EE3001 Machine Learning **Fall 2021**

Lecture 02. Bias-Variance Decomposition

Lecturer: Jie Wang Date: Sep 15, 2021

The major reference of this lecture is [1].

1 Introduction

Suppose that we would like to predict the value of a random variable X in the next experiment. As a random variable can take many different values, we will never know its exact value until we perform the experiment and observe the outcome. Nevertheless, we can still estimate the value of X in the next experiment. The question is, which value shall we pick to estimate X? Is there a *best* estimation?

To answer this question, first of all, we need a *measurement* that can determine how good our estimation is. One simple choice—which is widely used—is the average of the squares of errors between our estimation c and the observed values $\{x_i\}_{i=1}^n$ of X in a large number of experiments:

$$L(c) = \frac{1}{n}\sum_{i=1}^{n}(c - x_i)^2. \tag{1}$$

The *error function* $L(c)$ defined in Eq. (1) measures how accurate our estimation c of X is. Thus, the best estimation we should take is the one that can minimize the error $L(c)$. This is indeed the average of the sample values $\{x_i\}_{i=1}^n$ of X:

$$\underset{c}{\mathbf{argmin}}\ L(c) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Suppose that we can perform the experiment infinitely many times, i.e., $n \to \infty$. By the law of large numbers, we have

$$L(c) \to \mathbb{E}[(c - X)^2],$$

with probability 1. That is, the error function converges to the expectation of the square of error in probability, and the best estimation becomes the expectation of X, i.e.,

$$\underset{c}{\mathbf{argmin}}\,\mathbb{E}[(c - X)^2] = \mathbb{E}[X].$$

2 Loss Functions for Regression

Recall that, for supervised learning problems, each data instance consists of a D-dimensional input feature vector $X = (X_1, X_2, \ldots, X_D)^\top$ and the corresponding output $Y \in \mathbb{R}$. We would like to find a mapping $f(X)$ to estimate the value of Y given a sample of X. Let

$$\ell(y, f(\mathbf{x})) = (f(\mathbf{x}) - y)^2$$

be the square loss. Similar to the idea we introduced in the last section, we choose $f(X)$ by minimizing the expectation of the square loss:

$$\mathbb{E}[\ell] = \iint (y - f(\mathbf{x}))^2 p(\mathbf{x}, y)\, d\mathbf{x}\, dy, \tag{2}$$

where $p(\mathbf{x}, y)$ is the joint PDF. The expectation in Eq. (2) is the so-called **functional**, which is, roughly speaking, a mapping from **functions** to numbers.

Example 1. Let $C[a,b]$ be the set of continuous functions defined on the closed interval $[a,b]$. Notice that, the set $C[a,b]$ is an infinite dimensional vector space (or linear space). Different from the finite dimensional vector spaces we are familiar with, like $\mathbb{R}^n$ and $\mathbb{R}^{m\times n}$, the vectors/points/elements in $C[a,b]$ are functions. For any $h(x) \in C[a,b]$, the integration

$$\int_a^b h(x)dx$$

defines a functional on $C[a,b]$.

To emphasize that the expectation $\mathbb{E}[\ell]$ is a functional of the estimation f, **we denote $\mathbb{E}[\ell]$ by $J[f]$**. Then, how to find the function f^* that minimizes $J[f]$ in Eq. (2)? That is, we need to solve the optimization problem as follows:

$$\min_f \left\{ J[f] := \iint (y - f(\mathbf{x}))^2 p(\mathbf{x}, y) d\mathbf{x} dy \right\}, \tag{3}$$

where the minimizer—if exists—is a function. Calculus tells us that, we can first find the gradient of $J[f]$ and then set it to zero—with mild conditions—to solve for f^*. Although we know how to differentiate functions of multiple variables, how can we differentiate functions of functions?

Recall that, for functions of several variables, we can study its differentiability at a given point by its **directional derivatives** along all directions, which **reduces the problems with multiple variables to ones with a single variable**. Moreover, we can derive conditions for optimality by directional derivatives as well. Following this idea, we may define directional derivatives for functionals by mimicing their counterparts in the several variables cases, based on which we can derive optimality conditions to solve for the optimum. This leads to techniques named **calculus of variations**.

