ML01 – Spring 2019
Lab 6: Model selection

1 Prostate data

We consider again the prostate data (see Labs 2 and 5). We recall that lpsa is the response variable, and the other variables are predictors (except the variable train that is used to distinguish training and test data).

1. Using the training data, generate different regression models using the following methods :
   — Best subset selection
   — Forward and backward selection
   — Ridge
   — Lasso
   For subset selection methods, keep the best models according to adjusted R^2 and BIC. For ridge and lasso, select the best model using cross-validation.
2. Evaluate the models selected in the previous step using the test data.

2 Vowel data

We consider again the Vowel data. We recall that this dataset has six classes corresponding to the 6 vowels in English, and 10 predictors.

1. Split the data into a training set (approximately 2/3 of the data) and a test set.
2. Using the training data, estimate the error rates of the LDA, QDA, naive Bayes and logistic regression classifiers using 5-fold cross-validation. Select the classifier with the smallest cross-validation error rate.
3. Compute the test error rate of the best classifier selected in the previous step.