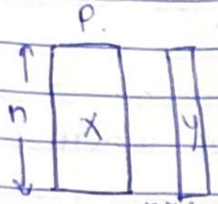
Lecture 25 05/07/2020



$y \in \mathbb{R}^n$ Linear models $\mathcal{H} = \{\vec{w} \cdot \vec{x} : \vec{w} \in \mathbb{R}^{p+1}\}$.

$X \in \mathbb{R}^{n \times (p+1)}$

$$\vec{b}_{OLS} = \underset{\vec{w} \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \{SSE\}$$

$$= \operatorname{argmin} \{(\vec{y} - X\vec{w})^T (\vec{y} - X\vec{w})\}$$

$$= \text{long proof} = (X^T X)^{-1} X^T \vec{y}$$

$X^T X \in \mathbb{R}^{(p+1) \times (p+1)}$ has same rank as X.

Let $p+1 > n$ $\quad X = n \underbrace{[\,\square \cdots\cdots \square\,]}_{\longleftarrow\, p+1 \,\longrightarrow}$ $\operatorname{rank}[X] = n < p+1$

$\Longrightarrow X^T X$ is not invertible

$\Longrightarrow \vec{b}_{OLS}$ is impossible to compute So $\mathcal{A} = OLS$ breaks down.

Hoerl and Kennard (1970) use $X^T X + \lambda I_p$ instead of $X^T X$. $\lambda \neq 0$ scalar constant

$X^T X + \lambda I$ is always invertible. Pf: Note $X^T X = V D V^{-1}$ (diagonalization)

has the eigenvector as column

$\operatorname{rank}[X^T X] = n < p+1$

$$D = \begin{bmatrix} \lambda_1 & & & & 0 \\ & \lambda_2 & & & \\ & & \ddots \lambda_n & & \\ & & & 0 & \\ 0 & & & & \ddots 0 \end{bmatrix}, \quad D + \lambda I = \begin{bmatrix} \lambda_1 + \lambda & & & & 0 \\ & \lambda_2 + \lambda & & & \\ & & \ddots \lambda_n + \lambda & & \\ 0 & & & & \ddots \lambda \end{bmatrix}$$

is diagonal with eigenvalues in the diagonal.

$$X^T X + \lambda I = V D V^{-1} + \lambda I = V D V^{-1} + \lambda I V V^{-1} = V D V^{-1} + V(\lambda I) V^{-1}$$

$$X^T X + \lambda I = V(D + \lambda I) V^{-1} \Longrightarrow (X^T X + \lambda I)^{-1}$$

$$= V(D + \lambda I)^{-1} V^{-1}$$

$\vec{b}_{ridge} = (X^T X + \lambda I)^{-1} X^T \vec{y}$ $\qquad \underbrace{\lambda \sum_{j=0}^{p} w_j^2}$ ← penalty term if $\lambda > 0$

Consider A: $\vec{b} = \underset{\vec{w} \in \mathbb{R}^{p+1}}{\text{argmin}} \left[SSE + \lambda \|\vec{w}\|^2 \right]$ L2 regularization (make it simpler)

$(\vec{y} - X\vec{w})^T (\vec{y} - X\vec{w}) + \lambda \vec{w}^T \vec{w}$
$= \vec{y}^T \vec{y} - 2\vec{w}^T X^T \vec{y} + \vec{w}^T X^T X \vec{w} + \vec{w}^T (\lambda I) \vec{w}$

$= \vec{y}^T \vec{y} - 2\vec{w}^T X^T \vec{y} + \vec{w}^T (X^T X + \lambda I) \vec{w}$

$= \frac{\partial}{\partial \vec{w}} \left[\quad \right] = -2X^T \vec{y} + 2(X^T X + \lambda I)\vec{w}$

$\overset{set}{=} 0 \implies (X^T X + \lambda I)\vec{w} = X^T \vec{y}$
$\implies \vec{b}_{ridge} = \vec{w} = (X^T X + \lambda I)^{-1} X^T \vec{y}$

Math 341 likelihood $\vec{Y} | X \sim N(X\vec{\beta}, \sigma^2 I)$, $B_j \overset{iid}{\sim} N(0, \tau^2)$
$\implies \hat{\beta}_{MAP} = b_{ridge}$ $\quad \lambda = \sigma^2 / \tau^2$.

