STT 873 HW3

This HW is due on Oct 30th.

Question 1(email the R code directly to zhang318@stt.msu.edu and xyy@egr.msu.edu): Compare the performance of LS, Best subset, Ridge Regression, LASSO, PCR and PLS on the Prostate Cancer Data. Use the same procedure on data preparation as used in the textbook (Sections 3.2.1 and 3.3.4). The Prostate Cancer Data are available from the book website www-stat.stanford.edu/ElemStatLearn. Report both the R code and the results.

Question 2 (email the R code directly to zhang318@stt.msu.edu and xyy@egr.msu.edu):

(1) Write a R code to generate the Lasso solution path and analyze the following dataset: https://goo.gl/h779vj (You can compare your result with the R packages 'Lars' or 'elasticnet'. Details of the algorithm can be found in Section 5.6 of the book "Statistical Learning with Sparsity").

(2) Now let's deal with some highly correlated X. Download data from the following link: https://goo.gl/xSHnWr
find the solution path for Lasso and Elastic net. Report the results and explain the differences.

(3) Report boxplots of 1000 bootstrap realizations of $\hat{\beta}^*(\hat{\lambda}_{\mathrm{CV}})$ similar to Figure 6.4 in the book "Statistical Learning with Sparsity".

(4) For $\lambda = 0.1$, run the post selection inference for Lasso using the second dataset. Report your result using figure similar to Figure 6.12 in the book "Statistical Learning with Sparsity".

Question 3 (Gaussian maximal inequalities): In class, we extensively used the following Gaussian maximal inequality to bound the slow rate, which you will prove here, over the next few parts. If $W_i \sim N(0, \sigma_i^2), i = 1, \ldots, p$ are Gaussian variates, not necessarily independent, then for any $\delta > 0$,

$$\mathbb{P}(\max_{i=1,\ldots,d} |W_i| \leq \sigma\sqrt{2\log(ed/\delta)}) \geq 1 - \delta, \tag{1}$$

where $\sigma = \max_{i=1,\ldots,d} \sigma_i$.

(1) Prove that, for any $t > 0$,

$$\mathbb{P}(\max_{i=1,\ldots,d} |W_i| \geq t) \leq 2d\frac{\phi(t/\sigma)}{t/\sigma},$$

where ϕ is the standard normal density. Hint: you may use Mills inequality.

(2) Using the result from the previous part, plug in $t = \sigma 2\sqrt{\log(ed/\delta)}$ and establish (1).

The result (1) is a high-probability bound on the maximum of Gaussians. We can also establish an expectation bound,

$$\mathbb{E}(\max_{i=1,\ldots,d} |W_i|) \leq \sigma\sqrt{2\log(2p)}. \tag{2}$$

(3) To prove that, for any $t > 0$

$$\mathbb{E}(\max_{i=1,\ldots,d} |W_i|) \leq \frac{\log(2p)}{t} + t\sigma^2/2.$$

Hint: use Jensens inequality to argue that $\exp(t\mathbb{E}(\max_{i=1,\ldots,d} W_i)) \leq \mathbb{E}(\exp(\max_{i=1,\ldots,d} tW_i))$; also, it will help to recall that the moment-generating function of a standard Gaussian variate. Then, you can bound the expectation for the absolute value.

(4) Using the result from the previous part, plug in an appropriate value of t and establish (2).

Question 4 (In-sample risk for the lasso): Assume that $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}, i = 1, \ldots, n$ are i.i.d. pairs satisfying $y_i = x_i^T\beta_0 + \epsilon_i$, where $\beta_0 \in \mathbb{R}^d$ is the unknown parameter. We also assume $x_i \sim P_X$ and $\epsilon_i \sim N(0, \sigma^2)$ with the predictors and errors being independent. Additionally, assume that P_X is a distribution supported on $[-M, M]^d$. Let $\hat{\lambda}$ be the lasso estimator in constrained form,

$$\hat{\beta} = \text{argmin}_{\beta\in\mathbb{R}^d}\|y - X\beta\|_2^2 \text{ subject to } \|\beta\|_1 \leq t,$$

where $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$ is the response vector and $X \in \mathbb{R}^{n\times p}$ is the matrix of predictors, with rows $x_i, i = 1, \ldots, n$.

(1) Prove that the lasso estimator, with $t = \|\beta_0\|_1 1$, has in-sample risk satisfying

$$\frac{1}{n}\mathbb{E}\|X\hat{\beta} - X\beta_0\|_2^2 \leq M\sigma\|\beta_0\|_1\sqrt{\frac{2\log(2d)}{n}},$$

where the expectation is taken over the training data $(x_i, y_i), i = 1, \ldots, n$. Hint: follow the same strategy we used in class to derive the slow rate for the lasso estimator in bound form. Take an expectation where appropriate (rather than invoking high-probability arguments as we did in class), and apply the result in (2).