Question 1:

We start with

P(A|B) = $\frac{P(A)\,P(B|A)}{P(B)}$

P(B) P(A|B) = P(A and B) = P(A) P(B|A)

We can make the two equal to each other:

P(B) P(A|B) = P(A) P(B|A)

thus

P(A|B) = $\frac{P(A)\,P(B|A)}{P(B)}$

Bayes theorem is important for machine learning because it helps us make predictions and decisions based on available data while considering uncertainty. It allows us to start with an initial guess and update our prediction as we get new data. You can update the prediction using Bayes theorem. You can essentially use bayes theorem to build models that learn from data and improve predictions over time. These models can not only make predictions but also tell you how uncertain those predictions are.

Question 2:

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

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

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

We take the derivative with respect to w:

$\frac{d}{dw}$ (($w^T X^T Xw$ - $y^T Xw$ - $w^T X^T$y - y^Ty) +λw^Tw

The derivative of $w^T X^T Xw$ with respect to w = $2X^T Xw$
The derivative of - $y^T Xw$ with respect to w is - $X^T y$
The derivative of - $w^T X^T y$ *with respect to w is* - $X^T y$
The derivative of - y^Ty with respect to w is 0 because it doesn't depend on w
The derivative of λw^Tw is $2\lambda w$

Summing all that up, we get

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

After combining like terms, we get:

$2X^T Xw$ - $2X^T y$ + $2\lambda w$ = 0

Now we can factor out 2

2 ($X^T Xw$ - $X^T y$ + λw) = 0

Now we can divide both sides by 2

To get

$X^T Xw$ - $X^T y$ + λw = 0

Moving terms involving w to the left

$X^T Xw + \lambda w - X^T y = -0$

We can factor out w

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

We can multiply both sides by the inverse of the matrix to get:

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

Question 3:

We are trying to derive the gradient of $J\Theta$ with respect to θk.

Using the multivariable chain rule we can deduce that

$$\frac{\partial J\Theta}{\partial \theta k} = \frac{\partial J\Theta}{\partial pki} * \frac{\partial \hat{pki}}{\partial sk(xi)} * \frac{\partial sk(xi)}{\partial \theta k}$$

$$\frac{\partial J\Theta}{\partial \theta} = \sum_{i=1}^{m} (\hat{pk}(i) - yk(i))\, x(i)$$

First we look at $\frac{\partial J\Theta}{\partial \hat{pki}}$

$$\frac{\partial J\Theta}{\partial pki} = \frac{\partial}{\partial \hat{pki}} - \frac{1}{m} \sum_{i=1}^{m} \sum_{k=1}^{k} yk(1) \log(\hat{pk}(i))$$

$$= -\frac{1}{m} \sum_{i=1}^{m} \sum_{k=1}^{k} \frac{yki}{\hat{pki}}$$

Next we look at $\frac{\partial \hat{pki}}{\partial sk(xi)}$

Here, we're calculating the derivative of the softmax with respect to its inputs where m is the number of inputs.

$$\frac{\partial \hat{pki}}{\partial sk(xm)} = \frac{\partial}{\partial sk(xm)} \frac{exp(sk(xi))}{\sum_{j=1}^{k} exp(sj(x))}$$

We can use the quotient rule which states that:

$$\frac{dy}{dx} = \frac{v * \frac{du}{dx} - u * \frac{dv}{dx}}{v^2}$$

f(x) = $exp(sk(xi))$ and g(x) = $\sum_{j=1}^{k} exp(sj(x))$

$$\frac{\sum_{j=1}^{k} exp(sj(x)) * \frac{\partial}{\partial sk} exp(sk(xi)) - exp(sk(xi)) * \frac{\partial}{\partial sk} \sum_{j=1}^{k} exp(sj(x))}{(\sum_{j=1}^{k} exp(sj(x)))^2}$$

The derivative of

$$\sum_{j=1}^{k} exp(sj(x)) = exp(sk(xm))$$

$$= \frac{\sum_{j=1}^{k} exp(sj(x)) * exp(sk(xi)) - exp(sk(xi)) * exp(sj(xm))}{(\sum_{j=1}^{k} exp(sj(x)))^2}$$

* In the above, we can get rid of the sum because all other terms other than sk are 0s. Only sk is 1.

