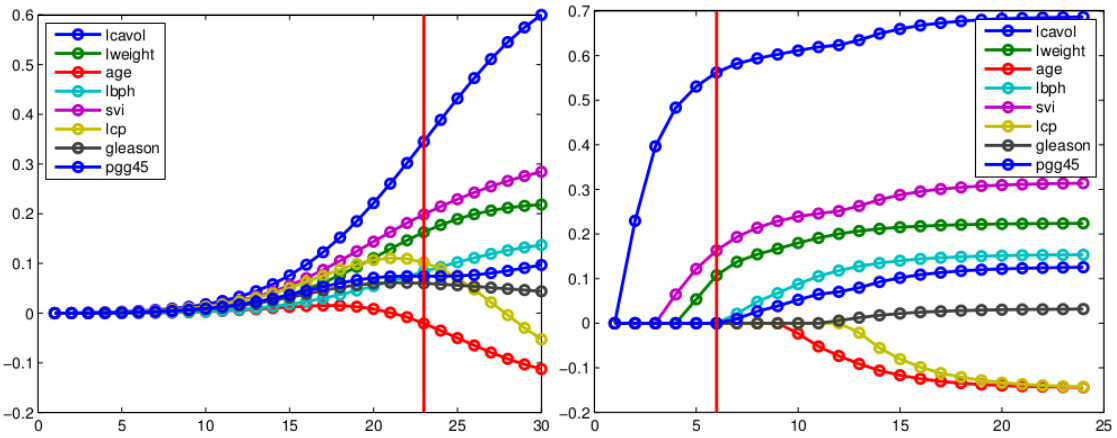
Download the data; and import to your regression environment. The data set contains 103 observations. Implement code to randomly select 85 observations for 5-fold cross validation (5 x 17; 4 x 17 training set; 17 validate set) and 18 observations for a test set. You will iterate this process ten times, reporting the average performance on the 18 observation test sets.
Your response variable (y) is *slump flow*. Your explanatory variables (x) are {*cement, fly ash, slag, water, superplasticizer, coarse aggregate, fine aggregate*}. Therefore, from the data file, on each of ten iterations, you will construct:
85 x 1 vector yTrain; 85 x 7 matrix xTrain; 18 x 1 vector yTest; and, 85 x 7 xTest.
yTrain and xTrain will be further split on each iteration for 5-fold validation to determine optimal regularization parameters.

**TASK 1 Model building 3 x 20 points (submit code; report results)**
Perform the following regressions; saving the best model as determined by cross validation. For the three methods below, compare the average performance against test data (*not used to train or validate your models!*).

1. An unregularized regression of slump flow values against the seven explanatory variables. What is the R-squared? Plot a graph evaluating each regression.
2. A regression regularized by L2 (equivalently, a ridge regression). You should estimate the regularization coefficient that produces the minimum error. Is the regularized regression better than the unregularized regression?
3. A regression regularized by L1 (equivalently, a lasso regression). You should estimate the regularization coefficient that produces the minimum error. How many variables are used by this regression? Is the regularized regression better than the unregularized regression?

**TASK 2 Graph the regularization paths (10 points)**

Graph the regularization paths, as illustrated in Murphy figure 13.7, below (ridge left, lasso right).




**TASK 3 Report (20 points)**

Write up a brief (2-4 pages) report of the machine learning environment you used; the summary comparison of test MSE for your best model in each class (MLE, ridge, lasso); your regularization path graphs for ridge and lasso. Please specify the time you spent on the assignment (to calibrate for future classes), the language you used, the libraries you used, other sources, and the classmates you collaborated with.

**TASK * extra (varies, up to 15 points)**

Consider improvements you could incorporate in the modeling process. For instance, perhaps you could expand the basis functions used for the regularized regressions. For instance, the paper cited above considered the 7 variables x(i), and the 21 interactions x(i)x(j) i!=j. Perhaps x(i)^2, x(i)^3, log(x(i)), … etc.

**Submission requirements**

Each student should submit a zip file name named *<ubperson#>**-project1.zip***

The zip file should contain two folders, *report* and *code*. *report* folder will contain a single pdf file containing your report. *code* will contain all code required to reproduce you analytic and graphical output.

*Remember: submit your own work; collaboration is ok; accessing libraries / other resources on web is ok; cut & paste, not ok! The report and code must be your own, meaning: if we ask you about your work, you must be able to explain it in detail to us.*

**Hints**

Start early. Raise issues on Piazza or in class. Dig into code performing similar tasks:

(matlab) http://people.cs.ubc.ca/~murphyk/MLbook/figReport-16-Aug-2012/pmlFigureCodeTable.html

(python) https://github.com/probml/pyprobml

(R) http://www-bcf.usc.edu/~gareth/ISL/code.html

Good luck!