Statistical Learning

Lecture 01c

ANU - RSFAS - AHW

Last Updated: Thu Feb 24 09:46:36 2022

ISL Equation 2.3

- ISL Equation 2.3 (which we saw in the slides) states:

$$E[(Y - \hat{Y})^2] \;\; = \;\; [f(X) - \hat{f}(X)]^2 + V(\epsilon)$$

How do we get this decomposition? The book states:

- Let $Y = f(X) + \epsilon$.
- Let $\hat{Y} = \hat{f}(X)$.
- Let f, $\hat{f}$ and X are assumed fixed here.

$$
\begin{aligned}
E[(Y-\hat{Y})^2] &= E[(f(X)+\epsilon-\hat{f}(X))^2] \\
&= E[(\underbrace{f(X)-\hat{f}(X)}_{a}+\underbrace{\epsilon}_{b})^2] \\
&= E[a^2+2ab+b^2] \\
&= E[(f(X)-\hat{f}(X))^2+2\epsilon(f(X)-\hat{f}(X))+\epsilon^2] \\
&= E[(f(X)-\hat{f}(X))^2]+2(f(X)-\hat{f}(X))E[\epsilon]+E[\epsilon^2] \\
&= [(f(X)-\hat{f}(X))^2]+V(\epsilon)
\end{aligned}
$$

- Seems we need an additional assumption!

$$E[\epsilon]=0$$

- Also, it is a bit strange that our estimator $\hat{f}(X)$ doesn't have any variability!

More Generally

- The mean squared error (MSE) of an estimator $\hat{\theta}$ of a parameter θ is the function

$$E[(\hat{\theta} - \theta)^2]$$

- θ is a fixed unknown parameter
- $\hat{\theta}$ is an estimator of θ and is random. We generally (almost never) assume that estimators are fixed!
- $\text{Bias}(\hat{\theta}) = E[\hat{\theta}] - \theta$
- $V(\hat{\theta}) = E\left[(\hat{\theta} - E[\hat{\theta}])^2\right]$

$$
\begin{aligned}
E[(\hat{\theta}-\theta)^2] &= E(\underbrace{\hat{\theta}-E(\hat{\theta})}_{a}+\underbrace{E(\hat{\theta})-\theta}_{b})^2 \\
&= E[(\hat{\theta}-E(\hat{\theta}))^2+2(\hat{\theta}-E(\hat{\theta}))(E(\hat{\theta})-\theta)+(E(\hat{\theta})-\theta)^2] \\
&= E[(\hat{\theta}-E(\hat{\theta}))^2]+2(E(\hat{\theta})-\theta)E[(\hat{\theta}-E(\hat{\theta}))]+E[(E(\hat{\theta})-\theta)^2] \\
&= E[(\hat{\theta}-E(\hat{\theta}))^2]+E[(E(\hat{\theta})-\theta)^2] \\
&= E[(\hat{\theta}-E(\hat{\theta}))^2]+(E(\hat{\theta})-\theta)^2 \\
&= V(\hat{\theta})+[\mathrm{Bias}(\hat{\theta})]^2
\end{aligned}
$$

Loss Function Optimality

- A loss function is a non-negative function that generally increases as the distance between $f(X)$ and Y increases.
  - For example we can consider squared-error loss:

$$L(Y, f(X)) = (Y - f(X))^2$$

- Why should we set $f(x) = E[Y|X = x]$ if we are interested in minimizing mean squared error loss?

- We want to choose c such that:

$$E[(Y-c)^2|X=x]$$

is as small as possible. For ease of notation let's drop the conditioning:

$$E[(Y-c)^2] = \underbrace{E[(Y-E[Y])^2]}_{V(Y)} + \underbrace{(E[Y]-c)^2}_{Bias(Y)^2}$$

- We want to minimize this function wrt to c:

$$\min_c \; E[(Y-c)^2] = \min_c \left\{ \underbrace{E[(Y-E[Y])^2]}_{V(Y)} + \underbrace{(E[Y]-c)^2}_{Bias(Y)^2} \right\}$$

- Set $c = E[Y] \Rightarrow E[Y|X=x]$.

Back to ISL

- Let's consider ISL Equation 2.7. Here we are interested in examining the difference between a y_0 in our **Test data** using an x_0 in our **Test data**.

- $\hat{f}$ was estimated using our training data. Here we will assume it has variability and is not constant!

- Under squared-error loss we find:

$$E[(y_0 - \hat{f}(x_0))^2] \quad = \quad V(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + V(\epsilon)$$

- Bias $= E[\hat{f}(x_0)] - f(x_0)$
- $V(\hat{f}(x_0)) = E[(\hat{f}(x_0) - E[\hat{f}(x_0)])^2]$

- Let's derive the result:

$$
\begin{aligned}
E[(y_0 - \hat{f}(x_0))^2] &= E\left[\left(f(x_0) + \epsilon - \hat{f}(x_0)\right)^2\right] \\
&= E\left[\left(f(x_0) + \epsilon - \hat{f}(x_0) + E[\hat{f}(x_0)] - E[\hat{f}(x_0)]\right)^2\right] \\
&= E\left[\left(\underbrace{E[\hat{f}(x_0)] - \hat{f}(x_0)}_{a} + \underbrace{f(x_0) - E[\hat{f}(x_0)]}_{b} + \underbrace{\epsilon}_{c}\right)^2\right] \\
&= E\left[a^2 + b^2 + c^2 + 2ab + 2ac + 2bc\right] \\
&= E\left[\left(\hat{f}(x_0) - E[\hat{f}(x_0)]\right)^2\right] \\
&\quad + E\left[\left(E[\hat{f}(x_0)] - f(x_0)\right)^2\right] \\
&\quad + E\left[\epsilon^2\right] + \cdots
\end{aligned}
$$

$$
\begin{aligned}
E[2ab] &= 2E\left[\left(E[\hat{f}(x_0)] - \hat{f}(x_0)\right)\left(f(x_0) - E[\hat{f}(x_0)]\right)\right] \\
&= 2\left(f(x_0) - E[\hat{f}(x_0)]\right) E\left[\left(E[\hat{f}(x_0)] - \hat{f}(x_0)\right)\right] \\
&= 2\left(f(x_0) - E[\hat{f}(x_0)]\right)\left[\left(E[\hat{f}(x_0)] - {\color{red}E[\hat{f}(x_0)]}\right)\right] = 0
\end{aligned}
$$

$$
\begin{aligned}
E[2bc] &= 2E\left[\left(f(x_0) - E[\hat{f}(x_0)]\right)\epsilon\right] \\
&= 2(f(x_0) - E[\hat{f}(x_0)])E[\epsilon] = 0
\end{aligned}
$$

$$
\begin{aligned}
E[2ac] &= 2E\left[\left(E[\hat{f}(x_0)] - \hat{f}(x_0)\right)\epsilon\right] \\
&= 2E[\hat{f}(x_0)]E[\epsilon] - 2E[\hat{f}(x_0)\epsilon] \\
&= 0 - 2E[\hat{f}(x_0))]E[\epsilon] \\
&= 0 - 0 = 0
\end{aligned}
$$

- So we have:

$$
\begin{aligned}
E[(y_0 - \hat{f}(x_0))^2] &= E\left[\left(\hat{f}(x_0) - E[\hat{f}(x_0)]\right)^2\right] \\
&\quad + E\left[\left(E[\hat{f}(x_0)] - f(x_0)\right)^2\right] \\
&\quad + E\left[\epsilon^2\right] \\
&= V(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + V(\epsilon)
\end{aligned}
$$