Aashita Pathwari

~~RS2~~

0048107708

(a)

x_1	x_2	y
1	1	1
1	-1	1
-1	1	1
-1	-1	-1

a valid perception is

$\theta_1 x + \theta_2 x + b$ where - if $\theta_1 x + \theta_2 x + b \geq 0$ predict 1

$\theta = (1,1)$ and $b = 0$

Another valid perception is

$\theta = (1,1)$ and $b = 1$.

b)

x_1	x_2	y
1	-1	1
-1	1	1
1	1	-1
-1	-1	-1

There is no valid solution as the data is not linearly separable.

$h_\theta(x_n) = \sigma(\theta^T x_n)$

$\frac{\partial \sigma(x)}{\partial (x)} = \sigma(x)(1 - \sigma(x))$

2a) $\frac{\partial J}{\partial \theta_j} = -\sum_{n=1}^{N} \frac{y_n}{h_\theta(x_n)} (\sigma(\theta^T x_n))(1 - \sigma(\theta^T x_n)) \frac{\partial(\theta^T x_n)}{\partial \theta_j}$

$+ \frac{(1-y_n)}{(1-h_\theta(x_n))} - \left[(\sigma(\theta^T x_n))(1 - \sigma(\theta^T x_n)) \frac{\partial(\theta^T x_n)}{\partial \theta_j} \right]$

$= -\sum_{n=1}^{N} y_n (1 - \sigma(\theta^T x_n))(x_{n,i}) - (1-y_n)(\sigma(\theta^T x_n))(x_{n,i})$

$= -\sum_{n=1}^{N} x_{n,i} \left(y_n - y_n h_\theta(x_n) - h_\theta(x_n) + y_n h_\theta(x_n) \right)$

$= -\sum_{n=1}^{N} x_{n,i} (y_n - h_\theta(x_n))$

b) $\frac{\partial J}{\partial \theta_j \partial \theta_k} = \frac{\partial \left(-\sum_{n=1}^{N} x_{n,j} (y_n - h_\theta(x_n)) \right)}{\partial \theta_k}$

$= \sum_{n=1}^{N} x_{n,j} (h_\theta(x_n))(1 - h_\theta(x_n))(x_{n,k})$

$= \sum_{n=1}^{N} h_\theta(x_n)(1 - h_\theta(x_n)) x_{n,i} x_{n,k}$

$$H = \begin{bmatrix} \frac{\partial J}{\partial \theta_1 \partial \theta_1} & \frac{\partial J}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial J}{\partial \theta_1 \partial \theta_D} \\ \vdots & & & \vdots \\ \frac{\partial J}{\partial \theta_D \partial \theta_1} & \cdots & \cdots & \frac{\partial J}{\partial \theta_D \partial \theta_D} \end{bmatrix}$$

$= \sum_{n=1}^{N} h_\theta(x_n)(1 - h_\theta(x_n)) \vec{X}_n \vec{X}_n^T$

$z^T H z$

$= z^T X D X^T z$

$= (z^T X) D (z^T X)^T$

$= \|z^T D X\|^2 \geq 0$

since $D > 0$ and

$\|z^T X\| \geq 0.$

$\therefore$ H is PSD.

$X D X^T$

$= \sum_{n=1}^{N} h_\theta(x_n)(1 - h_\theta(x_n)) \vec{x}_n \vec{x}_n^T$

where D is a digonal matrix

where $D_{n,n} = h_\theta(x_n)(1 - h_\theta(x_n))$

and X is a $d \times n$ matrix

such that $\sum_{i=1}^{n} x_i x_i^T = X X^T \quad (x_i \in \mathbb{R}^n)$

$\star \sum_{i=1}^{n} x_i x_i^T = X X^T$

3) a) $\frac{\partial J}{\partial \theta_0} = \sum_{n=1}^{N} w_n 2(\theta_0 + \theta_1 x_{n,1} - y_n)$

$\frac{\partial J}{\partial \theta_1} = \sum_{n=1}^{N} w_n 2(\theta_0 + \theta_1 x_{n,1} - y_n) x_{n,1}$

b) $\frac{\partial J}{\partial \theta_0} = 0 \Rightarrow \sum_{n=1}^{N} w_n(\theta_0 + \theta_1 x_{n,1} - y_n) = 0$

$\frac{\partial J}{\partial \theta_1} = 0 \Rightarrow \sum_{n=1}^{N} w_n(\theta_0 + \theta_1 x_{n,1} - y_n) x_{n,1} = 0$

$\theta_0 \sum_{n=1}^{N} w_n + \theta_1 \sum_{n=1}^{N} x_{n,1} w_n = \sum_{n=1}^{N} w_n y_n \Rightarrow \theta_0 = \frac{\sum_{n=1}^{N} w_n y_n - \theta_1 \sum_{n=1}^{N} x_{n,1} w_n}{\sum_{n=1}^{N} w_n}$

