MLE Exercise-5

July 9, 2024

1 Bias and Variance of Ridge Regression

Ridge regression solves the regularized least squares problem:

$$\hat{\beta}_\tau = \arg\min_\beta (y - X\beta)^T (y - X\beta) + \tau \beta^T \beta$$

with regularization parameter $\tau \geq 0$. Regularization introduces some bias into the solution in order to achieve a potentially large gain in variance. Assume that the true model is $y = X\beta^* + \epsilon$ with zero-mean Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and centered features $\frac{1}{N}\sum_i X_i = 0$ (note that these assumptions imply that y is also centered in expectation). Prove (e.g. using the SVD of X) that expectation and covariance matrix of the regularized solution (both taken over all possible training sets of size N) are then given by:

$$\mathbb{E}[\hat{\beta}_\tau] = S_\tau^{-1} S \beta^*$$

$$\mathrm{Cov}[\hat{\beta}_\tau] = S_\tau^{-1} S S_\tau^{-1} \sigma^2$$

where S and S_τ are the ordinary and regularized scatter matrices:

$$S = X^T X \quad \text{and} \quad S_\tau = X^T X + \tau I_D$$

Notice that expectation and covariance reduce to the corresponding expressions of ordinary least squares (as derived in the lecture) when $\tau = 0$:

$$\mathbb{E}[\hat{\beta}_{\tau=0}] = \beta^* \quad \text{and} \quad \mathrm{Cov}[\hat{\beta}_{\tau=0}] = S^{-1}\sigma^2$$

Since S_τ is greater than S (in any norm), regularization has a shrinking effect on both expectation and covariance.

Solution

The ridge regression estimator is given by:

$$\hat{\beta}_\tau = (X^T X + \tau I_D)^{-1} X^T y$$

Given $y = X\beta^* + \epsilon$, we have:

$$\hat{\beta}_\tau = (X^T X + \tau I_D)^{-1} X^T (X\beta^* + \epsilon)$$

$$\hat{\beta}_\tau = (X^T X + \tau I_D)^{-1} (X^T X\beta^* + X^T \epsilon)$$

Since $X^T X = S$,

$$\hat{\beta}_\tau = (S + \tau I_D)^{-1} (S\beta^* + X^T \epsilon)$$

Taking the expectation:

$$\mathbb{E}[\hat{\beta}_\tau] = (S + \tau I_D)^{-1} S\beta^* + (S + \tau I_D)^{-1} \mathbb{E}[X^T \epsilon]$$

Since $\mathbb{E}[\epsilon] = 0$,

$$\mathbb{E}[\hat{\beta}_\tau] = (S + \tau I_D)^{-1} S\beta^*$$

For the covariance:

$$\mathrm{Cov}(\hat{\beta}_\tau) = \mathbb{E}[(\hat{\beta}_\tau - \mathbb{E}[\hat{\beta}_\tau])(\hat{\beta}_\tau - \mathbb{E}[\hat{\beta}_\tau])^T]$$

$$\hat{\beta}_\tau - \mathbb{E}[\hat{\beta}_\tau] = (S + \tau I_D)^{-1} X^T \epsilon$$

$$\mathrm{Cov}(\hat{\beta}_\tau) = \mathbb{E}[(S + \tau I_D)^{-1} X^T \epsilon\epsilon^T X (S + \tau I_D)^{-1}]$$

Since $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$,

$$\mathrm{Cov}(\hat{\beta}_\tau) = (S + \tau I_D)^{-1} X^T \mathbb{E}[\epsilon\epsilon^T] X (S + \tau I_D)^{-1}$$

$$\mathrm{Cov}(\hat{\beta}_\tau) = \sigma^2 (S + \tau I_D)^{-1} X^T X (S + \tau I_D)^{-1} = \sigma^2 (S + \tau I_D)^{-1} S (S + \tau I_D)^{-1}$$

Thus, we have proved:

$$\mathbb{E}[\hat{\beta}_\tau] = S_\tau^{-1} S\beta^* \quad \text{and} \quad \mathrm{Cov}[\hat{\beta}_\tau] = S_\tau^{-1} S S_\tau^{-1} \sigma^2$$

2 LDA-Derivation from the Least Squares Error

We aim to show that the decision rule from minimizing the squared loss gives the same decision boundary as Linear Discriminant Analysis (LDA).

Given Definitions

Decision rule for LDA:

$$\hat{y} = \text{sign}(X_i \cdot \hat{\beta}_{\text{DDA}})$$

where

$$\hat{\beta}_{\text{DDA}} = \Sigma^{-1}(\mu_1 - \mu_{-1})^T$$

Class means:

$$\mu_{-1} = \frac{1}{N_{-1}} \sum_{i:y_i=-1} X_i \quad \text{and} \quad \mu_1 = \frac{1}{N_1} \sum_{i:y_i=1} X_i$$

Covariance matrix:

$$\Sigma = \frac{1}{N} \left[\sum_{i:y_i=-1} (X_i - \mu_{-1})^T (X_i - \mu_{-1}) + \sum_{i:y_i=1} (X_i - \mu_1)^T (X_i - \mu_1) \right]$$

Goal

Show that the equivalent decision rule arises from minimizing the squared loss:

$$\hat{\beta}_{\text{OLS}} = \arg\min_{\beta} \sum_{i=1}^{N} (y_i^* - X_i \cdot \beta)^2 \implies \hat{\beta}_{\text{OLS}} = \tau \Sigma^{-1}(\mu_1 - \mu_{-1})^T$$

for some $\tau > 0$.

Steps

1. **Define the Squared Loss Function:**

$$L(\beta) = \sum_{i=1}^{N} (y_i^* - X_i \cdot \beta)^2$$

2. **Take the Gradient of the Loss Function:**

$$\frac{\partial}{\partial \beta} \sum_{i=1}^{N} (y_i^* - X_i \cdot \beta)^2 = 0$$

$$-2 \sum_{i=1}^{N} X_i^T (y_i^* - X_i \cdot \beta) = 0$$

$$\sum_{i=1}^{N} X_i^T y_i^* = \sum_{i=1}^{N} X_i^T X_i \beta$$

$$\beta = (X^T X)^{-1} X^T y^*$$

3. **Balanced and Centered Data Assumptions:** Since classes are balanced and centered:

$$y_i^* = \begin{cases} 1 & \text{if } y_i = 1 \\ -1 & \text{if } y_i = -1 \end{cases}$$

Implies:

$$\sum_{i=1}^{N} y_i^* = 0$$

4. **Expectation and Covariance:**

$$\hat{\beta}_{\text{OLS}} = \tau \Sigma^{-1} (\mu_1 - \mu_{-1})^T$$

5. **Derivation (Equating to LDA):**

$$\hat{\beta}_{\text{DDA}} = \Sigma^{-1} (\mu_1 - \mu_{-1})^T$$

Given:

$$\Sigma \cdot \beta + \frac{1}{4} (\mu_1 - \mu_{-1})^T (\mu_1 - \mu_{-1}) \cdot \beta = \frac{1}{2} (\mu_1 - \mu_{-1})^T$$

For scalar τ', bringing second term to the right:

$$\Sigma \cdot \beta = \left(\frac{1}{2} - \frac{\tau'}{4} \right) (\mu_1 - \mu_{-1})^T$$

Therefore:

$$\hat{\beta}_{\text{OLS}} = \tau \Sigma^{-1} (\mu_1 - \mu_{-1})^T$$

with $\tau = \frac{1}{2} - \frac{\tau'}{4}$.

Thus, the equivalent decision rule from minimizing the squared loss is shown to yield the same result as LDA.