CS 189 Introduction to Machine Learning
Spring 2023 Jonathan Shewchuk

DIS7

1 Simple Bias-Variance Tradeoff

Consider a random variable X, which has unknown mean μ and unknown variance σ^2. Given n i.i.d. realizations of training points $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ from the random variable, we wish to estimate the mean of X. We will call our estimate of X the random variable $\hat{X}$, which has mean $\hat{\mu}$. There are a few ways we can estimate μ given the realizations of the n samples:

1. Average the n sample points: $\dfrac{x_1 + x_2 + \ldots + x_n}{n}$.
2. Average the n sample points and one sample point of 0: $\dfrac{x_1 + x_2 + \ldots + x_n}{n+1}$.
3. Average the n sample points and n_0 sample points of 0: $\dfrac{x_1 + x_2 + \ldots + x_n}{n+n_0}$.
4. Ignore the sample points: just return 0.

In the parts of this question, we will measure the *bias* and *variance* of each of our estimators. The *bias* is defined to be

$$E[\hat{X} - \mu]$$

and the *variance* is defined to be

$$\mathrm{Var}[\hat{X}].$$

(a) What is the bias of each of the four estimators above?

(b) What is the variance of each of the four estimators above?

(c) Suppose we have constructed an estimator $\hat{X}$ from some samples of X. We now want to know how well $\hat{X}$ estimates a fresh (new) sample of X. Denote this fresh sample by X'. Note that X' is an i.i.d. copy of the random variable X.

Derive a general expression for the expected squared error $E[(\hat{X} - X')^2]$ in terms of σ^2 and the bias and variance of the estimator $\hat{X}$. Similarly, derive an expression for the expected squared error $E[(\hat{X} - \mu)^2]$. Compare the two expressions and comment on the differences between them, if any.

(d) For the following parts, we will refer to expected total error as $E[(\hat{X} - \mu)^2]$. It is a common mistake to assume that an unbiased estimator is always "best." Let's explore this a bit further.

Compute the expected squared error for each of the estimators above.

(e) Demonstrate that the four estimators are each just special cases of the third estimator, but with different instantiations of the hyperparameter n_0.

(f) What happens to bias as n_0 increases? What happens to variance as n_0 increases?

(g) Say that $n_0 = \alpha n$. Find the setting for α that would minimize the expected total error, assuming you secretly knew μ and σ. Your answer will depend on σ, μ, and n.

(h) For this part, let's assume that we had some reason to believe that μ *should be small* (close to 0) and σ *should be large*. In this case, what happens to the expression in the previous part?

(i) In the previous part, we assumed there was reason to believe that μ *should be small*. Now let's assume that we have reason to believe that μ is not necessarily small, but *should be close to some fixed value* μ_0.

In terms of X and μ_0, how can we define a new random variable X' such that X' is expected to have a small mean? Compute the mean and variance of this new random variable.

(j) Draw a connection between α in this problem and the regularization parameter λ in the ridge-regression version of least-squares.

What does this problem suggest about choosing a regularization coefficient and handling our data-sets so that regularization is most effective? This is an open-ended question, so do not get too hung up on it.

2 The Ridge Regression Estimator

Recall the ridge regression estimator for $\lambda > 0$,

$$\widehat{\theta}_\lambda := \arg\min_\theta \|X\theta - y\|_2^2 + \lambda\|\theta\|_2^2.$$

Let

$$X = UDV^\top = \sum_i d_i u_i v_i^\top$$

be the singular value decomposition of X. Here U and V are orthogonal matrices, meaning that $U^\top U = I$ and $V^\top V = I$. D is a diagonal matrix.

(a) Show that the optimal weight vector $\widehat{\theta}_\lambda$ can be expressed in the form

$$\widehat{\theta}_\lambda = V\Sigma U^\top y$$

where Σ is a diagonal matrix with $\Sigma_{ii} = \dfrac{d_i}{d_i^2 + \lambda}$. Equivalently, we can write $\widehat{\theta}_\lambda$ as

$$\widehat{\theta}_\lambda = \sum_{i=1}^{d} \frac{d_i}{d_i^2 + \lambda} v_i \langle u_i, y\rangle.$$

(b) Show that

$$\|\widehat{\theta}_\lambda\|_2^2 = \sum_{i:d_i>0} \left(\frac{d_i}{d_i^2 + \lambda}\right)^2 (u_i^\top y)^2.$$

(c) Recall the least-norm least-squares solution is $\widehat{\theta}_{LN,LS}$ from Discussion Section 6. Show that if $\widehat{\theta}_{LN,LS} = 0$, then $\widehat{\theta}_\lambda = 0$ for all $\lambda > 0$. *Hint*: Recall that in Discussion 6 we showed that $\widehat{\theta}_{LN,LS} = \sum_{i:d_i>0} d_i^{-1}\langle u_i, y\rangle v_i$. This shows that in the case where the least-norm least square solution is zero, the ridge regression solution is also zero.

