MLE 5

Sonntag, 18. Juni 2023 11:07

1 Bias and variance of ridge regression (8 points)

Ridge regression solves the regularized least squares problem

$$\hat{\beta}_\tau = \text{argmin}_\beta (y - X\beta)^\top (y - X\beta) + \tau\, \beta^\top \beta$$

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

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

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

$$S = X^\top X \qquad S_\tau = X^\top X + \tau \mathbb{1}_D$$

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

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

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

$\mathbb{E}[\hat{\beta}_\tau] = (X^TX + \tau \mathbb{1}_D)^{-1} X^T y$

$= (X^TX + \tau \mathbb{1}_D)^{-1} X^T (X\beta^* + \varepsilon)$

$= \underbrace{(X^TX + \tau \mathbb{1}_D)^{-1}}_{S_\tau^{-1}} \underbrace{X^TX}_{S} \beta^* + \underbrace{X^T\varepsilon}_{\mathbb{E}=0}$

$\Rightarrow \mathbb{E}(\hat{\beta}_\tau) = S_\tau^{-1} S\, \beta^*$

I seem to have started with a wrong formulation, hence, all my steps are different. However, as far as I can tell, the algebraic transformations are correct and the solution is also correct

$\text{bias} = \hat{\beta}_\tau - \beta^* = (X^TX + \tau \mathbb{1}_D)^{-1} \cdot (X^TX\beta^* + X^T\varepsilon) - \beta^*$ $\Rightarrow$ In the lecture, this step was introduced and therefore I used it again. But I should have just used the above proven term instead.

$\text{cov}(\hat{\beta}_\tau) = \mathbb{E}\left((\hat{\beta}_\tau - \mathbb{E}(\hat{\beta}_\tau))(\hat{\beta}_\tau - \mathbb{E}(\hat{\beta}_\tau))^T\right)$

$= \mathbb{E}\left[\left((X^TX + \tau \mathbb{1}_D)^{-1}(X^TX\beta^* + X^T\varepsilon) - \beta^*\right)\left(\beta^{*T}XX^T + \varepsilon^TX\right)\left(X^TX + \tau \mathbb{1}_D\right)^{-1} - \beta^{*T}\right)\right]$ This is why I could not progress in the calculation.

Most probably, there is something wrong here which I can't seem to find/solve

Number 2:

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

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

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

$\Rightarrow \sum_{i=1}^{N} \left(-2y_i^* X_i + 2X_i^2\beta\right) = 0$

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

$\Rightarrow \sum_{i=1}^{N} (X_i^2\beta) = \sum_{i=1}^{N} y_i^* X_i$ ⟶ up to this step quite correct, but instead of X_i^2 $X_iX_i^T$ and instead of $y_i^*X_i$, $y_i^*X_i^T$ would have been correct

$\Rightarrow \beta \cdot \sum_{i=1}^{N} X_i^2 = \sum_{i=1}^{N} y_i^* X_i$

$\Rightarrow \Sigma\beta \sum_{i=1}^{N} X_i^2 = \Sigma \sum_{i=1}^{N} y_i^* X_i$?

The above error leaves me at this point as I am unable to rewrite the solution as explained in the sample-solution

I have no clue how to go on from this

or if it is even remotely correct so far