STAT 499: Undergraduate Research

Week 7: Tutorial on the Lasso simulation experiment

Facheng Yu Compiled on: 2024-02-27, 05:08:07

Assume that $(Y_i, X_i) \in \mathbb{R} \times \mathbb{R}^d$ are i.i.d. observed from P_{XY} for $i = 1, \ldots, n$. Throughout this lecture, we assume that $\mathbb{E}[X_i X_i^\top]^{-1}$ exsits.

7.1 Start from the correctly specified model

In this section, we consider that the correctly specified model:

$$Y_i = X_i^\top \theta^* + w_i$$

where $\mathbb{E}[w_i|X_i] = 0$. That is, $\mathbb{E}[Y_i|X_i] = X_i^\top \theta^*$ is a linear function.

1. Under the correctly specified model, consider such a program:

   $$\theta_{ols} = \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}[(Y_i - X_i^\top \theta)^2].$$

   Show that $\theta_{ols} = \theta^*$. That is, the minimizer of MSE is unique and is also the true regression coefficient. (Hint: differentiating in terms of θ and use the fact $\mathbb{E}[Y_i|X_i] = X_i^\top \theta^*$.)

2. Generate a dataset under the correctly specified model setting. In specific, use `datasets.make_regression` from `sklearn` to generate the dataset with `n_samples` $= 10,000$, `n_features` $= 1,000$, `n_informative` $= 100$ and `noise` $= 10$.

   Explain what are the meanings of `n_samples`, `n_features`, `n_informative` and `noise`? How can you keep the true coefficient $\theta*$ when using `datasets.make_regression`?

3. Now verify that $\mathbb{E}[X_i X_i^\top]$ is invertible using plug-in principle. That is, calculate the smallest eigenvalue λ_{min} of $N^{-1} \sum_{i=1}^N X_i X_i^\top$, and draw the plot of λ_{min} changing with N.

   For simplicity, for each $N = 500, 1,000, \ldots, 10,000$, calculate the corresponding λ_{min} and draw the plot. Can you say $\mathbb{E}[X_i X_i^\top]$ is invertible from your result?

4. Now we do Lasso program on the generated dataset:

   $$\hat{\theta}_n = \arg\min_{\theta \in \mathbb{R}^d} \quad \frac{1}{2n} \|Y - \mathbf{X}\theta\|_2^2 + \lambda_n \|\theta\|_1,$$

   where $Y = (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n$, $\mathbf{X} = (X_1, \ldots, X_n)^\top \in \mathbb{R}^{n \times d}$, and $\lambda_n > 0$ is the regularization parameter.

   Explain what is the optimal λ_n according to the lasso consistency theorem? For each $N = 500, 600, \ldots, 10,000$, draw the plot of the optimal λ_N changing with N.

5. For each $N = 500, 600, \ldots, 10,000$, use `Lasso` from `sklearn.linear_model` to solve the Lasso program with the optimal λ_N already obtained, i.e., for each N, use the fist N samples from the dataset and take λ_N as the regularization parameter.

   Draw the plot of how $\|\hat{\theta}_N - \theta^*\|_2$ changes with N increasing. On the same plot, draw the theoretical bound for $\|\hat{\theta}_N - \theta^*\|_2$ according to the Lasso consistency theorem.

6. For $N = 10,000$, draw the histogram of $\hat{\theta}_N$. Is $\hat{\theta}_N$ has similar sparsity as θ^*?
7. Is there any N such that $\|\hat{\theta}_N - \theta^*\|_2$ exceeds the theoretical bound? You do not need to explain for this if such a N exists.