① for Non-parametric estimator / parametric estimator

$\hat{\theta} \longrightarrow$ estimator of $\theta$ $\quad \hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$

$\longrightarrow$ $\boxed{\text{MSE decomposition}}$ $\longrightarrow$ $\hat{\theta}$ depends on $X_1, \dots, X_n$

$$\mathbb{E}[(\hat{\theta}-\theta)^2] = \mathbb{E}[(\hat{\theta} - \mathbb{E}(\hat{\theta}) + \mathbb{E}[\hat{\theta}] - \theta)^2]$$

$$= \mathbb{E}[(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2] + (\mathbb{E}[\hat{\theta}] - \theta)^2 + 0$$

$$= \underbrace{\text{var}[\hat{\theta}]}_{\text{variance}} + \underbrace{(\mathbb{E}[\hat{\theta}] - \theta)^2}_{\text{bias}^2}$$

$\longrightarrow$ Similar result holds for Density Estimation

$$\mathbb{E}[(\hat{p}_n(x) - p(x))^2] = \underbrace{(\mathbb{E}[\hat{p}_n(x)] - p(x))^2}_{\text{Bias}^2} + \underbrace{\text{Var}[\hat{p}_n(x)]}_{\text{variance}}$$

② for some $\boxed{\text{Machine Learning Task}}$ (Non-parametric Regression included)

Generally speaking, things are more Complicated!

$\longrightarrow$ $\boxed{\text{Setting}}$

have $(X, Y)$, find $f \in \mathcal{F}$ s.t $f(X) \approx Y$

$\longrightarrow$ $\boxed{\text{Starting Frome Decision Theory}}$

a) Assume we have $(X, Y) \sim P$ $\quad \boxed{R(f)}$

Then consider $f^* = \underset{f \in \mathcal{F}}{\text{argmin}}\ \mathbb{E}_{(X,Y)}[L(f(X), Y)]$

Assume of Bias & Variance Tradeoff

Specially, if $L(t_1, t_2) = (t_1 - t_2)^2$

then $\mathbb{E}_{(X,Y)}[L(f(X), Y)]$

Correspond to the CMU Wass... Lecture 26

$\boxed{\text{Optimal Regression Function}}$

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

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

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

$\mathbb{E}[Var[Y|X]]$

$$+ 2\boxed{\mathbb{E}_{(X,Y)}[(f(X) - \mathbb{E}[Y|X])(\mathbb{E}[Y|X] - Y)]} = 0$$

consider $\mathbb{E}_Y[\underbrace{(f(X) - \mathbb{E}[Y|X])}_{\text{const}}(\mathbb{E}[Y|X] - Y)|X]$

$$= \text{Const} \cdot \underbrace{\mathbb{E}_Y[(\mathbb{E}[Y|X] - Y)|X]}_{0}$$

$\boxed{\text{Conclusion}}$

$\Rightarrow$ If we choose $L(t_1, t_2) = (t_1 - t_2)^2$

then the optimal model $f^* = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \boxed{\mathbb{E}_{(X,Y)}[L(f(X), Y)]}$ $R(f)$

$\Leftrightarrow \boxed{f^*(x) = \mathbb{E}[Y|X=x]}$

If we use different $L(t_1, t_2)$, then the corresponding $f^*(\cdot)$ is different

$\Downarrow$

they are all optimal in different aspects

Also, $R(f) = \underbrace{\mathbb{E}_{X \sim P_X}[(f(X) - f^*(X))^2]}_{\text{MSE}} + \underbrace{\mathbb{E}_{(X,Y)}[(Y - f^*(X))^2]}_{\text{noise (variance)}}$

$\mathbb{E}_X[Var[Y|X]]$

b) After the first step, we consider the randomness of $(X, Y) \Longrightarrow R(f)$ to achieve the 'optimal model' $\boxed{f^*(x)}$

$\Downarrow$

can be viewed as 'oracle'

$$\boxed{\text{We want to use } \underbrace{\hat{f}_n(x)}_{\text{RV on } \mathcal{D}_n} \text{ to estimate } f^*(x)}$$

ERM illustration

[e.g.] $\hat{f}_n = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \hat{R}_n(f)$

where $\hat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} L(f(X_i), Y_i)$

