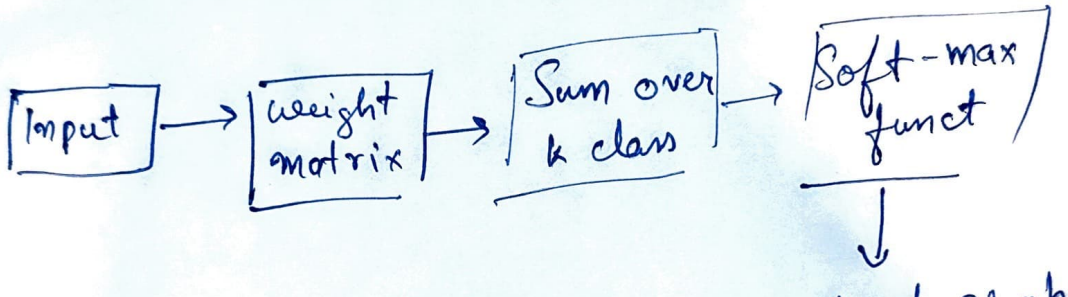
HW-2 Aman Gupta

Ques 2

$$\hat{p}_k = \delta(s_k(x))_k = \frac{\exp(s_k(x))}{\sum_{j=1}^{K} \exp(s_j(x))}$$

(a) $s_k(x) = x^T(\theta)^{(k)}$

we need $\theta^{(k)}$ parameters. All these are stored in parameter matrix.



The input vect is fed into the weight matrix and the bais vector is added to the weighted sum. the result is then passed through the Softmax function, which produces the output probability distribution over the K class.

(b) $$J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} \sum_{k=1}^{K} y_k^{(i)} \log(\hat{p}_k^{(i)})$$

we have to derive gradient of $J(\theta)$

$$\partial(J(\theta)) \Rightarrow -\frac{1}{m} \sum_{i=1}^{m} x_i \cdot \frac{1(e^{\theta_j^T, x^i})}{\sum_{j=1} e^{(\theta_j - x_j)}}$$

$$= \frac{1}{m} \sum_{i=1}^{m} (\hat{P}(x)^{(i)} - y_k^{(i)}) x^{(i)})$$

$$-2 X^T y + 2 (X^T X \lambda I) w = 0$$

$$-2 X^T y = -2 (X^T X \lambda I) w$$

$$X^T y = (X^T X \lambda I) w$$

$$(X^T X + \lambda) w = X^T y$$

multypiy both sides by

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

$$(X^T X + \lambda))^{-1} (X^T X + \lambda I) w$$

$$\Rightarrow (X^T X + \lambda I)^{-1} X^T y$$

$$-2 X^T y + 2 (X^T X \lambda I) w = 0$$

$$X^T y (X^T X \lambda I) w$$

now multiply by $(X^T X + \lambda I)^{-1}$

$$(X^T X + \lambda I)^{-1} (X^T X - \lambda I) w$$

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

$$w = (X^T X + \lambda I)^{-1} X^T y$$

which is ridge regression.