CS342 Machine Learning: Lab #3
Ridge and Lasso regression

Labs on January 28 & 29, 2016

Week 3 of Term 2

Office Hours:
CS 3.07, Monday & Friday 10:00-11:00

Instructor: **Dr Theo Damoulas** (T.Damoulas@warwick.ac.uk)
Tutors: **Helen McKay** (H.McKay@warwick.ac.uk), **Shan Lin** (Shan.Lin@warwick.ac.uk)

In the third Lab we will explore the use and implementation of Ridge Regression and the Lasso. Refer to the module slides and the Hastie & Tibshirani book for supporting material (module website).

1 Ridge Regression and Lasso

The material here builds on lecture 6. We will be providing the Python (unoptimised) "solutions" a week after each Lab. We are here to assist with your learning experience so if you need help ask us.

Download the prostate dataset from `http://statweb.stanford.edu/~tibs/ElemStatLearn/datasets/prostate.data`. Our goal is to predict the variable **lpsa** given 8 attributes (lcavol, lweight, age, lbph, svi, lcp, gleason, pgg45). There are 97 observations in total. The last column is a train/predict flag that we will use to separate the observations into two sets. The"train = T" set will be used for model fitting and cross-validation and the "train=F" subset for final predictions once we finish with cross-validation and model selection.

→ As usual import your data into pandas data frame(s) and do any necessary pre-processing.

Scikit-learn has a plethora of linear models: `http://scikit-learn.org/stable/modules/classes.html#module-sklearn.linear_model`. It includes all the implementations you will need for this Lab, and then some. In fact, if you are too bored to run your own CV procedure for the next task, it offers implementations of our models (LassoCV, RidgeCV) with built-in CV procedures to choose the best alpha (the λ in our lectures that controls the regularisation strength) from a range of values.

→ Fit the following models to the "train=T" subset of the data: OLS, Ridge Regression, and Lasso

→ Use your fitted models from each type (OLS, Ridge Regression, Lasso) to predict the target of the observations in the train=F set. Compute the Coefficient of Determination, R^2, on this prediction set. Which models did the best? Why?

1.1 On your Own time

→ Fit a decision tree to the same data subset and compare predictive performance. Use what you learned in the lectures to choose the depth of the tree (how complex your model is) and avoid overfitting.