1) Prove that the closed form solution for Ridge regression:

$w = (\Lambda I + X^T X)^{-1} \cdot X^T \cdot Y$ $\quad\quad E(w) = \sum_{i=1}^{n} (w^T \cdot x^i - y^i)^2 + \lambda \sum_{i=1}^{n} w^2$

Assuming we already know the regression solution without the regularization term

$\beta = (X^T X)^{-1} X^T Y$ then prove that adding the L2 term gives us w.

ie objective function $L(\beta) = (Y - X\beta)^T \cdot (Y - X\beta)$

$L2 = \Lambda \|\beta\|_2^2$ so adding this $\rightarrow (Y - X\beta)^T (Y - X\beta) + \underline{\Lambda \beta^T \beta}$

Derive it w.r.t β

$$\rightarrow \frac{d}{d\beta} [\Lambda \beta^T \beta] = 2\lambda\beta = \Lambda I \beta$$

$$= [X^T X + \Lambda I]\beta = X^T Y$$

Putting this into original equation we get

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

Hence closed form solution for ridge regression is proved

(2) $P_K = \sigma(S_K(x))_K = \frac{\exp(S_K(x))}{\sum_{j=1}^{K} \exp(S_j(x))}$

1) We have $S_n(x) = \theta_n^T x$ ' x is a n-dimension vector this means all θ_n is an n-dimension vector. θ_n are vector parameter that are used to transfer the input feature to x to $Z_x(J_K)$. ~~which is used to get~~ Which the softmax is applied to get a pired prediction.

2) $$J(\theta) = -\frac{1}{m} \sum_{i=k}^{n} \sum_{k=1}^{K} y_K^{(i)} \log(p_K^{(i)})$$

$$P(K) = \sigma(S(x))_K = \frac{\exp(S_x(x))}{\sum_{j=1}^{K} \exp(S_j(x))}$$

where $S_K(x) = \theta_K^T x$)

for minimizing the cost function

$$J(\theta) = -1/m \sum_{K=1}^{K} y_K^{(i)} \log(\hat{p}_K^{(i)})$$

$$= -1/m \sum_{i=K}^{n} \sum_{K=1}^{K} y_K^{(i)} \log\left(\frac{\exp\ K_1(x)}{\sum_{j=1}^{K} \exp(S_j(x^{(i)}))}\right)$$

$$= -1/m \sum_{i=1}^{n} \{ (\sum_{K=1}^{K} y_K^{(i)} \log(\exp(S_K(x^{(i)}))) \}$$

$$\sum_{K=1}^{K} y_K^{(i)} \log(\sum_{j=1}^{K} \exp(S_j \cdot x^{(i)})))$$

$\llcorner$ ①

Now plug derivative of softmax in $\nabla J(\theta)$

$$\nabla J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left[\sum_{k=1}^{K} y_i\left(1\{k=j\} - P_i\right)\right]x^{(i)}$$

$$= \frac{-1}{m}\sum_{i=1}^{m}\left[y_k - \sum_{k=1}^{K} y_k P_i\right]x^{(i)}$$

$$= \frac{1}{m}\sum_{i=1}^{m}\left[y_k - P_i(1)\right]x^{(i)}$$

$$= \frac{1}{m}\sum_{i=1}^{m}\left[P_k^{(i)} - y_k^{(i)}\right]\cdot x^{(i)}$$