Another view on generalization

We have formalized the problem of supervised learning as finding a function h in a given set $\mathcal{H}$ that minimizes the risk $R(h)$. In the context of classification we hope to approximate the Bayes classifier, while in the context of regression we hope to approximate some true underlying function. We have already seen that the choice of $\mathcal{H}$ must strike a delicate tradeoff between two desirable characteristics:

- a more complex $\mathcal{H}$ leads to better chance of *approximating* ideal classifier/function;
- a less complex $\mathcal{H}$ leads to better chance of *generalizing* to unseen data.

The problem with a more complex $\mathcal{H}$ is that, while it can do a good job of approximating the ideal classifier/function, in the absence of a sufficient amount of data it is hard to determine the *correct* function from data and we run the risk of *overfitting*.

In the context of classification, we have already seen that the tradeoff can be precisely quantified in terms of the *VC generalization bound*, which takes the form

$$R(h) \leq \widehat{R}_n(h) + \epsilon(\mathcal{H}, n) \text{ with high probability.}$$

Here we develop an alternative method to quantify the tradeoff called the *bias-variance decomposition* which takes the form

$$R(h) \approx \text{noise} + \text{bias}^2 + \text{variance}.$$

Here, the *noise* represents fundamental/unavoidable error inherent in the problem, the *bias* captures how well $\mathcal{H}$ can approximate the optimal $h^\star$, while the *variance* captures how likely we are to pick a good $h \in \mathcal{H}$. This approach generalizes more easily to regression than the VC dimension approach developed for classification.

Georgia Tech ECE 6254 Spring 2024; Notes by M. Bloch, M. A. Davenport, and J. Romberg. Last updated 15:13, February 7, 2024

The bias-variance decomposition

We formalize the bias-variance tradeoff assuming the following:

- $h^\star : \mathbb{R}^d \to \mathbb{R}$ is the unknown target function that we are trying to learn;
- $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ is the dataset, where $(\boldsymbol{x}_i, y_i)$ are independent and identically distributed; specifically, $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i = f(\boldsymbol{x}_i) + \varepsilon_i \in \mathbb{R}$, where ε_i is a zero-mean noise random variable independent of $\boldsymbol{x}_i$ with variance σ_ε^2 (for instance $\epsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$);
- $h_\mathcal{D} : \mathbb{R}^d \to \mathbb{R}$ is our choice of function in $\mathcal{H}$, selected using $\mathcal{D}$;
- The performance of $h_\mathcal{D}$ is measured in terms of the mean squared error $R(h_\mathcal{D}) = \mathbb{E}_{XY}\left(h_\mathcal{D}(X) - Y\right)^2$;

Note that the random variables (X, Y) denote the data at *testing* and should not be confused with the random variables in $\mathcal{D}$ representing the *training* data, and that since $\mathcal{D}$ is itself random, the metric $R(h_\mathcal{D})$ is random and we will ultimately be interested in $\mathbb{E}_\mathcal{D}\left[R(h_\mathcal{D})\right]$.

Lemma 1 (Bias-variance decomposition)

$$\mathbb{E}_\mathcal{D}\left[R(h_\mathcal{D})\right] = \sigma_\varepsilon^2 + \mathbb{E}_X\left[\mathrm{var}\left(h_\mathcal{D}(X)\right)|X\right] + \mathbb{E}_X\left[\mathrm{bias}(h_\mathcal{D}(X))^2|X\right]$$

with

$$\mathrm{var}\left(h_\mathcal{D}(X)\right) = \mathbb{E}_\mathcal{D}\left[\left(h_\mathcal{D}(X) - \mathbb{E}_\mathcal{D}\, h_\mathcal{D}(X)\right)^2\right]$$

$$\mathrm{bias}(h_\mathcal{D}(X)) = \mathbb{E}_\mathcal{D}\left[h_\mathcal{D}(X)\right] - h^\star(X)$$

Proof To simplify notation, we set $\bar{h}(X) = \mathbb{E}_{\mathcal{D}}\left[h_{\mathcal{D}}(X)\right]$. Then,