Specifically, let h be a function of X. For a small number $\epsilon > 0$, we add ϵh—which can be understood as a small perturbation—to f, then

$$J[f + \epsilon h] = J[f] + \epsilon \int h(\mathbf{x}) \left\{ \int -2(y - f(\mathbf{x}))p(\mathbf{x}, y)dy \right\} d\mathbf{x} + \epsilon^2 \int (h(\mathbf{x}))^2 p(\mathbf{x}) d\mathbf{x}. \tag{4}$$

Notice that, at a given point (function) f with a fixed direction h, the RHS of Eq. (4) is a function—specifically, a quadratic function—of one single variable ϵ, which reduces the optimization problem in (3) with respect to functions to that of a single variable. Then, many techniques from elementary calculus apply. In view of Eq. (4), if f^* is a **local minimum** of $J[f]$ along the direction of h, we must have

$$\int h(\mathbf{x}) \left\{ \int -2(y - f^*(\mathbf{x}))p(\mathbf{x}, y)dy \right\} d\mathbf{x} \geq 0.$$

A similar argument to $-h$ concludes that

$$\int h(\mathbf{x}) \left\{ \int -2(y - f^*(\mathbf{x}))p(\mathbf{x}, y)dy \right\} d\mathbf{x} = 0.$$

Moreover, as h is arbitrary[1], we must have

$$\int -2(y - f^*(\mathbf{x}))p(\mathbf{x}, y)dy = 0, \tag{5}$$

[1] Indeed, we have mild conditions on h regarding to its measurability, which is out of the scopre of this course.

which leads to

$$f^*(\mathbf{x}) = \frac{\int y p(\mathbf{x}, y) dy}{p(\mathbf{x})} = \int y \frac{p(\mathbf{x}, y)}{p(\mathbf{x})} dy = \int y p(y|\mathbf{x}) dy = \mathbb{E}[y|\mathbf{x}]. \tag{6}$$

Eq. (6) implies that, given an observation $\mathbf{x}$ of the input variables X, the best estimation—at least in theory—we can make for the corresponding target variable Y is the conditional expectation $\mathbb{E}[y|\mathbf{x}]$. This results is unsurprising, as it is a counterpart of the result in Section 1.

Question 1. The above discussion derives Eq. (5) as a necessary condition for f^* being a local minimum. Is it sufficient for this problem?

***Remark* 1.** The second term on the RHS of Eq. (4) is a **linear** functional of the small perturbation ϵh. Similar to differentials of functions of several variables, the second term provides a linear approximation of the (nonlinear) functional $J[f]$ at f.

By letting $\epsilon \downarrow 0$ for both sides, we have

$$\lim_{\epsilon \downarrow 0} \frac{J[f + \epsilon h] - J[f]}{\epsilon} = \int h(\mathbf{x}) \left\{ \int -2(y - f(\mathbf{x})) p(\mathbf{x}, y) dy \right\} d\mathbf{x}. \tag{7}$$

The RHS of Eq. (7), if exists, is the so-called **Gateaux differential** of J at f with increment h, which is the counterpart of directional derivatives of differentiable functions of several variables.

The result in Eq. (6) may shed new light on our understanding of the expectation of the square loss in Eq. (2). We expand the square loss as follows

$$\begin{aligned}(f(\mathbf{x}) - y)^2 =& (f(\mathbf{x}) - f^*(\mathbf{x}) + f^*(\mathbf{x}) - y)^2 \\ =& \{f(\mathbf{x}) - \mathbb{E}[y|\mathbf{x}]\}^2 + 2\{f(\mathbf{x}) - \mathbb{E}[y|\mathbf{x}]\}\{\mathbb{E}[y|\mathbf{x}] - y\} + \{\mathbb{E}[y|\mathbf{x}] - y\}^2.\end{aligned}$$

Then,

$$\begin{aligned}\mathbb{E}[\ell] = & \int \{f(\mathbf{x}) - \mathbb{E}[y|\mathbf{x}]\}^2 p(\mathbf{x}) d\mathbf{x} + 2 \int \{f(\mathbf{x}) - \mathbb{E}[y|\mathbf{x}]\} \left\{ \int \{\mathbb{E}[y|\mathbf{x}] - y\} p(\mathbf{x}, y) dy \right\} d\mathbf{x} \\ & + \iint \{\mathbb{E}[y|\mathbf{x}] - y\}^2 p(\mathbf{x}, y) dy d\mathbf{x}.\end{aligned}$$

We can see that, the second term on the RHS of the above equation vanishes as

$$\begin{aligned}\int \{\mathbb{E}[y|\mathbf{x}] - y\} p(\mathbf{x}, y) dy =& \mathbb{E}[y|\mathbf{x}] \int p(\mathbf{x}, y) dy - \int y p(y|\mathbf{x}) p(\mathbf{x}) dy \\ =& \mathbb{E}[y|\mathbf{x}] p(\mathbf{x}) - p(\mathbf{x}) \mathbb{E}[y|\mathbf{x}] \\ =& 0.\end{aligned}$$

Therefore, we have

$$\mathbb{E}[\ell] = \int \{f(\mathbf{x}) - \mathbb{E}[y|\mathbf{x}]\}^2 p(\mathbf{x}) d\mathbf{x} + \iint \{\mathbb{E}[y|\mathbf{x}] - y\}^2 p(\mathbf{x}, y) dy d\mathbf{x}. \tag{8}$$

Given $\mathbf{x}$, our estimation $f(\mathbf{x})$ of y only appears in the first term on the RHS of Eq. (8). This implies that, the best estimation—in terms of the expectation of the square loss $\mathbb{E}[\ell]$—we can make is the conditional expectation $\mathbb{E}[y|\mathbf{x}]$. As the second term is independent with our estimation $f(\mathbf{x})$, it is the irreducible minimum value of the expected square loss.

