PRML Review: Bias-Variance Tradeoffs

2021. 02. 04.

1° For regression problems:

t: the target variables.

$D = \{(x_1, t_1), (x_2, t_2), \ldots, (x_n, t_n)\}$

Cost Function: (Recall: The expectation of a joint distribution: $E[g(x,y)] = \iint g(x,y) f(x,y)\,dx\,dy$)

The average/expected Loss: $E[L] = \iint L(t, y(x)) p(x,t)\,dx\,dt$

($y(x)$: the prediction of the model.)

L is the Loss function.

i.e. the squared loss function: $L(t, h(x)) = \{h(x) - t\}^2$

∵ Our Goal is to choose the best $h(\cdot)$ that minimize the Loss

∴ $\frac{\partial E[L]}{\partial y(x)} = 2\int \{h(x) - t\} p(x,t)\,dt = 0$ (1)

$\Rightarrow h^*(x) = \frac{\int t\, p(x,t)}{p(x)} = \int t\, p(t|x)\,dt = E_t[t|x]$ ∈ the best estimation is the conditional average of t on x.

Let $h(x)$ denotes the best estimation from the training data, which is:

$h(x) = \int t\, p(t|x)\,dx = E[t|x]$

$\Rightarrow E[L] = \iint \{y(x) - t\}^2 p(x,t)\,dx\,dt$

↗ $[y(x) - h(x) + h(x) - y]^2 = [y(x) - h(x)]^2 + 2[y(x) - h(x)][h(x) - y] + [h(x) - y]^2$

$= \iint \{y(x) - h(x) + h(x) - t\}^2 p(x,t)\,dx\,dt$

$= \iint [y(x) - h(x)]^2 p(x,t)\,dx\,dt + 2\iint [y(x) - h(x)][h(x) - t] p(x,t)\,dx\,dt + \iint [h(x) - t]^2 p(x,t)\,dx\,dt$

$= \int [y(x) - h(x)]^2 p(x)\,dx + 2\int [y(x) - h(x)] \int [h(x) - t] p(x,t)\,dt\,dx + \int [h(x) - t]^2 p(x)\,dx$

By equation (1): $h(x)$ is chosen to set $\int [h(x) - t] p(x,t)\,dt$ to zero.

$= \int [y(x) - h(x)]^2 p(x)\,dx + \int [h(x) - t]^2 p(x)\,dx$

$$= \int \{y(x) - \mathbb{E}[t|x]\}^2 p(x)\,dx + \int \{\mathbb{E}[t|x] - t\}^2 p(x)\,dx. \quad (2)$$

Remark:

$y(x)$: If we draw n inputs from the target distribution, forming a training data D. Based on the data D, we want to train an algorithm A to learn a hypothesis. The process is $y(x;D) = A(D)$.

$h(x) = \int t\,p(t|x)\,dt$ could be viewed as the intrinsic distribution of the target distribution that you want to model.

The second term in (2):

$\int \{h(x) - t\}^2 p(x)\,dx$ is the intrinsic noise in the data.

If we have unlimited supply of the data, we could have $y(x) = h(x) = \mathbb{E}[t|x]$

$\Rightarrow \mathbb{E}[L] = \int \{h(x) - t\}^2 p(x)\,dx \Rightarrow$ The loss of our model is just the noise of the data.

However, in reality, we can only derive the algorithm A from a finite data set D

$y(x;D) = A(D)$.

$$\Rightarrow \mathbb{E}[L] = \int \{y(x;D) - \mathbb{E}[t|x]\}^2 p(x)\,dx + \int \{\mathbb{E}[t|x] - t\}^2 p(x,t)\,dx\,dt.$$

For the first term:

$$\because \{y(x;D) - \mathbb{E}[t|x]\}^2$$

$$= \{y(x;D) - \mathbb{E}_D[y(x;D)] + \mathbb{E}_D[y(x;D)] - \mathbb{E}[t|x]\}^2$$

$$= \{y(x;D) - \mathbb{E}_D[y(x;D)]\}^2 + \{\mathbb{E}_D[y(x;D)] - \mathbb{E}[t|x]\}^2 + 2\{y(x;D) - \mathbb{E}_D[y(x;D)]\}\{\mathbb{E}_D[y(x;D)] - h(x)\}$$

$$\Rightarrow \mathbb{E}_D[\{y(x;D) - \mathbb{E}[t|x]\}^2] = \underbrace{\{\mathbb{E}_D[y(x;D)] - h(x)\}^2}_{\text{Bias}} + \underbrace{\mathbb{E}_D[\{y(x;D) - \mathbb{E}_D[y(x;D)]\}^2]}_{\text{variance}}$$

Remark:

$\mathbb{E}_D[y(x;D)] = \int_D y(x;D) \Pr(D)\,dD$: $\mathbb{E}_D[y(x;D)]$ can be viewed as the expected classifier; $\Pr(D)$ is the probability of drawing

D from the target distribution P.

Bias: $\mathbb{E}_D[y(x;D)] - h(x)$: the inherent error that one obtained from the model even with infinite training data (Often arised when you select the inappropriate model. i.e. model nonlinear models with linear model).

Variance: $\mathbb{E}_D\left[\{y(x;D) - \mathbb{E}_D[y(x;D)]\}^2\right]$: Captures how much your model changes if training on different data set.

