Informal Notes for Mathematics for Machine Learning
Imperial College London

Lecture 12: Bias-Variance Decomposition

Lecturer: Matthew Wicker

1 Learning Objectives

In our last lecture we studied further concentration inequalities and discussed how Hoeffding's inequality can be extended to give us a generalization error bound for a class of learning algorithms. We also discussed how cross-validation can help us select hyper-parameters while making the most of what limited data we have. In this lecture, we will complete our exposition of model validation by taking a look at the bias and variance of our risk/loss estimates from previous lectures.

2 Bias and Variance of Estimators

Previous lectures have involved us estimating the expected loss of one of our machine learning models, but we have not yet given a general definition of an estimator. Before giving this, however, we must first understand what a statistic is:

Definition 2.1. Statistic A statistic S is a random variable that is a function of some data $\mathcal{D}$, $S = g(\mathcal{D})$ where the data $\mathcal{D}$ is a collection of random variables.

Importantly, S is a random variable not because the function g is random, but the inputs to g are random. Now, an *estimator* is a statistic that is intended to estimate a property or parameter of the distribution from which $\mathcal{D}$ is drawn. We will often denote the estimator of a statistic dependent on a set of data with $|\mathcal{D}| = n$ with the notation $\hat{S}_n$. An example of the estimator that we have seen is: $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

2.1 Bias of an Estimator

We have worked mostly with mean estimators to this point (apart from when we extended Markov's inequality to Chebychev's inequality), and in previous exposure to statistics, you may have seen two different forms of estimators for the variance:

$$\hat{\sigma^2}_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

and you may also have seen:

$$\hat{\sigma^2}'_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu)^2$$

while both of these are estimators for the variance of our distribution, one of them is the *biased* estimator and one is the *unbiased* estimator. In order to determine which is which, let us defined bias. The bias of an estimator is given by $\text{bias}(\hat{S}_n) = \mathbb{E}[\hat{S}_n - S]$. We say an estimator is unbiased if its bias is 0. In fact, we showed in our proof of Chebychev's inequality that the mean is unbiased. But let us now reason about the two estimates that we have for our variance.

$$\begin{aligned}
\text{bias}(\hat{\sigma^2}_n) &= \mathbb{E}[\hat{\sigma^2}_n - \sigma^2] \\
\text{bias}(\hat{\sigma^2}_n) &= \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - \sigma^2] \\
&= \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2] - \sigma^2
\end{aligned}$$

Let us now turn our focus to the first term in this expectation as we can pull the σ^2 out.

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i^2 - 2X_i\mu + \mu^2)^2 \\
&\propto \sum_{i=1}^{n}X_i^2 - \sum_{i=1}^{n}2X_i\mu + \sum_{i=1}^{n}\mu^2 \\
&= \left(\sum_{i=1}^{n}X_i^2\right) - 2\mu\left(\sum_{i=1}^{n}X_i\right) + n\mu^2 \\
&= \left(\sum_{i=1}^{n}X_i^2\right) - 2n\mu^2 + n\mu^2 \\
&= \left(\sum_{i=1}^{n}X_i^2\right) - n\mu^2
\end{aligned}$$

Notice above we dropped the $1/n$ term and we will need to add it back in later. Now we need to place this back inside the expectation to get:

$$\mathbb{E}\left[\left(\sum_{i=1}^{n}X_i^2\right) - n\mu^2\right] = \left(\sum_{i=1}^{n}\mathbb{E}[X_i^2]\right) - \mathbb{E}[n\mu^2]$$

Here we will use a few expectation identities: $\mathbb{E}[X_i^2] = \sigma^2 + \mu^2$ and $\mathbb{E}[\mu] = \frac{\sigma^2}{n} + \mu^2$ to get that:

$$\begin{aligned}\sum_{i=1}^{n} &\sigma^2 + \mu^2 - \frac{\sigma^2}{n} + \mu^2 \\ &= n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2 \\ &= (n-1)\sigma^2\end{aligned}$$

Plutting back in the multiplication of $\frac{1}{n}$ and subtracting σ^2 we have that:

$$\text{bias}(\hat{\sigma^2}_n) = \frac{n-1}{n}\sigma^2 - \sigma^2$$

In other words the bias is non-zero, so the estimator we have looked at is a biased estimator of the variance. It is a very good practice exercise to, without looking at these notes, show that the other formula is unbiased. It follows the same exact derivation here save you will multiply by $n-1$ at the end.