Lec 19.Rmd

A: $\vec{b}_{Lasso} = \text{argmin} \left[SSE + \lambda \|\vec{w}\|_1 \right]$ $\quad$ L1 penalizal regression
$\lambda \sum_{j=0}^{p} |w_j|$ $\quad$ L1 regularization.

Generaly; there's no close form solution.

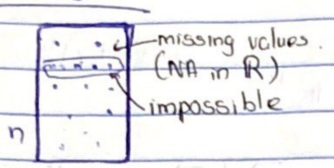
Math 341/621 same likelihood, $B_j \overset{iid}{\sim} \text{Laplace}(0, \tau) \Rightarrow \hat{\beta}_{map} = \vec{b}_{Lasso}$

Lasso sets many entries of $\vec{b}$ to be $= 0$. For those nonzero entries, you can consider those variable "selected". Lasso does "Variable selection". "Occam's Razouizor" (simplifies the model).

What if I want to combine ridge & lasso? $\lambda > 0, \alpha \in (0,1)$

A: $\vec{b}_{en} = \text{argmin} \left[SSE + \lambda (\alpha \|\vec{w}\|_1 + (1-\alpha) \|w\|_2^2) \right]$
"Elastic net" algorithm.

Missingness

$X =$ 

$Y =$ By def. there cannot be missingness in Y.

If there is missing value in X, all A's we discuss. will fail.

Missing Data Mechanism (MDM)

MDM	$P(M_j \mid X_{j,miss}, \vec{X}_{-j,miss}, Y, \vec{X}_{-j,obs}, \delta)$
MCAR - missing completely at random	$P(M_j \mid \delta)$
MAR - missing at random	$P(M_j \mid \vec{X}_{-j,miss}, \vec{X}_{-j,obs}, \delta)$
NMAR - NOT MAR	doesn't simplify - very difficult

(M_j: Bernoulli which is 1 if missing; δ: constar)

Strategies to fix X to be used in A's

(I) Listwise Deletion - drop all observation those have any missing values. Why is this bad?

(II) Impute (predict) missing values, build a prediction model for X_j's. Then fill in missing values with preditions.

My recommand imputation procedure: "Miss Forest"

(0) Fill in all missing values with ($\bar{x}_j$ the respective column averages).

(1) Fit $\vec{x}_1 \sim RF(\vec{x}_{-1})$ where $\vec{x}_1$ was present in original D. Then set missing values of $\vec{x}_1$ to be prediction from the RF

(2) Fit $\vec{x}_2 \sim RF(\vec{x}_{-2})$ " " & set missing value of $\vec{x}_2$ " " " " .

(p) Fit $\vec{x}_p \sim RF(\vec{X}_{-p})$. " " " .

(p+1) Repeat steps 1-p until " convergence " i.e. imputed value don't change significantly from iteration - iteration. Result: a D with no missingness

One more general recommendation

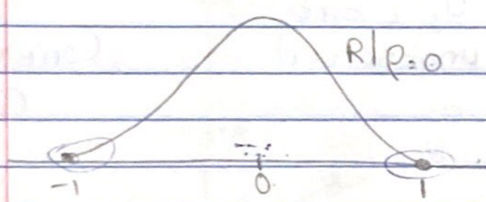
$$X_{final} = [\, X_{imp} \mid \underbrace{\vec{m}_1 \mid \vec{m}_2 \mid \dots \mid \vec{m}_p}_{m} \,]$$

dummy variable; binary vector indicating missingness by observation

$\longleftarrow 2p \longrightarrow$

Two r.v's x_1, x_2 has sample correlation.

$$\rho_{12} \approx r_{12} = \frac{\sum (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{\sqrt{\sum (x_{i1} - \bar{x}_1)^2 \sum (x_{i2} - \bar{x}_2)^2}}$$



You can imagine due to sampling error.

$|r_{12}| \approx 1$ but $\rho_{12} = 0$.

These situation are called "spurious correlation"