Linear Regression and Regularized Regression

COS424: Assignment # 3

Due : Thursday, April 12th, 2012

Turn in a hard copy of the assignment for all questions in class on Thurday, April 12th. Submit your code and data file (Name the file of your code as Question2.R and your data file as Question2.txt) for Question 2 to CS DropBox at
http://dropbox.cs.princeton.edu/COS424_S2012/Homework_3 before class.

Written Exercise

Question 1: (40 points) Regularized Regression

As is usual for linear regression, suppose we are given training data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$ where $y_i \in \mathbb{R}$ and $\mathbf{x}_i \in \mathbb{R}^n$ (with components x_{ij}). In this problem, we seek linear models of the form $\hat{f}(\mathbf{x}) = w_0 + \mathbf{w} \cdot \mathbf{x}$ where w_0 is the scalar intercept term, and $\mathbf{w} = \langle w_1, \ldots, w_n \rangle$ is a (column) vector of weights over the n inputs. Consider the problem in ridge regression of minimizing

$$\sum_{i=1}^{m} (w_0 + \mathbf{w} \cdot \mathbf{x}_i - y_i)^2 + \lambda \left\|\mathbf{w}\right\|_2^2 . \tag{1}$$

Here, as in Hastie et al. (but unlike in class), we include an explicit intercept term w_0, but omit this term from the regression penalty.

a. Suppose *for this part only* that $\sum_{i=1}^{m} x_{ij} = 0$ for all j. Let $\mathbf{X}$ be the $m \times n$ matrix of all inputs in which the i-th row is equal to (the transpose of) $\mathbf{x}_i$, and let $\mathbf{y}$ be the (column) vector whose i-th entry is y_i. Show that the solution of (1) is given by

$$\begin{aligned} \hat{w}_0 &= \frac{1}{m} \sum_{i=1}^{m} y_i \\ \hat{\mathbf{w}} &= (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y} \end{aligned}$$

where $\mathbf{I}$ is the $n \times n$ identity matrix.

b. Returning to the general case (in which the input vectors do not sum to zero), let

$$a_j = \frac{1}{m} \sum_{i=1}^{m} x_{ij}$$

and define $\mathbf{x}'_i$ by

$$x'_{ij} = x_{ij} - a_j$$

Note that, after centering in this fashion, the new input vectors sum to zero so that the technique in the last part can be applied. Show that minimizing (1) is equivalent to minimizing

$$\sum_{i=1}^{m} (w'_0 + \mathbf{w}' \cdot \mathbf{x}'_i - y_i)^2 + \lambda \left\|\mathbf{w}'\right\|_2^2 . \tag{2}$$

In other words, if $\{ \hat{w}_0, \hat{\mathbf{w}} \}$ is the solution that minimizes (1), and $\{ \hat{w}'_0, \hat{\mathbf{w}}' \}$ is the solution that minimizes (2), show that

$$\hat{w}_0 + \hat{\mathbf{w}} \cdot \mathbf{x} = \hat{w}'_0 + \hat{\mathbf{w}}' \cdot \mathbf{x}'$$

for any $\mathbf{x}$ and its transform $\mathbf{x}'$. Moreover, given a solution $\{ \hat{w}'_0, \hat{\mathbf{w}}' \}$ of (2), show explicitly how to transform it directly into a solution $\{ \hat{w}_0, \hat{\mathbf{w}} \}$ of (1).

c. Suppose that the inputs are both centered *and* scaled. In other words, suppose we instead define $\mathbf{x}'_i$ by

$$x'_{ij} = (x_{ij} - a_j)/s_j$$

for some constants s_j. Show that the minimization problems (1) and (2) need no longer be equivalent (in the sense described above). Show nevertheless how a solution $\{ \hat{w}'_0, \hat{\mathbf{w}}' \}$ of (2) can be transformed back into $\{ \hat{w}_0, \hat{\mathbf{w}} \}$, which is not necessarily a solution of (1), but for which

$$\hat{w}_0 + \hat{\mathbf{w}} \cdot \mathbf{x} = \hat{w}'_0 + \hat{\mathbf{w}}' \cdot \mathbf{x}'$$

for any $\mathbf{x}$ and its transform $\mathbf{x}'$.

Programming Exercises

Question 2: (60 points) Linear Regression

Collect at least 50 data points (x_i, y_i) of inputs x_i and a real-valued response y_i. You can measure the data yourself, or find an interesting data set on the web (including the UCI repository). Other places to look at are data.gov, kaggle and the many data sets built into R.

First, let us consider a single covariate. Choose one of the covariates in the data set.

a. Make a scatter plot of the data, with the covariate in the x axis.

b. Fit the data with a linear regression model and add the regression line to the scatter plot. In this part, please implement a function to fit the regression and turn in your code. We suggest centering the response variable and the covariate first (i.e., subtracting its mean) and omitting the intercept term from the regression.

c. Compute the mean predictive L1 distance with 5 fold cross validation.

d. Form the scatter plot of predicted responses versus observed responses. What do you notice? Why might they not lie on a single line as they would for a full in-sample fit?

Now, consider all the covariates. Here you can use the function lm() in R and are free to keep the responses and covariates uncentered.

a. Use the function summary() to identify which covariates are ”significant” and which are not. Do interesting patterns emerge? What does this suggest about the relationship between the covariates and the response?

b. Compute cross-validated mean predictive L1 distance (as in part c above) with the full model. How does it compare? For the same scatter plot as above, form a plot that shows both models(1-covariate and full model), in different colors.

c. Add an interaction term, one that might make sense. Again use summary() to determine if it makes a difference. Compute cross-validated mean predictive L1; does it make a practical difference? (Feel free to try more than one interaction term, if you like.)

Question 3: (20 points) Extra Credit

Use the packages available in R to examine regularized regression on your data, lasso and/or ridge. Again, compute cross-validated mean predictive L1. Plot the average value of this measure as a function of the regularization parameter. What is the best regularization parameter? How does this compare to the unregularized case?