2021 – 11 – 19

$$Y = f(X) + \varepsilon \quad , \quad E(\varepsilon) = 0 \quad , \quad Var(\varepsilon) = \sigma_\varepsilon^2$$

Συνάρτηση πρόβλεψης $\hat{f}(x)$

Σε νέο σημείο x_0: (δεν έχει χρησιμοποιηθεί για τον υπολογισμό του $\hat{f}$)

Πρόβλεψη: $\hat{f}(x_0)$, πραγμ. τιμή $Y|X=x_0$

$$Err(x_0) = \underset{Y|X=x_0}{E}\left[\left(Y - \hat{f}(x_0)\right)^2 \Big| X = x_0\right]$$

$$= E\left[\left(\underbrace{Y - f(x_0)} + \underbrace{f(x_0) - E\hat{f}(x_0)} + \underbrace{E\hat{f}(x_0) - \hat{f}(x_0)}\right)^2\right]$$

$$= E\left[\left(Y - f(x_0)\right)^2\right] + E\left[\left(\underbrace{f(x_0) - E\hat{f}(x_0)}_{\text{σταθερά}}\right)^2\right] +$$

$$+ E\left[\left(\hat{f}(x_0) - E\hat{f}(x_0)\right)^2\right] +$$

$$+ 2E\left[\left(Y - f(x_0)\right)\underbrace{\left(\overset{\text{σταθ.}}{f(x_0)} - \overset{\text{σταθ.}}{E\hat{f}(x_0)}\right)}\right] + \; = 0 = \left(f(x_0) - E\hat{f}(x_0)\right)\underbrace{E\left(Y - f(x_0)\right)}$$

$$+ 2E\left[\underbrace{\left(Y - f(x_0)\right)}\underbrace{\left(E\hat{f}(x_0) - \hat{f}(x_0)\right)}\right] + \; = 0$$

ανεξ

$$+ 2E\left[\underbrace{\left(f(x_0) - \hat{f}(x_0)\right)}_{\text{σταθ.}}\underbrace{\left(E\hat{f}(x_0) - \hat{f}(x_0)\right)}_{E\, = 0}\right] = 0$$

$$\mathrm{Err}(x_0) = \underbrace{E\left[(Y-f(x_0))^2\right]} + \left[E\hat{f}(x_0) - f(x_0)\right]^2 + E\left[\left(\hat{f}(x_0) - E\hat{f}(x_0)\right)^2\right]$$

$$= \sigma_\varepsilon^2 + \mathrm{Bias}^2 + \mathrm{Variance}\left(\hat{f}(x_0)\right)$$

Bias-Variance decomposition

Εφαρμογή $\underbrace{f_p(x) = x^T b}$ γραμμικό μοντέλο $b, x \in \mathbb{R}^p$

$$\boxed{\mathrm{Err}(x_0)} = E\left[(Y-\hat{f}(x_0))^2 \mid X = x_0\right] =$$

$$= \sigma_\varepsilon^2 + \underbrace{\left(E\hat{f}_p(x_0) - f(x_0)\right)^2}_{\text{bias}^2} + \mathrm{Var}\left(\hat{f}_p(x_0)\right)$$

Αν $f(x) = x^T b$

τότε για LSE $\text{bias} = 0$

διαφορετικά $\text{bias} \neq 0$

$$\hat{f}(x_0) = x_0^T \hat{b}^{LSE} \Rightarrow \mathrm{Var}(\hat{f}(x_0)) = x_0^T \mathrm{Var}(\hat{b})\, x_0$$

$$\hat{b} = (X^T X)^{-1} X^T y, \quad \mathrm{Var}(\hat{b}) = (X^T X)^{-1} \cdot \sigma_\varepsilon^2$$

$$\mathrm{Var}(\hat{f}(x_0)) = x_0^T \cdot (X^T X)^{-1} x_0 \, \sigma_\varepsilon^2$$

$$\hat{f}(x_0) = x_0^T \cdot \hat{b} = \underbrace{x_0^T \cdot (X^T X)^{-1} X^T}_{h(x_0)^T} \cdot y = h(x_0)^T \cdot y$$

$$\mathrm{Var}\,\hat{f}(x_0) = (h(x_0))^T \underbrace{\mathrm{Var}(y)}_{\sigma_\varepsilon^2 I_{N\times N}} \cdot h(x_0) = \|h(x_0)\|^2 \cdot \sigma_\varepsilon^2$$

In-Sample-Error

$$\frac{1}{N}\cdot\sum_{i=1}^{N} Err(x_i) \qquad (\neq \overline{err})$$

δεν χρησιμοποιούμε τα y_i του T.

$$= \sigma_\varepsilon^2 + \frac{1}{N}\sum_{i=1}^{N}\left[E\hat{f}(x_i) - f(x_i)\right]^2 + \frac{\sigma_\varepsilon^2}{N}\sum_{i=1}^{N}\|h(x_i)\|^2$$

$$h(x_i) = x_i^T (X^TX)^{-1} X^T = i \text{ γραμμή του } H \qquad X = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{pmatrix}$$

$$H = X(X^TX)^{-1}X^T \quad (\text{hat matrix}). \qquad H_{N\times N}$$

$$\|h(x_i)\|^2 = \sum_{j=1}^{N} H_{ij}^2 \qquad H_{i\cdot} = \left[H_{i1} \cdots \cdots H_{iN}\right] = h(x_i)^T$$

Όμως H : ταυτοδύναμος $\Rightarrow H\cdot H = H \Rightarrow$

$$\Rightarrow H_{i\cdot}\, H_{\cdot i} = H_{ii} \Rightarrow \|h(x_i)\|^2 = H_{ii}$$

$$\Rightarrow \sum_{i=1}^{N}\|h(x_i)\|^2 = \text{trace}(H) = P$$

$$\Rightarrow \frac{1}{N}\sum_{i=1}^{N} Err(i) = \sigma_\varepsilon^2 + \text{Average bias} + \frac{\sigma_\varepsilon^2 P}{N}$$

Variance $\nearrow P$ $\searrow N$

$$E_{x_0}\underbrace{\left(f(x_0)-E\hat{f}(x_0)\right)^2}_{\text{bias}^2} = \ldots$$

$$= \underbrace{E_{x_0}\left(f(x_0)-x_0^T b_*\right)^2}_{\text{model bias}} + \underbrace{E\left(x_0^T b_* - E x_0^T \hat{b}_a\right)^2}_{\text{estimation bias}}$$

Υποθέτουμε $E(Y|X=x) = f(x)$ άγνωστο.

Περιοριζόμαστε σε γρ. μοντέλα $\hat{f}_p = \underline{x^T \cdot b}$

b_* : το "καλύτερο" b στην $\underline{f(x)}$ [αριθ. αναφορών]

$\hat{b}_a$: εκτίμηση του b_* από ένα μοντέλο γενικά με regularization (αν. $a=0 \Rightarrow$ LSE, $a>0 \Rightarrow$ ridge)

Π.χ. ① αν $f(x) = x^T b_{true} \Rightarrow$ model bias $=0$.

$(b^* = b_{true})$

② αν LSE (χωρίς regularization)

$E(\hat{b}) = b_* \Rightarrow$ estimation bias $=0$.

Για ridge reg./lasso est bias >0.