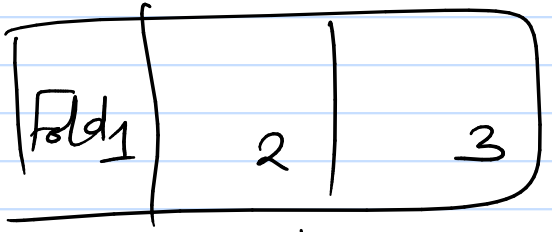
Model assessment

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

$$X^{full} = \begin{bmatrix} X \\ X^{test} \end{bmatrix} \quad \begin{matrix} \text{dcol} \\ \updownarrow \\ b \end{matrix}$$

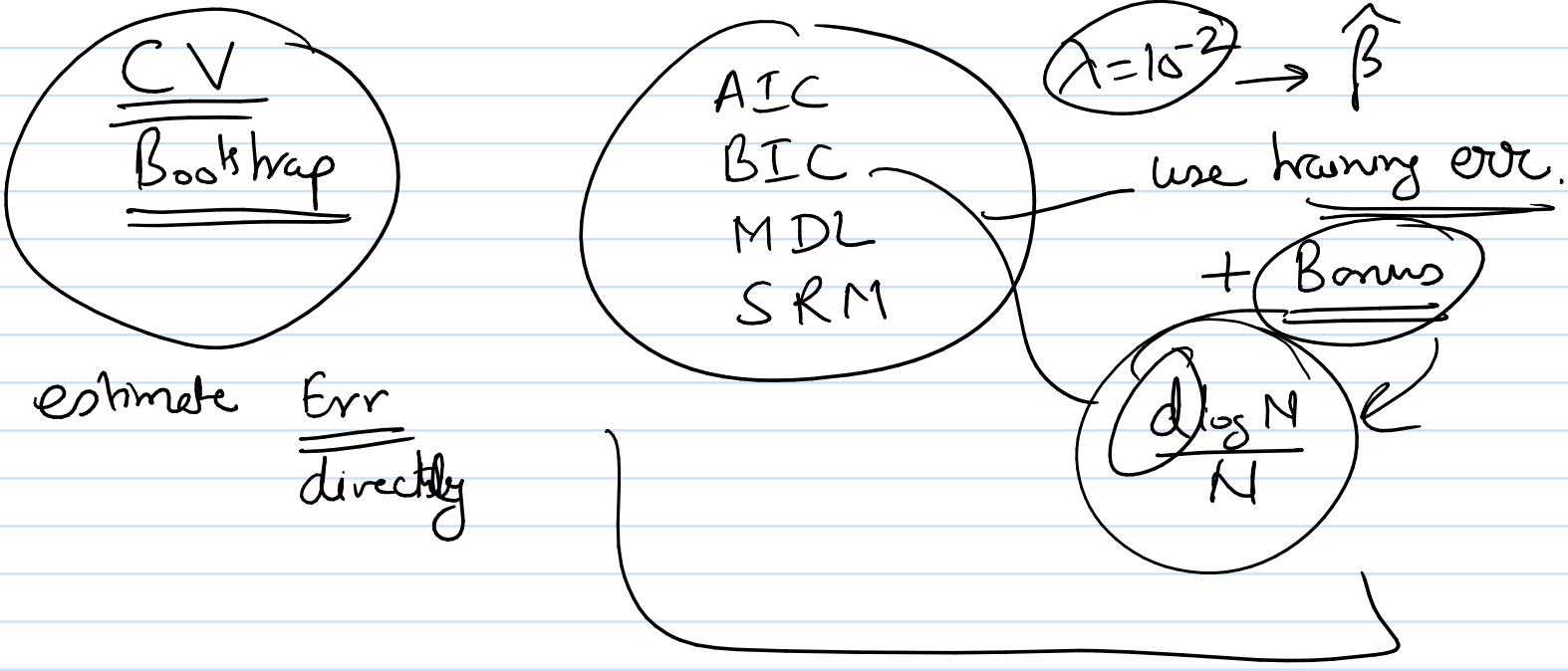
Selection CV



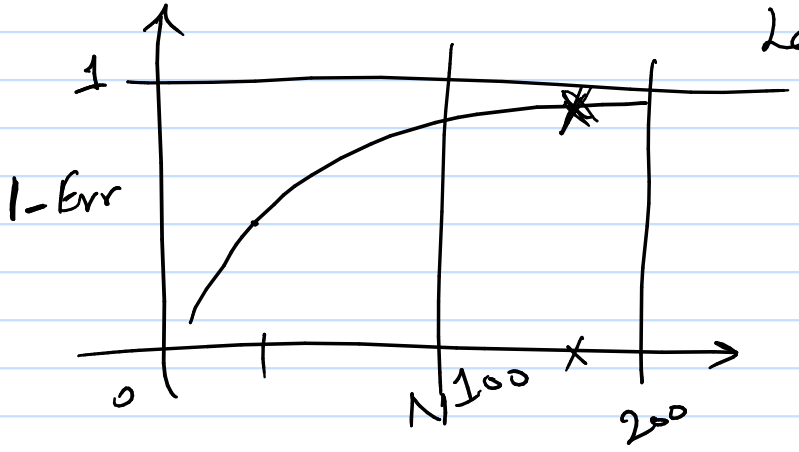

$\lambda = 10^{-2}$

$\hat{\beta}_{12}$

pey 1
pey 3
pey 2


CV
Bootstrap
estimate Err directly
AIC
BIC
MDL
SRM
$\lambda = 10^{-2} \rightarrow \hat{\beta}$
use training err.
+ Bonus
$\frac{d \log N}{N}$

Learning curve.

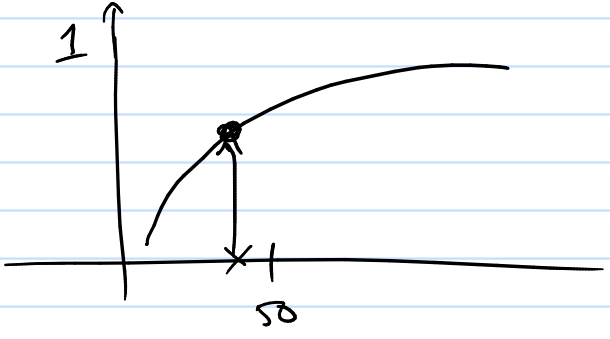



K- fold

5.

40 obs in each fold.

160 obs → $\hat{\beta}$



for every choice of model.

K fold
5

(4) → Pick 4 folds

$N = 50 \quad p = 5000 \quad X_i \sim N(0,1)$

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

$G = \mathbb{1}[Y > 0]$

Reality: 50%.

① Scatter $Y, X_i$

pick 100

build a classifier. → 3%

$K=1$
K-nn

Naive Bayes: $W = [W_1\ W_2\ W_3]^T \quad p = 3$

$$P_\theta(Y = \text{spam} \mid \underline{W} = (\omega_1, \omega_2, \omega_3)) > P_\theta(Y = \text{not spam} \mid W = \omega)$$

$$P_\theta(Y = y \mid W = \omega) \propto P_\theta(W = \omega \mid Y = y) \cdot P_\theta(Y = y)$$

$y$: spam

$$\star\ P(W_1 = \omega_1 \mid Y = y) \cdot P(W_2 = \omega_2 \mid Y = y) \cdot P(W_3 = \omega_3 \mid Y = y)$$

$y$: spam

$$= \frac{\#(\omega_1 \text{ appears \& spam})}{\#\text{spams}}$$

$$P(Y = \text{spam}) \approx \frac{\#\text{ spams}}{N} = \hat{\theta}_1 \quad : \text{MLE.}$$