(d) Show that if $\widehat{\theta}_{LN,LS} \neq 0$, then the function $f(\lambda) = \|\widehat{\theta}_\lambda\|_2^2$ is strictly decreasing and strictly positive on $(0, +\infty)$.

(e) Show that

$$\lim_{\lambda\to 0^+} \widehat{\theta}_\lambda \to \widehat{\theta}_{LN,LS}.$$

Note that just because the limit of the ridge-regression objective as $\lambda \to 0^+$ is the least-squares objective, this does not immediately guarantee that the limit of the ridge solution is the least-squares solution.

(f) In light of the above, why do you think that people describe the ridge regression as "controlling the complexity" of the solution $\widehat{\theta}_\lambda$?

3 The Bias-Variance Tradeoff for Ridge Regression

Recall the statistical model for ridge regression from lecture. We have a set of sample points $\{x_i, y_i\}_{i=1}^n$ and Gaussian noise z_i. Our model follows, where the rows of X are x_i.

$$Y = Xw^* + z$$

Throughout this problem, you may assume $X^\top X$ is invertible. Recall both least-squares estimators we studied.

$$w_{\text{ols}} = \min_{w \in \mathbb{R}^d} \|Xw - y\|_2^2$$

$$w_{\text{ridge}} = \min_{w \in \mathbb{R}^d} \|Xw - y\|_2^2 + \lambda \|w\|_2^2$$

(a) Write the solution for $w_{\text{ols}}, w_{\text{ridge}}$. No need to derive it.

(b) Let $\widehat{w} \in \mathbb{R}^d$ denote any estimator of w_*. In the context of this problem, an estimator $\widehat{w} = \widehat{w}(X, Y)$ is any function which takes the data X and a realization of Y, and computes a guess of w_*.

Define the MSE (mean squared error) of the estimator $\widehat{w}$ as

$$\text{MSE}(\widehat{w}) := E\left\|\widehat{w} - w_*\right\|_2^2 .$$

Above, the expectation is taken with respect to the randomness inherent in z. Define $\widehat{\mu} := E\widehat{w}$. Show that the MSE decomposes as

$$\text{MSE}(\widehat{w}) = \left\|\widehat{\mu} - w_*\right\|_2^2 + \text{Tr}(\text{Cov}(\widehat{w})) .$$

Hint: Expectation and trace commute, so $E[\text{Tr}(A)] = \text{Tr}(E[A])$ for any square matrix A.

(c) Show that

$$E[w_{\text{ols}}] = w_*, \qquad E[w_{\text{ridge}}] = (X^\top X + \lambda I_d)^{-1} X^\top X w_* .$$

That is, w_{ols} is an *unbiased* estimator of w_*, whereas w_{ridge} is a *biased* estimator of w_*.

(d) Let $\gamma_1 \geq \gamma_2 \geq ... \geq \gamma_d$ denote the d eigenvalues of the matrix $X^\top X$ arranged in non-increasing order. First, argue that the smallest eigenvalue, γ_d, is positive (i.e. $\gamma_d > 0$). Then, show that

$$\text{Tr}(\text{Cov}(w_{\text{ols}})) = \sigma^2 \sum_{i=1}^{d} \frac{1}{\gamma_i}, \qquad \text{Tr}(\text{Cov}(w_{\text{ridge}})) = \sigma^2 \sum_{i=1}^{d} \frac{\gamma_i}{(\gamma_i + \lambda)^2} .$$

Finally, use these formulas to conclude that

$$\text{Tr}(\text{Cov}(w_{\text{ridge}})) < \text{Tr}(\text{Cov}(w_{\text{ols}})) .$$

Hint: For the ridge variance, consider writing $X^\top X$ in terms of its eigendecomposition $U \Sigma U^\top$.