3 The Bias-Variance Decomposition

Section 2 shows that, the best estimation—**in theory**, as we put no constraints on f, which is usually not the case in practice—of y given $\mathbf{x}$ is the conditional expectation $\mathbb{E}[y|\mathbf{x}]$. However, in practice, there are two major reasons that keep us from finding the exact values of the conditional expectation.

1. The data we have to fit our model on is limited. This is why we never know the data distribution and the corresponding conditional expectation exactly. Moreover, even for the same task, different person may collect different data sets, leading to different estimation functions.

2. We choose our best estimation f^* from a set of candidate functions, i.e., the hypothesis set $\mathcal{H}$. For example, the hypothesis set for the linear regression problems only consists of all linear functions defined on input feature vectors, which probably do not include the conditional expectation.

Then, what is the best estimation we can make if we have to chose it from a hypothesis set with certain constraints? How can we properly evaluate the performance of a particular learning algorithm?

Notice that, the conditional expectation $\mathbb{E}[y|\mathbf{x}]$ is indeed a function of $\mathbf{x}$. For notational convenience, we denote $\mathbb{E}[y|\mathbf{x}]$ by $h(\mathbf{x})$. Furthermore, let $f_{\mathcal{D}}(\mathbf{x})$ be the prediction function (our estimation) returned by a particular learning algorithm on a given data set $\mathcal{D}$. Due to the randomness in $\mathcal{D}$, a reasonable way to assess the performance of the learning algorithm is to take average over many different data sets sampled from the same distribution.

Specifically, for a particular data set $\mathcal{D}$, the square loss between our estimation and the best possible estimation at a given $\mathbf{x}$ is

$$(f_{\mathcal{D}}(\mathbf{x}) - h(\mathbf{x}))^2.$$

The average (expectation) of our estimations over many data sets is $\mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})]$. Then,

$$\begin{aligned}(f_{\mathcal{D}}(\mathbf{x}) - h(\mathbf{x}))^2 =&(f_{\mathcal{D}}(\mathbf{x}) - \mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})] + \mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})] - h(\mathbf{x}))^2\\ =&(f_{\mathcal{D}}(\mathbf{x}) - \mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})])^2 + (\mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})] - h(\mathbf{x}))^2\\ &+2\{f_{\mathcal{D}}(\mathbf{x}) - \mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})]\}\{\mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})] - h(\mathbf{x})\}.\end{aligned}$$

We take expectation of both sides of the above expression with respect to $\mathcal{D}$, leading to

$$\mathbb{E}_{\mathcal{D}}[(f_{\mathcal{D}}(\mathbf{x}) - h(\mathbf{x}))^2] = \mathbb{E}_{\mathcal{D}}[(f_{\mathcal{D}}(\mathbf{x}) - \mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})])^2] + (\mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})] - h(\mathbf{x}))^2, \tag{9}$$

as

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})] - h(\mathbf{x}))^2] = (\mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})] - h(\mathbf{x}))^2,\\ &\mathbb{E}_{\mathcal{D}}[\{f_{\mathcal{D}}(\mathbf{x}) - \mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})]\}\{\mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})] - h(\mathbf{x})\}] = 0.\end{aligned}$$

We can see that, the first term—called the **variance**—on the RHS of Eq. (9) measures how spread out our estimation based on one single data set is. The second term—called the squared **bias**—how far on average our estimation on one single data is from that based on all data sets.

The decomposition in Eq. (9) only refers to a single input data instance $\mathbf{x}$. We substitute the integrand of the first term on the RHS of Eq. (8) by the LHS of Eq. (9), leading to the decomposition of the expected squared loss as follows:

$$\text{expected loss} = (\text{bias})^2 + \text{variance} + \text{noise}, \tag{10}$$

where

$$(\text{bias})^2 = \int (\mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})] - h(\mathbf{x}))^2 p(\mathbf{x}) d\mathbf{x}, \tag{11}$$

$$\text{variance} = \int \mathbb{E}_{\mathcal{D}}[(f_{\mathcal{D}}(\mathbf{x}) - \mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})])^2] p(\mathbf{x}) d\mathbf{x}, \tag{12}$$

$$\text{noise} = \iint (h(\mathbf{x}) - y)^2 p(\mathbf{x}, y) dy d\mathbf{x}. \tag{13}$$

References

[1] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.