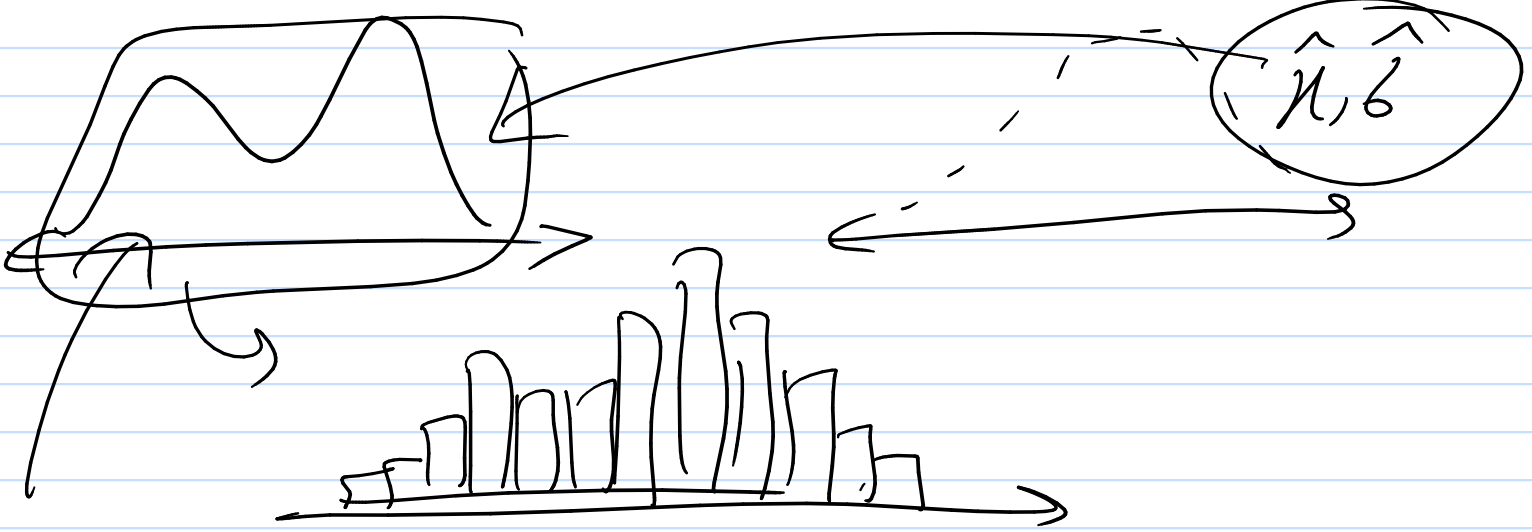
$$\frac{1}{\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

$$P_{\theta_1}(Y_1 = y_1) = \theta_1^{\mathbb{1}[y_1 = \text{spam}]} \cdot (1-\theta_1)^{\mathbb{1}[y_1 = \text{not spam}]}$$

$$\log P(Y_1 = y_1) + \log P(Y_2 = y_2) \dots + \log\left(P(Y_N = y_N)\right)$$

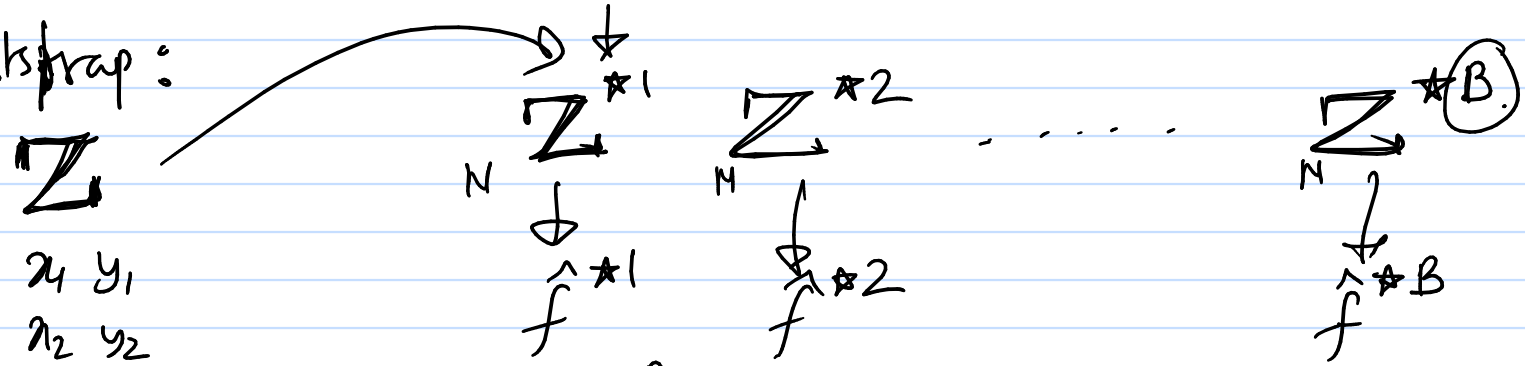
$$\sum_{i=1}^{N} \left( \mathbb{1}[y_i = \text{spam}] \log \theta_1 + \mathbb{1}[y_i = \text{not spam}] \log(1-\theta_1) \right)$$

