Math 342W Lecture 19

Bias-Variance Tradeoff in Regression Modeling ($y \subseteq \mathbb{R}$)

Assumptions

I) δ is a realization from Δ_1, a R.V $\overbrace{\text{mean independent of } \vec{x}}^{E[\Delta \mid \vec{x}] = \emptyset}$

$\hookrightarrow E[Y \mid \vec{x}] = E[f(\vec{x}) + \Delta] = E[f(\vec{x})] + E[\Delta \mid \vec{x}]$

$= E[f(\vec{x})] = f(\vec{x}) \rightarrow$ Conditional Expectation Formula (CEF)

II) Homoskedasticity $\approx$ Constant Variance

$\hookrightarrow \text{Var}[\Delta \mid \vec{x}] = E[\Delta^2 \mid \vec{x}] - E[\Delta \mid \vec{x}]^2 = E[\Delta^2 \mid \vec{x}] = \underbrace{\sigma^2}_{\text{Constant } \forall \vec{x}}$

Lets say we fit a model to $\mathbb{D}$ and obtain g

$y = g + e = g + (f - g) + \delta \Rightarrow e = (f - g) + \delta$

$Y = g + (f - g) + \Delta \Rightarrow E = (f - g) + \Delta$

Realization to Random Variable

Then we can define $\text{Bias}[\vec{x}_*] = E[Y_* - g(\vec{x}_*)] = E[E \mid \vec{x}_*]$

$= E[f - g + \Delta_* \mid \vec{x}_*] = f - g + E[\Delta_*] = f(\vec{x}_*) - g(\vec{x}_*)$

We can now define $MSE(\vec{x}_*) = E[(Y_* - g(\vec{x}_*))^2]$

$= E[Y_*^2 \mid \vec{x}_*] - 2g(\vec{x}_*) E[Y_* \mid \vec{x}_*] + E[g(\vec{x}_*)^2 \mid \vec{x}_*]$

$= f(\vec{x}_*)^2 + 2f(\vec{x}_*) E[\Delta_*] + E[\Delta_*^2] - 2g(\vec{x}_*)[f(\vec{x}_*) + E[\Delta]] + g(\vec{x}_*)^2$

$= f(\vec{x}_*)^2 + (-2g(\vec{x}_*) f(\vec{x}_*)) + g(\vec{x}_*)^2 + \sigma^2 = [f(\vec{x}_*) - g(\vec{x}_*)]^2 + \sigma^2$

Now we assume Randomness in $\Delta_1, \Delta_2, \ldots, \Delta_n, \Delta_*$

$$MSE(\vec{x}_*) = E_{\Delta_1, \Delta_2, \ldots, \Delta_n, \Delta_*}\left[\left(Y - G(\vec{x}_*)\right)^2 \middle| \vec{x}_*\right]$$

$$= E_{\Delta_1, \ldots, \Delta_*}[Y^2] - 2E_{\Delta_1, \ldots, \Delta_*}[Y_* G(\vec{x}_*)] + E_{\Delta_1, \ldots, \Delta_*}[G(\vec{x}_*)^2]$$

$$= [f(\vec{x}_*) + \sigma^2] - 2E[Y_* G(\vec{x}_*)] + Var[G(\vec{x}_*)] + E[G(\vec{x}_*)]^2$$

$$= [f(\vec{x}_*) - E[G(\vec{x}_*)]]^2 + Var[G(\vec{x}_*)] + \sigma^2$$

$$= \sigma^2 + Bias[G(\vec{x}_*)]^2 + Var[G(\vec{x}_*)]$$

Now, finally we assume Randomness in all previous + $X, \vec{x}_*$

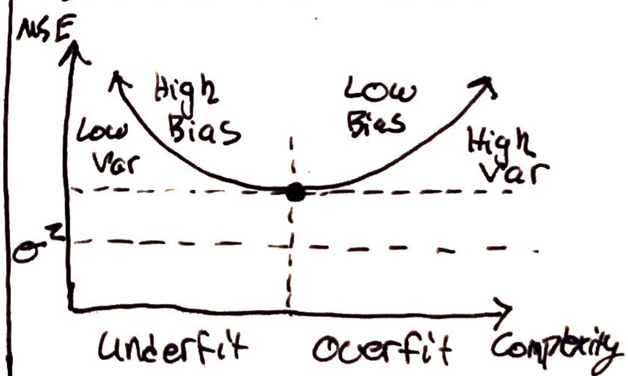
$\Rightarrow$ This is a random variable model producing $\vec{x}_1, \ldots, \vec{x}_n, \vec{x}_*$

$$E_{\vec{x}_1, \ldots, \vec{x}_n, \vec{x}_*}[MSE(\vec{x}_*)]$$

$$= E_{x\text{'s}}[Bias[G(\vec{x}_*)]]^2 + E_{x\text{'s}}[Var[G(\vec{x}_*)]] + \sigma^2$$

