1. According to the Product Rule $p(X,Y) = p(Y|X)p(X)$,

$$p(X,Y) = p(X|Y)\,p(Y)$$

thus, $p(X|Y)p(Y) = p(Y|X)p(X)$

$$p(Y|X) = \frac{P(X|Y)P(Y)}{p(X)}$$

And according to the Sum Rule: $p(X) = \sum_Y P(X,Y)$

$$p(X) = \sum_Y p(X|Y)p(Y)$$

Thus, the Bayes' Theorem is proved.

Most of decisions in life are made with incomplete information, and we have only limited information to training our model. Bayes Theorem can predict probabilities based on past data by converting posterior probability to likelihood and prior probability.

2. $MSE(w) = \frac{1}{m}\sum_{i=1}^{m}(w^T \cdot x^{(i)} - y^{(i)})^2$

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

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

in the linear regression, we let

$$\frac{\partial[(w^T\cdot x - y)^2]}{\partial w} = 0$$

$$\frac{\partial[(w^T\cdot x - y)^T(w^T\cdot x - y)]}{\partial w} = 0 \quad \text{so, } w^T X^T X - 2y^T X = 0$$

thus, $\hat{w} = (X^T\cdot X)^{-1}\cdot X^T\cdot y$

thus, in the ridge regression we let

$$\frac{\partial\,[(w^T\cdot x - y)(w^T\cdot x - y) + \frac{\lambda}{2}w^2]}{\partial w} = 0$$

so, $w^T x^T X - 2y^T X + \lambda I w = 0$

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

3\. 1) We need to ~~parameter~~ estimate one parameter, it is θ.

2) $$P_k = \frac{\exp(S_k(x))}{\sum_{j=1}^{k} \exp(S_j(x))} \qquad S_k(x) = \theta^T \cdot x$$

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{k} y_k^{(i)} \log(P_k^{(i)})$$

$$\Delta_{\theta_k} J(\theta) = \frac{\partial J(\theta)}{\partial P_k} \cdot \frac{\partial P_k}{\partial S_k} \cdot \frac{\partial S_k}{\partial \theta_k}$$

$$\frac{\partial J(\theta)}{\partial P_k} = -\frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{k} y_k^{(i)} \frac{1}{P_k^{(i)}}$$

$$\frac{\partial P_k}{\partial S_k} = \frac{\partial \left[\frac{\exp(S_k(x))}{\sum_{j=1}^{k}\exp(S_j(x))}\right]}{\partial S_k} = \frac{\exp(S_k(x)) \cdot \sum_{j=1}^{k}\exp(S_j(x)) - \exp^2(S_k(x))}{\left[\sum_{j=1}^{k}\exp(S_j(x))\right]^2}$$

$$= \frac{\exp(S_k(x))}{\sum_{j=1}^{k}\exp(S_j(x))} \cdot \left(1 - \frac{\exp(S_k(x))}{\sum_{j=1}^{k}\exp(S_j(x))}\right)$$

$$= P_k \cdot (1 - P_k)$$

$$\frac{\partial S_k}{\partial \theta_k} = x^{(i)}$$

$$\therefore \Delta_{\theta_k} J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{k} y_k^{(i)} \frac{1}{P_k^{(i)}} \left[P_k^{(i)} \cdot (1 - P_k^{(i)})\right] \cdot x^{(i)}$$

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

$\because y_k^{(i)} = 1$ if the ith instance belongs to class k,

$$\therefore \Delta_{\theta_k} J(\theta) = \frac{1}{m}\sum_{i=1}^{m} (P_k^{(i)} - y_k^{(i)}) \cdot x^{(i)}$$