Homework 4
10/36-702
Due Wednesday April 26 at 3:00 pm

1 In-sample and out-of-sample risk for least squares

Assume that $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \ldots, n$ are i.i.d. pairs satisfying a linear relationship

$$y_i = x_i^T \beta_0 + \epsilon_i, \quad i = 1, \ldots, n$$

where $\beta_0 \in R^p$ is the unknown regression parameter to be estimated, $x_i \sim P_X$, $i = 1, \ldots, n$, and $\epsilon_i \sim N(0, \sigma^2)$, $i = 1, \ldots, n$ with the predictors and errors being independent. Write $\widehat{\beta}$ for the least squares estimator trained on (x_i, y_i), $i = 1, \ldots, n$.

In class we showed that the in-sample risk satisfies

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(x_i^T \widehat{\beta} - x_i^T \beta_0)^2 = \sigma^2 \frac{p}{n},$$

where the expectation is taken over the i.i.d. training points (x_i, y_i), $i = 1, \ldots, n$. We also showed the out-of-sample risk satisfies

$$\mathbb{E}(x_0^T \widehat{\beta} - x_0^T \beta_0)^2 \geq \sigma^2 \frac{p}{n},$$

where the expectation is taken over the i.i.d. training points (x_i, y_i), $i = 1, \ldots, n$ as well as the independent draw $x_0 \sim P_X$.

Prove or disprove: there is a distribution P_X such that the out-of-sample risk is equal to $\sigma^2 p/n$. You should consider the cases $p = 1$ and $p > 1$ separately. Hint: you may use the fact that if U is random variable that is not almost surely constant, and f is a strictly convex function, then $\mathbb{E}[f(U)] > f(\mathbb{E}[U])$, i.e., we get a strict inequality in Jensen's inequality.

2 Gaussian maximal inequalities

In class we extensively used the following Gaussian maximal inequality, 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 $\delta > 0$,

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

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

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

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

where ϕ is the standard normal density. Hint: you may use Mill's inequality, which states that $P(|Z| > t) \leq 2\phi(t)/t$ for a standard Gaussian variate Z.

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

The result (1) is a high-probability bound on the maximum or Gaussians; of interest is also an expectation bound,

$$\mathbb{E}\Big(\max_{i=1,\dots,p} |W_i|\Big) \le \sigma\sqrt{2\log(2p)}. \tag{2}$$

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

$$\mathbb{E}\Big(\max_{i=1,\dots,p} |W_i|\Big) \le \frac{\log(2p)}{t} + \frac{t\sigma^2}{2}.$$

Hint: use Jensen's inequality to argue that $\exp(t\mathbb{E}(\max_{i=1,\dots,p}|W_i|)) \le \mathbb{E}(\exp(\max_{i=1,\dots,p} t|W_i|))$; also, it will help to recall that the moment-generating function of a standard Gaussian variate Z is $\mathbb{E}(e^{tZ}) = e^{t^2/2}$.

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

(e) Suppose now, instead of Gaussianity, we assume that $\mathbb{E}(e^{tW_i}) \le e^{\sigma^2 t^2/2}$, $i = 1,\dots,n$. Argue that the same result as in (2) still holds.

3 In-sample and out-of-sample risk for the lasso

Assume the same model as in Problem 1, except additionally assume that P_X is a distribution supported on $[-M, M]^p$. Now write $\widehat{\beta}$ for the lasso estimator in constrained form,

$$\widehat{\beta} = \underset{\beta\in\mathbb{R}^p}{\operatorname{argmin}}\ \|y - X\beta\|_2^2 \ \text{ subject to } \ \|\beta\|_1 \le t,$$

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

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

$$\frac{1}{n}\mathbb{E}\|X\widehat{\beta} - X\beta_0\|_2^2 \le 4M\sigma\|\beta_0\|_1\sqrt{\frac{2\log(2p)}{n}},$$

where the expectation is taken over the training data (x_i, y_i), $i = 1,\dots,n$. 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).

(b) Let Σ denote the predictor covariance matrix, i.e., $\Sigma = \mathbb{E}(x_0 x_0^T)$, and $\widehat{\Sigma} = (1/n)X^T X$. Let $V = \widehat{\Sigma} - \Sigma$. Prove that

$$\mathbb{E}\Big(\max_{j,k=1,\dots,p} |V_{jk}|\Big) \le 2M^2\sqrt{\frac{2\log(2p^2)}{n}}.$$

Hint: you may use the following fact, which is a consequence of Lemma 4 and Theorem 5 (i.e., Hoeffding's inequality) in our concentration of measure class notes. If Z_i, $i = 1,\dots,n$ are i.i.d. mean zero random variables lying in $[a,b]$, then

$$\mathbb{E}\Big[\exp\Big(\frac{t}{n}\sum_{i=1}^n Z_i\Big)\Big] \le e^{t^2(b-a)^2/(8n)}.$$

Write $V_{jk} = (1/n)\sum_{i=1}^n x_{ij}x_{ik} - \mathbb{E}(x_{0j}x_{0k})$. Apply the above fact to bound the moment generating function of each V_{jk}, $j,k = 1,\dots,p$. Then apply Problem 2 part (e) to conclude the result.

