

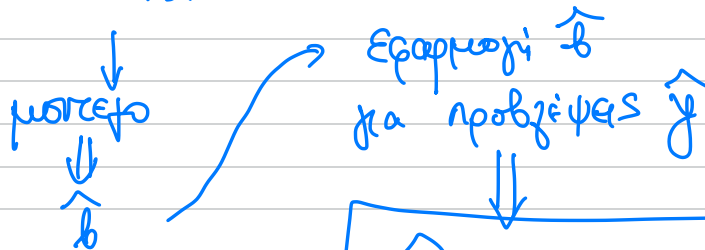
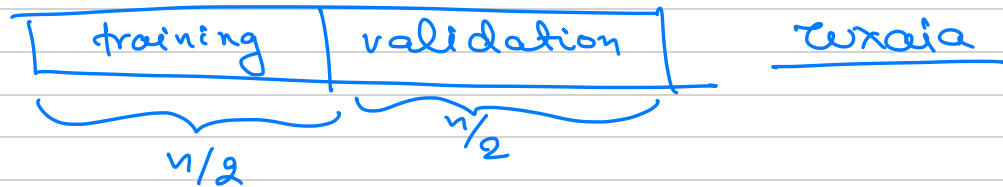
28-3-2025

Resampling methods

Cross Validation

Data Set n παραμορφώσεις
Validation Set

①



$$\hat{MPE} = \frac{1}{n} \sum (y_i - \hat{y}_i)^2$$

μεγάλο
variance

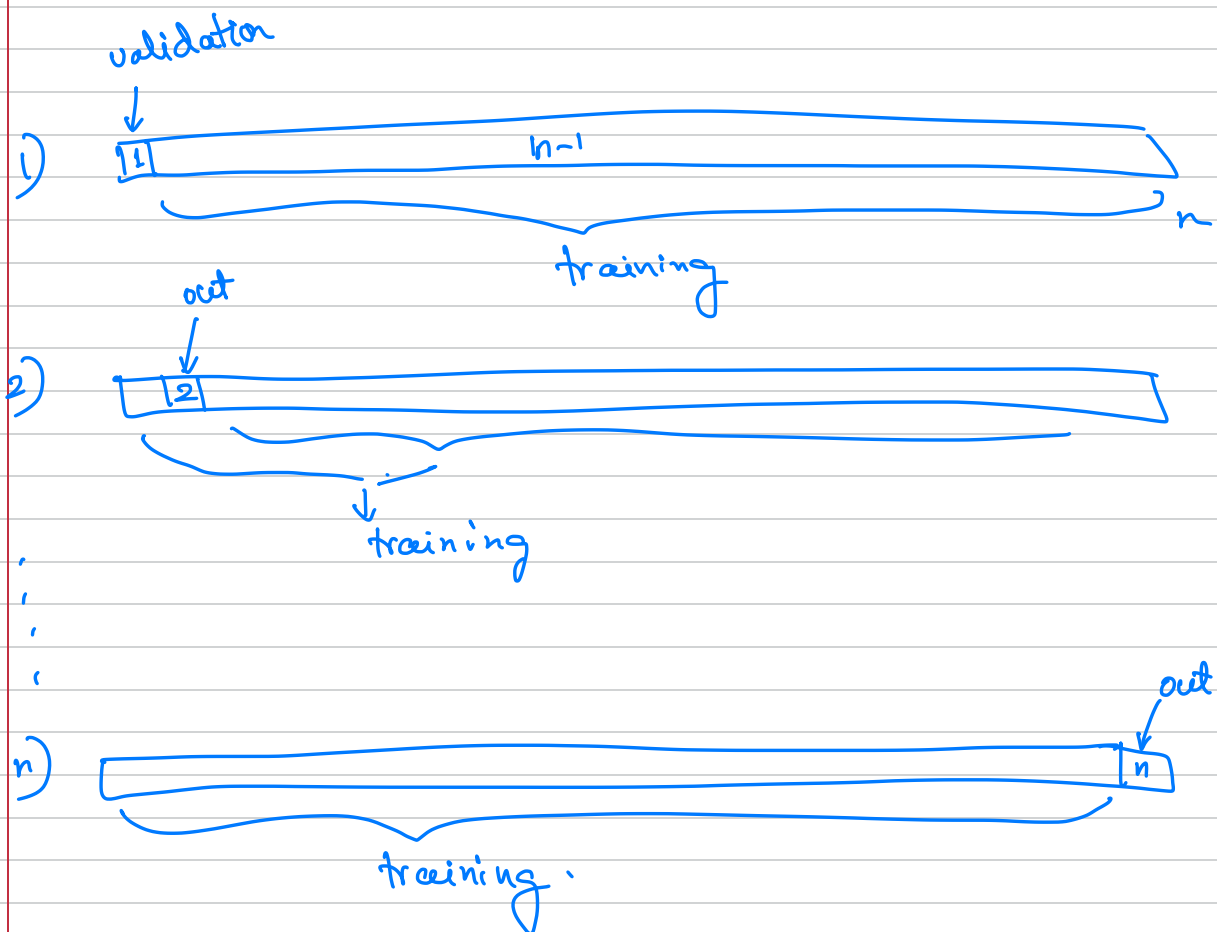
(ως προς

την τυχαία
του validation
set)

Το $\hat{MPE}$ μετράει

υπερεκτιμά το πραγματικό MPE
επειδή το training set έχει το πιο
μικρότερο από το αρχικό data set.

2) LOOCV (Leave One Out Cross Validation)



η εκτίμηση του MPE