$\#$ spams $\qquad$ $(N - \#\text{ spams})$




Gaussian
mixture
models

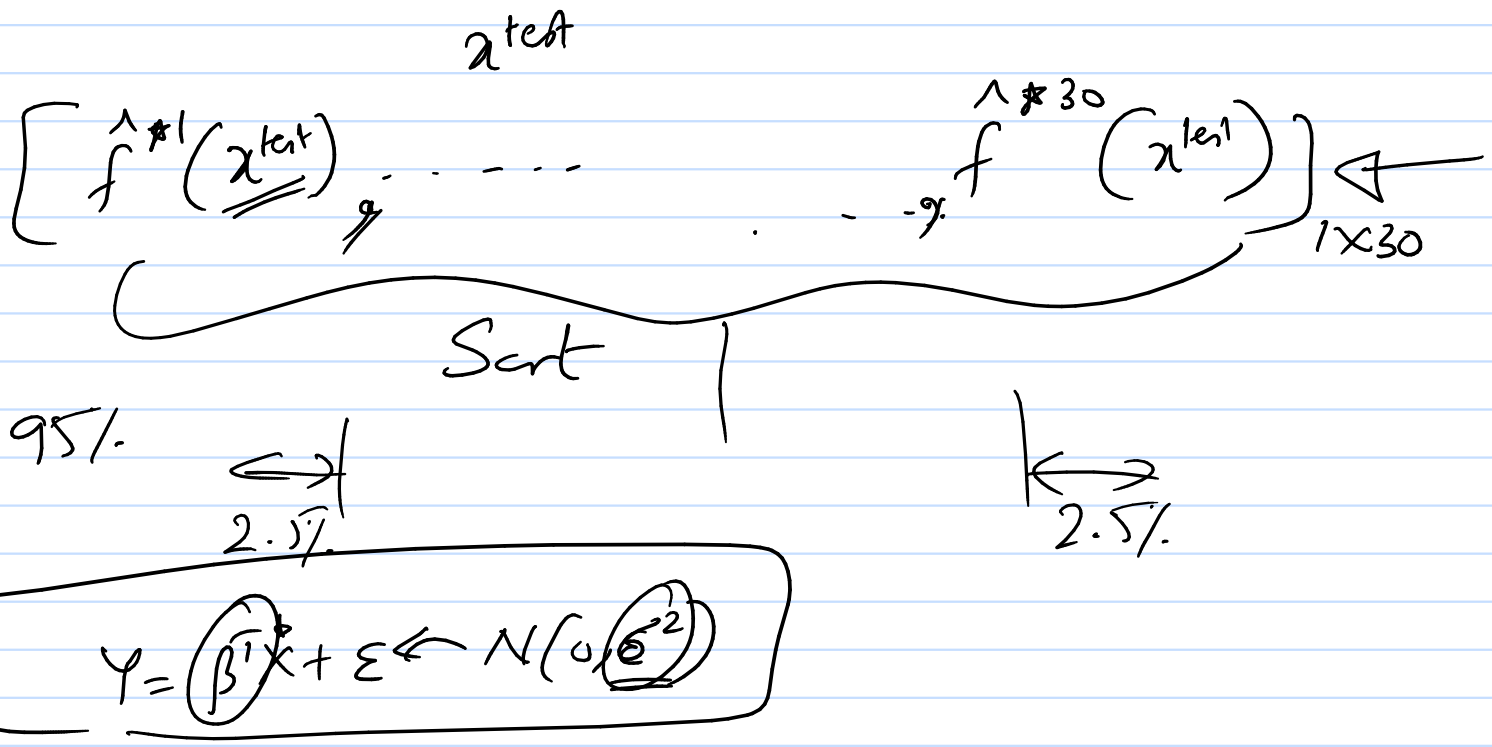
Bootstrap:




