4a) Derivation of the closed form solution for parameter θ that minimizes the loss function $J(\theta)$ in ridge regression.

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{N}(h_\theta(x_i) - y_i)^2 + \frac{\lambda}{2}\sum_{j=1}^{d}\theta_j^2$$

The term d in regularization part denotes amount of covariables used in the model.

we can rewrite $J(\theta)$ in matrix-notation and further break it down.

$$J(\theta) = \frac{1}{2}(X\theta - y)^T(X\theta - y) + \frac{\lambda}{2}(\theta^T\theta)$$

the term $\frac{\lambda}{2}\sum_{j=1}^{d}\theta_j^2$ represents regularization we apply on coefficients.

$$\Rightarrow \frac{1}{2}\left[X^T\theta^T X\theta - X^T\theta^T y - X\theta y^T + y^T y + \lambda\theta^T\theta\right]$$

$\left[(X\theta)^T y = (X\theta)y^T\right]$ the a transposed scalar is same scalar.

$$\Rightarrow \frac{1}{2}\left[X^T\theta^T X\theta - 2X^T\theta^T y + y^T y + \theta^T\lambda I\,\theta\right]$$

$[I$ = Identity matrix$]$

$$\Rightarrow \frac{1}{2}\left[y^T y - 2X^T\theta^T y + \theta^T(X^T X + \lambda I)\theta\right]$$

θ should minimize $J(\theta)$.

By matrix differentiation rule $\frac{\partial x^T A x}{\partial x} = (A + A^T)x = 2Ax$

we can apply it in here as

$$\frac{\partial J}{\partial \theta} \Rightarrow \frac{1}{2}\left[0 - 2X^T y + 2(X^T X + \lambda I)\theta\right] = 0.$$

$$\Rightarrow -X^T y + (X^T X + \lambda I)\theta = 0$$

$$\Rightarrow (X^T X + \lambda I)\theta = X^T y.$$

$$\boxed{\theta = (X^T X + \lambda I)^{-1} + X^T y.}$$ is a closed form eq.

5) $$X = \begin{bmatrix} 1 & x_{i1} & x_{j1} \\ 1 & x_{i2} & x_{j2} \\ 1 & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ 1 & x_{in} & x_{jn} \end{bmatrix}_{n\times 3}$$

$$X^T = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_{i1} & x_{i2} & \dots & x_{in} \\ x_{j1} & x_{j2} & \dots & x_{jn} \end{bmatrix}_{3\times n.}$$

$x_{j1} = 2x_{i1}$

So $X = \begin{bmatrix} 1 & x_{i1} & 2x_{i1} \\ 1 & x_{i2} & 2x_{i2} \\ \vdots & \vdots & \vdots \\ 1 & x_{in} & 2x_{in} \end{bmatrix}_{n\times 3}$

$X^T = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{i1} & x_{i2} & \cdots & x_{in} \\ 2x_{i1} & 2x_{i2} & \cdots & 2x_{in} \end{bmatrix}_{3\times n}$.

$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{i1} & x_{i2} & \cdots & x_{in} \\ 2x_{i1} & 2x_{i2} & \cdots & 2x_{in} \end{bmatrix} \begin{bmatrix} 1 & x_{i1} & 2x_{i1} \\ 1 & x_{i2} & 2x_{i2} \\ \vdots & \vdots & \vdots \\ 1 & x_{in} & 2x_{in} \end{bmatrix}$

$X^TX = \begin{bmatrix} 1 & \sum_{i=1}^{n} x_i & 2\sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 & 2\sum_{i=1}^{n} x_i^2 \\ 2\sum_{i=1}^{n} x_i & 2\sum_{i=1}^{n} x_i^2 & 4\sum_{i=1}^{n} x_i^2 \end{bmatrix}$ Row 1 and Row 2 are linearly dependent

Rank $(X^TX) = 1$ (using row echelon form).

The columns of X^TX are linearly dependent. col 1 $\times \sum_{i=1}^{n} x_i =$ col 2 and etc.

As per invertible matrix theorem linearly dependent matrix do not have inverses.

So, X^TX is non-invertible.

6 b) Laplace PDF

$$P(x) = \frac{e^{-|x|/b}}{2b}$$

hypothesis be,

$$y_i = h_\theta(x_i) + \epsilon_i$$

ϵ_i - random noise generated from laplace distribution.

$$f(y_i|x_i;\theta,b) = \frac{1}{2b} e^{\frac{-|y_i - (\theta_0+\theta_1 x_i)|}{b}}$$

(func of y_i given x_i and parameters for laplace distribution)

For, Assuming that the points are independent.

$f(y_1 \dots y_n | x_1 \dots x_n, \theta, b)$. $P[y|x,\theta] = \underset{\theta}{\text{Max}}\ L(\theta)$

MLE $L(\theta) = \prod_{i=1}^{n} f(y_i|x_i,\theta)$

$$\log(L(\theta)) = \sum_{i=1}^{n} \log(f(y_i|x_i,\theta))$$

Scanned by CamScanner

$$\Rightarrow \sum_{i=1}^{n} \log\left(\frac{1}{2b} \cdot e^{-|y_i - (\theta_0 + \theta_1 x_i)|/b}\right)$$

$$\Rightarrow \sum_{i=1}^{n} \left[\log\left(\frac{1}{2b}\right) - \frac{|y_i - (\theta_0 + \theta_1 x_i)|}{b} \right]$$

Minimize all negative terms to maximize the MLE function.

(ie) $$J(\theta) = \sum_{i=1}^{n} \frac{|y_i - (\theta_0 + \theta_1 x_i)|}{b}$$

MAE.

$$\therefore J(\theta) = \frac{1}{b} \sum_{i=1}^{n} |y_i - (\theta_0 + \theta_1 x_i)|$$