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

$$\mathbb{E}(x_0^T\widehat{\beta} - x_0^T\beta_0)^2 \le 4M\sigma\|\beta_0\|_1\sqrt{\frac{2\log(2p)}{n}} + 8M^2\|\beta_0\|_1^2\sqrt{\frac{2\log(2p^2)}{n}},$$

where the expectation is taken over the training data (x_i, y_i), $i = 1,\dots,n$ and independent draw $x_0 \sim P_X$. Hint: first, argue that the in-sample risk and out-of-sample risk can be written as

$$\mathbb{E}\Big[(\widehat{\beta}-\beta_0)^T\widehat{\Sigma}(\widehat{\beta}-\beta_0)\Big] \text{ and } \mathbb{E}\Big[(\widehat{\beta}-\beta_0)^T\Sigma(\widehat{\beta}-\beta_0)\Big],$$

respectively, where the expectations above are each taken with respect to the training samples (x_i, y_i), $i = 1,\dots,n$ only. Next, argue that

$$\begin{aligned}(\widehat{\beta}-\beta_0)^T\Sigma(\widehat{\beta}-\beta_0) - (\widehat{\beta}-\beta_0)^T\widehat{\Sigma}(\widehat{\beta}-\beta_0) &\le \sum_{j,k=1}^p (\widehat{\beta}-\beta_0)_j(\widehat{\beta}-\beta_0)_k|V_{jk}| \\ &\le 4\|\beta_0\|_1^2 \max_{j,k=1,\dots,p}|V_{jk}|,\end{aligned}$$

where recall $V_{jk} = (\widehat{\Sigma}-\Sigma)_{jk}$, $j,k = 1,\dots,p$. Then, apply the previous parts, (b) and (a), to conclude the result.

4 Bonus: high-dimensional regression simulation

Produce a convincing simulation where the lasso estimator has smaller out-of-sample risk than forward stepwise regression. Produce a convincing simulation where forward stepwise regression has smaller out-of-sample risk than the lasso.

Note: the descriptor "convincing" is of course kind of ambiguous here. But a good answer will involve averaging simulation results over multiple runs, tuning the lasso and forward stepwise estimators in a reasonable way, etc.

5 Graphical models

(a) Let $X = (X_1,\dots,X_d) \in \mathbb{R}^d$. Suppose that $X \sim N(\mu,\Sigma)$. Let $\Omega = \Sigma^{-1}$. Let j,k be integers among $1,\dots,d$ such that $j \ne k$. Let $Z = (X_s : s \ne j,k)$.

i. Show that the distribution of $(X_j, X_k)|Z$ is $N(a, B)$ and find a and B explicitly.

ii. Show that $X_j \perp\!\!\!\perp X_k | Z$ if and only if $\Omega_{jk} = 0$.

(b) Let $X = (X_1, \ldots, X_5)$ be a random vector distributed as $X \sim N(0, \Sigma)$ where the covariance matrix Σ is given by

$$\Sigma = \frac{1}{15}\begin{pmatrix} 9 & -3 & -3 & -3 & -3 \\ -3 & 6 & 1 & 1 & 1 \\ -3 & 1 & 6 & 1 & 1 \\ -3 & 1 & 1 & 6 & 1 \\ -3 & 1 & 1 & 1 & 6 \end{pmatrix}.$$

i. What is the graph for X, viewed as an undirected graphical model?

ii. Which of the following independence statements are true?

1. $X_2 \perp\!\!\!\perp X_3 | X_1$
2. $X_3 \perp\!\!\!\perp X_4$
3. $X_1 \perp\!\!\!\perp X_3 | X_2$
4. $X_1 \perp\!\!\!\perp X_5$

iii. List the local Markov properties for this graphical model.

iv. Find the conditional density $p(x_4 | X_1 = 2)$.

(c) Let $X = (X_1, \ldots, X_4)$ where each variable is binary. Suppose the probability function is

$$\log p(x) = \psi_\varnothing + \psi_{12}(x_1, x_2) + \psi_{13}(x_1, x_3) + \psi_{24}(x_2, x_4) + \psi_{34}(x_3, x_4).$$

i. Draw the implied graph.

ii. Write down all the independence and conditional independence relations implied by the graph.

iii. Is the model graphical? Is the model hierarchical?