Introduction to Computational Statistics INSH 5301

Homework 12

04/06/2020

1. For this example, we are trying to predict the price of a diamond. You can get the data from the ggplot2 package using data(diamonds) – you need to load ggplot2 first with library(ggplot2) –.
1.a. Divide the dataset into two equal-sized samples, the in-sample and the out-sample, the samples have to be random. Estimate an elastic net model using the in-sample data for three different values of α (0, 0.5, and 1), using cv.glmnet to find the best lambda level for each value of α. Remember that glmnet prefers that data be in a numeric matrix, therefore, you need to transform any factor into dummies manually

1.b. Choose the value of α and λ that minimize the MSE, and then test the model using the out-sample data. That is, compute the MSE using the out-of-sample data.

1.c. Compare your out-of-sample results to regular a multiple regression using all the variables in the dataset. That is; (1) fit the standard regression model using the in-sample data and using all the variables, (2) predict the out-of-sample using the estimated paramters in (1), and, (3) compute MSE. Which model works best out-of sample, the multivariate regression or the one estimated in (b)?

2. For this example, we are going to predict the quality of the diamond. For that you need to create a dummy that is equal to one if the quality is Premium or ideal and zero otherwise.

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

2.b. Chose the kernel and cost with the best in-sample performance, and then test that model out-ofsample using the out-sample data. That is, compute the percentage of correct predictions using the out-of-sample data.

2.c. Compare your out-of-sample results with a logistic regression using all the variables in the dataset. That is; (1) fit the standard logistic regression model using the in-sample data and using all the variables, (2) predict the out-of-sample data using the estimated paramters in (1), and, (3) compute the model accuracy as the percentage of correct predictions. Which model works best out-of sample, the logistic regression or the one estimated in (b)?