$\boxed{\hat{R}_n(f) \xrightarrow{P} R(f)}$ $\quad (X_i, Y_i) \sim P$

Then, $R(f) = \mathbb{E}_{(X,Y)}[L(f(X), Y)] \to$ generalization error

useful for Bias & Variance Tradeoff

$$\mathrm{EPE}(\hat{f}_n) = \mathbb{E}_{(X,Y,\mathcal{D}_n)}[L(\hat{f}_n(X), Y)] \to \text{expected prediction error}$$

① Square Loss
② use previous result

$$\mathbb{E}_{(X,\mathcal{D}_n)}[(\hat{f}_n(X) - f^*(X))^2] + \text{noise}$$

$$= \mathbb{E}_X \mathbb{E}_{\mathcal{D}_n}[(\hat{f}_n(X) - f^*(X))^2 \mid X] + \text{noise}$$

$$= \mathbb{E}_X\left[\left(\mathbb{E}_{\mathcal{D}_n}[\hat{f}_n(X)] - f^*(X)\right)^2 + \mathrm{Var}_{\mathcal{D}_n}[\hat{f}_n(X)]\right] + \text{noise}$$

$$= \mathbb{E}_X[\underbrace{\mathrm{bias}^2(X)}_{\text{w.r.t estimator } \hat{f}_n}] + \mathbb{E}_X[\underbrace{\mathrm{var}(X)}_{\text{w.r.t estimator } \hat{f}_n}] + \underline{\text{noise}}$$

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

$\mathbb{E}_X[\mathrm{Var}[Y|X]]$

noise w.r.t $Y$ & $f^*(X)$

Practically speaking, $\begin{cases} R(f) \to \text{infeasible} \\ EPE(\hat{f}_n) \to \text{infeasible} \end{cases}$ → but can use to explain the BIAS & VARIANCE TRADE-OFF

→ But can do some kind of simulation:

In our formulation, we require:

$$\begin{cases} Y = f(X) + \varepsilon \\ \boxed{\begin{cases} \mathbb{E}[Y \mid X=x] = f(x) \leftarrow \mathbb{E}[\varepsilon]=0 \\ Var[Y \mid X=x] = \sigma^2 \text{ for simplicity} \leftarrow Var[\varepsilon]=\sigma^2 \end{cases}} \end{cases}$$

use this to generate data & estimate

$$\begin{cases} ① \ bias(\hat{f}_n(x)) \\ ② \ var(\hat{f}_n(x)) \end{cases}$$

---

What is feasible?

- 1. $\hat{R}_n(f) \longrightarrow$ unbiased estimator for $EPE(\hat{f}_n) \Rightarrow$ waste data

ML Model

- 2. $CV(f) \longrightarrow$ unbiased estimator for $EPE(\hat{f}_n) \Rightarrow$ computationally expensive

Statistical Model

$$\boxed{EPE(\hat{f}_n) = \mathbb{E}_{(X,Y,\mathcal{D})}[L(f(X),Y)]}$$

What we are interested? → $\hat{R}_n(f)$

- ERM framework
- Our interest: $|R(\hat{f}) - R(f^*)|$

idea is to make sure that $\hat{f}$ is good enough

$$= |\underbrace{R(\hat{f}) - \hat{R}_n(\hat{f})}_{①} + \underbrace{\hat{R}_n(\hat{f}) - \hat{R}_n(f^*)}_{②} + \underbrace{\hat{R}_n(f^*) - R(f^*)}_{③}|$$

$$= |①| + |②| + |③|$$

→ ② can be controlled since $\hat{f} = \underset{f \in \mathcal{F}}{\arg\min}\, \hat{R}_n(\hat{f})$

(L function should be good enough)

→ ③ can be controlled due to concentration ineq.

$$\mathbb{P}(|\hat{R}_n(f^*) - R(f^*)| \geq \varepsilon) \leq e^{-n \cdot C(\varepsilon)}$$

since $\hat{R}_n(f^*) = \frac{1}{n} \sum_{i=1}^{n} L(f^*(X_i), Y_i)$

→ ① Requires uniform bound !

$$\mathbb{P}\left( \boxed{\sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)|} \geq \varepsilon \right) \leq C(\varepsilon) \qquad \Leftrightarrow \qquad \boxed{\|\hat{R}_n - R\|_{f \in \mathcal{F}}}$$

Conceptual idea

VC dimension

Rademacher Theory

More tight bound