Homework #2 (BST 263, Spring 2019)

Part A: Book problems

Problems 1, 2, 3, and 5 from Chapter 2 of the ISL book.

Part B: Bias-variance decomposition

In this part, you will derive the bias-variance decomposition.

Suppose the training data set is $\mathcal{D} = ((x_1, Y_1), \ldots, (x_n, Y_n))$. (The x_i's are fixed, whereas the Y_i's are random variables.) Suppose $\hat{f}_{\mathcal{D}}(x)$ is the prediction function generated by some algorithm using $\mathcal{D}$. Suppose x_0 is a fixed test point, and we want to predict the true unobserved Y_0. Define $\hat{Y}_0 = \hat{f}_{\mathcal{D}}(x_0)$. Suppose $Y_0 = f(x_0) + \varepsilon$, where $\varepsilon \perp\!\!\!\perp \mathcal{D}$ and $\mathrm{E}(\varepsilon) = 0$.

You can use the following facts without justification:

(i) The probability basics (problems 1-7) in Homework #1 apply to any real-valued random variables (not just discrete real-valued r.v.s).

(ii) $\varepsilon \perp\!\!\!\perp \mathcal{D}$ implies that $Y_0 \perp\!\!\!\perp \hat{Y}_0$. (This is because if $X \perp\!\!\!\perp Y$ then $g(X) \perp\!\!\!\perp h(Y)$ for any functions g and h. The notation $X \perp\!\!\!\perp Y$ means that X and Y are independent.)

(iii) If $X, Y \in \mathbb{R}$ are independent random variables, then $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$. (This is a special case of problem 11 from Homework #1.)

Define $Z = \hat{Y}_0 - Y_0$. Justify steps (a)-(f) below by citing one or more of the Homework #1 problems or the assumptions and definitions above. Each step can be justified in one sentence.

$$
\begin{aligned}
\mathrm{E}\big((\hat{Y}_0 - Y_0)^2\big) &\overset{\text{(a)}}{=} \mathrm{E}(Z^2) \\
&\overset{\text{(b)}}{=} \mathrm{Var}(Z) + \mathrm{E}(Z)^2 \\
&\overset{\text{(c)}}{=} \mathrm{Var}(\hat{Y}_0 - Y_0) + \mathrm{E}(\hat{Y}_0 - Y_0)^2 \\
&\overset{\text{(d)}}{=} \mathrm{Var}(\hat{Y}_0 - Y_0) + \big(\mathrm{E}(\hat{Y}_0) - \mathrm{E}(Y_0)\big)^2 \\
&\overset{\text{(e)}}{=} \mathrm{Var}(\hat{Y}_0) + \mathrm{Var}(Y_0) + \big(\mathrm{E}(\hat{Y}_0) - \mathrm{E}(Y_0)\big)^2 \\
&\overset{\text{(f)}}{=} \mathrm{Var}(\hat{Y}_0) + \mathrm{Var}(\varepsilon) + \big(\mathrm{E}(\hat{Y}_0) - f(x_0)\big)^2.
\end{aligned}
$$

(g) Which term is the “bias squared”, which is the “variance”, and which is the “noise”?