ST790-003, Homework 4 (Updated Feb 26)

Due Friday, Mar 16, 2015 @ 11:59PM

Linear Programming

This homework explores applications of linear programming (LP) in statistics. Please do Q1, Q2 or Q3, Q4 or Q5, and Q6. This is a *solo* homework. Discussion with fellow students is allowed but you have to write your code and report independently.

1. (Prostate cancer data) Read in the prostate cancer data from the `prostate.txt` file and consult the `prostate.info` file for description. Section 3.2.1 of the book *Elements of Statistical Learning* by Hastie et al. (2009) also describes this data set. The response variable is log of prostate-specific antigen (`lpsa`). The predictors are log cancer volume (`lcavol`), log prostate weight (`weight`), `age`, log of amount of benign prostatic hyperplasia (`lbph`), seminal vesicle invasion (`svi`), log of capsular penetration (`lcp`), Gleason score (`gleason`), and percent of Gleason scores 4 or 5 (`pgg45`).

   (a) Center and standardize all predictors to have mean 0 and variance 1.

   (b) Split data into the training and test sets, according to the labels in the last column in `prostate.txt`.

   (c) Fit a linear regression (with intercept) on the training set. Report the estimated regression coefficients and plot the residuals versus observation number. What is the prediction error on the test set, defined as $\sum_{i \in I_{\text{test}}} (y_i - \hat{y}_i)^2 / n_{\text{test}}$?

2. (ℓ_1 and ℓ_∞ regressions)

   (a) Make the last observation in the prostate cancer training set an outlier by multiplying its response value y_i by 10.

   (b) Re-fit linear regression on the perturbed training data set. Plot residuals versus observation number.

   (c) Fit ℓ_1 regression (using an LP solver) on the perturbed training data set. Plot residuals versus observation number.

   (d) Fit ℓ_∞ regression (using an LP solver) on the perturbed training data set. Plot residuals versus observation number.

   (e) Compare the estimated regression coefficients and prediction errors of ℓ_1, ℓ_2 and ℓ_∞ regressions obtained in this question to those obtained in Q1. Summarize your findings.

3. (Quantile regression)

   (a) Make the last observation in the prostate cancer training set an outlier by multiplying its response value y_i by 10.

(b) Re-fit linear regression on the perturbed training data set. Plot residuals versus observation number.

(c) Fit quantile regressions with $\tau = 0.25$, 0.5, and 0.75 on the perturbed training data set.

(d) Compare the estimated regression coefficients and prediction errors of the three quantile regressions to those obtained in Q1. Summarize your findings.

4. (Lasso penalized ℓ_1 regression) Assume the first column of $\boldsymbol{X}$ is intercept. Solve, using an LP solver, the Lasso penalized ℓ_1 regression

$$\widehat{\boldsymbol{\beta}}(\lambda) = \arg\min \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_1 + \lambda \sum_{j=2}^{p} |\beta_j|$$

on the original (unperturbed) prostate cancer training data at $\lambda = 0, 1, 2, \ldots, 45$ and plot the solution path (excluding intercept). Also plot the prediction errors of $\widehat{\boldsymbol{\beta}}(\lambda)$ on the test set over λ.

5. (Dantzig selector) Candes and Tao (2007) propose a variable selection method called the Dantzig selector that solves

$$\begin{aligned} \text{minimize} \quad & \|\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\|_\infty \\ \text{subject to} \quad & \sum_{j=2}^{p} |\boldsymbol{\beta}_j| \le t, \end{aligned}$$

assuming the first column of $\boldsymbol{X}$ is intercept. Use an LP solver to obtain Dantzig selector solutions at $t = 0, 0.1, 0.2, \ldots, 2.5$ on the original (unperturbed) prostate cancer training data and plot the solution path (excluding solution path). Also plot the prediction errors of $\widehat{\boldsymbol{\beta}}(t)$ on the test set over t.

6. (1-norm svm) In two-class classification problems, we are given training data $(\boldsymbol{x}_i, y_i)$, $i = 1, \ldots, n$, where $\boldsymbol{x}_i \in \mathbb{R}^p$ are feature vectors and $y_i \in \{-1, 1\}$ are class labels. Zhu et al. (2004) propose the 1-norm support vector machine (svm) that achieves the dual purpose of classification and feature selection. Denote the solution of the optimization problem

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{n} \left[1 - y_i \left(\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j \right) \right]_+ \\ \text{subject to} \quad & \|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j| \le t \end{aligned}$$

by $\hat{\beta}_0(t)$ and $\widehat{\boldsymbol{\beta}}(t)$. 1-norm svm classifies a future feature vector $\boldsymbol{x}$ by the sign of fitted model

$$\hat{f}(\boldsymbol{x}) = \hat{\beta}_0 + \boldsymbol{x}^T\widehat{\boldsymbol{\beta}}.$$

(a) Read in the South African heart disease data set in the `saheart.txt` file. The response variable is the presence or absence of myocardial infarction (MI) at the time of surgery.

The predictors are `sbp`, `tobacco`, `ldl`, `famhist`, `obesity`, `alcohol`, and `age`. Section 4.4.2 of the book *Elements of Statistical Learning* by Hastie et al. (2009) also describes this data set. Standardize predictors to have mean 0 and variance 1 and add intercept. Then split data into the training and test sets, according to the labels in the last column in `saheart.txt`.

(b) Obtain, using an LP solver, the 1-norm svm solutions $\hat{\beta}_0(t)$, $\widehat{\boldsymbol{\beta}}(t)$ at $t = 0, 0.1, 0.2, \ldots, 2.5$ on the training set. Plot the solution path of $\widehat{\boldsymbol{\beta}}(t)$ and misclassification rates on the test set over t.

References

Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer Series in Statistics. Springer, New York, second edition.

Zhu, J., Rosset, S., Tibshirani, R., and Hastie, T. J. (2004). 1-norm support vector machines. In Thrun, S., Saul, L., and Schölkopf, B., editors, *Advances in Neural Information Processing Systems 16*, pages 49–56. MIT Press.