San Diego State University **CS549 Machine Learning**

Assignment -3

Due: October 15 2021

- Please type the solutions using a word processor such as MS Word, Latex, or write by hand neatly and upload the scanned copy of it.
- **I, Andrick (sign your name here), guarantee that this homework is my independent work and I have never copied any part from other resources. Also, I acknowledge and agree with the plagiarism penalty specified in the course syllabus.**
- Turn in your assignment before the deadline. Penalty will be applied to late submission.

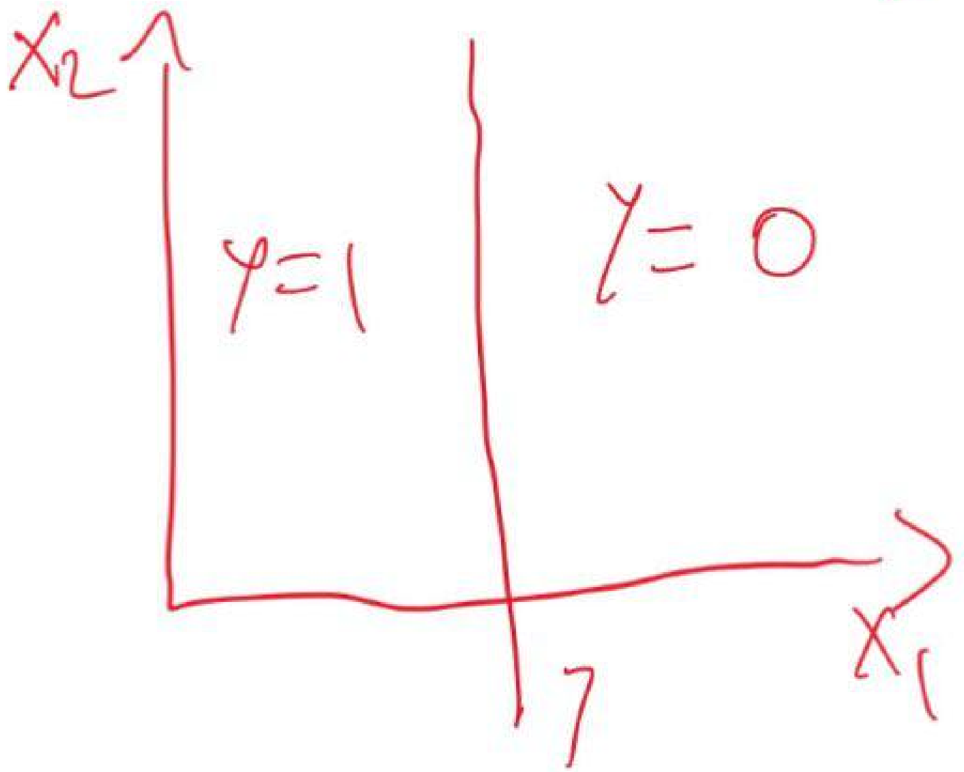
a_e

1. $P(y=0 \mid x;\theta) = 1 - .19 = .81$

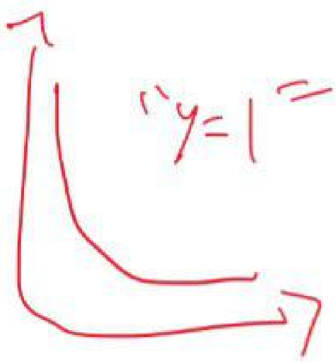
2. $P(y=1 \mid x;\theta) = .19$

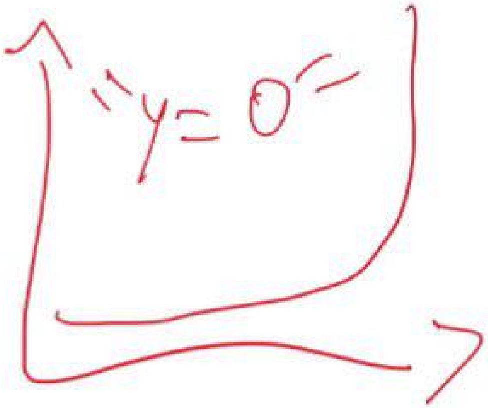
because in logistic regression we have numbers ranging from 0 to 1, and we get that the probability of yes is .19

b. $\theta_0 = 7,\ \theta_1 = -1,\ \theta_2 = 0$



C.