$\theta_0 \sum x_{n,1} w_n + \theta_1 \sum_{n=1}^{N} (x_{n,1})^2 w_n = \sum_{n=1}^{N} y_n w_n x_{n,1}$

$\frac{\sum w_n y_n - \theta_1 x_{n,1} w_n}{\sum w_n} \cdot \sum x_{n,1} w_n + \theta_1 \sum (x_{n,1})^2 w_n = \sum y_n w_n x_{n,1}$

$$\left(\sum w_n y_n - \theta_1 \sum x_{n,1} w_n\right) \cdot \sum x_{n,1} w_n + \theta_1 \sum (x_{n,1})^2 w_n \cdot \sum w_n = \sum y_n w_n x_{n,1} \cdot \sum w_n$$

$$\theta_1 \left(\sum (x_{n,1})^2 w_n \cdot \sum w_n - \left(\sum x_{n,1} w_n\right)^2 \right) + \sum w_n y_n \cdot \sum x_{n,1} w_n = \sum y_n x_{n,1} w_n \cdot \sum w_n$$

$$\theta_1 = \frac{\sum y_n x_{n,1} w_n \cdot \sum w_n - \sum w_n y_n \cdot \sum x_{n,1} w_n}{\sum (x_{n,1})^2 w_n \cdot \sum w_n - \left(\sum x_{n,1} w_n\right)^2}$$

($\sum$ and $\sum_{n=1}^{N}$ used interchangeably here)

$$\theta_0 = \frac{\sum_{n=1}^{N} w_n y_n - \theta_1 \sum_{n=1}^{N} x_{n,1} w_n}{\sum_{n=1}^{N} w_n}$$

where θ_1 is as stated.

c) $d(\theta)$

$$= \sum_{n=1}^{N} w_n (\theta_0 + \theta_1 x_{n,1} - y_n)^2$$

$$= \sum_{n=1}^{N} w_n (\vec{x}_n \vec{\theta} - y_n)^2$$

$$= \sum_{n=1}^{N} w_n \left(\vec{x}_n \vec{\theta} - y_n\right)^T \left(\vec{x}_n \vec{\theta} - y_n\right)$$

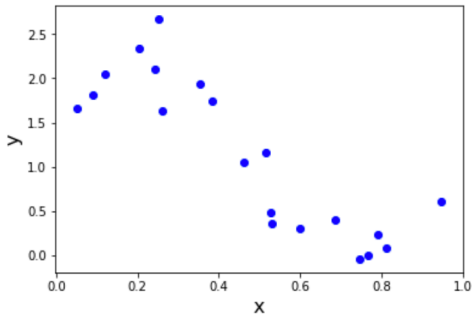
$$= (X\theta - y)^T W (X\theta - y)$$ where W is a $n \times n$ diagonal matrix where $W_{n,n} = w_n$

$\theta_0 + \theta_1 x_{n,1} = \vec{x}_n \vec{\theta}$ where $\vec{x}_n = (1 \quad x_{n,1})$

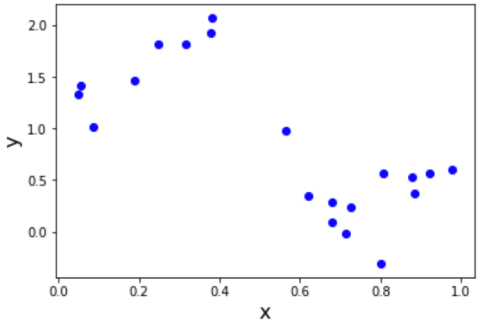
and $\theta = \begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix}$

4a) From the training and testing data it seems like there is a negative relationship between x and y and hence a linear regression model might be effective in predicting the data.

training data:



test data:



d)

Step size	Number of iterations	Value of coefficients	Time taken to converge	Model cost
0.0001	10000	[2.27044798 -2.46064834]	0.18908	4.086
0.001	7021	[2.4464068 -2.816353]	0.194689	3.91257
0.01	765	[2.44640703 -2.81635347]	0.02116	3.91257
0.0407	10000	[-9.40470931e+18 -4.65229095e+18]	0.167212	2.710916

As the step size increases from 0.0001 to 0.01, time taken to converge decreases. Model cost and number of iterations decrease too, while the value of the coefficients increases. However, when the step size is 0.0407, time taken to converge and number of iterations increases and the value of the coefficient is significantly different from the other 3 values. This suggests that the step size is too large.

e) The closed form solution is [2.44640709 -2.81635359]. The time taken is 0.0004570484 and the model cost is 3.91257. The algorithm runs 50 times faster than the fastest step size in the previous table. The coefficients and cost is the same as that obtained by GD for step size = 0.01 and step size = 0.001

f) It takes 2094 iterations and the time taken is 0.04934501647949219

h) RMSE averages the error over the amount of training data making it more of an accurate measure. Taking square root also ensures that the error scales more linearly.

i) m =5 best fits the data. Yes, when the model underfits (for lower m values), both training and test errors are relatively high as seen from the plot. When the model overfits (for values above 6), training error decreases but there is a steep increase in test error.