CS 446 / ECE 449 — Homework 4

William Lee (wl72)

1. Bias–Variance in Ridge Regression

We have fixed scalar inputs $\{x^{(i)}\}_{i=1}^{N}$ and labels

$$y^{(i)} = w^\star x^{(i)} + \varepsilon^{(i)}, \qquad \varepsilon^{(i)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2),$$

independent of $x^{(i)}$. Ridge regression solves

$$w_D = \arg\min_w \frac{1}{N} \sum_{i=1}^{N} \bigl(w x^{(i)} - y^{(i)}\bigr)^2 + \lambda w^2,$$

whose closed form is

$$w_D = \frac{\frac{1}{N}\sum_{i=1}^{N} x^{(i)} y^{(i)}}{\lambda + s^2}, \quad s^2 := \frac{1}{N} \sum_{i=1}^{N} \bigl(x^{(i)}\bigr)^2.$$

(a) Expected label and noise

Conditional mean:

$$\bar{y}(x) = \mathbb{E}[y \mid x] = \mathbb{E}[w^\star x + \varepsilon \mid x] = w^\star x,$$

since $\mathbb{E}[\varepsilon] = 0$. Therefore

$$\text{Noise} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\bigl[(\bar{y}(x^{(i)}) - y^{(i)})^2\bigr] = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}[(\varepsilon^{(i)})^2] = \sigma^2.$$

(b) Mean predictor $\bar{w}$

$$\frac{1}{N} \sum_i x^{(i)} y^{(i)} = \frac{1}{N} \sum_i x^{(i)} \bigl(w^\star x^{(i)} + \varepsilon^{(i)}\bigr) = w^\star s^2 + \underbrace{\frac{1}{N} \sum_i x^{(i)} \varepsilon^{(i)}}_{\text{mean } 0}.$$

Hence

$$\boxed{\bar{w} := \mathbb{E}_D[w_D] = \frac{s^2}{\lambda + s^2}\, w^\star.}$$

(c) Squared bias

$$\text{Bias}^2 = \frac{1}{N} \sum_{i=1}^{N} \bigl(\bar{w}\, x^{(i)} - w^\star x^{(i)}\bigr)^2 = (w^\star - \bar{w})^2 s^2 = \Bigl(\frac{\lambda}{\lambda + s^2}\Bigr)^2 (w^\star)^2 s^2.$$

(d) Variance

From the closed form solution:

$$w_D = \frac{\frac{1}{N}\sum_{i=1}^{N} x^{(i)} y^{(i)}}{\lambda + s^2}$$

Substituting $y^{(i)} = w^* x^{(i)} + \varepsilon^{(i)}$:

$$w_D = \frac{\frac{1}{N}\sum_{i=1}^{N} x^{(i)}(w^* x^{(i)} + \varepsilon^{(i)})}{\lambda + s^2} = \frac{w^* s^2 + \frac{1}{N}\sum_{i=1}^{N} x^{(i)}\varepsilon^{(i)}}{\lambda + s^2}$$

Let $Z := \frac{1}{N}\sum_{i=1}^{N} x^{(i)}\varepsilon^{(i)}$. Then:

$$w_D = \frac{w^* s^2}{\lambda + s^2} + \frac{Z}{\lambda + s^2} = \bar{w} + \frac{Z}{\lambda + s^2}$$

Since $\varepsilon^{(i)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$:

$$\text{Var}(Z) = \text{Var}\left(\frac{1}{N}\sum_{i=1}^{N} x^{(i)}\varepsilon^{(i)}\right) = \frac{1}{N^2}\sum_{i=1}^{N}(x^{(i)})^2 \text{Var}(\varepsilon^{(i)}) = \frac{\sigma^2 s^2}{N}$$

Therefore:

$$\text{Var}(w_D) = \text{Var}\left(\frac{Z}{\lambda + s^2}\right) = \frac{\text{Var}(Z)}{(\lambda + s^2)^2} = \frac{\sigma^2 s^2}{N(\lambda + s^2)^2}$$