2.2 Variance of an Estimator

The variance of an estimator allows us to quantify the how far, on average, our estimators deviate from the true value they seek to estimate. Let $\hat{S}$ be an estimator for a parameter S, and let $\mathbb{E}(\hat{S})$ represent the expected value of the estimator. The variance of $\hat{S}$, denoted as $\text{Var}(\hat{S})$, is defined as the average squared difference between the individual estimates and their expected value:

$$\text{Var}(\hat{S}) = \mathbb{E}\Big[(\hat{S} - \mathbb{E}(\hat{S}))^2\Big]$$

Even without formally defining this, the above equation intuitively follows from the fact that $\hat{S}$ is a random variable and the equation above is simply the definition of the variance for any random variable. A lower variance indicates that the estimator is more precise and tends to provide estimates close to the true parameter.

3 Bias-variance decomposition

Above, we have stated the bias and variance of our estimator and discussed what it means for an estimator to be unbiased. We also saw the variance of an estimator is given by the definition of variance for a random variable. The bias-variance decomposition simply states that error of an estimate can be totally accounted for with these two quantities. Formally, we have that:

$$\text{Err}(\hat{S}) = \mathbb{E}[(\hat{S} - S)^2] = \text{Bias}^2 + \text{Var}$$

. This middle quantity should look very familiar to us. To more clearly identify this recall that a statistic S is just some function of our data $S = g(\mathcal{D})$. Plugging this in above we have:

$$\text{Err}(\hat{S}) = \mathbb{E}[(g(\mathcal{D}) - S)^2]$$

Now, it should be clear that this is just the mean squared error that we have worked with several times throughout our course. So, what is left to do is prove the first equation in this section. In order to do so, all we have to do is expand the mean squared error of our estimator, letting $\mu = \mathbb{E}[\hat{S}]$:

$$\begin{aligned}
\mathbb{E}[(\hat{S} - S)^2] &= \mathbb{E}\left[\left(\hat{S} - \mu + \mu - S\right)^2\right] \\
&= \mathbb{E}\left[\left((\hat{S} - \mu) + (\mu - S)\right)^2\right] \\
&= \mathbb{E}\left[(\hat{S} - \mu)^2 + 2(\hat{S} - \mu)(\mu - S) + (\mu - S)^2\right] \\
&= \mathbb{E}\left[(\hat{S} - \mu)^2 + (\mu - S)^2\right] \qquad (\star) \\
&= \mathbb{E}\left[(\hat{S} - \mu)^2\right] + \mathbb{E}\left[(\mu - S)^2\right] \\
&= \mathbb{E}\left[(\hat{S} - \mu)^2\right] + (\mu - S)^2
\end{aligned}$$

where the step $(\star)$ is due to the fact that $\hat{S} - \mu$ is 0 and therefore the middle term drops out. Observing our final value, we can see that this is precisely the definition of the variance of our estimator plus the bias squared.

3.1 Bias-variance trade-off in model selection

A natural question after seeing this fact is: how can I apply this knowledge to machine learning? The bias-variance decomposition gives us one way of reasoning about what we call the *bias-variance trade-off* in machine learning. To observe this clearly, let us take a look at two models from lecture that we used to introduce the idea of overfitting.

The bias-variance trade-off for machine learning can be intuitively understood as saying that for models (fit using the mean squared error) their error can be accounted for totally by the bias and the variance of their estimation and that overly-simple model will incur high bias while overly-complex models will incur high variance. In our lecture slides, we illustrate this affects a single prediction from these two models pictorially.

