6-10-2021 | Παλινδρόμηση

$X = (X_1, \dots, X_p)$ διάνυσμα τυχαίων μεταβλητών
inputs - ανεξ. μεταβλητές

Y = τυχαία μεταβλητή (outcome - εξαρτημένη)

συνάρτηση πρόβλεψης του Y : $f(X)$

$$f(X) = E(Y|X)$$

Στατιστική: Υποθέτουμε $E(Y|X) = b_0 + b_1 X + \dots + \dots b_p X_p$

$\boxed{\text{εκτίμηση των } b, \sigma^2}$

b άγνωστες παράμετροι

$Y|X \sim N(b_0 + b_1 X_1 + \dots b_p X_p, \sigma^2)$

Γενική Προσέγγιση (Αρχικά καμμία στατιστική υπόθεση)

Θεωρούμε την κλάση συναρτήσεων πρόβλεψης

$$f(X) = b_0 + b_1 X_1 + \dots + b_p X_p \leftarrow$$

Θέτω $X_0 = 1$, $b = (b_0, b_1, \dots, b_p) \in \mathbb{R}^{p+1}$

$\boxed{f(X) = \underline{X}^T b}$ $\quad [\underline{X} = (X_0, X_1, \dots, X_p)]$

Training set $(x_1, y_1), \dots, (x_N, y_N)$

$\underline{x}_1, x_2$: γνωστά διανύσματα $\in \mathbb{R}^p$

(1, ...

$x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$, $i = 1, \dots, N$

$\underline{y} = (y_1, \dots, y_N)^T$

$$X = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{pmatrix} = \begin{array}{cccc|l} \text{μετ.}0 & \text{μετ.}1 & & \text{μετ.}p & \\ \hline \downarrow & \downarrow & & \downarrow & \\ x_{10} & x_{11} & \cdots & x_{1p} & \leftarrow \text{παρ.}1 \\ x_{20} & x_{21} & \cdots & x_{2p} & \leftarrow \text{παρ.}2 \\ \vdots & & & & \\ \underbrace{x_{N0}}_{''\,1} & x_{N1} & \cdots & x_{Np} & \cdots \end{array}$$

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$$

X_j = ποσοτικές μεταβλητές (π.χ. ύψος, βάρος, υγεία κτλ)

μετασχηματισμοί π.χ. $X_2 = \log(X_1)$

$X = \log(\text{υγείας})$

$X_1, X_1^2, X_1^3, \dots$ ων. βάση

Δυαδικές μεταβλητές που μετατρέπουν κατηγορικές

Όροι αλληλεπίδρασης π.χ. $X_3 = X_1 X_2$

$$X = \text{χρώμα} \in \{R, G, B\}$$

$$\left.\begin{aligned} x_R &= 1(X=R) \\ x_G &= 1(X=G) \\ X \longrightarrow x_B &= 1(X=B) \end{aligned}\right\} \Rightarrow \tilde{X} = (x_R, x_G, x_B)$$

$$X = R \Rightarrow \tilde{X} = (1, 0, 0)$$

$$B \Rightarrow \tilde{X} = (0, 0, 1)$$

$$G \Rightarrow \tilde{X} = (0, 1, 0)$$

← επίπεδο αναφοράς

Interaction / Αλληλεπίδραση

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \underbrace{b_3 X_1 X_2}_{\text{interaction term}}$$

$$\frac{\partial Y}{\partial X_1} = b_1 + b_3 X_2$$ = η επίπτωση του X_1 σ

στην Y εξαρτάται από τη X_2

$$\frac{\partial}{\partial X_2}\left(\frac{\partial Y}{\partial X_1}\right) = b_3$$ ← αλληλεπίδραση

<u>Ερώτημα</u> Να βρεθεί το $\hat{b}$ τέτοιο

ώστε η $f(x) = x^T \hat{b}$ να κάνει

την <u>καλύτερη προσαρμογή</u> για δεδομένα

του training set!

<u>Κριτήριο</u> Άθροισμα τετραγωνικών καταλοίπων

Residual Sum of Squares (RSS)

$$RSS(b) = \sum_{i=1}^{N} \left[y_i - \underline{f(x_i)} \right]^2 \qquad f(x_i) = x_i^T b$$

$$\underline{e} = \begin{pmatrix} y_1 - x_1^T b \\ y_2 - x_2^T b \\ \vdots \\ y_N - x_N^T b \end{pmatrix} \qquad RSS(b) = \underline{e}^T \underline{e}$$

