1) Bayes eq.: $P(A|B) = \dfrac{P(B|A) \cdot P(A)}{P(B)}$

Proof: $P(A \cap B) = P(A) \cdot P(B|A)$

$P(B \cap A) = P(B) \cdot P(A|B)$

$P(A \cap B) = P(B \cap A) \Rightarrow P(A) \cdot P(B|A) = P(B) \cdot P(A|B)$

$P(A) \cdot P(B|A) = P(B) \cdot P(A|B)$

Solve for $P(A|B)$: $P(A|B) = \dfrac{P(A) \cdot P(B|A)}{P(B)}$

Solve for $P(B|A)$: $P(B|A) = \dfrac{P(B) \cdot P(A|B)}{P(A)}$

There are many reasons why Bayes Theorem is and can be useful in machine learning applications. One way it could be helpful is with handling uncertainty. With Bayes theorem, we can meticulously update our hypotheses with the introduction of new information. An example of this would be as follows: Say you predict there to be a 20% of snow tomorrow. You then see that it is actually 50%. You can use the new evidence and your prior claim to update your hypothesis.

There are many other applications where Bayes Theorem is useful. Some examples could be classification tasks using Naive Bayes Classifier, updating models when new info arrives, Bayesian networks, and estimating parameters of a model.

2)

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

$$= E(w) = (Xw-y)^T (Xw-y) + \lambda w^T w$$

$$\Rightarrow (Xw-y)^T (Xw-y)$$

$$= (X^T w^T - y^T)(Xw-y) = X^T w^T X w - X^T w^T y - y^T X w + y^T y$$

$$= w^T X^T X w - 2y^T X w + y^T y$$

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

$$= w^T X^T X w + \lambda w^T w - 2y^T X w + y^T y$$

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

$$\frac{\partial E(w)}{\partial w} = 2(X^T X + \lambda I) w - 2X^T y = 0$$

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

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

3) 1) θ_k is a vector of weights corresponding to input vector x with n dimensions, $\theta \in R^n$

Therefore, for K classes, we have K weight vectors, each of n dimensions. Hence the total number of parameters to estimate is $K \times n$.

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

$$\hat{p}_k^{(i)} = \frac{\exp(\theta_k^T x_i)}{\sum_{j=1}^{K} \exp(\theta_j^T x_i)}$$

$$\frac{\partial J(\theta)}{\partial \theta_k} = \frac{-1}{m} \sum_{1}^{m} \left[y_k^{(i)} \frac{\partial}{\partial \theta_k} \log\left(\hat{p}_k^{(i)}\right) \right]$$

$$\log\left(\hat{p}_k^{(i)}\right) = \frac{\log(\exp(\theta_k^T x_i))}{\sum_{j=1}^{K} \exp(\theta_j^T x_i)} = \theta_k^T x_i - \log\left(\sum_{j=1}^{K} \exp(\theta_j^T x_i)\right)$$

$$\frac{\partial}{\partial \theta_k} \log\left(\hat{p}_k^{(i)}\right) = x_i - \frac{\sum_{j=1}^{K} \exp\left(\theta_j^T x_i\right) x_i}{\sum_{j=1}^{K} \exp\left(\theta_j^T x_i\right)} = x_i - \sum_{j=1}^{K} \hat{p}_j^{(i)} x_i$$

$$\hookrightarrow \frac{\partial J(\theta)}{\partial \theta_k} = -\frac{1}{m} \sum_{i=1}^{m} \left[y_k^{(i)} x_i - y_k^{(i)} \sum_{j=1}^{K} \hat{p}_j^{(i)} x_i \right]$$

$$\frac{\partial J(\theta)}{\partial \theta_k} = -\frac{1}{m} \sum_{i=1}^{m} \left[y_k^{(i)} x_i - y_k^{(i)} \sum_{j=1}^{K} \hat{p}_j^{(i)} x_i \right]$$

$$\hookrightarrow \frac{\partial J(\theta)}{\partial \theta_k} = -\frac{1}{m} \sum_{i=1}^{m} \left[y_k^{(i)} x_i - \hat{p}_k^{(i)} x_i \right]$$

$$\hookrightarrow \frac{\partial J(\theta)}{\partial \theta_k} = \boxed{-\frac{1}{m} \sum_{i=1}^{m} x_i \left[y_k^{(i)} - \hat{p}_k^{(i)} \right]}$$