CS 76140 - Spring 2017 Olga Vitek

Homework 3

Due on Blackboard before 5pm on Thursday February 24, 2017.

Note: Use any tool of your choice (including Word, latex, Markdown, or pencil+paper) to prepare the solutions. The answers should be easy to find and grade. Unreadable hand-written solutions will be given 0 points, at the grader's discretion, regardless of the correctness of the answer. For each problem, use the appropriate notation for random variables, probabilities etc. State the full formula in addition to the numerical conclusions. For data analysis problems, use reproducible research tools such as R Markdown whenever possible.

For this homework we will consider the dataset "South African Heart Disease" available on the HTF website, and used as an example in HTF Section 4. The goal is to predict the binary variable `chd`.

We will compare four classification procedures: Logistic regression, linear discriminant analysis, logistic regression with Lasso regularization, and nearest shrunken centroids.

1. **(10pts)** Data preparation and exploration
   (a) **Select the training set:** Download the data. Partition the dataset into a training and a validation subsets of equal size, by randomly selecting rows in the training set.
   (b) **Data exploration:** Consider the training set only. Report one-variable summary statistics, two-variable summary statistics, and discuss your findings (e.g., presence of highly correlated predictors, categorical predictors, missing values, outliers etc).
2. **(10pts)** Fit logistic regression on the training set. Perform variable selection using all subsets selection and AIC or BIC criteria. [Hint: it may be interesting to also consider statistical interactions]
3. **(10pts)** Fit LDA on the training set, using the standard workflow.
4. **(10pts)** Fit logistic regression with Lasso regularization on the training set.
   (a) Produce and interpret the plot of paths of the individual coefficients.
   (b) Produce the plot of regularized parameter versus cross-validated predicted error.
   (c) Select regularization parameter, and refit the model with this parameter.
   (d) Fit the model with the selected predictors only on the full raining set.
5. **(10pts)** Fit the nearest shrunken centroids model on the training set.
   (a) Use cross-validation to select the best regularization parameter.

(b) Refit the model with the selected regularization parameter

(c) Visualize the centroids of the selected model

6. **(15pts)** Evaluate the performance of the classifiers

(a) Evaluate the performance of the classifiers using ROC curves on the training set.

(b) Evaluate the performance of the classifiers using ROC curves on the validation set.

(c) Summarize your findings. How do the results differ between the training and the validation set? Which approach(es) perform(s) better on the validation set? What is are the reasons for this difference in performance? Which models are more interpretable?

7. **(10pts)** KM problem 4.20