The cost function for logistic regression is

$$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$$

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

$$\frac{d(\sigma(x))}{dx} = \frac{(e^{-x})}{(1+e^{-x})^2} = \frac{1+e^{-x}}{(1+e^{-x})^2}$$

$$= \frac{1}{(1+e^{-x})^2} = \frac{1}{1+e^{-x}} \cdot \left(1 - \frac{1}{1+e^{-x}}\right)$$

$$= \sigma(x)(1-\sigma(x))$$

hence by the derivative of the sigmoid function

$$\partial(J(\theta)) = -\frac{1}{M} \cdot \sum_{i=1}^{M} \left(y^{(i)} \cdot \frac{1}{h_\theta(x^{(i)})} \cdot \frac{\partial(h_\theta(x^{(i)}))}{\partial(\theta j)} \right.$$

$$\left. + \sum_{i=1}^{M} \left((1-y^{i}) \cdot \frac{1}{(1-h_\theta(x^{i})} \cdot \frac{\partial(1-h_\theta(x^{i}))}{\partial(\theta j)} \right) \right)$$

$$= -\frac{1}{M} \cdot \left(\sum_{i=1}^{M} \left(y^{i} \cdot \frac{1}{h_\theta(x^{i})} \cdot G(z)(1-G(z)) \cdot \frac{\partial(\theta^T x)}{\partial(\theta j)} \right. \right.$$

$$+ \sum_{i=1}^{M} \left((1-y^{i}) \cdot \frac{1}{(1-h_\theta(x^{i})} \cdot (-G(z)(1-G(z)) \cdot \frac{\partial(\theta^T x)}{\partial(\theta j)} \right.$$

$$= -\frac{1}{M} \cdot \left(\sum_{i=1}^{M} \left(y^{i} \cdot \frac{1}{h_\theta(x^{i})} \cdot G(z)(1-G(z)) \cdot \frac{\partial(\theta^T x)}{\partial(\theta j)} \right. \right.$$

$$+ \sum_{i=0}^{M} \left((1-y^{i}) \cdot \frac{1}{1-h_\theta(x^{i})} \cdot (-G(z)(1-G(z)) \cdot \frac{\partial(\theta^T x)}{\partial(\theta j)} \right)$$

$$= -\frac{1}{M} \cdot \left(\sum_{i=0}^{M} \left(y^{i} \cdot \frac{1}{h_\theta(x^{i})} \cdot h_\theta(x^{i}) \cdot (1-h_\theta(x^{i})) \cdot x_j^{i} \right) \right.$$

$$+ \sum_{i=1}^{M} \left((1-y^{i}) \cdot \frac{1}{1-h_\theta(x^{i})} \cdot (-h_\theta(x^{i})(1-h_\theta(x^{i})) \cdot x_j^{i} \right)$$

$$= -\frac{1}{m} \cdot \left(\sum_{i=1}^{m} y^i \cdot (1 - h_\theta(x^i)) \cdot x_j^i - (1 - y^i) \cdot h_\theta(x^i) \cdot x_j^i \right)$$

$$= -\frac{1}{m} \cdot \left(\sum_{i=1}^{m} \left(y^i - y^i \cdot h_\theta(x^i) - h_\theta(x^i) + y^i \cdot h_\theta(x^i) \right) \cdot x_j^i \right)$$

$$= \frac{-1}{m} \cdot \left(\sum_{i=1}^{m} \left(y^i - h_\theta(x^i) \right) \cdot x_j^i \right)$$

hence, $\frac{\partial (J(\theta))}{\partial (\theta)} = \frac{1}{m} X^T [h_\theta(X) - Y]$

* using matrix form instead of sum

d. Gradient descent is an algorithm that finds the optimal graph based on the optimal values of coefficients that reduce the cost as much as possible.

first lets called the minimum function $J(\theta)$

Next, $\theta_j \leftarrow \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$

$$\frac{\partial}{\partial \theta} J(\theta) = \frac{1}{m} \sum_{i=1}^{m} (h\theta(x^i) - y^i) x_j^i$$

$$\theta_j \leftarrow \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h\theta(x^i) - y^i) x_j^i$$

c). • α is a hyper parameter that is used to mamase the rate of updates the algorithm to learn new values of the parameter.

• Changing the value can alter the learning speed and or decide wheter the cost function is minimized or not.

ee the smaller the α the slower the learning rate of the algorithm, and gradient descent can be slow, while if α is large it may fail to converge and or diverge, due to the learning rate being fast.