Parameter Estimation

MSE and bias variance trade-off

Parameter Estimation

Main Idea

- Set of data points $(y_n, \mathbf{x}_n)$ for $n = 1, 2, \ldots, N$.
- Functional form $f_{\boldsymbol{\theta}}(\cdot)$:
  - Linear: $f_{\boldsymbol{\theta}}(x) = \theta_0 + \theta_1 x$
  - Polynomial: $f_{\boldsymbol{\theta}} = \theta_0 + \theta_1 x + \theta_2 x^2$
- Minimize cost function $\mathbf{J}(\boldsymbol{\theta}) = \sum_{n=1}^{N} \mathcal{L}(y_n, f_{\boldsymbol{\theta}}(\mathbf{x}_n))$
  - $\mathcal{L}$ is a loss function
  - Squared error loss: $\mathcal{L}(y_n, f_{\boldsymbol{\theta}}(\mathbf{x}_n)) = (y - f_{\boldsymbol{\theta}}(\mathbf{x}))^2$

Linear Regression I

Consider the linear regression model:

$$y = \boldsymbol{\theta}^T \mathbf{x} + \eta$$

The matrix form of the model can be written as:

$$\mathbf{y} = \mathbf{X}\theta + \boldsymbol{\eta}$$

The objective is to minimize the sum of squared residuals. The sum of squared residuals (SSR) is given by:

$$SSR = (\mathbf{y} - \mathbf{X}\theta)^\top (\mathbf{y} - \mathbf{X}\theta)$$

Expanding this expression:

$$SSR = \mathbf{y}^\top \mathbf{y} - 2\theta^\top \mathbf{X}^\top \mathbf{y} + \theta^\top \mathbf{X}^\top \mathbf{X}\theta$$

Linear Regression II

To minimize the SSR, take the derivative with respect to θ and set it to zero:

$$\frac{\partial}{\partial \theta} SSR = -2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\theta = 0$$

This results in the normal equations:

$$\mathbf{X}^\top \mathbf{X}\theta = \mathbf{X}^\top \mathbf{y}$$

Solving for θ gives the ordinary least squares (OLS) estimator:

$$\theta = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

Biased and Unbiased Estimation I

Bias of an Estimator ($\hat{\theta}$):

$$\text{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$$

Variance of an Estimator ($\hat{\theta}$):

$$\text{Variance}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2]$$

Biased and Unbiased Estimation II

MSE and bias variance trade-off

$$\text{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2]$$

This expression can be expanded and decomposed as follows:

$$\text{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}] + \mathbb{E}[\hat{\theta}] - \theta)^2]$$
$$\text{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2 + 2(\hat{\theta} - \mathbb{E}[\hat{\theta}])(\mathbb{E}[\hat{\theta}] - \theta) + (\mathbb{E}[\hat{\theta}] - \theta)^2]$$
$$\text{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2] + 2\mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])(\mathbb{E}[\hat{\theta}] - \theta)] + \mathbb{E}[(\mathbb{E}[\hat{\theta}] - \theta)^2]$$

Since $\mathbb{E}[\hat{\theta} - \mathbb{E}[\hat{\theta}]] = 0$, the middle term vanishes:

$$\text{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2] + (\mathbb{E}[\hat{\theta}] - \theta)^2$$

Derivation of Optimal Bias for MSE Reduction I

- Unbiased Estimator: $\hat{\theta}_u$ with $\mathbb{E}[\hat{\theta}_u] = \theta_o$
- Biased Estimator: $\hat{\theta}_b := (1+\alpha)\hat{\theta}_u$, where $\alpha \in \mathbb{R}$

Calculate MSE of Biased Estimator:

$$\text{MSE}(\hat{\theta}_b) = \mathbb{E}\left((1+\alpha)\hat{\theta}_u - \theta_o\right)^2$$

Expand and Simplify:

$$\text{MSE}(\hat{\theta}_b) = (1+\alpha)^2\text{MSE}(\hat{\theta}_u) + \alpha^2\theta_o^2$$