Bias Variance Decomposition

Bias Variance Tradeoff



• – $\sigma^2 + \min\{bias^2 + var\}$

Overfit : Low Bias & High Variance

Underfit : High Bias & Low Variance

Generalized Linear Models (GLM)

Independently Realized

$$P(\mathbb{D}) = P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n \mid \vec{X}_1 = \vec{x}_1, \ldots, \vec{X}_n = \vec{x}_n)$$

$$= \prod_{i=1}^{n} P(Y_i = y_i \mid \vec{X}_i = \vec{x}_i) = \prod_{i=1}^{n} \underbrace{f_{pr}(\vec{x}_i)^{y_i} (1 - f_{pr}(\vec{x}_i))^{1-y_i}}$$

$\{ V \sim \text{Bern}(\theta) = \theta^{v} (1-\theta)^{1-v} \}$

$\mathcal{A}$: Maximize $P(\mathbb{D})$

Assume $\mathcal{H}_{pr} = \{ \phi(\vec{w} \cdot \vec{x}) : \vec{w} \in \mathbb{R}^{p} \}$

Assume $\phi : \mathbb{R} \to (0,1)$, called a <u>Link Function</u>

If your algorithm uses the space $\mathcal{H} = \{ \phi(\vec{w} \cdot \vec{x}) : \vec{w} \in \mathbb{R}^{p} \}$, then your model is called a generalized linear model

<u>Common Link Functions</u>

① Logistic Link Function $\phi(u) = \dfrac{e^{u}}{1+e^{u}} = \dfrac{1}{1+e^{-u}}$

② Probit Link Function $\phi(u) = \Phi(u)$, where Φ is CDF of $N(0,1)$

③ Complementary Log-Log Link Function $\phi(u) = 1 - e^{-e^{-u}}$

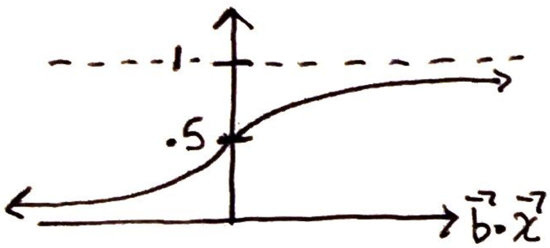
Assume Logistic Link

$$= \underbrace{\prod_{i=1}^{n} \left(\frac{1}{1+e^{-(\vec{w} \cdot \vec{x}_i)}} \right)^{y_i} \left(\frac{1}{1+e^{(\vec{w} \cdot \vec{x}_i)}} \right)^{1-y_i}} \Rightarrow \mathcal{A} : \vec{b} = \operatorname{argmax} \{ \; \}$$

NOTE: No closed form sln' to $\vec{b}$. Numerical Methods to approximate

$$\vec{\nabla} P(\mathbb{D}) := \vec{0}_{p+1}$$

$$g_{pr}(\vec{x}) = \frac{1}{1+e^{-(\vec{b}\cdot\vec{x})}} \Rightarrow \text{estimates } P(Y_* = 1 \mid \vec{x}_*) = \hat{p}$$



$$\Rightarrow \frac{1}{\hat{p}} = 1 + e^{-(\vec{b}\cdot\vec{x})}$$

$$\Rightarrow \frac{1}{\hat{p}} - 1 = e^{-(\vec{b}\cdot\vec{x})}$$

$$\Rightarrow \underbrace{\frac{1-\hat{p}}{\hat{p}}}_{\text{odds against}} = e^{-(\vec{b}\cdot\vec{x})}$$

$$\Rightarrow \underbrace{\frac{\hat{p}}{1-\hat{p}}}_{\text{odds}} = e^{\vec{b}\cdot\vec{x}} \qquad \Rightarrow \vec{b}\cdot\vec{x} = \underbrace{\ln\left(\frac{\hat{p}}{1-\hat{p}}\right)}_{\text{log odds}}$$

Now the question is, how do we validate this sort of model?

We have to validate $g_{pr}(\vec{x}_i)$ with y_i, but

$$g_{pr}(\vec{x}_i) = \hat{p} \in (0,1) \quad \text{and} \quad y_i \in \{0,1\}$$

$\neq$

<u>Proper Scoring Rules</u>

$S(\hat{p}_i, y_i)$ satisfies $\forall_i\ f_{pr}(\vec{x}_i) = \arg\max \{S(\hat{p}_i, y_i)\}$

① Brier Score: $S_i = -(y_i - \hat{p}_i)^2 \leq 0$

② Log Score: $S_i = y_i \ln(\hat{p}_i) + (1-y_i)\ln(1-\hat{p}_i) \leq 0$

These scoring rules deal with the different spaces for $\hat{p}_i$ and y_i, so we can judge the model performance in a way that makes sense.