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}\left[R(h_{\mathcal{D}})\right] &= \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{XY}\left[(h_{\mathcal{D}}(X) - Y)^2\right]\right] \\
&= \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{X\varepsilon}\left[(h_{\mathcal{D}}(X) - h^\star(X) - \varepsilon)^2\right]\right] \\
&= \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{X\varepsilon}\left[(h_{\mathcal{D}}(X) - \bar{h}(X) + \bar{h}(X) - h^\star(X) - \varepsilon)^2\right]\right] \\
&= \mathbb{E}_{\mathcal{D}}\mathbb{E}_X\mathbb{E}_\varepsilon\left[(h_{\mathcal{D}}(X) - \bar{h}(X))^2 + (\bar{h}(X) - h^\star(X))^2 + \varepsilon^2\right. \\
&\qquad + 2(h_{\mathcal{D}}(X) - \bar{h}(X))(\bar{h}(X) - h^\star(X)) \\
&\qquad \left. -2(h_{\mathcal{D}}(X) - \bar{h}(X))\varepsilon - 2(\bar{h}(X) - h^\star(X))\varepsilon\right]
\end{aligned}$$

Note that in the final equality above we have used the fact that $\mathcal{D}$, X, and ε are independent. Notice that

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}\mathbb{E}_X\mathbb{E}_\varepsilon\left[(h_{\mathcal{D}}(X) - \bar{h}(X))^2\right] &= \mathbb{E}_X\left[\mathrm{var}\left(h_{\mathcal{D}}(X)|X\right)\right] \\
\mathbb{E}_{\mathcal{D}}\mathbb{E}_X\mathbb{E}_\varepsilon\left[(\bar{h}(X) - h^\star(X))^2\right] &= \mathbb{E}_X\left[\mathrm{bias}(h_{\mathcal{D}}(X))^2\right] \\
\mathbb{E}_{\mathcal{D}}\mathbb{E}_X\mathbb{E}_\varepsilon\left[\varepsilon^2\right] &= \sigma_\varepsilon^2.
\end{aligned}$$

The last three terms turn out to be zero since

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\mathbb{E}_X\left[(h_{\mathcal{D}}(X) - \bar{h}(X))(\bar{h}(X) - h^\star(X))\right] \\
&\quad = \mathbb{E}_X\mathbb{E}_{\mathcal{D}}\left[(h_{\mathcal{D}}(X) - \bar{h}(X))(\bar{h}(X) - h^\star(X))\right] \\
&\quad = \mathbb{E}_X\left[(\mathbb{E}_{\mathcal{D}}\left[h_{\mathcal{D}}(X)\right] - \bar{h}(X))(\bar{h}(X) - h^\star(X))\right] \\
&\quad = 0
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}\mathbb{E}_X\mathbb{E}_\varepsilon\left[(h_{\mathcal{D}}(X) - \bar{h}(X))\varepsilon\right] &= \mathbb{E}_X\left[\mathbb{E}_{\mathcal{D}}\left[h_{\mathcal{D}}(X) - \bar{h}(X)\right]\right]\mathbb{E}_\varepsilon[\varepsilon] = 0 \\
\mathbb{E}_{\mathcal{D}}\mathbb{E}_X\mathbb{E}_\varepsilon\left[(\bar{h}(X) - h^\star(X))\varepsilon\right] &= \mathbb{E}_X\left[\bar{h}(X) - h^\star(X)\right]\mathbb{E}_\varepsilon[\varepsilon] = 0.
\end{aligned}$$

■

Example

We now consider a concrete example. Suppose that $h^\star(X) = \sin(\pi X)$ where $X \sim \text{Uniform}[-1, 1]$ and that we are given $n = 2$ noise-free training examples. We will consider two possible hypothesis sets:

- $\mathcal{H}_0 = \{h : h(x) = b, b \in \mathbb{R}\}$,
- $\mathcal{H}_1 = \{h : h(x) = ax + b, a, b \in \mathbb{R}\}$.

Since the observations are noise-free (meaning each observation is of the form $Y = \sin(\pi X)$ with no ε term), the noise component of $R(h_{\mathcal{D}})$ is zero and all we need to compute are the bias and variance terms for each hypothesis set. We will perform these computations for each hypothesis set in turn.

Bias-variance decomposition for $\mathcal{H}_0$

We will begin, for reference, by computing

$$h^\sharp = \operatorname*{arg\,min}_{h \in \mathcal{H}_0} \ \mathbb{E}_X \left[(h - h^\star(X))^2 \right].$$

Since any $h \in \mathcal{H}_0$ is just a constant, this reduces to determining

$$b^\star = \operatorname*{arg\,min}_{b \in \mathbb{R}} \ \mathbb{E}_X \left[(b - \sin(\pi X))^2 \right].$$

