Bias-Variance Tradeoffs

compiled by Alvin Wan from Professor Benjamin Recht's lecture

It isn't immediately obvious that bias and variance are inextricably related. As it turns out, as a matter of fact, the two have an inverse relationship, meaning that the accuracy of a model is highly contingent on the balance you strike, between bias and variance. In other words, lowering bias often implies raising variance and vice versa. As we will find, any regression problem also has irreducible error, meaning that the accuracy of our model is actually limited.

1 Single Sample MLE

We will begin by analyzing the most simple of cases, which is not used in practice for soon-to-be-apparent reasons. Suppose that a single sample $x \sim \mathcal{N}(\mu, \sigma^2 I)$ is a d-dimensional vector.

1.1 Basic Estimator

We start with the maximum likelihood estimate, and then show that there are far better estimators. For any set of $d \times 1$ column vectors x_i, the maximum-likelihood estimate for $\hat{\mu}_{ML} = \frac{1}{n}\sum_i^n x_i$. Given that $n = 1$ in this case

$$\hat{\mu}_{ML} = x$$

What's more, we have that this maximum-likelihood estimate, in expectation, returns the true μ, since $x \sim \mathcal{N}(\mu, \sigma^2 I)$.

$$\mathbb{E}[\hat{\mu}_{mL}] = \mathbb{E}[x] = \mu$$

Let us consider the mean squared error.

$$\begin{aligned}\mathbb{E}[\|\hat{u}_{ML} - \mu\|^2] &= \mathbb{E}[\|x - \mu\|^2] \\ &= \mathbb{E}[(x-\mu)^T(x-\mu)]\end{aligned}$$

Since, the l2 norm is a scalar, we can take the trace of this quantity.

$$\begin{aligned}&= \mathbb{E}[\mathrm{Tr}((x-\mu)^T(x-\mu))] \\ &= \mathbb{E}[\mathrm{Tr}((x-\mu)(x-\mu)^T]\end{aligned}$$

We note first that our matrix $M = (x-\mu)(x-\mu)^T$ is a $d \times d$ matrix, making d entries along the diagonal. Then, we recall the definition of variance, which is $var(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$. We can see that every entry along the diagonal is

$$M_{ii} = (x_i - \mu_i)(x_i - \mu_i) = (x_i - \mu_i)^2$$

Considering that a random multivariate gaussian vector is identical to a vector of gaussian random variables, we know each $x_i \sim \mathcal{N}(\mu_i, \sigma^2)$. Thus, $\mu_i = \mathbb{E}[x_i]$ and $M_{ii} = (x_i - \mathbb{E}[x_i])^2$ giving us that

$$\mathbb{E}[M_{ii}] = \mathbb{E}[(x_i - \mathbb{E}[x_i])^2] = \sigma^2$$

Applying this to all entries along the diagonal of M, we have that

$$\mathbb{E}[\|\hat{u}_{ML} - \mu\|^2] = d\sigma^2$$

1.2 Scaled Estimator

Let us now consider a new estimator for $\hat{\mu}_2$. For some $0 < \alpha < 1$, take the following.

$$\hat{\mu}_2 = \alpha x$$

We find that the expectation is not μ, by linearity of expectation.

$$\mathbb{E}[\hat{\mu}_2] = \mathbb{E}[\alpha x] = \alpha \mathbb{E}[x] = \alpha\mu$$

We now compute the mean squared error. Here, we apply a common trick: Add 0 and expand the l2-norm so that $(a+b)^2 = a^2 + 2ab + b^2$. Below, let $a = \hat{\mu}_2 - \alpha\mu$ and $b = \mu(\alpha - 1)$.

$$\begin{aligned}
\mathbb{E}[\|\hat{\mu}_2 - \mu\|^2] &= \mathbb{E}[\hat{\mu}_2 - \alpha\mu + \alpha\mu - \mu\|^2] \\
&= \mathbb{E}[\|(\hat{\mu}_2 - \alpha\mu) + \mu(\alpha - 1)\|^2] \\
&= \mathbb{E}[\|\hat{\mu}_2 - \alpha\mu\|^2 + 2(\hat{\mu}_2 - \alpha\mu)^T \mu(\alpha - 1) + \|\mu(\alpha - 1))\|^2] \\
&= \mathbb{E}[\|\hat{\mu}_2 - \alpha\mu\|^2] + \mathbb{E}[2(\hat{\mu}_2 - \alpha\mu)^T \mu(\alpha - 1)] + \mathbb{E}[\|\mu(\alpha - 1))\|^2]
\end{aligned}$$

We consider the first term $\mathbb{E}[\|\hat{\mu}_2 - \alpha\mu\|^2]$. We see that $\hat{\mu}_2 = \alpha x$.

$$\mathbb{E}[\|(x - \mu)\alpha\|^2] = \alpha^2 \mathbb{E}[\|x - \mu\|^2]$$

By the argument in the single-variable case, we know that $\mathbb{E}[(x - \mu)^T (x - \mu)] = d\sigma^2$, so we have that the first term is $d\alpha^2\sigma^2$.

