IEOR 165 – Lecture 15
Bias-Variance Tradeoff

1 Bias-Variance Tradeoff

Consider the case of parametric regression with $\beta \in \mathbb{R}$, and suppose that we would like to analyze the error of the estimate $\hat{\beta}$ in comparison to the true parameter β. There are a number of ways that we could characterize this error. For mathematical and computational reasons, a popular choice is the *squared loss*: The difference between the estimate and the true parameter is quantified as $\mathbb{E}\big((\hat{\beta} - \beta)^2\big)$.

By doing some algebra, we can better characterize the nature of this error measure. In particular, we have that

$$\begin{aligned}
\mathbb{E}\big((\hat{\beta} - \beta)^2\big) &= \mathbb{E}\big((\hat{\beta} - \mathbb{E}(\hat{\beta}) + \mathbb{E}(\hat{\beta}) - \beta)^2\big) \\
&= \mathbb{E}\big((\mathbb{E}(\hat{\beta}) - \beta)^2\big) + \mathbb{E}\big((\hat{\beta} - \mathbb{E}(\hat{\beta}))^2\big) + 2\mathbb{E}\big((\mathbb{E}(\hat{\beta}) - \beta)(\hat{\beta} - \mathbb{E}(\hat{\beta})\big) \\
&= \mathbb{E}\big((\mathbb{E}(\hat{\beta}) - \beta)^2\big) + \mathbb{E}\big((\hat{\beta} - \mathbb{E}(\hat{\beta}))^2\big) + 2(\mathbb{E}(\hat{\beta}) - \beta)(\mathbb{E}\big(\hat{\beta}) - \mathbb{E}(\hat{\beta})) \\
&= \mathbb{E}\big((\mathbb{E}(\hat{\beta}) - \beta)^2\big) + \mathbb{E}\big((\hat{\beta} - \mathbb{E}(\hat{\beta}))^2\big).
\end{aligned}$$

The term $\mathbb{E}\big((\hat{\beta} - \mathbb{E}(\hat{\beta}))^2\big)$ is clearly the variance of the estimate $\hat{\beta}$. The other term $\mathbb{E}\big((\mathbb{E}(\hat{\beta}) - \beta)^2\big)$ measures how far away the "best" estimate is from the true value, and it is common to define $\text{bias}(\hat{\beta}) = \mathbb{E}\big(\mathbb{E}(\hat{\beta}) - \beta\big)$. With this notation, we have that

$$\mathbb{E}\big((\hat{\beta} - \beta)^2\big) = (\text{bias}(\hat{\beta}))^2 + \text{var}(\hat{\beta}).$$

This equation states that the expected estimation error (as measured by the squared loss) is equal to the bias-squared plus the variance, and in fact there is a tradeoff between these two aspects in an estimate.

It is worth making three comments. The first is that if $\text{bias}(\hat{\beta}) = \mathbb{E}\big(\mathbb{E}(\hat{\beta}) - \beta\big) = 0$, then the estimate $\hat{\beta}$ is said to be *unbiased.* Second, this bias-variance tradeoff exists for vector-valued parameters $\beta \in \mathbb{R}^p$, for nonparametric estimates, and other models. Lastly, the term *overfit* is sometimes used to refer to an model with low bias but extremely high variance.

2 Example: Estimating Variance

Suppose $X_i \sim \mathcal{N}(\mu, \sigma^2)$ for $i = 1, \ldots, n$ are iid random variables, where μ, σ^2 are both unknown. Recall that we have considered two estimators for the variance

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \\ \hat{\sigma^2} &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \tfrac{n-1}{n}s^2. \end{aligned}$$

And recall that in the homework, we showed that

$$\begin{aligned} \mathbb{E}(s^2) &= \sigma^2 \\ \mathbb{E}(\hat{\sigma^2}) &= \tfrac{n-1}{n}\sigma^2. \end{aligned}$$

In this example, we would say that s^2 is an unbiased estimate of the variance because

$$\text{bias}(s^2) = \mathbb{E}\big(\mathbb{E}(s^2) - \sigma^2\big) = \mathbb{E}\big(\sigma^2 - \sigma^2\big) = 0.$$

However, the $\hat{\sigma^2}$ is not an unbiased estimate of the variance because

$$\text{bias}(\hat{\sigma^2}) = \mathbb{E}\big(\mathbb{E}(\hat{\sigma^2}) - \sigma^2\big) = \mathbb{E}\big(\tfrac{n-1}{n}\sigma^2 - \sigma^2\big) = \sigma^2/n.$$

It is tempting to say that s^2 is a better estimate of the variance than $\hat{\sigma^2}$ becuase the former is unbiased and the later is biased. However, lets examine the situation more closely. In particular, suppose we compute the variance of both estimators. From the part of the course on null hypothesis testing, we know that $s^2 \sim \sigma^2\chi^2(n-1)/n-1$, where $\chi^2(n-1)$ denotes a χ^2 distribution with $(n-1)$ degrees of freedom. The variance of a $\chi^2(n-1)$ distribution is $2(n-1)$ by properties of the χ^2 distribution. Hence, by properties of variance, we have that

$$\begin{aligned} \text{var}(s^2) &= 2\sigma^4/(n-1) \\ \text{var}(\hat{\sigma^2}) &= 2\tfrac{n-1}{n^2}\sigma^4. \end{aligned}$$

Since $n^2 > (n-1)^2$, dividing by $(n-1)n^2$ gives the inequality that $1/(n-1) > (n-1)/n^2$. This means that $\text{var}(s^2) > \text{var}(\hat{\sigma^2})$: In words, the variance of $\hat{\sigma^2}$ is lower than the variance of s^2.