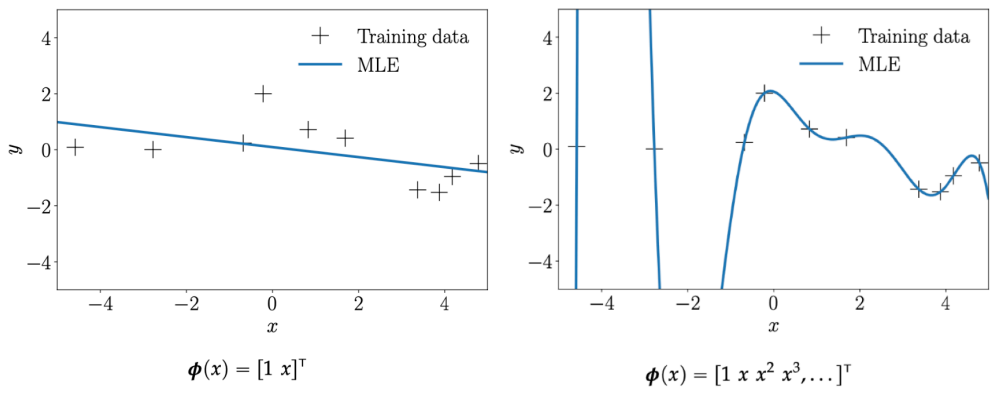

Figure 1: Left: linear regression fitting an affine basis expansion. Right: linear regression fitting a high-degree polynomial basis expansion.

3.2 Bias-variance in linear regression

Now, let us decompose the prediction error of an arbitrary linear regression model with parameters θ fit to dataset a dataset $\boldsymbol{X}, \boldsymbol{y} \sim \pi$ into its bias and variance. To simplify our analysis (since we need to analytically quantify our error) we will assume there is no modelling error. That is, there is some $\theta^\star$ such that generated our data. The predictive error of our model is given as:

$$\text{P.Err}(\theta) = \mathbb{E}_\pi[||\boldsymbol{y} - \boldsymbol{X}\theta + \epsilon||_2^2] \tag{1}$$

Above, as before, we are simply measuring the error of our model as its predictive mean squared error this time stated in terms of the ℓ_2-norm we have that:

$$\text{P.Err}(\theta) = \mathbb{E}_{\boldsymbol{X}}[\boldsymbol{X}^\top \text{Err}(\theta)\boldsymbol{X}] + \sigma^2$$

which is the bias variance decomposition but stated for our linear model (assuming no model missmatch) which is dependent on our model's error:

$$\begin{aligned}\text{Err}(\theta) &= \mathbb{E}_{\mathcal{D}\sim\pi}[(\theta(\mathcal{D}) - \theta^\star)(\theta(\mathcal{D}) - \theta^\star)^\top] \\ &= \mathbf{b}(\theta)\mathbf{b}(\theta)^\top + \mathbb{V}(\theta)\end{aligned}$$

Where the last equality is simply our bias variance decomposition (bias squared plus variance), and by definition we have that $\mathbf{b}(\theta) = \mathbb{E}_\mathcal{D}[\theta(\mathcal{D})] - \theta^\star$ and the variance is $\mathbb{V}_\mathcal{D}(\theta(\mathcal{D}))$ So what is left is to understand the bias and variance of *our* models, that is, the models that we have built up to in the course so far. We have seen that the ridge regression estimate (equivalent to the MAP) the best parameter is given by:

$$\theta^{\text{Ridge}} = (\sigma^2\lambda\mathbf{I} + \boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^{-1}\boldsymbol{y}$$

and that the OLS estimator is given by:

$$\theta^{\text{OLS}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$$

Now, under the assumption that there is some $\theta^\star$ such that generated our data (i.e., no modelling error), we know that the OLS estimator is the optimal parameter setting. Considering our observational noise we have that the OLS estimator becomes:

$$\theta^{\text{OLS}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top (\boldsymbol{X}\theta^\star + \epsilon)$$

where all we have done is substituted in the true model for $\boldsymbol{y}$ since there is no mismatch. Thus, proving that the OLS is an unbiased estimator of $\theta^\star$ is shown by verifying:

$$\mathbb{E}_\pi[(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top (\boldsymbol{X}\theta^\star + \epsilon)] = \theta^\star$$

this follows given that our noise model has mean 0. But what about for our ridge regression estimator? Well, by the same logic we get:

$$\mathbb{E}_\pi[(\sigma^2\lambda\mathbf{I} + \boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top (\boldsymbol{X}\theta^\star + \epsilon)] = \mathbb{E}_{\boldsymbol{X}}[(\sigma^2\lambda\mathbf{I} + \boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{X}]\theta^\star$$

which shows us that our ridge regression estimator is actually a biased estimator of the true parameter in this case, to completely show this we should of course use the definition of bias of an estimator. Here, we will consider the bias as a function of the parameter λ as it's introduction was the only thing that has changed between our ridge regression estimate and the OLS estimator which we showed to be biased.