$$\widehat{MPE}_j = \underbrace{(y_j - \hat{y}_j)^2}_{\substack{\hat{y}_j \text{ από training set} \\ \text{με υπόλοιπων } n-1 \\ \text{εξώς του } j \text{ παρατηρήσεων}}}$$

$$\widehat{MPE} = \frac{1}{n} \sum \widehat{MPE}_j$$

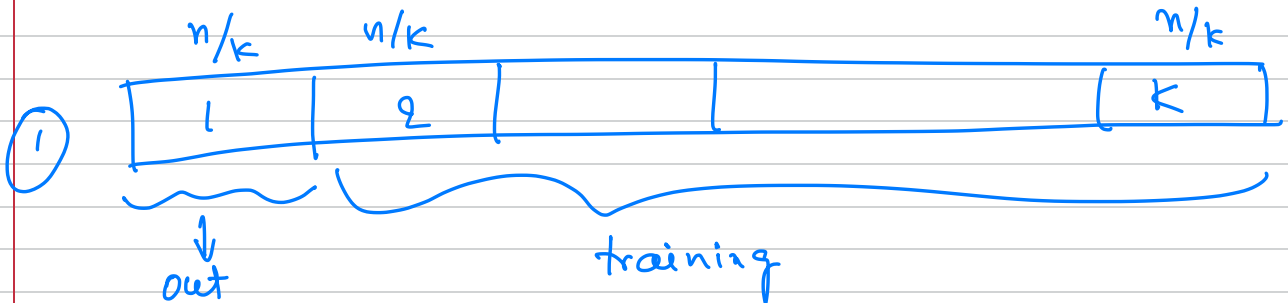
Όταν το ποσοστό έξιμων είναι μικρότερο
π.χ. ελάχιστο τετράγωνο τότε

$$\widehat{MPE}_j \rightarrow \text{προκύπτει από} \\ \text{το ποσοστό } \frac{n}{n} \text{ παρατηρήσεις} \\ \sim \frac{MSE}{\sqrt{1-h_j}}$$

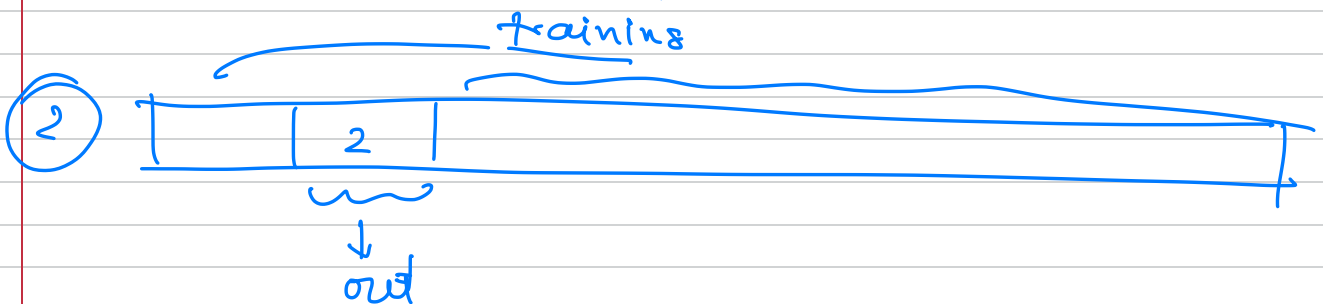
επίσης ποσοστό
μεγαλύτερο
υποδηλώνει
χρόνος