We consider the second term $\mathbb{E}[2(\hat{\mu}_2 - \alpha\mu)^T \mu(\alpha - 1)]$. Start off by plugging in $\hat{\mu}_2 = \alpha x$. We then ignore lower-order terms.

$$\begin{aligned}
&= 2\mathbb{E}[(\alpha x - \alpha\mu)^T \mu(\alpha - 1)] \\
&= 2\mathbb{E}[\alpha(\alpha - 1)(x - \mu)^T \mu] \\
&= 2\alpha(\alpha - 1)\mu\mathbb{E}[x - \mu]
\end{aligned}$$

We now that $\mathbb{E}[x - \mu] = 0$, so the above is 0.

We consider the third term $\mathbb{E}[\|\mu(\alpha - 1))\|^2]$. Since $\mu(\alpha - 1)$ is a scalar, we can pull it out of the expectation.

$$\mathbb{E}[(\|\mu(\alpha - 1))\|^2] = (\alpha - 1)^2\|\mu\|^2$$

We can multiply the term by $1 = (-1)^2$ to get our final, desired form for the third term.

$$(1 - \alpha)^2\|\mu\|^2$$

In sum, we then have the following expression

$$\mathbb{E}[\|\hat{\mu}_2 - \mu\|^2] \approx d\alpha^2\sigma^2 + (1 - \alpha)^2\|\mu\|^2$$

With this equation, we have a surprising result. Suppose that $\|\mu\| < \sigma^2 d$. Then, since $0 < \alpha < 1$, we have the following.

$$d\alpha^2\sigma^2 + (1 - \alpha)^2\|\mu\|^2 < d\sigma^2$$

This means that $\forall a \in [0, 1], \mathbb{E}[\|\hat{\mu}_2 - \mu\|^2] < \mathbb{E}[\|\hat{\mu}_{ML} - \mu\|^2]$. Thus, $\hat{\mu}_2$ may have lower mean squared error than the maximum likelihood estimate. Moreover, if we have the following value of a,

$$a = \frac{\|\mu\|^2}{\sigma^2 d + \|\mu\|^2}$$

then we *always* have that $\mathbb{E}[\|\hat{\mu}_2 - \mu\|^2] \leq \mathbb{E}[\|\hat{\mu}_{ML} - \mu\|^2]$. There actually exists a far better estimator, called the **James-Stein Estimator**, which always has lower mean squared error than $\hat{\mu}_{ML}$ when $d \geq 3$.

$$\hat{\mu}_{JS} = (1 - \frac{(d - 2)\sigma^2}{\|x\|^2})x$$

1.3 Single Sample Bias-Variance Tradeoff

We can now consider the bias-variance tradeoff in general terms. We will call the **bias** the following

$$\text{bias}(\hat{\mu}) = \|\mathbb{E}[\hat{\mu} - \mu\|^2$$

We will then call the **variance** the following

$$\text{var}(\hat{\mu}) = \mathbb{E}[\|\hat{\mu} - \mathbb{E}[\hat{\mu}]\|^2]$$

As it turns out, our mean squared error can be expressed as a sum of our bias and variance. We will employ the same tricks used in previous parts, where we add 0 and expand the squared term.

$$\begin{aligned}\mathbb{E}[\|\hat{\mu} - \mu\|] &= \mathbb{E}[\|(\hat{\mu} - \mathbb{E}[\hat{\mu}]) - (\mu - \mathbb{E}[\hat{\mu}])\|^2 \\ &= \mathbb{E}[\|\hat{\mu} - \mathbb{E}[\hat{\mu}]\|^2] - 2\mathbb{E}[\langle\|\hat{\mu} - \mathbb{E}[\hat{\mu}], \mu - \mathbb{E}[\hat{\mu}]\rangle] + \mathbb{E}[\|\mu - \mathbb{E}[\hat{\mu}]\|^2]\end{aligned}$$