Note that since $X \sim \text{Uniform}[-1, 1]$ we have that the probability density function for X is

$$f_X(x) = \begin{cases} \frac{1}{2} & x \in [-1, 1], \\ 0 & \text{otherwise}, \end{cases}$$

and thus

$$\mathbb{E}_X\left[(b-\sin(\pi X))^2\right] = \int_{-1}^{1} \frac{1}{2}(b-\sin(\pi x))^2 dx$$
$$= b^2 + \frac{1}{2}.$$

Taking a derivative with respect to b and setting this equal to zero yields the requirement that $2b^\star = 0$, and hence $b^\star = 0$, i.e., $h^\sharp(x) = 0$. We can also now easily compute

$$R(h^\sharp) = \mathbb{E}_X\left[(h^\sharp(X)-\sin(\pi X))^2\right] = \mathbb{E}_X\left[(\sin(\pi X))^2\right] = \frac{1}{2}.$$

The next step in our calculations is to determine $\bar{h}(X) = \mathbb{E}_{\mathcal{D}}\left[h_{\mathcal{D}}(X)\right]$. To do this we must consider what $h_{\mathcal{D}}$ will be for any dataset $\mathcal{D} = \{(x_1, y_1), (x_2, y_2)\}$, where $y_i = \sin(\pi x_i)$ and $X_1, X_2 \sim \text{Uniform}[-1, 1]$. Given any particular $\mathcal{D}$, a natural strategy is to choose $h_{\mathcal{D}}(x)$ to minimize the empirical squared error, i.e., to choose b to minimize

$$(b-y_1)^2 + (b-y_2)^2.$$

It is straightforward to show that the optimal b is given by

$$b = \frac{y_1+y_2}{2},$$

and thus

$$h_{\mathcal{D}}(X) = \frac{y_1+y_2}{2} = \frac{\sin(\pi x_1)+\sin(\pi x_2)}{2}.$$

From this we have

$$\bar{h}(X) = \mathbb{E}_{X_1,X_2}\left[\frac{\sin(\pi X_1)+\sin(\pi X_2)}{2}\right]$$
$$= \mathbb{E}_{X_1}\left[\frac{\sin(\pi X_1)}{2}\right] + \mathbb{E}_{X_2}\left[\frac{\sin(\pi X_2)}{2}\right]$$
$$= 0.$$

The last equality can be seen either by symmetry or by manually calculating the expectation using an integral. Finally, we can conclude that

$$\begin{aligned}
\text{bias}^2 &= \mathbb{E}_X\left[\text{bias}(h_{\mathcal{D}}(X))^2\right] \\
&= \mathbb{E}_X\left[(\bar{h}(X) - h^\star(X))^2\right] \\
&= \mathbb{E}_X\left[(\sin(\pi X_2))^2\right] \\
&= \frac{1}{2}.
\end{aligned}$$

The final step is to compute the variance term. Specifically, we must compute

$$\mathbb{E}_X\left[\text{var}\left(h_{\mathcal{D}}(X)|X\right)\right] = \mathbb{E}_X\left[E_{\mathcal{D}}\left[(h_{\mathcal{D}}(X) - \bar{h}(X))^2\right]\right]$$

Note that

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}\left[(h_{\mathcal{D}}(X) - \bar{h}(X))^2\right] &= \mathbb{E}_{X_1,X_2}\left[\left(\frac{\sin(\pi X_1) + \sin(\pi X_2)}{2} - 0\right)^2\right] \\
&= \mathbb{E}_{X_1}\left[\frac{\sin^2(\pi X_1)}{4}\right] + \mathbb{E}_{X_2}\left[\frac{\sin^2(\pi X_2)}{4}\right] \\
&\quad + \mathbb{E}_{X_1,X_2}\left[\sin(\pi X_1)\sin(\pi X_2)\right] \\
&= \frac{1}{8} + \frac{1}{8} + \mathbb{E}_{X_1}\left[\sin(\pi X_1)\right]\mathbb{E}_{X_2}\left[\sin(\pi X_2)\right] \\
&= \frac{1}{4}
\end{aligned}$$

Thus

$$\mathbb{E}_X\left[\text{var}\left(h_{\mathcal{D}}(X)|X\right)\right] = \mathbb{E}_X\left[\frac{1}{4}\right] = \frac{1}{4}.$$

Putting this all together, we have

$$\mathbb{E}_{\mathcal{D}}\left[R(h_{\mathcal{D}})\right] = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}.$$

Bias-variance decomposition for $\mathcal{H}_1$

As before, we will begin by computing

$$h^\sharp = \operatorname*{arg\,min}_{h \in \mathcal{H}_1} \ \mathbb{E}_X\left[(h - h^\star(X))^2\right].$$

