IEOR 165 – Course Project
Due Wednesday, May 3, 2017

The course project must be submitted on bCourses as a PDF file. You are allowed to work in groups of up to 5 total students, and each group only needs to turn-in a single writeup. The project will be graded on the basis of the quality of the modeling approach.

1. In the paper: P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, Modeling wine preferences by data mining from physicochemical properties, *Decision Support Systems*, vol. 47, no. 4:547-553, 2009. the authors considered the problem of modeling wine preferences. Wine can be evaluated by experts who give a subjective score, and the question the authors of this paper considered was how to build a model that relates objective features of the wine (e.g., pH values) to its rated quality. For this homework, we will use the data set available at: http://ieor.berkeley.edu/~ieor165/homeworks/winequality-red.csv

   Use the following methods to identify the coefficients of a linear model relating wine quality to different features of the wine: (1) ordinary least squares (OLS), (2) ridge regression (RR), (3) lasso regression, (4) elastic net. Make sure to include a constant (intercept) term in your model, and choose the tuning parameters using cross-validation. You may use any programming language you would like to. For your solutions, please include (i) plots of tuning parameters versus cross-validation error, (ii) coefficients (labeled by the feature) computed by each method, (iii) the minimum cross-validation error for each method, and (iv) the source code used to generate the plots and coefficients. Some hints are below:

   - a constant (intercept) term can be included in OLS by solving

   $$\begin{bmatrix}\hat{\beta}_0 \\ \hat{\beta}\end{bmatrix} = \arg\min_{\beta_0,\beta} \left\| Y - \begin{bmatrix}\mathbf{1}_n & X\end{bmatrix} \begin{bmatrix}\beta_0 \\ \beta\end{bmatrix} \right\|_2^2$$

   - RR and lasso have one tuning parameter, while elastic net has two tuning parameters
   - RR (with an intercept term) can be formulated as

   $$\begin{bmatrix}\hat{\beta}_0 \\ \hat{\beta}\end{bmatrix} = \arg\min_{\beta_0,\beta} \left\| \begin{bmatrix}Y \\ 0\end{bmatrix} - \begin{bmatrix}\mathbf{1}_n & X \\ 0 & \mu \cdot \mathbb{I}\end{bmatrix} \begin{bmatrix}\beta_0 \\ \beta\end{bmatrix} \right\|_2^2,$$

     where μ is a tuning parameter.

2. In a data set derived from N. Abreu, Analise do perfil do cliente Recheio e desenvolvimento de um sistema promocional, Mestrado em Marketing, ISCTE-IUL, Lisbon, 2011., the data

consists of annual spending on different categories of products by each retailer, along with corresponding data on the sales channel (Hotel/Restaurant/Cafe = 1 vs. Retail = 2) and the region (Lisbon = 1, Oporto = 2, or Other = 3) of the retailer. For this homework, we will use the data set available at:
http://ieor.berkeley.edu/~ieor165/homeworks/wholesale-customers.csv

Use the following methods to identify the coefficients of an SVM that uses spending on different categories of products to predict the sales channel: (1) linear SVM, (2) SVM with polynomial kernel, (3) SVM with Gaussian kernel. Choose the tuning parameters using cross-validation. You may use any programming language you would like to. For your solutions, please include (i) plots of tuning parameters versus cross-validation error, (ii) coefficients computed by each method, (iii) the minimum cross-validation error for each method, and (iv) the source code used to generate the plots and coefficients. For the linear SVM, relate the coefficients to the features.