The result can be factored

So we get

$$= \frac{exp(sk(xi))(\sum_{j=1}^{k} exp(sj(x)) * -exp(sk(xm)))}{(\sum_{j=1}^{k} exp(sj(x)))^2}$$

$$= \frac{exp(sk(xi))}{\sum_{j=1}^{k} exp(sj(x))} * \frac{\sum_{j=1}^{k} exp(sj(x) - exp(sk(xm))}{\sum_{j=1}^{k} exp(sj(x))}$$

$$= \frac{exp(sk(xi))}{\sum_{j=1}^{k} exp(sj(x))} * (1 - \frac{exp(sk(xm))}{\sum_{j=1}^{k} exp(sj(x))})$$

$= \hat{p}ki * (1 - \hat{p}km)$

Since i = m

The above equals $\hat{p}ki * (1 - \hat{p}ki)$

When i is not equal to m:

$$\frac{\partial \hat{p}ki}{\partial sk(xm)} = \frac{\partial}{\partial sk(xm)} \frac{exp(sk(xi))}{\sum_{j=1}^{k} exp(sj(x))}$$

$$= \frac{\sum_{j=1}^{k} exp(sj(x)) * \frac{\partial}{\partial sk} exp(sk(xi)) - exp(sk(xi)) * \frac{\partial}{\partial sk} \sum_{j=1}^{k} exp(sj(x))}{(\sum_{j=1}^{k} exp(sj(x)))^2}$$

$\frac{\partial}{\partial sk(xi)} exp(sk(xi))$ = 0 now since it doesn't depend on sj(xm)

$$= \frac{\sum_{j=1}^{k} exp(sj(x)) * \frac{\partial}{\partial sk} exp(sk(xi)) - exp(sk(xi)) * \frac{\partial}{\partial sk} \sum_{j=1}^{k} exp(sj(x))}{(\sum_{j=1}^{k} exp(sj(x)))^2}$$

$$= \frac{0 - exp(sk(xi)) * \frac{\partial}{\partial sk} \sum_{j=1}^{k} exp(sj(x))}{(\sum_{j=1}^{k} exp(sj(x)))^2}$$

$$\frac{\partial}{\partial sk} \sum_{j=1}^{k} exp(sj(x)) = exp(sk(xm))$$

So

$$= \frac{- exp(sk(xi))}{\sum_{j=1}^{k} exp(sj(x))} * \frac{exp(sk(xm))}{\sum_{j=1}^{k} exp(sj(x))}$$

= - $\hat{p}$ki * $\hat{p}$km

Lastly

$$\frac{\partial}{\partial \theta k^T x^i} = \theta k^T x^i = x^i$$

If we put it all together, we get:

$$\frac{\partial J\Theta}{\partial \theta k} = \frac{1}{m} \sum_{i=1}^{m} -\frac{yki}{\hat{p}ki} * \frac{\partial \hat{p}ki}{\partial sk(xm)}$$

$$= \frac{1}{m} \sum_{i=1}^{m} (- \frac{yki}{pki} * \hat{p}ki (1 - \hat{p}ki) + \sum_{m != i} - ykm * \frac{1}{\hat{p}km} (- \hat{p}km * \hat{p}ki))$$

$= \frac{1}{m} \sum_{i=1}^{m}$ (-yki (1 - $\hat{p}$ki) + $\sum_{m\,!=\,i}^{m}$ ykm * $\hat{p}ki$)

$= \frac{1}{m} \sum_{i=1}^{m}$ (-yki + (yki *pki)) + $\sum_{m\,!=\,i}^{m}$ ykm * $\hat{p}ki$)

$= \frac{1}{m} \sum_{i=1}^{m}$ (-yki + $\sum_{m}$ykm * $\hat{p}$ki)

$= \frac{1}{m} \sum_{i=1}^{m}$ ($\hat{p}$ki $\sum_{m}$ ykm - yki

In the above, the summation of ykm is equal to 1 so we get:

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

Lastly we must multiply this by $\frac{\partial}{\partial \theta k^T x^i} = \theta k^T x^i = x^i$

To get $\frac{1}{m} \sum_{i=1}^{m}$ ($\hat{p}$ki - yki) * x^i