Bias/Variance Tradeoff

A parameter is some quantity about a distribution that we would like to know. We'll estimate the parameter θ using an estimator $\hat{\theta}$. The bias of estimator $\hat{\theta}$ for parameter θ is defined as:

• $\text{Bias}(\hat{\theta},\theta) := \mathbf{E}(\hat{\theta}) - \theta$, where the expectation is with respect to the distribution that $\hat{\theta}$ is constructed from.

An estimator whose bias is 0 is called *unbiased*. Contrast bias with:

• $\text{Var}(\hat{\theta}) = \mathbf{E}(\hat{\theta} - \mathbf{E}(\hat{\theta}))^2$.

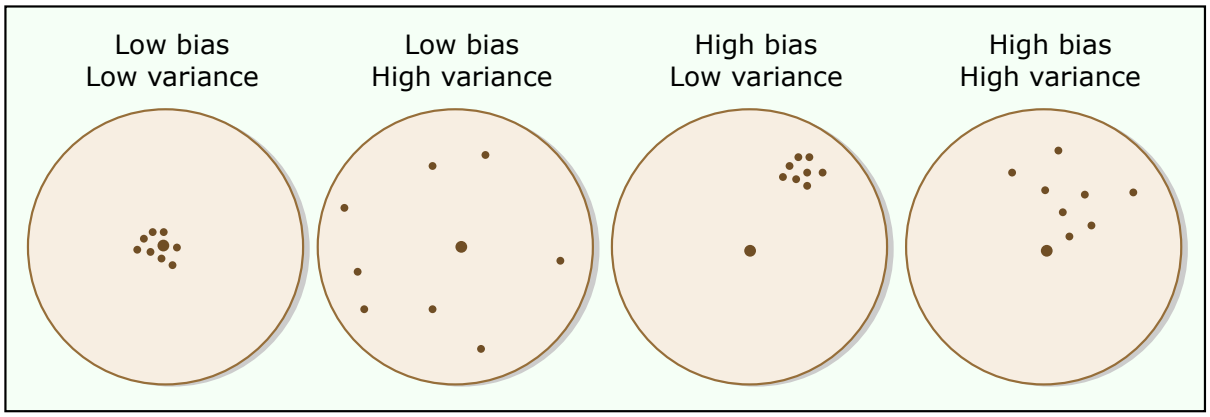

Image by MIT OpenCourseWare.

Of course, we'd like an estimator with low bias and low variance.

A little bit of decision theory

(The following is based on notes of David McAllester.)

Let's say our data come from some distribution D on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y} \subset \mathbf{R}$. Usually we don't know D (we instead only have data) but for the moment, let's say we know it. We want to learn a function $f : \mathcal{X} \to \mathcal{Y}$.

Then if we could choose any f we want, what would we choose? Maybe we'd choose f to minimize the least squares error:

$$\mathbf{E}_{x,y\sim D}[(y - f(x))^2].$$

It turns out that the f^* that minimizes the above error is the conditional expectation!

Draw a picture

Proposition.

$$f^*(x) = \mathbf{E}_y[y|x].$$

Proof. Consider each x separately. For each x there's a marginal distribution on y. In other words, look at $\mathbf{E}_y[(y - f(x))^2|x]$ for each x. So, pick an x. For this x, define $\bar{y}$ to be $\mathbf{E}_y[y|x]$. Now,

$$\begin{aligned}
&\mathbf{E}_y[(y - f(x))^2|x] \\
&= \mathbf{E}_y[(y - \bar{y} + \bar{y} - f(x))^2|x] \\
&= \mathbf{E}_y[(y - \bar{y})^2|x] + \mathbf{E}_y[(\bar{y} - f(x))^2|x] + 2\mathbf{E}_y[(y - \bar{y})(\bar{y} - f(x))|x] \\
&= \mathbf{E}_y[(y - \bar{y})^2|x] + (\bar{y} - f(x))^2 + 2(\bar{y} - f(x))\mathbf{E}_y[(y - \bar{y})|x] \\
&= \mathbf{E}_y[(y - \bar{y})^2|x] + (\bar{y} - f(x))^2
\end{aligned}$$

where the last step follows from the definition of $\bar{y}$.

So how do we pick $f(x)$? Well, we can't do anything about the first term, it doesn't depend on $f(x)$. The best choice of $f(x)$ minimizes the second term, which happens at $f(x) = \bar{y}$, where remember $\bar{y} = \mathbf{E}_y[y|x]$.

So we know for each x what to choose in order to minimize $\mathbf{E}_y[(y - f(x))^2|x]$. To complete the argument, note that:

$$\mathbf{E}_{x,y}[(y - f(x))^2] = \mathbf{E}_x[\mathbf{E}_y[(y - f(x))^2|x]]$$

and we have found the minima of the inside term for each x. ■

Note that if we're interested instead in the absolute loss $\mathbf{E}_{x,y}[|y - f(x)|]$, it is possible to show that the best predictor is the conditional median, that is, $f(x) = \text{median}[y|x]$.

Back to Bias/Variance Decomposition

Let's think about a situation where we created our function f using data S. Why would we do that of course?

We have training set $S = (x_1, y_1), \ldots, (x_m, y_m)$, where each example is drawn iid from D. We want to use S to learn a function $f_S : \mathcal{X} \to \mathcal{Y}$.

