1. Ridge Regression

$$E(w) = \sum_{i=1}^{m} (w^T \cdot x^{(i)} - y^{(i)})^2 + \lambda \sum_{i=1}^{m} w_i^2$$

Closed-form solution: $w = (\lambda I + \Phi^T \Phi)^{-1} \Phi^T t$

$$E(w) = MSE(w) + \frac{\lambda}{2} \sum_{i=1}^{m} w_i^2$$

$$= \frac{1}{m} \sum_{i=1}^{m} (h_w(x^i) - y_i)^2 + \frac{\lambda}{2} \sum_{i=1}^{m} w_i^2$$

$$E(w) = \frac{1}{m} (xw - y)^T (yw - y) + \frac{\lambda}{2} w^T w$$

$$\Rightarrow ((xw)^T - y^T)(yw - y) + \lambda w^T w$$

$$= (xw)^T (yw) - (yw)^T (y)$$
$$- (yw)^T (y) + y^T y + \lambda w^T w$$

$$= w^T x^T \quad xw - 2 (xw)^T y + y^T y + \lambda w^T w$$

For minimizing of w

$$\frac{\partial \hat{w}}{\partial w} = 0$$

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

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

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

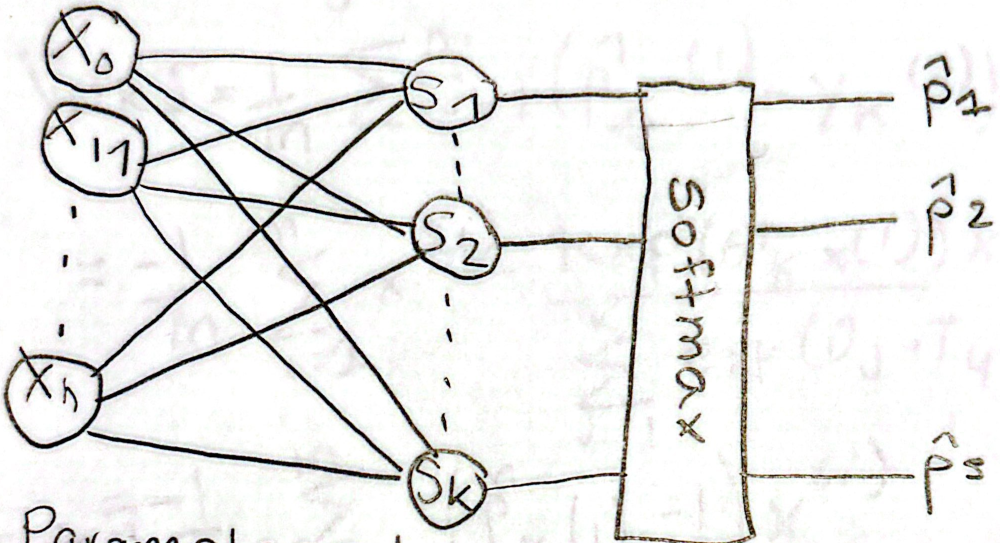
$$(X^T X + \lambda I) w = X^T y$$

Now Multipling Both sides with $(X^T X + \lambda I)^{-1}$

2.

1. $s_k(x) = \theta_k^T \cdot x$ class $1 \leq k \leq K$

$\hat{p}_k = \sigma(s_k(x))_k = \dfrac{\exp(s_k(x))}{\sum_{j=1}^{K}(s_j(x))}$ where θ_k is the vector param. of input feat for s_k.



Parameterestimation is $n+1$ and the parameters are $\theta_1, \theta_2 \ldots \theta_n$.

2. m training samples $\{(x_i, y_i)\}_{i=1,2..m}$
Derive the gradient of $J(\theta)$ regarding θ_k.

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

where $y_k^{(i)} = 1$ if the i^{th} instance belongs to k.

$$= \frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{K} y_k^{(i)} \log\left(\frac{\exp(s_k(x))}{\sum_{j=1}^{K}(s_j(x))}\right)$$

$$= \frac{1}{m} \sum_{i=1}^{m} \left(\sum_{k=1}^{K} y_k^{(i)} \log(\exp(s_k(x^{(i)})) - \right.$$

$$\sum_{k=1}^{K} y_k^{(i)} \log\left(\sum_{j=1}^{K} \exp(s_j(x^{(i)})\right))$$

$$y_k^{(i)} = 1$$

Softmax Regression Entropy

$$\nabla_{\theta_k} J = \frac{1}{m} \sum_{i=1}^{m} (\hat{p}_k^{(i)} - y_k^{(i)}) x^{(i)}$$

$$= \frac{-1}{m} \sum_{i=1}^{m} x^{(i)} - \frac{1 \exp(\theta_k^T x^{(i)}) x^{(i)}}{\sum_{j=1}^{K} \exp(\theta_j^T y^{(i)})}$$

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

$$= \frac{1}{m} \sum_{i=1}^{m} (\hat{p}_k(i) - y_k^{(i)}) x^{(i)} = \boxed{\nabla_{\theta_k} J(\theta)}$$