Given that one estimate has lower bias and higher variance than the other, the natural question to ask is which estimate has lower estimation error? Using the bias-variance tradeoff equation, we have

$$\begin{aligned} \mathbb{E}((s^2 - \sigma^2)^2) &= (\text{bias}(s^2))^2 + \text{var}(s^2) = 0^2 + 2\sigma^4/(n-1) = 2\sigma^4/(n-1) \\ \mathbb{E}((\hat{\sigma^2} - \sigma^2)^2) &= (\text{bias}(\hat{\sigma^2}))^2 + \text{var}(\hat{\sigma^2}) = \tfrac{\sigma^4}{n^2} + 2\tfrac{n-1}{n^2}\sigma^4 = \tfrac{2n-1}{n^2}\sigma^4 = \tfrac{(2n-1)(n-1)}{2n^2} \cdot 2\sigma^4/(n-1). \end{aligned}$$

But note that $(2n-1)(n-1) < 2(n-1)(n-1) < 2n^2$. As a result, $\frac{(2n-1)(n-1)}{2n^2} < 1$. This means that

$$\mathbb{E}((\hat{\sigma^2} - \sigma^2)^2) \leq \mathbb{E}((s^2 - \sigma^2)^2),$$

and so $\hat{\sigma^2}$ has lower estimation error than s^2 when measuring error with the squared loss.

3 Stein's Paradox

The situation of estimating variance for a Gaussian where a biased estimator has less estimation error than an unbiased estimator is not an exceptional case. Rather, the general situation is that biased estimators have lower estimation errors. Such a conclusion holds in even surprising cases such as the one described below.

Suppose we have jointly independent $X_i \sim \mathcal{N}(\mu_i, 1)$ for $i = 1, \dots, n$, with $\mu_i \neq \mu_j$ for $i \neq j$. This is a model where we have n measurements from Gaussians, where the mean of each measurement is different. It is worth emphasizing that in this model we only have one measurement from each different Gaussian. Now for this model, it is natural to consider the problem of estimating the mean of each Gaussian. And the obvious choice is to use the estimator

$$\hat{\mu} = \begin{bmatrix} X_1 & X_2 & \dots & X_n \end{bmatrix}'.$$

This is an obvious choice because

$$\mathbb{E}(\hat{\mu}) = \begin{bmatrix} \mu_1 & \mu_2 & \dots & \mu_n \end{bmatrix}'.$$

It is hard to imagine that any other estimator could do better (in terms of estimation error) than this estimator, because (i) we only have one data point for each Gaussian, and (ii) each Gaussian is jointly independent.

However, this is the exact conclusion from Stein's paradox. It turns out that if $n \geq 3$, then the following *James-Stein estimator*

$$\hat{\mu}_{JS} = \Big(1 - \frac{n-2}{\|\hat{\mu}\|_2^2}\Big) \cdot \hat{\mu},$$

where $\hat{\mu}$ is as defined above, has strictly lower estimation error under the squared loss than that of $\hat{\mu}$. This is paradoxical because the estimation of the means of Gaussians can be impacted by the number of total Gaussians, even when the Gaussians are jointly independent. The intuition is that as we try to jointly estimate more means, we can reduce the estimation error by purposely adding some bias to the estimate. In this case, we bias the estimates of the means towards zero. (This is sometimes called *shrinkage* in statistics, because we are shrinking the values towards zero.) The error introduced by biasing the estimate is compensated by a greater reduction in the variance of the estimate.

4 Proportional Shrinkage for OLS

Given that it can be the case that adding some bias can improve the estimation error, it is interesting to consider how we can bias OLS estimates. Recall that the OLS estimate with data (x_i, y_i) for $x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$, $i = 1, \dots, n$ is given by

$$\hat{\beta} = \arg\min \|Y - X\beta\|_2^2,$$

where the matrix $X \in \mathbb{R}^{n\times p}$ and the vector $Y \in \mathbb{R}^n$ are such that the i-th row of X is x_i' and the i-th row of Y is y_i. One popular approach to adding shrinkage to OLS is known as Ridge Regression, Tikhonov regularization, or L2-regularization. It is given by the solution to the following optimization problem

$$\hat{\beta} = \arg\min \|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2,$$

where $\lambda \geq 0$ is a tuning parameter. (We will return to the question of how to choose the value of λ in a future lecture.)

To understand why this method adds statistical shrinkage, consider the one-dimensional special case where $x_i \in \mathbb{R}$. In this case, the estimate is given by

$$\begin{aligned}\hat{\beta} &= \arg\min \sum_{i=1}^{n} (y_i - x_i \cdot \beta)^2 + \lambda\beta^2 \\ &= \arg\min \sum_{i=1}^{n} y_i^2 - 2y_i x_i \beta + x_i^2\beta^2 + \lambda\beta^2 \\ &= \arg\min (\textstyle\sum_{i=1}^n -2y_i x_i)\beta + (\lambda + \sum_{i=1}^n x_i^2)\beta^2 + \sum_{i=1}^n y_i^2\end{aligned}$$

Next, setting the derivative of the objective equal to zero gives

$$\frac{dJ}{d\beta} = (\textstyle\sum_{i=1}^n -2y_i x_i) + 2(\lambda + \sum_{i=1}^n x_i^2)\beta = 0 \Rightarrow \hat{\beta} = (\sum_{i=1}^n y_i x_i)/(\lambda + \sum_{i=1}^n x_i^2).$$

When $\lambda = 0$, this is simply the OLS estimate. If $\lambda > 0$, then the denominator of the estimate is larger; hence, the estimate will have a smaller absolute value (i.e., it will shrink towards zero). If we can choose a good value of λ, then the estimate will have lower estimation error. We will consider how to choose the value of λ in a future lecture.