Supplement: Bias-variance decomposition for classification

Yongdai Kim

Seoul National University
1. Introduction

- We have learned that the overfitting phenomenon can be (partly) explained by the bias-variance decomposition for regression problems.

- The aim of this supplement is to give an example of classification where the bias-variance decomposition does not work for explaining overfitting.

- This supplement is based on an example in Friedman (1997, On bias, variance, 0-1 loss and the curse of dimensionality).
**Example**

- Fix \( x \).
- \( P(y = 1|x) = 0.6 \).
- \( \mathcal{L} : \) training data

Suppose we have two estimators \( f_1(x, \mathcal{L}) \) and \( f_2(x, \mathcal{L}) \) such that
  - \( f_1(x, \mathcal{L}) = 0.4 \) with probability 1 (wrt \( \mathcal{L} \))
  - \( f_2(x, \mathcal{L}) = 0.1 \) with probability 0.5 and \( = 0.7 \) with probability 0.5.

- Note that
  - \( \mathbb{E}_{\mathcal{L}}(f_1(x, \mathcal{L})) = \mathbb{E}_{\mathcal{L}}(f_2(x, \mathcal{L})) = 0.4 \)
  - \( \text{Var}_{\mathcal{L}}(f_1(x, \mathcal{L})) \leq \text{Var}_{\mathcal{L}}(f_2(x, \mathcal{L})) \)

That is, \( f_1 \) has a smaller variance than \( f_2 \) while the biases are the same. Hence, we may expect that \( f_1 \) is better than \( f_2 \).
• Let $g_k(x, \mathcal{L}) = I(f_k(x, \mathcal{L}) > 0.5)$ be the corresponding classifiers.

• The (average) misclassification errors are

$$E_{\mathcal{L}, Y} I(Y \neq g_1(x, \mathcal{L})) = 0.6 > 0.5 = E_{\mathcal{L}, Y} I(Y \neq g_2(x, \mathcal{L})).$$

That is, $g_2$ is better in prediction accuracy.
Remark

- Within my knowledge, there is no nice story in explaining overfitting for classification.

- Empirical evidences suggest that overfitting emerges very slowly in classification, which is partly because the property of 0-1 loss.

- Often, deliberately overfitted models work well.

- “Estimating a decision boundary” can be qualitatively different from “estimating the probability.”