$$\begin{aligned}\mathbf{b}(\lambda) := \mathbb{E}_\pi[\theta^{\text{Ridge}}] - \theta^\star &= (\sigma^2\lambda\mathbf{I} + \boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{X}\theta^\star - \theta^\star \\ &= -\mathbb{E}_{\mathbf{X}_{\text{train}}}[\sigma^2\lambda(\sigma^2\lambda\mathbf{I} + \boldsymbol{X}^\top \boldsymbol{X})^{-1}]\theta^\star\end{aligned}$$

Notice of course that when $\lambda = 0$ we recover the unbiased case! We can also state the variance of the estimator when $(\lambda > 0)$ using the definition:

$$\mathbb{V}(\lambda) := \mathbb{E}_{\mathbf{X}}[\sigma^2(\sigma^2\lambda\mathbf{I} + \boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{X}(\sigma^2\lambda\mathbf{I} + \boldsymbol{X}^\top \boldsymbol{X})^{-1}]$$

Variance of linear regression estimator $(\lambda = 0)$:

$$\mathbb{V}(0) = \mathbb{E}_{\mathbf{X}_{\text{train}}}[\sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}]$$

Discussion In previous lectures, we have shown that regularization helps us generalize, yet here we have shown that the ridge regression estimator is biased. So, if regression does indeed help in reducing generalization error, then it must be the case that the improvement comes from the variance term in our bias variance trade-off. This is exactly what we ask you to show in your exercises for today.

A Lecture 12: Bias-variance tradeoff

Question 1 (Variance reduction for Ridge regression)**.** In our notes we use θ^{ridge} here to make notation more compate we wull denote the ridge regression estimator $\theta_R^*(\mathcal{D})$. Consider the covariance matrix of the ridge regression estimator which depends on λ:

$$\mathbf{V}(\lambda) = \mathbb{V}_{\mathcal{D}\sim\pi^N}[\theta_R^*(\mathcal{D})], \quad \theta_R^*(\mathcal{D}) := \arg\min_\theta \frac{1}{2\sigma^2}||\boldsymbol{y} - \boldsymbol{X}\theta||_2^2 + \frac{\lambda}{2}||\theta||_2^2.$$

Show that $\mathbf{V}(\lambda) \preceq \mathbf{V}(0)$ for all $\lambda > 0$. (We assume $\boldsymbol{X}^\top \boldsymbol{X}$ is invertible.)

Question 2 (Bias-Variance tradeoff in Ridge regression)**.** Continuing Question 1, let us write the bias of the ridge regression estimator $\theta_R^*(\mathcal{D})$ as (under no model error assumption and assume the ground-truth parameter is $\theta^\star$)

$$\mathbf{b}(\theta_R^*) = \mathbb{E}_{\mathcal{D}\sim p_{data}^N}[\theta_R^*(\mathcal{D})] - \theta^\star.$$

Show that when $0 \leq \lambda \leq \frac{2}{||\theta^\star||_2^2}$ we have

$$\mathbf{b}(\lambda)\mathbf{b}(\lambda)^\top + \mathbf{V}(\lambda) \preceq \mathbf{V}(0).$$

This result is immediately useful to show that the expected test error of Ridge regression can be smaller than the usual linear regression (i.e., MLE estimate).

Question 1 (Variance reduction for Ridge regression)**.** In our notes we use θ^{ridge} here to make notation more compate we wull denote the ridge regression estimator $\theta_R^*(\mathcal{D})$. Consider the covariance matrix of the ridge regression estimator which depends on λ:

$$\mathbf{V}(\lambda) = \mathbb{V}_{\mathcal{D}\sim\pi^N}[\theta_R^*(\mathcal{D})], \quad \theta_R^*(\mathcal{D}) := \arg\min_\theta \frac{1}{2\sigma^2}||\boldsymbol{y} - \boldsymbol{X}\theta||_2^2 + \frac{\lambda}{2}||\theta||_2^2.$$

Show that $\mathbf{V}(\lambda) \preceq \mathbf{V}(0)$ for all $\lambda > 0$. (We assume $\boldsymbol{X}^\top\boldsymbol{X}$ is invertible.)