Homework 12 – Intro. to Computational Statistics

Using one or two high-dimensional datasets of your choice, estimate a shrinkage and an SVM model and test them out-of-sample.

A large number of high-dimensional datasets can be found here: http://archive.ics.uci.edu/ml/datasets.html . Be sure to choose those that make your life easier, rather than something that takes a lot of manipulation to get into shape. But feel free to use other data or a dataset you have already used, as long as they have at least 10 independent variables and a continuous dependent variable (for lasso/ridge) and/or a binary dependent variable (for SVM). You can also convert a continous dependent variable to a binary for the SVM stage, as we did in the lesson.

1. Use your dataset with a continuous dependent variable:

a. Divide your data into two equal-sized samples, the in-sample and the out-sample. Estimate the elastic net model using at least three levels of alpha (ie, three positions in between full lasso and full ridge; eg, alpha = 0, 0.5, and 1), using cv.glmnet to find the best lambda level for each run. (Remember that glmnet prefers that data be in a numeric matrix format rather than a data frame.)

b. Choose the alpha (and corresponding lambda) with the best results (lowest error), and then test that model out-of-sample using the out-sample data.

c. Compare your out-of-sample results to regular multiple regression: fit the regression model in-sample, predict yhat out-of-sample, and estimate the error. Which works better?

d. Which coefficients are different between the multiple regression and the elastic net model? What, if anything, does this tell you substantively about the effects of your independent variables on your dependent variable?

2. Repeat the same process using your dataset with a binary dependent variable:

a. Divide your data into an in-sample and out-sample as before, and estimate an SVM using at least two different kernels and `tune` to find the best cost level for each.

b. Chose the kernel and cost with the best results, and then test that model out-of-sample using the out-sample data.

c. Compare your results to a logistic regression: fit the logit in-sample, predict yhat out-of-sample, and estimate the accuracy. Which works better?

d. Can you make any guesses as to why the SVM works better (if it does)? Feel to speculate, or to research a bit more the output of `svm`, the meaning of the support vectors, or anything else you can discover about SVMs (no points off for erroneous speculations!).