(If we have the best model from the training data, how far we can from the average model?)

Combine these terms into equation 3:

$$\mathbb{E}[L] = \int \{\mathbb{E}_D[y(x;D)] - h(x)\}^2 p(x)\,dx \quad \rightarrow \text{(Bias)}^2$$

$$+ \int \mathbb{E}_D\left[\{y(x;D) - \mathbb{E}_D[y(x;D)]\}^2\right] p(x)\,dx \quad \text{variance}$$

$$+ \int \{\mathbb{E}[t|x] - t\}^2 p(x,t)\,dx \quad \text{noise.}$$

For Classification problems:

$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ $x_i \in X$ $y_i \in Y$, X and Y are two countable spaces.

$y_i \in \{C_1, \ldots, C_k\}$ ← classification problems.

The target f: $f = P\{Y_F = y_F | x\}$

one-hot: $P\{Y_F = y_F | x\} = 1$ for true value of Y_F. and 0 for all others.

The hypothesis h: c generated from a learning algorithm)

$h = P\{Y_H = y_H | x\}$

Similarly: $P\{Y_H = y_H | x\} = 1$ for true value of Y_H, and 0 for all others.

Cost Function,

$$E[L] = \sum_{y_H, y_F} l(y_H, y_F) P\{y_H, y_F\}$$

We use zero-one loss.

$$\Rightarrow l(y_F, y_H) = 1 - \delta(y_F, y_H)$$

$$\delta(y_F, y_H) = \begin{cases} 1 & \text{if } y_F = y_H \\ 0 & \text{otherwise.} \end{cases}$$

$$\Rightarrow E[L] = \sum_{y_H, y_F} [1 - \delta(y_H, y_F)] P\{y_H, y_F\}$$

$$= 1 - \sum_{y_F, y_H} \delta(y_H, y_F) P\{y_H, y_F\}$$

The misclassification rate

$$\Rightarrow L = 1 - \sum_{y \in Y} P\{Y_H = Y_F = y | x\}$$

$$= \sum_{y \in Y} -P\{Y_H = y_F = y | x\} + \sum_{y \in Y} P\{Y_H = y | x\} P\{Y_F = y | x\} + \sum_{y \in Y} \left[-P\{Y_H = y | x\} P\{Y_F = y | x\} + \frac{1}{2} P\{Y_F = y | x\}^2 + \frac{1}{2} P\{Y_H = y | x\}^2 \right]$$

$$+\left[\frac{1}{2}-\frac{1}{2}\sum_{y\in Y}P\{Y_H=y|x\}^2\right]+\left[\frac{1}{2}-\frac{1}{2}\sum_{y\in Y}P\{Y_F=y|x\}^2\right]$$

$$=\sum_{y\in Y}\left[P\{Y_H=y|x\}P\{Y_F=y|x\}-P\{Y_F=Y_H=y|x\}\right] \quad \text{covariance}$$

$$+\frac{1}{2}\sum_{y\in Y}\left[P\{Y_F=y|x\}-P\{Y_H=y|x\}\right]^2 \quad \text{bias}^2$$

$$+\frac{1}{2}\left[1-\sum_{y\in Y}P\{Y_H=y|x\}^2\right] \quad \text{variance}$$

$$+\frac{1}{2}\left[1-\sum_{y\in Y}P\{Y_F=y|x\}^2\right] \quad \text{noise.}$$

Proposition: Y_F and Y_H are conditionally independent given f and test points x

Sketch of proof: $\because Y_F$ depends only on the target f and the test point of x

$$\therefore P\{Y_F, Y_H|f,x\}=P\{Y_F|y_H,f,x\}P\{Y_H|f,x\}$$

$$=P\{Y_F|f,x\}P\{Y_H|f,x\}$$

$\Rightarrow$ The covariance term:

$$\sum_{y\in Y}\left[P\{Y_H=y|x\}P\{Y_F=y|x\}-P\{Y_F=Y_F=y|x\}\right]=0$$

$$\Rightarrow \mathbb{E}[L]=\sum_{x\in X}P(x)\,L(x)$$

$$=\sum_{x\in X}P(x)\left\{\frac{1}{2}\sum_{y\in Y}\left[P\{Y_F=y_F|x\}-P\{Y_H=y|x\}\right]^2\right. \quad \Rightarrow \text{bias}_x^2$$

$$+\frac{1}{2}\left[1-\sum_{y\in Y}P\{Y_H=y|x\}^2\right] \quad \Rightarrow \text{variance}_x$$

$$\left.+\frac{1}{2}\left[1-\sum_{y\in Y}P\{Y_F=y|x\}^2\right]\right\} \quad \Rightarrow \text{noise}_x$$

Remark:

$P\{Y_F = y \mid x\}$: the intrinsic distribution of the target data

$P\{Y_H = y \mid x\}$: The full expression is:

$$P\{Y_H = y \mid f, m, x\}$$

$$= \sum_d P\{d \mid f, m, x\}\, P\{Y_H = y \mid d, f, m, x\}$$

$$= \sum_d \underbrace{P\{d \mid f, m\}}\ \underbrace{P\{Y_H = y \mid d, x\}}$$

generate training set d from the target distribution f

$\Rightarrow$ the learning algorithm trained on data d to make a guess to the Y value.