classmate

Date

Page

THEORY QUESTIONS

Ans.1 The cost function is given as follows.

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{N} (h_\theta(x_i') - y_i)^2$$

where $x_i' = \begin{bmatrix} 1 \\ x_i \end{bmatrix}$ $\forall$ i = 1 to N.

And,

$$h_\theta(x_i') = \sum_{j=0}^{d} \theta_j x_{ij}' = \theta^T x_i'$$

Here, $x_{i0} = 1$, $x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{id} \end{bmatrix}$

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_d \end{bmatrix}$$

Given data can be written as

$$X = \begin{bmatrix} — (x_1')^T — \\ — (x_2')^T — \\ \vdots \\ — (x_N')^T — \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

$$X\theta - y = \begin{bmatrix} (x_1')^T \theta \\ \vdots \\ (x_n')^T \theta \end{bmatrix} - \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

$$
= \begin{bmatrix} h_\theta(x_1^{1}) - y_1 \\ \vdots \\ h_\theta(x_N^{1}) - y_N \end{bmatrix}
$$

Now, we will use the fact that for a given vector z, we have that $z^T z = \sum_i z_i^2$

$$\frac{1}{2}(X\theta - y)^T(X\theta - y) = \frac{1}{2}\sum_{i=1}^{N}(h_\theta(x_i) - y_i)^2$$

$$= J(\theta)$$

To minimize J, we need its derivative w.r.t. θ.

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

$$= \frac{1}{2}\nabla_\theta\left((X\theta)^T X\theta - (X\theta)^T y - y^T(X\theta) + y^T y\right)$$

$$= \frac{1}{2}\nabla_\theta\left(\theta^T(X^T X)\theta - y^T(X\theta) - y^T(X\theta)\right)$$

$\left[\because a^T b = b^T a \text{ where } a^T b \text{ is a scalar}\right]$

$$= \frac{1}{2}\nabla_\theta\left(\theta^T(X^T X)\theta - 2(X^T y)^T\theta\right)$$

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

$\left[\because \nabla_x x^T A x = 2Ax\right]$ for symmetric matrix A $\left[\because \nabla_x b^T x = b\right]$

$$= X^T X\theta - X^T y$$

To minimize ~~θ~~ J, its derivative w.r.t θ should be 0.

$$X^T X \theta = X^T y$$

$$\theta = (X^T X)^{-1} X^T y$$

θ exists under 'suitable conditions' ~~given~~ mentioned in Ans 2.

Ans.2 $X^T X$ should be inventible

Ans. 3 Closed form solution for a linear regression problem with d independent variables requires us to find inverse of the matrix $X^T X$.

Dimensions of $X^T X$ = $(d+1) \times (d+1)$

The general algorithm that finds inverse has time complexity $O((d+1)^3) \sim O(d^3)$

~~Answer~~ 2

Also, forming the ~~equations~~ matrix $X^T X$ takes $O(d^2 N)$.

Gradient descent performs $O(dn)$ operations in 1 iteration.

We can decide how many iterations we want to perform based on error tolerance and time constraint. Gradient descent takes less time than closed normal form if no. of iterations are small.

When we have less time, we can use gradient descent.

Ans. 4 $(X^T X)\theta = X^T y$

Since, we have simple linear regression, $d = 1$.

$$X = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{12} \\ \vdots & \\ 1 & x_{1N} \end{bmatrix}, \quad X^T = \begin{bmatrix} 1 & - - - - & 1 \\ x_{11} & - - - - - & x_{1N} \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 1 & \sum_{j=1}^{N} x_{1j} \\ \sum_{j=1}^{N} x_{1j} & \sum_{j=1}^{N} (x_{1j})^2 \end{bmatrix}$$

Date
Page

$$\begin{bmatrix} N & \sum_{j=1}^{N} x_{1j} \\ \sum_{j=1}^{N} x_{1j} & \sum_{j=1}^{N} (x_{1j})^2 \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 1 \\ x_{11} & \cdots & x_{1N} \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} \theta_0 N + \theta_1 \sum_{j=1}^{N} x_{1j} \\ \theta_0 \sum_{j=1}^{N} x_{1j} + \theta_1 \sum_{j=1}^{N} (x_{1j})^2 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_{1i} y_i \end{bmatrix}$$

