Cross Validation (K)

$\lambda \in [10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100]$

$\hat{Err}$

① K may impact

K = 5

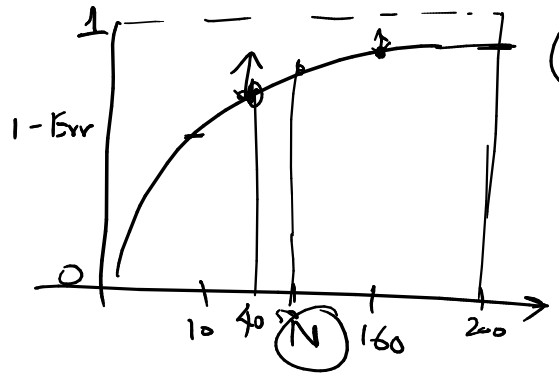
A. N = 200

B. K = 5

N = 50

K = N.

$\{x_i, y_i\}_{i=1}^{N} = D$



P_{XY}

models

2). Wrong way: Preprocessing the data using both x_is and y_is. (training)

Right way: $K = 5$

for each choice:

for each fold:

preprocessing here. on the remaining folds]

~~X~~ [$N = 50$, $P = 5000$, X_i are Gaussians, $X_i \perp Y$]

$Y = 0 \cdot X_1 + 0 X_2 + \cdots 0 X_{5000} + \varepsilon$

Bootstrap

$$\frac{\hat{\beta}_j}{\sqrt{v_i}\,\sigma} \sim N(0,1)$$

1. Procedure
2. Understanding

(1) $\{z_i\}_{i=1}^{N} = Z$

Create bootstrap samples.

5 times 1st row was sampled.

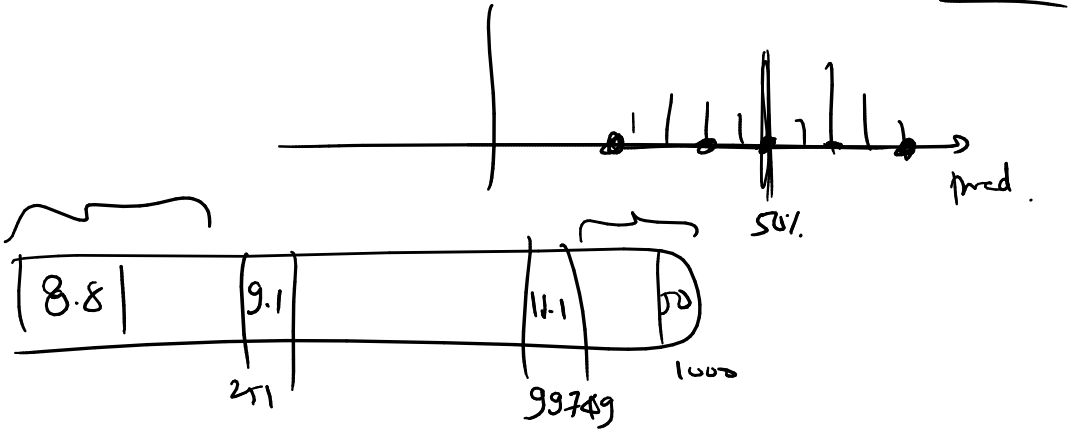
10,000

$Z^{*1}, \ldots\ldots Z^{*B}$.

N obs each.

$\hat{f}^{*1}$

$\hat{f}^{*1}(x^{test})$, $\hat{f}^{*2}(x^{test})$, ----- B = 10000



Estimate Err

$$E_{\tau} E_{P_{XY}}\left[\left(\hat{f}_{\tau}(x) - Y\right)^2\right] \qquad \tau \sim (P_{XY})^N$$

for each obs i

pick the ones it is not part of.

$$C_i = \{1, \ldots, 100\}$$

$$Err \approx \frac{1}{N}\sum_{i=1}^{N} \underbrace{\frac{1}{|C_i|}\sum_{b \in C_i} \left(\hat{f}^{*b}(x_i) - y_i\right)^2}$$

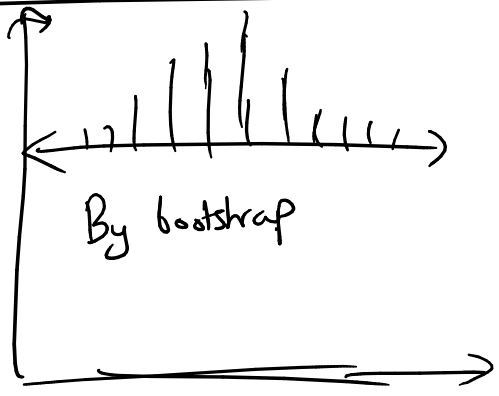
$$\{x_i, y_i\}_{i=1}^{N} \longrightarrow \hat{f}$$

$$E_{P_{XY}}\left[\left(\hat{f}(X) - Y\right)^2\right]$$

Understanding:

$$\frac{(P_{XY})^N \quad \{x_i, y_i\}^N \longrightarrow \hat{f}}{(P_Z)^N \quad \{z_i\}^N \longrightarrow \hat{\theta} = \frac{1}{N}\sum_{i=1}^{N} z_i}$$