The prediction variance is:

$$\text{Variance} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_D[(w_D x^{(i)} - \bar{w} x^{(i)})^2] = s^2 \text{Var}(w_D) = \boxed{\frac{s^4 \sigma^2}{N(\lambda + s^2)^2}}$$

(e) Behavior as $\lambda \to 0, \infty$

From parts (c) and (d), we have:

$$\text{Bias}^2 = \left(\frac{\lambda}{\lambda + s^2}\right)^2 (w^*)^2 s^2 \tag{1}$$

$$\text{Variance} = \frac{s^4 \sigma^2}{N(\lambda + s^2)^2} \tag{2}$$

As $\lambda \to 0$ (no regularization):

- $\text{Bias}^2 \to \left(\frac{0}{s^2}\right)^2 (w^*)^2 s^2 = 0$ (unbiased estimation)
- $\text{Variance} \to \frac{s^4\sigma^2}{Ns^4} = \frac{\sigma^2}{N}$ (high variance)

As $\lambda \to \infty$ (strong regularization):

- $\text{Bias}^2 \to \left(\frac{\lambda}{\lambda}\right)^2 (w^*)^2 s^2 = (w^*)^2 s^2$ (high bias)
- $\text{Variance} \to \frac{s^4\sigma^2}{N\lambda^2} = 0$ (low variance)

This demonstrates the classical **bias-variance tradeoff**: as λ increases, bias increases monotonically while variance decreases monotonically. Since we don't know w^* or the true noise distribution in practice, we use model selection techniques (like cross-validation) to find the optimal λ that balances this tradeoff.

(f) Constraint diameter

Constraint $\|w\|^2 \leq R$ implies $|w_D - \bar{w}| \leq |w_D| + |\bar{w}| \leq 2\sqrt{R}$, hence $\boxed{|w_D - \bar{w}|^2 \leq 4R}$.

(g) Variance bound

$$(w_D x^{(i)} - \bar{w}\, x^{(i)})^2 \leq x^{(i)2} |w_D - \bar{w}|^2 \leq 4R\, x^{(i)2}.$$

Averaging and taking expectation gives $\boxed{\text{Variance} \leq 4Rs^2}$.

2. Optimal Classifier under Squared Loss

Given loss $L(h) = \mathbb{E}_{(x,y),D}\big[(h(x) - y)^2\big]$.

(a) Optimal predictor

Since we seek an optimal predictor $h_{\text{opt}}(x)$ that is independent of any specific dataset D, we can write the loss as:

$$L(h) = \mathbb{E}_{(x,y)\sim P}[(h(x) - y)^2]$$

Apply the law of total expectation:

$$L(h) = \mathbb{E}_x\left[\mathbb{E}_{y|x}[(h(x) - y)^2]\right]$$

For each fixed x, we need to minimize the inner expectation $\mathbb{E}_{y|x}[(h(x) - y)^2]$.
Using the bias-variance decomposition for squared error:

$$\mathbb{E}_{y|x}[(h(x) - y)^2] = \text{Var}(y|x) + (h(x) - \mathbb{E}[y|x])^2$$

Since $\text{Var}(y|x)$ is independent of our choice of $h(x)$, the minimum is achieved when the second term equals zero:

$$h(x) - \mathbb{E}[y|x] = 0$$

Therefore: $\boxed{h_{\text{opt}}(x) = \mathbb{E}[y|x]}$

(b) Minimum achievable error

Substituting the optimal predictor back into the loss function:

$$L_{\min} = \mathbb{E}_x\left[\mathbb{E}_{y|x}[(\mathbb{E}[y|x] - y)^2]\right]$$

Since $\mathbb{E}[y|x]$ is constant with respect to the inner expectation over $y|x$:

$$L_{\min} = \mathbb{E}_x\left[\mathbb{E}_{y|x}[(y - \mathbb{E}[y|x])^2]\right] = \mathbb{E}_x[\text{Var}(y|x)]$$

Therefore: $\boxed{L_{\min} = \mathbb{E}_x[\text{Var}(y|x)]}$
This represents the **irreducible error** or **Bayes risk** - the fundamental limit on prediction accuracy due to the inherent noise in the relationship between x and y.