Mathematical techniques in data science

Lecture 13: Model consistency the lasso estimator

March 18th, 2019

Schedule

Week	Chapter
1	Chapter 2: Intro to statistical learning
3	Chapter 4: Classification
4	Chapter 9: Support vector machine and kernels
5, 6	Chapter 3: Linear regression
7	Chapter 8: Tree-based methods + Random forest
8	
9	Neural network
12	PCA $\rightarrow$ Manifold learning
11	Clustering: K-means $\rightarrow$ Spectral Clustering
10	Bootstrap + Bayesian methods + UQ
13	Reinforcement learning/Online learning/Active learning
14	Project presentation

Chapter 3 & 6: Topics on Linear regression

- Linear regression
- Subset selection
- Shrinkage methods
- Model consistency of lasso

Note: Homework 2 is uploaded. Due on 03/29 at 5pm.

Settings

- We start with the simple linear regression problem

$$Y = \beta_1 X^{(1)} + \beta_2 X^{(2)} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

- Sparsity: assume that the data is generated using the "true" vector of parameters $\beta^* = (\beta_1^*, 0)$.
- We assume that $E[X^{(1)}] = E[X^{(2)}] = 0$.

Matrix form

- we observe a dataset $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- use the same notations as in the previous lectures

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} x_1^{(1)} & x_1^{(2)} \\ \ldots & \ldots \\ x_n^{(1)} & x_n^{(2)} \end{bmatrix}$$

Goal

The lasso estimator solves the optimization problem

$$\hat{\beta} = \min_{\beta} \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda(|\beta_1| + |\beta_2|).$$

We want to investigate the conditions under which we can verify that

$$sign(\hat{\beta}_1) = sign(\beta_1^*) \quad \text{and} \quad \hat{\beta}_2 = 0$$

Sub-gradient

Issue: the penalty of lasso is non-differentiable

Definition

We say that a vector $s \in \mathbb{R}^k$ is a subgradient for the ℓ_1-norm evaluated at $\beta \in \mathbb{R}^k$, written as $s \in \partial\|\beta\|$ if for $i = 1, \dots, k$ we have

$$s_i = sign(\beta_i) \quad \text{if} \quad \beta_i \neq 0 \quad \text{and} \quad s_i \in [-1, 1] \quad \text{otherwise.}$$

Properties of lasso solutions

Theorem

(a) *A vector $\hat{\beta}$ solve the lasso program if and only if there exists a $\hat{z} \in \partial\|\hat{\beta}\|$ such that*

$$X^T(Y - X\hat{\beta}) - \lambda\hat{z} = 0 \tag{0.1}$$

(b) *Suppose that the subgradient vector satisfies the strict dual feasibility condition*

$$|\hat{z}_2| < 1$$

then **any** *lasso solution $\tilde{\beta}$ satisfies $\tilde{\beta}_2 = 0$.*

(c) *Under the condition of part (b), if $X^{(1)} \neq 0$, then $\hat{\beta}$ is the unique lasso solution.*

The primal-dual witness method.

The primal-dual witness (PDW) method consists of constructing a pair of $(\tilde{\beta}, \tilde{z})$ according to the following steps:

- First, we obtain $\tilde{\beta}_1$ by solving the restricted lasso problem

$$\tilde{\beta}_1 = \min_{\beta=(\beta_1,0)} \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda(|\beta_1|).$$

  Choose a subgradient $\tilde{z}_1 \in \mathbb{R}$ for the ℓ_1-norm evaluated at $\tilde{\beta}_1$
- Second, we solve for a vector $\tilde{z}_2$ satisfying equation (0.1), and check whether or not the dual feasibility condition $|\tilde{z}_2| < 1$ is satisfied
- Third, we check whether the *sign consistency condition*

$$\tilde{z}_1 = sign(\beta_1^*)$$

  is satisfied.

PDW

- This procedure is not a practical method for solving the ℓ_1-regularized optimization problem, since solving the restricted problem in Step 1 requires knowledge about the sparsity of β^*
- Rather, the utility of this constructive procedure is as a proof technique: it succeeds if and only if the lasso has a optimal solution with the correct signed support.

A more detailed computation

We note that the matrix form of equation (0.1) can be written as

$$[X^{(1)}]^T(Y - X^{(1)}\beta_1 - X^{(2)}\beta_2) - \lambda\hat{z}_1 = 0$$

$$[X^{(2)}]^T(Y - X^{(1)}\beta_1 - X^{(2)}\beta_2) - \lambda\hat{z}_2 = 0$$

To simplify the notation, we denote

$$C_{ij} = [X^{(i)}]^T[X^{(j)}]$$

Step 1

- we find $\tilde{\beta}_1$ and $\tilde{z}_1$ that satisfies

$$[X^{(1)}]^T(Y - X^{(1)}\tilde{\beta}_1) - \lambda\tilde{z}_1 = 0$$

- Moreover, to make sure that the sign consistency in Step 3 is satisfied, we impose that

$$\tilde{z}_1 = sign(\beta_1^*) \quad \text{and} \quad \tilde{\beta}_1 = C_{11}^{-1}([X^{(1)}]^T Y - \lambda sign(\beta_1^*)).$$

This is acceptable as long as $\tilde{z}_1 \in \partial|\tilde{\beta}_1|$. That is,

$$sign(\tilde{\beta}_1) = sign(\beta_1^*)$$

Step 2

- Step 2:
$$[X^{(2)}]^T(Y - X^{(1)}\tilde{\beta}_1) - \lambda\hat{z}_2 = 0$$
- Choose
$$\tilde{z}_2 = \frac{1}{\lambda}[X^{(2)}]^T(Y - X^{(1)}\tilde{\beta}_1).$$

We want $|\tilde{z}_2| < 1$.

Conditions

In principle, we want two conditions:

- $sign(\tilde{\beta}_1) = sign(\beta_1^*)$
- $|\tilde{z}_2| < 1$

Recalling that $Y = X^{(1)}\beta_1^* + \epsilon$, we have

$$\begin{aligned} \tilde{\beta}_1 &= C_{11}^{-1}([X^{(1)}]^T(X^{(1)}\beta_1^* + \epsilon) - \lambda sign(\beta_1^*)) \\ &= \beta_1^* + C_{11}^{-1}([X^{(1)}]^T\epsilon - \lambda sign(\beta_1^*))) \end{aligned}$$

Conditions

Thus if we denote

$$\Delta = C_{11}^{-1}([X^{(1)}]^T \epsilon - \lambda sign(\beta_1^*))$$

then the first condition can be further simplified as $sign(\beta_1^*) = sign(\beta_1^* + \Delta)$.
Similarly,

$$\begin{aligned} \tilde{z}_2 &= \frac{1}{\lambda}[X^{(2)}]^T (X^{(1)}\beta_1^* + \epsilon - X^{(1)}\tilde{\beta}_1) \\ &= \frac{1}{\lambda}[X^{(2)}]^T (X^{(1)}\Delta + \epsilon) \end{aligned}$$

Zero-noise setting

- we assume that the observations are collected with no noise ($\epsilon = 0$).
- Then

$$\Delta = -C_{11}^{-1}\lambda sign(\beta_1^*)$$

and

$$\tilde{z}_2 = \frac{-1}{\lambda}C_{21}\Delta = C_{21}C_{11}^{-1}sign(\beta_1^*)$$

Conditions

- Mutual incoherence: $|C_{21}C_{11}^{-1}| < 1$.
- Minimum signal: $|\beta_1^*| > \lambda C_{11}^{-1}$