$$X_{N\times(p+1)} \quad b_{(p+1)\times 1}$$

$$\underline{e} = \underline{y} - X \underline{b}$$

$$RSS(b) = (y - Xb)^T \cdot (y - Xb)$$

$$\min_{b} RSS(b) = \min_{b} (y - Xb)^T (y - Xb)$$

$$= \min_{b} \left\{ y^T y - \underbrace{\overbrace{y^T X}^{a^T} b}_{a^T b} - \underbrace{b^T X^T y}_{b^T a} + b^T X^T X b \right\}$$

$$= \min_{b} \left(y^T y - 2\, y^T X \cdot b + \underline{b^T X^T X b} \right)$$

$RSS(b)$: κυρτή $(\underline{b})$

$D^2 RSS(b) = 2 X^T X$

$X^T X$ θετικά ορισμένος

$\Longleftrightarrow$ X πλήρους βαθμού

Επομένως $RSS(b) \leftarrow \min$ για $\nabla_b RSS(b) = 0$.

$$RSS(b) = y^T y - 2 \overbrace{y^T X}^{a^T} \cdot b + b^T \underbrace{X^T X}_{A} b,$$

$$\nabla_b RSS(b) = \begin{pmatrix} \frac{\partial RSS}{\partial b_0} \\ \vdots \\ \frac{\partial RSS}{\partial b_p} \end{pmatrix} = \underline{-2X^T y + 2(X^T X) \cdot b}$$

Παράδειγμα $g(\underline{b}) = \alpha^T \cdot b = \alpha_0 b_0 + \alpha_1 b_1 + \cdots + \alpha_p b_p$

$$\frac{\partial g}{\partial b_i} = \alpha_i \quad \Rightarrow \quad \nabla g = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_p \end{pmatrix} = \alpha.$$

$$h = b^T A b = \sum_{ij} \alpha_{ij} b_i b_j$$, A : συμμετρικός

Δ.Ο. $\nabla_b b^T A b = 2Ab$

$$\nabla_\beta RSS(\beta) = 0 \Rightarrow -2X^T(y - X\beta) = 0 \Leftarrow$$

$$\Rightarrow X^T y = X^T X \cdot \beta \Rightarrow \boxed{\hat{\beta} = (X^T X)^{-1} \cdot X^T \cdot y}$$

$$\hat{y} = \begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_N \end{pmatrix} = X \cdot \hat{\beta} = \underbrace{X (X^T X)^{-1} \cdot X^T}_{H} \cdot y = H \cdot y$$

$$\Rightarrow \boxed{\hat{y} = H \cdot y}$$

H : hat matrix $\quad \begin{pmatrix} H^T = H \\ H^2 = H \end{pmatrix}$

Νέα παρατήρηση $\quad x_0^T = (x_{01}, \dots, x_{0p})$

$$\hat{y} = x_0^T \cdot \hat{\beta}$$

Στατιστικές Ιδιότητες του $\hat{b}$

(1) Υποθέτουμε $y_1, \dots, y_N$: ανεξάρτητα (παρατηρήσεις από πληθυσμό)

~~ομοσκεδαστικότητα~~ $Var(y_1) = \dots = Var(y_N) = \sigma^2$

X : όχι τυχαίος

$\hat{b}$ = τυχαία μεταβλητή $\in \mathbb{R}^{p+1}$

$$\hat{b} = (X^T X)^{-1} \cdot X^T \underline{y}$$

Παρατ. 1 : $x_1^T = (x_{10}, \dots, x_{1p})$ $\quad Y_1 | X_1 = x_1 \longrightarrow y_1$

$x_2^T = (x_{20}, \dots, x_{2p})$ $\quad Y_2 | X_2 = x_2 \longrightarrow y_2$

$$E(\hat{b}) = (X^T X)^{-1} \cdot X^T \cdot \underbrace{E(\underline{y})}$$

$$Var(\hat{b}) = Var(G \cdot \underline{y}) = G\, Var(y)\, G^T, \quad G = \underbrace{(X^T X)^{-1} \cdot X^T}$$

$$\underline{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} \Rightarrow Var(\underline{y}) = \sigma^2 \underline{I} = \begin{pmatrix} \sigma^2 & & 0 \\ & \ddots & \\ 0 & & \sigma^2 \end{pmatrix}$$

$\underline{X} : \quad \Sigma_X$

$Y = A \cdot X \Rightarrow \Sigma_Y = A \Sigma_X A^T$ (πιθανότητα).

$$\Rightarrow Var(\hat{b}) = G \cdot \sigma^2 I \; G^T = \sigma^2 \cdot G G^T$$

$$\left.\begin{aligned} G &= (X^T X)^{-1} \cdot X^T \\ G^T &= X (X^T X)^{-1} \end{aligned}\right\} \Rightarrow G G^T = \underbrace{(X^T X)^{-1} \cdot (X^T \cdot X)}_{I} (X^T X)^{-1} = (X^T X)^{-1}$$

$$\Rightarrow \boxed{Var(\hat{b}) = (X^T X)^{-1} \cdot \sigma^2} \leftarrow$$

Ιδιότητες

① Αν εκτιμήσουμε το σ^2 : $\hat{\sigma}^2 = \frac{1}{N-p-1} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$

$\hat{\sigma}_2 = \frac{RSS(\hat{b})}{N-p-1}$

$E(\hat{\sigma}^2) = \sigma^2$ $\hat{\sigma}^2$: αμερόληπτη

[στην ορολογία λεγόταν MSE]

(δείξτε το !!)