(A)



$z_1, \dots z_N \sim (P_Z)^N$ ← population dist

$\hat{\theta}(z_1, \dots z_N)$ ← sampling dist

$(z_1 \ldots z_N)^{*1} \sim (P_z)^N \to \hat{\theta}_1$

$(z_1 \ldots z_N)^{*1} \sim (P_z)^N \to \hat{\theta}_2$

$\vdots$

$\hat{\theta}_B$

$\approx$ Sampling dist $\hat{\theta}(z_1, \ldots z_N$

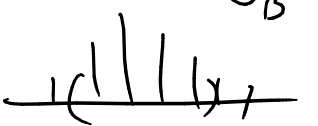
$\{z_1 \ldots z_N\} \sim (P_z)^N$

↳ w/R. B times

$\hat{\theta}_1$

$\vdots$

$\hat{\theta}_B$



MLE.

$z_1, \ldots z_N \sim N(\mu, 1)$

likelihood of data for a given parameter.

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} z_i$$

$$P_\mu(z_1 \ldots z_N) = \prod_{i=1}^{N} P_\mu(z_i)$$

$$\propto e^{-\frac{(z_i-\mu)^2}{2}}$$

$$LL = \sum_{i=1}^{N} \left(-\frac{(z_i-\mu)^2}{2} \right) + C$$

least squares : MLE interpretation.

P_{XY} [X is not random.
$Y = \beta^T X + \varepsilon$
$\varepsilon \sim N(0, \sigma^2)$

$$P(Y \mid X = x) \sim N(\beta^T x, \sigma^2)$$

β, σ

$\{x_i, y_i\}_{i=1}^N$

$$LL \propto \sum_{i=1}^{N} \log\left(\exp\left(-\frac{(\beta^T x_i - y_i)^2}{2\sigma^2}\right)\right) + C(\sigma)$$

$$\max_{\beta} LL \equiv \min_{\beta} \sum_{i=1}^{N} (y_i - \beta^T x_i)^2$$

Ridge regression.

$$\min_{\beta} \left(\sum_{i=1}^{N} (y_i - x_i^T \beta)^2 + \lambda \|\beta\|_2^2 \right)$$

MAP estimation

$$\max_{\beta} P(\beta \mid \text{data}) \propto P(\text{data} \mid \beta) \cdot P(\beta)$$

$$\Rightarrow \max_{\beta} \log P(\beta \mid \text{data}) \propto \log P(\text{data} \mid \beta) + \log P(\beta)$$

$$\Rightarrow \text{MAP est} \equiv \min_{\beta} \sum_{i=1}^{N} (y_i - x_i^T \beta)^2 + \frac{1}{\tau^2} \beta^T \beta$$

$$P(\beta \mid \text{data}) \propto P(\text{data} \mid \beta) \cdot P(\beta)$$

$$P(\beta) \sim N(0, \tau^2 I)$$

$$-\beta^T \Sigma^{-1} \beta$$

$$p(\beta) \sim N(0, \tau^2 I)$$

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \tau^2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\frac{1}{\sqrt{2\pi \det(\tau^2 I)}} \cdot \exp\left(-\frac{1}{2}(\beta - 0)^T \underset{\sim}{\Sigma^{-1}} (\beta - 0)\right)$$

$$(\tau^2 I)^{-1}$$

$$\frac{1}{\tau^2} I$$

$$-\frac{1}{2} \beta^T \left(\frac{1}{\tau^2} I\right) \beta = -\frac{1}{2\tau^2} \beta^T \beta$$

① LASSO: Belief on β each coordinate is Laplace distributed.
↳ with 0 mean.

$$\lambda \|\beta\|_1$$

$$\equiv \lambda \sum_{j=1}^{p} |\beta_j|$$

$$P(\beta_j) \propto \exp\left(-\frac{|\beta_j|}{c}\right)$$

Naive Bayes for classification : MLE

$$p(x_i \mid g_i) = \prod_{j=1}^{p} p(x_{ij} \mid g_i)$$

$G = \{\underset{1}{\text{Spam}}, \underset{2}{\text{No-spam}}\}$

x^{test} : $\underset{k=\{1,2\}}{\arg\max}\ P(G=k \mid x^{test})$

$X_1 \ldots\ldots X_p$

$P(x^{test} \mid G=k) \cdot P(G=k)$

$\hat{P}(G=1) = \frac{\#\ \text{class } 1}{N}$

$\hat{P}(X_j=1 \mid G=k) = ?$ for $j=1,\ldots p$, $k=1,2$.