Equating the first rows,

$$\theta_0 N + \theta_1 \sum_{j=1}^{N} x_{1j} = \sum_{i=1}^{N} y_i$$

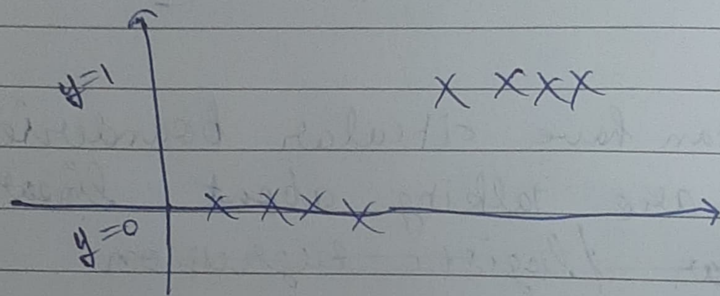
$$\theta_0 + \theta_1 \frac{\sum_{j=1}^{N} x_{1j}}{N} = \frac{\sum_{i=1}^{N} y_i}{N}$$

$$\Rightarrow \theta_0 + \theta_1 \bar{X} = \bar{Y}$$

All the notations have been defined in Ans.1

Ans. 5

We can use linear regression by fixing a threshold.
Consider this example of 8 datapoints, 4 of which



are having label 0 and ~~four~~ ~~other~~ rest 4 are having label 1.

We need to learn θ_0, θ_1 such that

$$y = \theta_0 + x\theta_1.$$

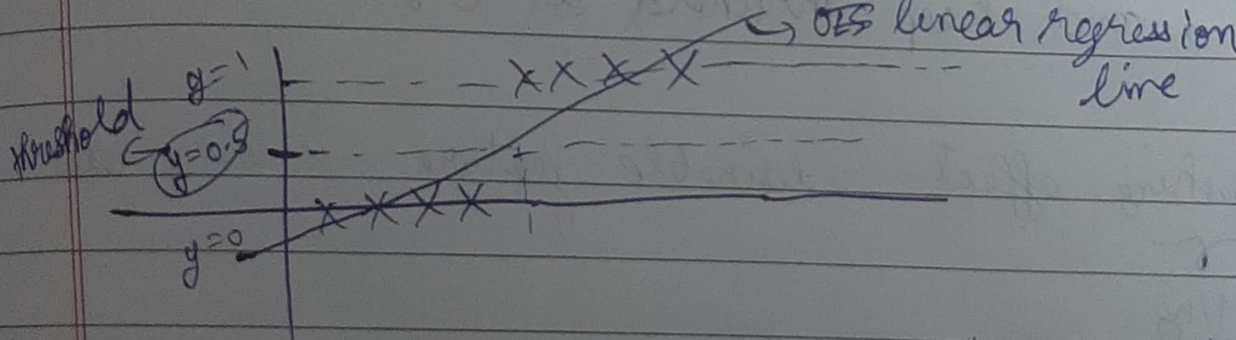
We can set threshold at 0.5.

If $\hat{\theta}_0 + x\hat{\theta}_1 > 0.5$, then we classify the point as 1

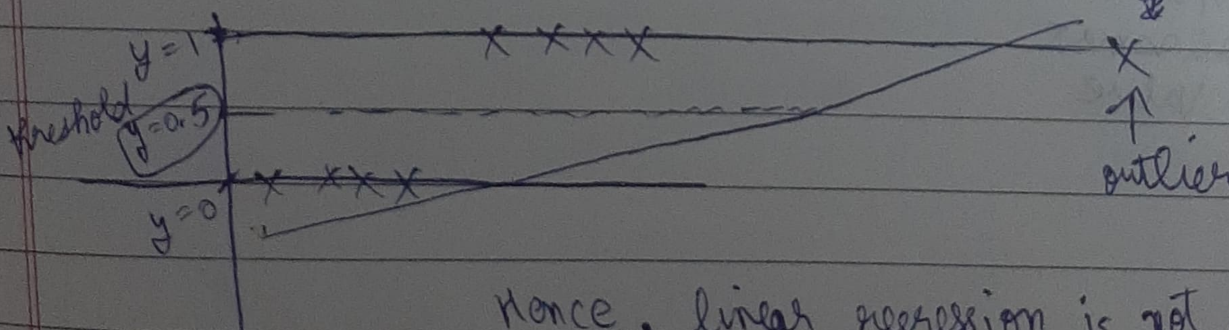
else 0

But it does not perform well when outliers exist.

regression line



When there is an outlier, we classify ~~many~~ points on ~~any~~ y=1 ~~outlier~~ wrongly



Hence, linear regression is not advisable to be used