Condition for MSE Reduction:

$$(1+\alpha)^2\text{MSE}(\hat{\theta}_u) + \alpha^2\theta_o^2 < \text{MSE}(\hat{\theta}_u)$$

Derivation of Optimal Bias for MSE Reduction II

Solution for α:

$$\alpha^2 \mathrm{MSE}(\hat{\theta}_u) + 2\alpha \mathrm{MSE}(\hat{\theta}_u) + \alpha^2 \theta_o^2 < 0$$

$$\Rightarrow -\frac{2\mathrm{MSE}(\hat{\theta}_u)}{\mathrm{MSE}(\hat{\theta}_u) + \theta_o^2} < \alpha < 0$$

$$\Rightarrow -2 < \alpha < 0$$

Conclusion: The optimal bias for reducing MSE lies within the interval $-2 < \alpha < 0$.

Rigde Regression I

- Biased estimator
- Norm shrinking loss:

$$L(\boldsymbol{\theta}, \lambda) = \sum_{n=1}^{N} (y_n - \boldsymbol{\theta}^T \boldsymbol{x}_n)^2 + \lambda \|\boldsymbol{\theta}\|^2$$

- Cost-function:

$$J(\boldsymbol{\theta}) = (y - X\boldsymbol{\theta})^T (y - X\boldsymbol{\theta}) + \lambda \boldsymbol{\theta}^T \boldsymbol{\theta}$$

- Differentiate:

$$\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -2X^T y + 2X^T X\boldsymbol{\theta} + 2\lambda I\theta$$

Rigde Regression II

- Setting the derivative to 0:

$$-2X^T\mathbf{y} + 2X^T X\theta + 2\lambda I\theta = 0$$

$$(2X^T X + 2\theta I)\theta = 2X^T\mathbf{y}$$

$$(X^T X + \theta I)\theta = X^T\mathbf{y}$$

- Solution:

$$\hat{\theta}_{\text{ridge}} = (X^T X + \lambda\mathbf{I})^{-1}X^T\mathbf{y}$$

Exercise 2.3.2 I

Purpose of Exercise:

- Illustrate the effects of the regularization parameter λ on
- Show how λ can balance bias and variance, using the mean squared error (MSE) as a measure.

Mathematical Framework: The generalization error at point x is given by:

$$\mathbb{E}_D\left[(f(x;\mathcal{D})-\mathbb{E}[y|x])^2\right]=\mathbb{E}_\mathcal{D}\left[(f(x;\mathcal{D})-\mathbb{E}_D[f(x;\mathcal{D})])^2\right]+(\mathbb{E}_D[f(x;\mathcal{D})]-\mathbb{E}[y|x])^2$$

Implementation:

```
A = Xtrain.T @ Xtrain + lambda_ * np.eye(d)
b = Xtrain.T @ Ttrain
theta = np.linalg.solve(A,b)
```

Exercise 2.3.2 II

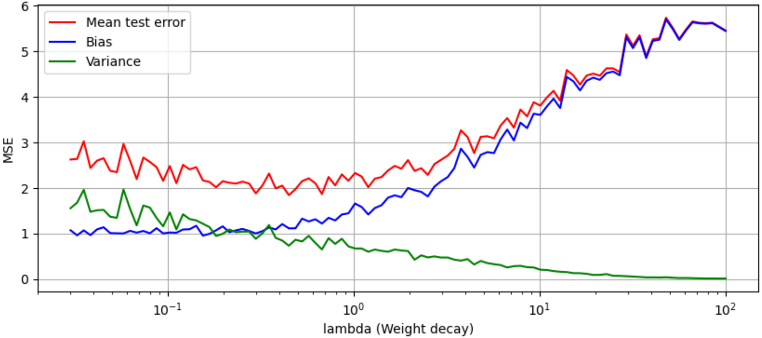

Figure: The effect of regularization parameter λ on bias, variance, and mean test error.