This is somewhat more complicated since now $h \in \mathcal{H}_1$ is an arbitrary affine function, thus we must solve

$$(a^\star, b^\star) = \operatorname*{arg\,min}_{a,b \in \mathbb{R}} \ \mathbb{E}_X\left[(aX + b - \sin(\pi X))^2\right].$$

Now, note that

$$\begin{aligned}\mathbb{E}_X\left[(aX + b - \sin(\pi X))^2\right] &= \int_{-1}^{1} \frac{1}{2}(ax + b - \sin(\pi x))^2 dx \\ &= \frac{a^2}{3} - \frac{2a}{\pi} + b^2 + \frac{1}{2}.\end{aligned}$$

Taking a derivative with respect to b and setting it equal to zero again yields the requirement that $2b^\star = 0$, and hence $b^\star = 0$. Repeating with a yields the requirement that $\frac{2}{3}a^\star - \frac{2}{\pi} = 0$, and hence $a^\star = \frac{3}{\pi}$. Thus $h^\sharp(x) = \frac{3}{\pi}x$. We can also now easily compute

$$\begin{aligned}R(h^\sharp) &= \mathbb{E}_X\left[\left(h^\sharp(X) - \sin(\pi X)\right)^2\right] \\ &= \mathbb{E}_X\left[\left(\frac{3}{\pi}X - \sin(\pi X)\right)^2\right] \\ &= \frac{(3/\pi)^2}{3} - \frac{2(3/\pi)}{\pi} + \frac{1}{2} = \frac{1}{2} - \frac{3}{\pi^2} \approx 0.196.\end{aligned}$$

Note that this is substantially smaller than $R(h^\sharp)$ for $\mathcal{H}_0$.

The next step in our calculations is to determine $\bar{h}(X) = \mathbb{E}_{\mathcal{D}}\left[h_{\mathcal{D}}(X)\right]$. As before, we must consider what $h_{\mathcal{D}}$ will be for any dataset $\mathcal{D} = \{(x_1, y_1), (x_2, y_2)\}$, where $y_i = \sin(\pi x_i)$ and $X_1, X_2 \sim \text{Uniform}[-1, 1]$. If we follow the same approach of minimizing the empirical error, this will result in the a and b that *interpolate* the dataset. The formula for this will be given by

$$a = \frac{y_2 - y_1}{x_2 - x_1} \qquad b = \frac{x_2 y_1 - x_1 y_2}{x_2 - x_1}.$$

The formula for a is simply the standard formula for the slope of a line passing through two points, and the formula for b can be obtained by simply constraining $y_1 = ax_1 + b$ for this choice of a. Thus

$$\begin{aligned} h_{\mathcal{D}}(X) &= \frac{y_2 - y_1}{x_2 - x_1} X + \frac{x_2 y_1 - x_1 y_2}{x_2 - x_1} \\ &= \frac{(\sin(\pi x_2) - \sin(\pi x_1))X + x_2 \sin(\pi x_1) - x_1 \sin(\pi x_2)}{x_2 - x_1}. \end{aligned}$$

From linearity we have

$$\begin{aligned} \bar{h}(X) &= \mathbb{E}_{X_1, X_2}\left[h_{\mathcal{D}}(X)\right] \\ &= \bar{a}X + \bar{b}, \end{aligned}$$

where

$$\begin{aligned} \bar{a} &= \mathbb{E}_{X_1, X_2}\left[\frac{(\sin(\pi X_2) - \sin(\pi X_1))}{X_2 - X_1}\right] \\ &= \int_{-1}^{1}\int_{-1}^{1} \frac{1}{4} \cdot \frac{(\sin(\pi x_2) - \sin(\pi x_1))}{x_2 - x_1} dx_1 dx_2 \\ &\approx 0.7759 \end{aligned}$$

and

$$\begin{aligned}\bar{b} &= \mathbb{E}_{X_1,X_2}\left[\frac{X_2\sin(\pi X_1) - X_1\sin(\pi X_2)}{X_2 - X_1}\right] \\ &= \int_{-1}^{1}\int_{-1}^{1}\frac{1}{4}\cdot\frac{(x_2\sin(\pi x_1) - x_1\sin(\pi x_2))}{x_2 - x_1}dx_1dx_2 \\ &= 0.\end{aligned}$$