The second term becomes 0, so we have the following.

$$\mathbb{E}[\|\hat{\mu} - \mu\|] = \mathbb{E}[\|\hat{\mu} - \mathbb{E}[\hat{\mu}]\|^2] + \mathbb{E}[\|\mu - \mathbb{E}[\hat{\mu}]\|^2] = \text{bias}(\hat{\mu}) + \text{var}(\hat{\mu})$$

2 Regression MLE

Generate $x_1, ...x_n$ using random process, where the function f is unknown and the noise ϵ_i are unknown. Let us assume, however, that the ϵ_i are independent of x_i and that the expected noise is 0, or$E[\epsilon] = 0$.

$$y_i = f(x_i) + \epsilon_i$$

Let our estimator be

$$\hat{y} = h(x)$$

We will now compute the mean squared error. Start by plugging in $y = f(x) + \epsilon$.

$$\begin{aligned}
\mathbb{E}[(\hat{y} - y)^2] &= \mathbb{E}[(h(x) - y)^2] \\
&= \mathbb{E}[(h(x) - f(x) - \epsilon)^2] \\
&= \mathbb{E}[(h(x) - f(x))^2] - 2\mathbb{E}[\langle h(x) - f(x), \epsilon \rangle] + \mathbb{E}[\epsilon^2]
\end{aligned}$$

Since ϵ is independent of all x and $\epsilon \sim \mathcal{N}(0, \sigma^2)$, meaning $\mathbb{E}[\epsilon] = 0$,

$$2\mathbb{E}[\langle h(x) - f(x), \epsilon \rangle] = 2\mathbb{E}[h(x) - f(x)]E[\epsilon] = 0$$

Again, $\mathbb{E}[\epsilon] = 0$, so $\text{var}(\epsilon) = \mathbb{E}[\epsilon^2]$. This means that the mean squared error is

$$\begin{aligned}
&= \mathbb{E}[(h(x) - f(x))^2] + \text{var}(\epsilon) \\
&= \mathbb{E}[(h(x) - \mathbb{E}[h(x)]) + (\mathbb{E}[h(x)] - f(x))^2] + \text{var}(\epsilon) \\
&= \mathbb{E}[(h(x) - \mathbb{E}[h(x)])^2] + 2\mathbb{E}[\langle h(x) - \mathbb{E}[h(x)], \mathbb{E}[h(x)] - f(x) \rangle] + \mathbb{E}[(\mathbb{E}[h(x)] - f(x))^2] + \text{var}(\epsilon)
\end{aligned}$$

All x are independent, so the middle term becomes 0. Keep in mind that we only control h, thus we cannot control $\text{var}(\epsilon)$ which is only a function of our noise. This means that the mean squared error for any regression problem is composed of three segments:

$$\mathbb{E}[(\hat{y} - y)^2] = \mathbb{E}[(h(x) - \mathbb{E}[h(x)])^2] + \mathbb{E}[(\mathbb{E}[h(x)] - f(x))^2] + \text{var}(\epsilon)$$

1. **bias**: fitting linear functions to polynomials
2. **variance**: fitting high degree polynomials to data
3. **Irreducible error**: noisy or incorrect labels

3 Conclusions

Let f_* be an unknown function that truly models our data. We can in theory consider the space of all functions and minimize the loss over all functions.

$$\text{minimize}_f \mathbb{E}[\text{loss}(f(x_i), y_i)] \to f_*$$

We find, of course, that minimizing over uncountably many functions isn't feasible. As a result, we consider a class of functions and minimize over those. Let f_c be the best function of this particular class.

$$\text{minimize}_{f \in \text{class}} \mathbb{E}[\text{loss}(f(x_i), y_i)] \to f_c$$

We then find that this is likewise infeasible, as computing the expectation of our function over all inputs is too computationally expensive. Thus, we end up with our common formulation of the regression problem.

$$\text{minimize}_{f \in \text{class}} \sum_{i=1}^{n} \text{loss}(f(x_i), y_i)] \to f_s$$

We note that for this problem, we have a general form for our loss.

$$\begin{aligned} R[f_s] &= R[f_s] - R[f_c] \text{ (variance)} \\ &+ R[f_c] - R[f_*] \text{ (bias)} \\ &+ R[f_*] \text{ (irreducible error)} \end{aligned}$$

As a result, all problems of this form have risk composed of bias, variance, and irreducible error.