We want to know what the error of f_S is on average. In other words, we want to know what $\mathbf{E}_{x,y,S}[(y - f_S(x))^2]$ is. That will help us figure out how to minimize it. This is going to be a neat result - it's going to decompose into bias and variance terms!

First, let's consider some learning algorithm (which produced f_S) and its expected prediction error:

$$\mathbf{E}_{x,y,S}[(y - f_S(x))^2].$$

Remember that the estimator f_S is random, since it depends on the randomly drawn training data. Here, the expectation is taken with respect to a new randomly drawn point $x, y \sim D$ and training data $S \sim D^m$.

Let us define the mean prediction of the algorithm at point x to be:

$$\bar{f}(x) = \mathbf{E}_S[f_S(x)].$$

In other words, to get this value, we'd get infinitely many training sets, run the learning algorithm on all of them to get infinite predictions $f_S(x)$ for each x. Then for each x we'd average the predictions to get $\bar{f}(x)$.

We can now decompose the error, at a fixed x, as follows:

$$\begin{aligned}
&\mathbf{E}_{y,S}[(y - f_S(x))^2] \\
&\quad = \mathbf{E}_{y,S}[(y - \bar{y} + \bar{y} - f_S(x))^2] \\
&\quad = \mathbf{E}_y(y - \bar{y})^2 + \mathbf{E}_S(\bar{y} - f_S(x))^2 + 2\mathbf{E}_{y,S}[(y - \bar{y})(\bar{y} - f_S(x))].
\end{aligned}$$

The third term here is zero, since $\mathbf{E}_{y,S}[(y-\bar{y})(\bar{y}-f_S(x))] = \mathbf{E}_y(y-\bar{y})\mathbf{E}_S(\bar{y}-f_S(x))$, and the first part of that is $\mathbf{E}_y(y - \bar{y}) = 0$.

The first term is the variance of y around its mean. We don't have control over that when we choose f_S. This term is zero if y is deterministically related to x.

Let's look at the second term:

$$\begin{aligned}
&\mathbf{E}_S(\bar{y} - f_S(x))^2 \\
&\quad = \mathbf{E}_S(\bar{y} - \bar{f}(x) + \bar{f}(x) - f_S(x))^2 \\
&\quad = \mathbf{E}_S(\bar{y} - \bar{f}(x))^2 + \mathbf{E}_S(\bar{f}(x) - f_S(x))^2 + 2\mathbf{E}_S[(\bar{y} - \bar{f}(x))(\bar{f}(x) - f_S(x))]
\end{aligned}$$

The last term is zero, since $(\bar{y} - \bar{f}(x))$ is a constant, and $\bar{f}(x)$ is the mean of $f_S(x)$ with respect to S. Also the first term isn't random. It's $(\bar{y} - \bar{f}(x))^2$.

Putting things together, what we have is this (reversing some terms):

$$\mathbf{E}_{y,S}[(y - f_S(x))^2] = \mathbf{E}_y(y - \bar{y})^2 + \mathbf{E}_S(\bar{f}(x) - f_S(x))^2 + (\bar{y} - \bar{f}(x))^2.$$

In this expression, the second term is the variance of our estimator around its mean. It controls how our predictions vary around its average prediction. The third term is the bias squared, where the bias is the difference between the average prediction and the true conditional mean.

We've just proved the following:

Theorem.

For each fixed x, $\mathbf{E}_{y,S}[(y - f_S(x))^2] = \mathrm{var}_{y|x}(y) + \mathrm{var}_S(f_S(x)) + \mathrm{bias}(f_S(x))^2$.

So

$$\mathbf{E}_{x,y,S}[(y - f_S(x))^2] = \mathbf{E}_x[\mathrm{var}_{y|x}(y) + \mathrm{var}_S(f_S(x)) + \mathrm{bias}(f_S(x))^2].$$

That is the bias-variance decomposition.

The "Bias-Variance" tradeoff: want to choose f_S to balance between reducing the second and third terms in order to make the lowest MSE. We can't just minimize one or the other, it needs to be a balance. Sometimes, if you are willing to inject some bias, this can allow you to substantially reduce the variance. E.g., modeling with lower degree polynomials, rather than higher degree polynomials.

Question: Intuitively, what happens to the second term if f_S fits the data perfectly every time (overfitting)?

Question: Intuitively, what happens to the last two terms if f_S is a flat line every time?

The bottom line: In order to predict well, you need to strike a balance between bias and variance.

- The variance term controls wiggliness, so you'll want to choose simple functions that can't yield predictions that are too varied.
- The bias term controls how close the average model prediction is close to the truth, $\bar{y}$. You'll need to pay attention to the data in order to reduce the bias term.
- Since you can't calculate either the bias or the variance term, what we usually do is just impose some "structure" into the functions we're fitting with, so the class of functions we are working with is small (e.g., low degree polynomials). We then try to fit the data well using those functions. Hopefully this strikes the right balance of wiggliness (variance) and capturing the mean of the data (bias).
- One thing we like to do is make assumptions on the distribution D, or at least on the class of functions that might be able to fit well. Those assumptions each lead to a different algorithm (i.e. model). How well the algorithm works or not depends on how true the assumption is.
- Even when we're not working with least squares error, we hope a similar idea holds (and will work on proving that later in the course). We'll use the same type of idea, where we impose some structure, and hope it reduces wiggliness and will still give accurate predictions.

Go back to the other notes!

15.097 Prediction: Machine Learning and Statistics
Spring 2012