① $\frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(x^{test})$

② Estimate "Err": average( held out data for each $\hat{f}^{*b}$ performance on )

Confidence bands around estimates: $\hat{\beta}$

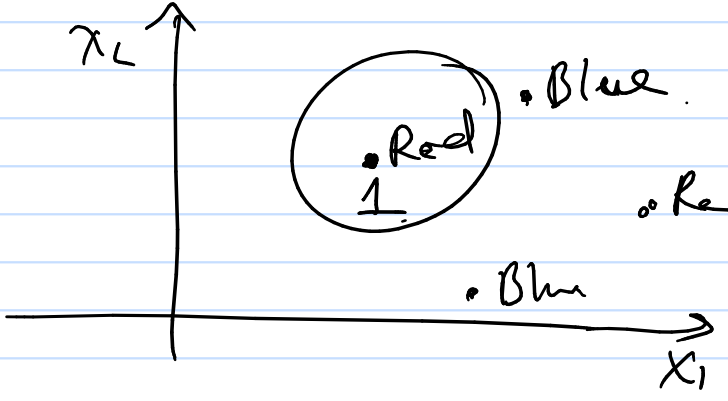


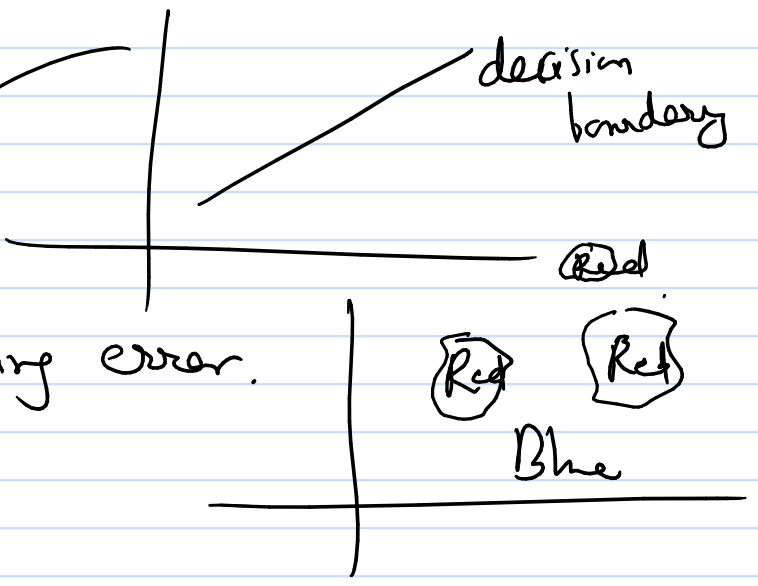
① PCA & Ridge regression.

② Why $\hat{\beta} = (X^T X)^{-1} X^T Y$

L1 : k-nn & linear model.

- k- Choice k=1 gives 0 training error.

