Bias-Variance Trade-off and Binary Classification

Shirong Xu

October 5, 2024

1 Basics

For a random variable X, the expectation and variance are given as

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xP(x)\, dx,$$

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \int_{-\infty}^{\infty} (x - \mathbb{E}[X])^2 P(x)\, dx,$$

where $P(x)$ is the probability density function of X. Usually, if we use notation $\mathbb{E}_X(\cdot)$, it means our expectation is **taken with respect to** the randomness of X.

For a pair of random variable (X, Y), the conditional expectation $\mathbb{E}(Y|X)$ is given as

$$\mathbb{E}(Y|X) = \int_{-\infty}^{\infty} yP(y|X)\, dy,$$

where $P(y|X)$ is the conditional probability density function of Y given X. It is worth noting that conditional expectation $\mathbb{E}(Y|X)$ is a random variable of X. This is because the randomness of Y is eliminated through the expectation.

2 Bias-Variance Tradeoff

Assume the response can be represented as:

$$Y = f^{\star}(X) + \epsilon$$

where ϵ is the noise term, and it satisfies $\mathbb{E}[\epsilon|X] = 0$.

The bias-variance tradeoff pattern appears when I analyze the **testing performance** of a general model f. Consider a prediction function $f(X)$. We define the prediction error as:

$$\text{True Performance of } f : \mathbb{E}[(Y - f(X))^2].$$

The true performance evaluates f via the squared-loss of over all possible pairs of (X, Y) (The expectation).

(Objective) :We aim to minimize the expected prediction error:

$$\mathbb{E}_{X,Y}[(Y - f(X))^2]$$

We first consider the decomposition:

$$R(f) = \mathbb{E}_{X,Y}[(Y - f(X))^2] = \underbrace{\mathbb{E}_X[(f(X) - \mathbb{E}[Y|X])^2]}_{\text{Model Error}} + \underbrace{\mathbb{E}_{X,Y}[(Y - E[Y|X])^2]}_{\text{Noise}}$$

(1) $f(X) - \mathbb{E}[Y|X]$: the difference between the **prediction value** $f(X)$ and the **expected value** $\mathbb{E}[Y|X]$ (or the optimal prediction value).

(2) $Y - \mathbb{E}[Y|X]$: By the assumption $Y = f^\star(X) + \epsilon$, we have $Y - \mathbb{E}[Y|X] = \epsilon$. This error is **irreducible**. Therefore, finding a f to minimize $R(f)$ is equivalent to finding a f to minimize $\mathbb{E}_X[(f(X) - \mathbb{E}[Y|X])^2]$.

Conclusion: If $f(X)$ is closer to $\mathbb{E}[Y|X]$, then $f(X)$ will have better testing performance.

In practice, the closeness between $f(X)$ and $\mathbb{E}[Y|X]$ depends on **two factors** if f is obtained from a dataset $D = \{(x_i, y_i)\}_{i=1}^n$. We can consider a model specification $\mathcal{F} = \{f(x) = \beta x : \beta \in \mathbb{R}\}$. The following two are **equivalent**.

$$\widehat{f} = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^n (y_i - f(x_i))^2 \Leftrightarrow \widehat{\beta} = \arg\min_{\beta \in \mathbb{R}} \sum_{i=1}^n (y_i - \beta x_i))^2$$

Then, these two factors are

(1) **The stability of** $\widehat{\beta}$: This is because $\widehat{\beta}$ is estimated from a finite dataset. Consequently, the estimated $\widehat{\beta}$ may deviate significantly from the true value. This situation can be allieviated by a larger training dataset (More samples are used for estimation, our estimator $\widehat{\beta}$ is more accurate).

(2) **Whether $\mathcal{F}$ correctly speifies** $\mathbb{E}[Y|X]$: For example, if $\mathbb{E}[Y|X] = x + x^2$, then using $\mathcal{F} = \{f(x) = \beta x : \beta \in \mathbb{R}\}$ actually mis-specifies the groud truth model. Even if we can estimate β accurately, the estimated model $\widehat{f} = \widehat{\beta} x$ still deviates from $\mathbb{E}[Y|X] = x + x^2$ since x^2 term is missing.

Therefore, we can consider the following analysis:

$$\text{Model Error at } x_0: \Big(\widehat{f}(x_0) - \mathbb{E}[Y|X = x_0]\Big)^2$$

Then we take the expectation with respect to the dataset $D = \{(x_i, y_i)\}_{i=1}^n$, we have

$$\mathbb{E}_D\Big(\widehat{f}(x_0) - \mathbb{E}[Y|X = x_0]\Big)^2 = \mathbb{E}_D\Big(\widehat{f}(x_0) - \mathbb{E}(\widehat{f}(x_0)) + \mathbb{E}(\widehat{f}(x_0)) - \mathbb{E}[Y|X = x_0]\Big)^2$$
$$= \underbrace{\mathbb{E}_D\Big(\widehat{f}(x_0) - \mathbb{E}(\widehat{f}(x_0))\Big)^2}_{\text{Variance of } \widehat{f}(x_0)} + \underbrace{\Big(\mathbb{E}(\widehat{f}(x_0)) - \mathbb{E}[Y|X = x_0]\Big)^2}_{Bias^2}. \tag{1}$$