③ Cross-Validation

Χωρίζουμε το dataset σε k ισομεγέθη υποσύνολα με τυχαίο τρόπο (συνήθως $k \approx 5$ ή 10)



$$\hat{MPE}_1 = \frac{1}{n/k} \sum_{i=1}^{n/k} (y_i - \hat{y}_i)^2 \quad (\text{στο group 1})$$



$$\hat{MPE}_2 = \frac{1}{n/k} \sum \quad (\text{στο group 2})$$

Οι ηφισότερες μέθοδοι εφίπησης (regression
glm ...)
έχουν και μια CV. version

πχ. $\text{glm}(y \sim x, \dots)$

$\text{cv.glm}(y \sim x, \dots, k=5)$

② Bootstrap

Παράδειγμα $(X_1, X_2, \dots, X_n)$ από $\text{Exp}(\lambda)$
 $\hat{\lambda} = ?$

$$EX = \frac{1}{\lambda} = \mu$$

$\hat{\mu} = \overline{X}_n \rightarrow$ αμερόληπτη εκτίμ. του $\mu = \frac{1}{\lambda}$

$$E(\overline{X}_n) = \frac{1}{\lambda}$$

συνεπώς $\lim_{n \rightarrow \infty} \overline{X}_n = \frac{1}{\lambda}$ π.Α.Ι.

$$\hat{\lambda} = \frac{1}{\overline{X}_n}$$

συνεπώς $\frac{1}{\overline{X}_n} \xrightarrow{\text{π.Α.Ι.}} \frac{1}{\mu} = \lambda$

$$E\left(\frac{1}{\overline{X}_n}\right) \neq \lambda \quad \underline{\text{(biased)}}$$

$$\bullet \lambda = \frac{1}{E(\overline{X}_n)} \neq E\left(\frac{1}{\overline{X}_n}\right)$$

Ποιό είναι το bias του $1/\overline{X}_n$?

$$\text{Bias}(\hat{\lambda}) = E(\hat{\lambda} - \lambda) = b(\lambda) \quad \left(\begin{array}{l} \text{μπορεί να} \\ \text{υπαρξήσει} \\ \text{καρπύζα} \\ \text{συναρτήσει του } \lambda \end{array} \right)$$

Ενδιαφέροντα $\left\{ \begin{array}{l} \text{Προσομοίωση } N \text{ δειγμάτων μεγ. } n \\ \text{με } \text{Exp}(\lambda) \\ \forall \text{ δείγμα} \rightarrow (\overline{X}_n)_j, j=1, \dots, N \end{array} \right.$

$$\hat{b} = \frac{1}{N} \sum \left(\frac{L}{(\bar{x}_n)_j} - 1 \right) \in \mathcal{A}$$

Εστω συγκρισιμότητα δείγμα από $\text{Exp}(\lambda)$

όπου λ : άγνωστο

$$\hat{\lambda} = \frac{1}{\bar{x}_n} \quad \left(\begin{array}{c} \text{έγκλ} \\ \text{αριθμός} \end{array} \right) \in \text{ποσό είναι το bias?}$$