Thus,

$$\bar{h}(X) \approx 0.7759X.$$

Finally, we can conclude that

$$\begin{aligned}\text{bias}^2 &= \mathbb{E}_X\left[\text{bias}(h_{\mathcal{D}}(X))^2\right] \\ &= \mathbb{E}_X\left[(\bar{h}(X) - h^\star(X))^2\right] \\ &= \mathbb{E}_X\left[(0.7759X - \sin(\pi X))^2\right] \\ &= \frac{(0.7759)^2}{3} - \frac{2\cdot 0.7759}{\pi} + \frac{1}{2} \approx 0.207\end{aligned}$$

The final step is to compute the variance term. Specifically, we must compute

$$\mathbb{E}_X\left[\text{var}\left(h_{\mathcal{D}}(X)|X\right)\right] = \mathbb{E}_X\left[E_{\mathcal{D}}\left[(h_{\mathcal{D}}(X) - \bar{h}(X))^2\right]\right]$$

Note that if we let

$$A_{\mathcal{D}} = \frac{\sin(\pi X_2) - \sin(\pi X_1)}{X_2 - X_1}$$

and

$$B_{\mathcal{D}} = \frac{X_2\sin(\pi X_1) - X_1\sin(\pi X_2)}{X_2 - X_1},$$

then we can write

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\left[(h_{\mathcal{D}}(X)-\bar{h}(X))^2\right] &= \mathbb{E}_{\mathcal{D}}\left[\left(A_{\mathcal{D}}X+B_{\mathcal{D}}-0.7759X\right)^2\right] \\ &= X^2\,\mathbb{E}_{\mathcal{D}}\left[(A_{\mathcal{D}}-0.7759)^2\right] \\ &\qquad + 2X\,\mathbb{E}_{\mathcal{D}}\left[(A_{\mathcal{D}}-0.7759)B_{\mathcal{D}}\right]+\mathbb{E}_{\mathcal{D}}\left[B_{\mathcal{D}}^2\right].\end{aligned}$$

Note that

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\left[(A_{\mathcal{D}}-0.7759)^2\right] &= \left(\mathbb{E}_{\mathcal{D}}\left[A_{\mathcal{D}}^2\right]-0.7759\cdot\mathbb{E}_{\mathcal{D}}\left[A_{\mathcal{D}}\right]+0.7759^2\right) \\ &\approx \left(\mathbb{E}_{\mathcal{D}}\left[A_{\mathcal{D}}^2\right]-0.7759^2\right),\end{aligned}$$

where the final (approximate) equality follows from the fact that $\mathbb{E}_{\mathcal{D}}\left[A_{\mathcal{D}}\right]=\bar{a}\approx 0.7759$. We can compute

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\left[A_{\mathcal{D}}^2\right] &= \int_{-1}^{1}\int_{-1}^{1}\frac{1}{4}\cdot\left(\frac{(\sin(\pi x_2)-\sin(\pi x_1))}{x_2-x_1}\right)^2 dx_1dx_2 \\ &\approx 2.8981.\end{aligned}$$

Similarly, we can also compute

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\left[B_{\mathcal{D}}^2\right] &= \int_{-1}^{1}\int_{-1}^{1}\frac{1}{4}\cdot\left(\frac{(x_2\sin(\pi x_1)-x_1\sin(\pi x_2))}{x_2-x_1}\right)^2 dx_1dx_2 \\ &\approx 0.9109.\end{aligned}$$

We will not bother to compute

$$\rho=\mathbb{E}_{\mathcal{D}}\left[(A_{\mathcal{D}}-0.7759)B_{\mathcal{D}}\right]$$

since, as we will see below, it does not matter. Putting this together yields

$$\begin{aligned}\mathbb{E}_X\left[\mathrm{var}\left(h_{\mathcal{D}}(X)|X\right)\right] &\approx \mathbb{E}_X\left[(2.8981-0.7759^2)X^2+2\rho X+0.9109\right] \\ &\approx 2.2961\,\mathbb{E}_X\left[X^2\right]+2\rho\,\mathbb{E}_X[X]+0.9109 \\ &= \frac{2.2961}{3}+0.9109\approx 1.676,\end{aligned}$$

where above we have used that for $X \sim \text{Uniform}[-1,1]$, $\mathbb{E}[X] = 0$ and $\mathbb{E}[X^2] = \frac{1}{3}$. Putting this all together, we have

$$\mathbb{E}_{\mathcal{D}}\left[R(h_{\mathcal{D}})\right] \approx 0.207 + 1.676 = 1.883.$$

Thus, despite the fact that $\mathcal{H}_1$ resulted in a much smaller $R(h^\sharp)$, we see that $\mathbb{E}_{\mathcal{D}}\left[R(h_{\mathcal{D}})\right]$ is *substantially* larger than before because of the much larger variance.