Bias and Variance

David Rosenberg

New York University

March 17, 2015

Approximation Error and Estimation Error

- Recall the excess risk decomosition for any $f \in \mathcal{F}$:

$$\textbf{Excess Risk}(f) = \underbrace{R(f) - R(f_{\mathcal{F}}^*)}_{\text{estimation error}} + \underbrace{R(f_{\mathcal{F}}^*) - R(f^*)}_{\text{approximation error}}$$

- Restricting the hypothesis space $\mathcal{F}$
  - leads to approximation error
  - but helps to reduce estimation error (i.e. $\hat{f}$ is closer to $f_{\mathcal{F}}^*$).
- Now, we'll switch to the bias/variance terminology more common when discussing the topics of this lecture.

Bias and Variance

- Restricting the hypothesis space $\mathcal{F}$ "**biases**" the fit
  - **towards** a simpler model and
  - **away** from the best possible fit of the training data.
- Full, unpruned decision trees have very little bias.
- Pruning decision trees introduces a bias.
- **Variance** describes how much the fit changes across different random training sets.
- Decision trees are found to be high variance.

Bias and Variance for Square Loss

- Input space $\mathcal{X}$
- Output space $\mathcal{Y}$
- $(X, Y) \sim P_{\mathcal{X} \times \mathcal{Y}}$
- From Homework #1, recall that for square loss, the bayes prediction function is
$$f^*(x) = \mathbb{E}[Y \mid X = x]$$
- Let's consider a prediction function $\hat{f}$ trained on a random set of data.
- $\hat{f}$ is random because training data is random.

Excess Risk for Square Error

- Excess risk of $f \in \mathcal{F}$, conditional on $X = x$:

$$\begin{aligned}\text{ExcessRisk}(f \mid X = x) &= \underbrace{\mathbb{E}\left[(Y - f(x))^2 \mid X = x\right]}_{\text{Risk of } f} \\ &\quad - \underbrace{\mathbb{E}\left[(Y - f^*(x))^2 \mid X = x\right]}_{\text{Risk of } f^*}\end{aligned}$$

- Can show

$$\text{ExcessRisk}(f \mid X = x) = (f(x) - f^*(x))^2.$$

- In words: excess risk at x is the square difference between the prediction and the Bayes prediction.

Random Training Data $\implies$ Random Prediction Function

- A learning algorithm produces $\hat{f}$ based on training data.
- The training data is a random sample from $P_{\mathcal{X}\times\mathcal{Y}}$.
- Since the training data is random, so is $\hat{f}$.
- Thus for any fixed x, the prediction $\hat{f}(x)$ is a random variable.
- As a random variable, $\hat{f}(x)$ has an expectation and variance.
- As an estimator of $f^*(x)$, $\hat{f}(x)$ may have a bias.
- We now compute these things.

Bias-Variance Decomposition for Excess Risk

- Prediction $\hat{f}(x)$ for any fixed input x has bias and variance:

$$
\begin{aligned}
\mathrm{Bias}(\hat{f}(x)) &= \mathbb{E}\left[\hat{f}(x)\right] - f^*(x) \\
\mathrm{Var}\left(\hat{f}(x)\right) &= \mathbb{E}\left[\left(\hat{f}(x) - \mathbb{E}\left[\hat{f}(x)\right]\right)^2\right]
\end{aligned}
$$

  where the expectations are taken over the training data.
- Can show **bias-variance decomposition** for excess risk at x:

$$
\mathbb{E}\left[\left(\hat{f}(x) - f^*(x)\right)^2\right] = \left[\mathrm{Bias}(\hat{f}(x))\right]^2 + \mathrm{Var}\left(\hat{f}(x)\right)
$$

- Could we reduce variance without increasing bias?

Variance of a Mean

- Let $Z_1, \ldots, Z_n$ be independent r.v's with mean μ and variance σ^2.
- Suppose we want to estimate μ.
- We could use any single Z_i to estimate μ.
- Variance of estimate would be σ^2.
- Let's consider the average of the Z_i's.
- Average has the same expected value but smaller variance:

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Z_i\right] = \mu \qquad \operatorname{Var}\left[\frac{1}{n}\sum_{i=1}^{n} Z_i\right] = \frac{\sigma^2}{n}.$$

- Can we apply this to reduce variance of prediction models?

Averaging Independent Prediction Functions

- Suppose we have B independent training sets.
- Let $\hat{f}_1(x), \hat{f}_2(x), \ldots, \hat{f}_B(x)$ be the prediction models for each set.
- Define the average prediction function as:

$$\hat{f}_{\text{avg}}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b(x).$$

- The average prediction function has lower variance than an individual prediction function.
- But in practice we don't have B independent training sets...
- Instead, we can use **the bootstrap**.... next lecture.