Στατιστική Συμπερασματολογία για τα $\hat{b}$

Επιπλέον υποθέσεις $E(Y|X=x) = \boxed{x^T \cdot b}$ ① $= b_0 + b_1 x_1 + \dots + b_p x_p$

όπου $b_0, b_1, \dots, b_p$ άγνωστες παράμετροι

$Y|X=x \;=\; x^T b + \varepsilon$, $\varepsilon \sim \boxed{\mathcal{N}}(0, \sigma^2)$ ②

$\varepsilon_1, \varepsilon_2, \dots, \varepsilon_N$ ανεξάρτητες

$y_1, \dots\ y_N$ ανεξάρτητα

$\hat{b} = G \cdot y$

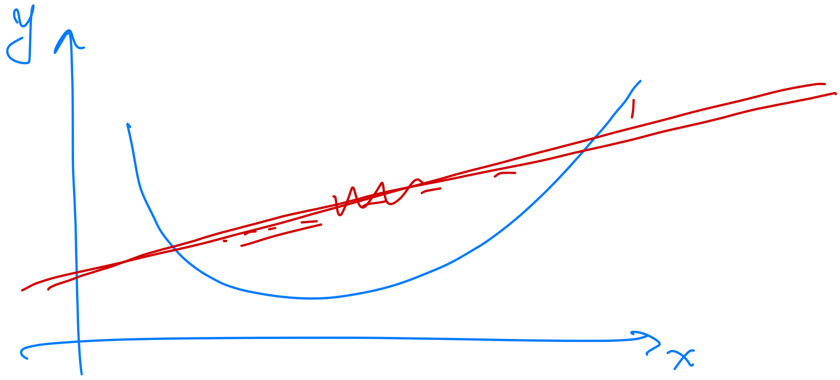
$E\,\hat{b} = \underbrace{(X^T X)^{-1} \cdot X^T}_{G} \cdot E y =$

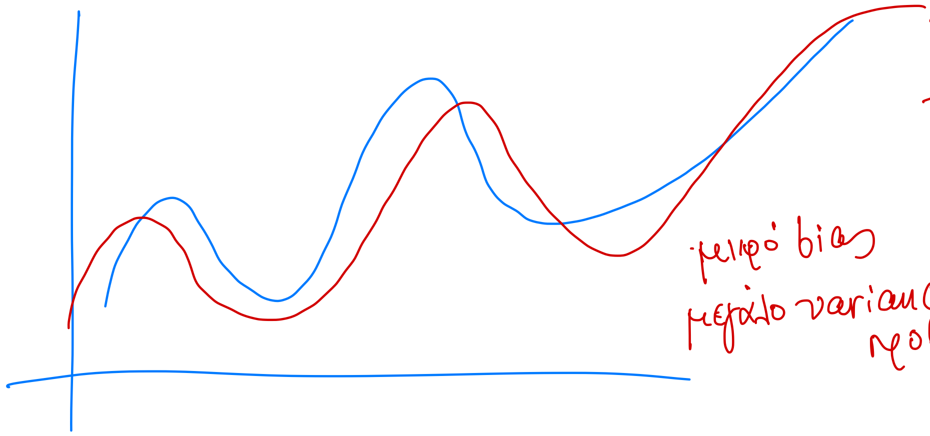
$= \underbrace{(X^T X)^{-1}} \cdot \underbrace{X^T \cdot X} b = b \Rightarrow \boxed{E(\hat{b}) = b}$ αμερόληπτη

$\hat{y} = X \cdot \hat{b} \Rightarrow E(\hat{y}) = E(y)$: και οι προβλέψεις είναι αμερόληπτες !!

δεν έχουν συστηματικό σφάλμα

Αν δεν ισχύει $E(Y|x) = x^T b$ κ' εμείς εκτιμήσουμε $\hat{b}$, $\hat{y} = X\hat{b} \Rightarrow \boxed{E(\hat{y}) \neq E(y)}$ bias





$b_0 + b_1 X_1 + b_2 X_1^2 + b_3 X_1^3 + b_4 X_1^4$

μικρό bias
μεγάλο variance στην πρόβλεψη