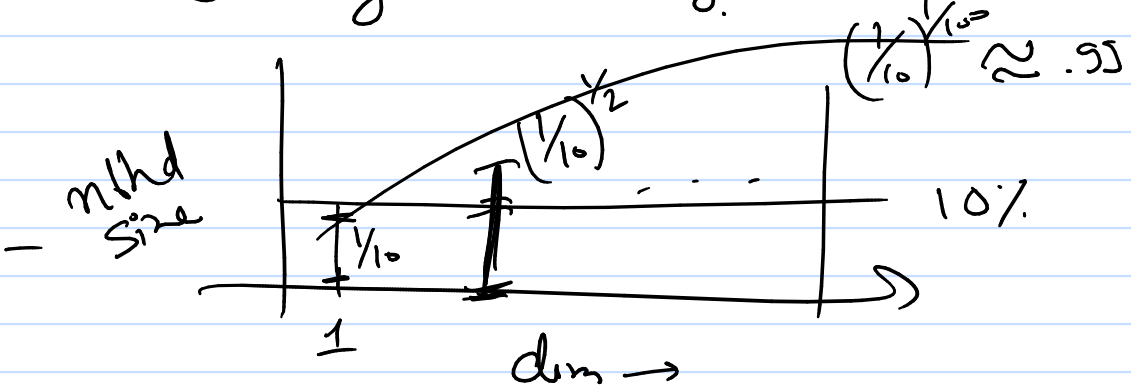


L2: Regression function $E[Y|X=x]$ $\approx$ K-nn : in theory

$\approx \beta^T x$

Curse of dimensionality.




- Bayes classifier: $P_{G|X}$

$$RSS(\beta) = \frac{1}{N}\underbrace{(Y - X\beta)^T}_{1\times N}\underbrace{(Y - X\beta)}_{N\times 1} = \frac{1}{N}\sum_{i=1}^{N}(y_i - x_i^T\beta)^2.$$

L3: Bias-Variance tradeoff. kNN $\hat{f} \propto \frac{1}{k}$

Sampling properties of $\hat{\beta}$

↳ $Y = \beta^T X + \varepsilon$

- $\varepsilon \sim N(0, \sigma^2)$ — variance known; independence
- $X$ — not random.

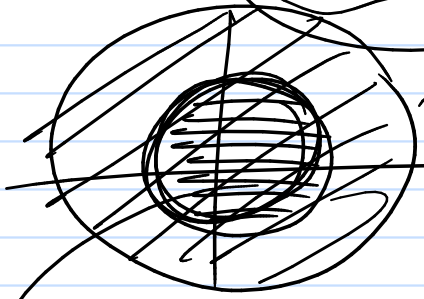
$Z \propto \frac{\hat{\beta}_i - \beta_i}{\sigma}$ — Normal.

Ways to add bias: subset, Ridge & lasso.

CV

L4: Ridge & lasso geometric likelihood view.

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - x_i^T\beta)^2 + \lambda\sum_j \beta_j^2$$




λ large.

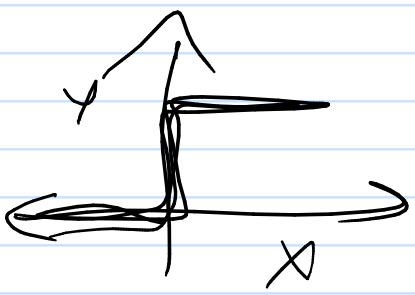
$$\hat{\beta}_{ridge} = (X^TX + \lambda I)^{-1} X^T Y$$

$$RSS(\beta) + \lambda \|\beta\|_2^2$$

$$(y - X\beta)^T (y - X\beta) + \lambda \beta^T I \beta$$

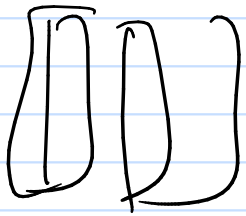
$$(y^T - \beta^T X^T)(y - X\beta) + \lambda \beta^T I \beta$$

$$y^T y - y^T X\beta - \beta^T X^T y + \beta^T X^T X \beta + \lambda \beta^T I \beta$$



Reg for classification.

$$\underset{p\times 3}{\hat{\beta}} = \underbrace{\underbrace{(X^T X)^{-1}}_{p\times p} \underbrace{X^T}_{p\times N}}_{A_{p\times N}} \underbrace{Y}_{N\times 3}$$



LDA vs LReg.

Normality

$y_i = [0 \; 0 \; 0 \; 1 \; 0 \; 0 \; 0]$ (1 × 10)

"7"

$\beta_\ell \; \ell = 1, \ldots, K-1$

Cross entropy loss (i) $= \sum_{\ell=1}^{K} y_{ij} \cdot \frac{e^{\beta_\ell^T x}}{1 + \sum_{k=1}^{K-1} e^{\beta_k^T x}}$

L6:

CV, Bootstrap: Model selection, assessment.