Ευαισθησία $\text{MSE}(\hat{\lambda}) = E(\hat{\lambda} - \lambda)^2$

Αν unbiased $\text{MSE} = \text{Var}(\hat{\lambda}) \rightarrow \text{εξέλιξη?}$

Bootstrap

Εστω δείγμα $x_1, \dots, x_n$ iid από F

$$F(x) = P(X \leq x)$$

Από δείγμα $(x_1, \dots, x_n) \Rightarrow$ εμπειρική κατανομή

$$\hat{F}_n(a) = \% j : x_j \leq a = \frac{\sum_{j=1}^n 1(x_j \leq a)}{n}$$

Θεώρημα $\forall a \in \mathbb{R} : \hat{F}_n(a) \rightarrow F(a) \text{ μ.π.τ}$

$$1(x \leq a) \sim \text{Ber}(F(a))$$

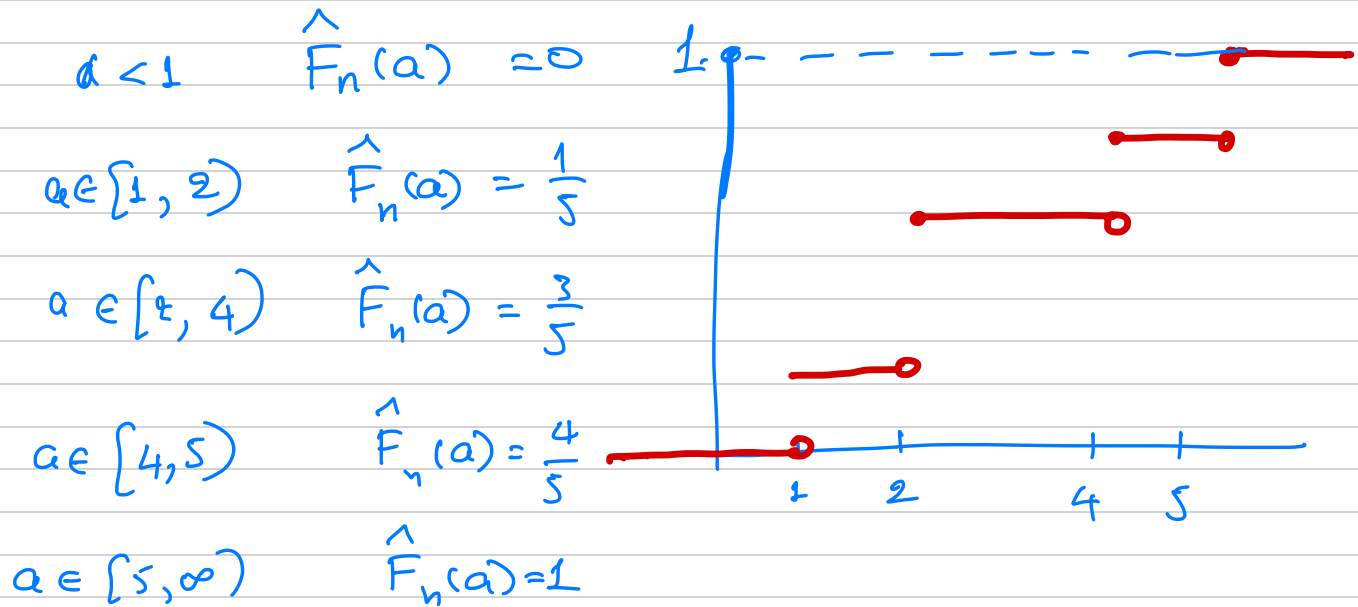
$$\frac{\sum 1(x_j \leq a)}{n} \rightarrow F(a)$$

Θεωρ 2

$\hat{F}_n(a) \rightarrow F(a)$ $n \rightarrow \infty$ μ.π.1
ομοιότητα ως προς a

$\hat{F}_n(\cdot) \leftarrow$ εμπειρική κατανομή
των δεδομένων

π.χ. $x = (1, 2, 2, 4, 5)$



$X = \left\{ \begin{array}{ll} 1 & \mu_n \frac{1}{5} \\ 2 & \mu_n \frac{2}{5} \\ 4 & \mu_n \frac{1}{5} \\ 5 & \mu_n \frac{1}{5} \end{array} \right\}$ συχνότητες των τιμών
στο δείγμα.

Για να προσομοιώσει την $\hat{F}$

αρκεί να προσομοιώσω την X

$\hat{F}$ εκτίμηση της $F \leftarrow$

Θα μπορούσαμε να δημιουργήσουμε N προσομ.
δείγματα από τον $\hat{F}$ & να πάρουμε
μια εκτίμηση του bias όπως παραπάνω

Πείραμα τύχης

Από το δείγμα $(x_1, \dots, x_n)$ παίρνω τυχαία
η ανεξάρτητες παρατηρήσεις με επανάδειξη (!!)
Έστω $(y_1, \dots, y_n)$ $y_j \in \{x_1, \dots, x_n\} \quad \forall j$

Αντικ $(y_1, \dots, y_n) \sim \hat{F}$

R : δείγμα με επανάδειξη

$f = \text{sample}(x, n, \text{replace} = \text{True})$