Here we can view $\widehat{f}(x_0)$ as a random variable. Because the training dataset consists of random samples and $\widehat{f}$ is obtained from the dataset (can be understood as a kind of transformation). If we let $A = \widehat{f}(x_0)$, then the first term of (1) can be written as

$$\mathbb{E}_D\Big(\widehat{f}(x_0) - \mathbb{E}(\widehat{f}(x_0))\Big)^2 = \mathbb{E}_A\Big(A - \mathbb{E}(A)\Big)^2.$$

The general idea about bias and variance is that

(1) In general, if $\mathcal{F}$ has more parameters, the corresponding model complexity will increase. However, this is not always the case. For example, $f(x) = \beta_1 \times \beta_2 \times x$ and $g(x) = \beta_3 \times x$. Here f has two parameters and g has one parameter, but they have the same model complexity since they are both linear function of x.

(2) If model complexity of $\mathcal{F}$ increases, then bias will decrease. This is because a more complicated $\mathcal{F}$ allows us find a function within $\mathcal{F}$ to approximate $\mathbb{E}[Y|X = x_0]$ better.

(3) If $\mathcal{F}$ has more parameters, then $\widehat{f}$ will be more unstable (the variance of $\widehat{f}$ will increase).

From the above analysis, we know there exists a bias-variance tradeoff on the choice of $\mathcal{F}$. Therefore, a common practice is choosing a model with medium model complexity.

2.1 Why the cross-term is zero

Let us consider the cross term in the decomposition of $R(f)$:

$$\mathbb{E}_{X,Y}\Big[\Big(f(X) - \mathbb{E}[Y|X]\Big) \cdot \Big(Y - \mathbb{E}[Y|X]\Big)\Big].$$

Since $Y = f^\star(X) + \epsilon = \mathbb{E}[Y|X] + \epsilon$, we have

$$\mathbb{E}_{X,Y}\Big[\Big(f(X) - \mathbb{E}[Y|X]\Big)\cdot\Big(Y - \mathbb{E}[Y|X]\Big)\Big] = \mathbb{E}_{X,Y}\Big[\underbrace{\Big(f(X) - \mathbb{E}[Y|X]\Big)}_{\text{Not related to } Y}\cdot\epsilon\Big]$$
$$= \mathbb{E}_X\Big[\Big(f(X) - \mathbb{E}[Y|X]\Big)\cdot\mathbb{E}_Y(\epsilon)\Big] = \mathbb{E}_X\Big[\Big(f(X) - \mathbb{E}[Y|X]\Big)\cdot\mathbb{E}_\epsilon(\epsilon)\Big]$$
$$= \mathbb{E}_X\Big[\Big(f(X) - \mathbb{E}[Y|X]\Big)\cdot 0\Big] = 0.$$

3 Binary Classification

Consider a binary classification problem where:

- X represents the GPA of an applicant.
- Y is a binary random variable indicating whether the applicant is accepted by UCLA (1 if accepted, -1 otherwise).

If $X = 3.5$, then the conditional probability is

$$\eta(X) = \eta(3.5) = \mathbb{P}(Y = 1 : \text{Being accepted by UCLA}|\text{GPA} = 3.5)$$

Then we assume that the acceptance probability $\eta(X)$ has the following form

$$\eta(X) = \frac{X}{4}.$$

This means that if GPA=4, then you will be 100% accepted by UCLA. If GPA=2, then the probability you will be accepted by UCLA is 2/4=0.5=50%.

Let us consider the Bayes classifier in this problem. The risk is given as

$$R(f) = \mathbb{E}(I(f(X) \neq Y)) = \int P(x)\Big[\eta(x)I(f(X) \neq 1) + (1 - \eta(x))I(f(X) \neq -1)\Big]dx.$$

Suppose a student GPA is 3, then $\eta(3) = 3/4 = 0.75$ (the probability of being accepted is 75%). If the classifier predict **accepted** ($f(3) = 1$), the probability of wrong prediction is 0.25, that is

$$\eta(3)I(f(3) \neq 1) + (1 - \eta(3))I(f(3) \neq -1)$$
$$= \eta(3)I(1 \neq 1) + (1 - \eta(3))I(1 \neq -1) = 1 - \eta(3) = 0.25$$

If the classifier predict **rejecteded** ($f(3) = -1$), then the probability of wrong prediction is 0.25, that is

$$\begin{aligned}&\eta(3)I(f(3) \neq 1) + (1-\eta(3))I(f(3) \neq -1)\\=&\eta(3)I(-1 \neq 1) + (1-\eta(3))I(-1 \neq -1) = \eta(3) = 0.75\end{aligned}$$

Therefore, for achieving minimal prediction error (minimal $R(f)$), the optimal classifier will be based on $\eta(x) > 1/2$ or not, that is $f^{\star}(x) = \text{sign}(\eta(X) - 1/2)$.