Solutions:

1. Let A and B be two events,
if the conditional probability of A given B has already occurred is $P(A|B)$

Baye's Theorem states $P(B|A) = \frac{P(A|B)P(B)}{P(A)}$

$$\boxed{= \frac{P(A|B)\,P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}} \rightarrow eq^n ①$$

Proof:

we know $P(A|B) = \frac{P(A\cap B)}{P(B)} \Rightarrow \boxed{P(A\cap B) = P(A|B)P(B)} \rightarrow eq^n ②$.

Now, $P(A) = P((A\cap B) \cup (A\cap B^c))$

$= P(A\cap B) + P(A\cap B^c) \rightarrow ③$

from eq^n ② and ③, we can say that

$P(A) = P(A|B)P(B) + P(A|B^c)\cdot P(B^c)$ [substituting results from ②]

Now, ~~if A and B are~~ $\because (A\cap B)$ & $(A\cap B^c)$ are mutually exclusive, we have,

$$P(B|A) = \frac{P(B\cap A)}{P(A)} = \frac{P(A\cap B)}{P(A)} \quad \left[\because A\cap B = B\cap A\right]$$

from ②, we have

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

Hence, $\boxed{P(B|A) = \frac{P(A|B)\,P(B)}{P(A)}}$ QED.

Bayes Theorem provides a method to visualize relationship between data and a model. A machine Learning algorithm is aimed at deciphering the structured relationship between the data and a model.

Thus, Bayes Theorem ~~of~~ ~~cos~~ for computing conditional probability is very useful in ML, in ~~others~~ determining the methods which will make more mathematically accurate prediction

2.

Using MSE as the cost function, and minimizing it, we found the normal equation.

Similarly, to find the Ridge Regression closed form solution, we must
↳ Determine the value of w for which corresponding cost is minimum.

Given: $X = \begin{bmatrix} x^1 \\ x^2 \\ \vdots \\ x^m \end{bmatrix}$ is the i/p data matrix,

$x^i = (1, x_1, x_2, \dots, x_n) \rightarrow i^{th}$ sample

$y = (y^1\ y^2, \dots y^m) \rightarrow y^j$ = measurement of hyp, $h_w(x)$ for j^{th} sample

$h_w(x) = w_0 + w_1 x_1 + \dots w_n x_n$

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

Finding partial derivate wrt w.

$$\frac{\partial E(w)}{\partial w} = \frac{\partial}{\partial w}\left[\sum_i (w^T x^i - y^i)^2 + \lambda \sum_i w_i^2\right]$$

$$= 2\sum_i (w^T x^i - y^i) + 2\lambda \sum_i w_i$$

Premultiplication with X^T $= 2(X^T X)w - X^T y) + 2\lambda I w$ [$\because \lambda$ is a scalar]

Now, taking second derivative, we find.

$$\frac{\partial^2(E(w))}{\partial w^2} = 2X^TX + 2\lambda I$$
$$= 2|X|^2 + 2\lambda I \quad [\because A^TA = |A|^2]$$

$$\therefore \frac{\partial^2(E(w))}{\partial w^2} > 0 \text{ for } \lambda > 0. \quad [\because I \text{ is } > 0]$$

Hence, equating first derivative wrt 'w' will give us w for which $E(w)$ or the cost is minimum.

$$\therefore \frac{\partial E(w)}{\partial w} = 0 = 2X^TXw + 2\lambda Iw - 2X^Ty$$

By the commutative property.
$$\Rightarrow X^Ty = X^TXw + 2\lambda Iw.$$

Premultiplication with $(X^TX + \lambda I)^{-1}$ on both sides

we get $(\lambda I + X^TX)^{-1}.X^Ty = (\lambda I + X^TX)^{-1}.(\lambda I + X^TX)w$
$$= I\,w$$
$$= w.$$

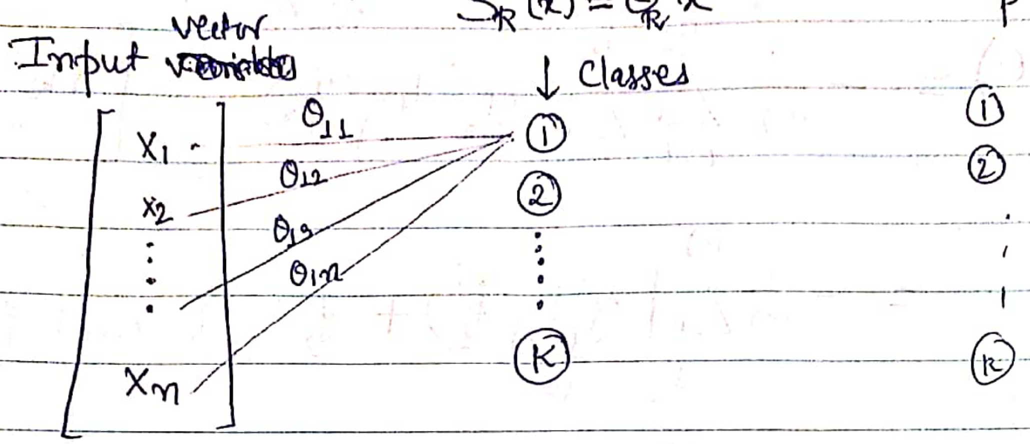
$$\therefore w_{\text{(for min cost)}} = (\lambda I + X^TX)^{-1}.X^Ty.$$

$\therefore$ for Ridge regn, $(\lambda I + X^TX)X^Ty$ is the closed form solution.

Scanned with CamScanner

3.

$S_k(x) = \theta_k^T x$



$$P_k = \frac{e^{s_k(x)}}{\sum_{j=1}^{k} e^{s_j(x)}}$$

1) So for each class, in order to learn this softmax Regression model, we need to estimate 'n' parameters, for the n corresponding input.

$\therefore$ In total, we need $(n \times k)$ parameters (for all k classes).

2) $S_k(x) = \theta_k^T \cdot x$; $\hat{p}_k = \dfrac{e^{(S_k(x))}}{\sum_{j=1}^{k} e^{(S_k(x))}}$

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{k} y_k^i \log(\hat{p}_k)$$ [$y_k^i = 1$, if i belongs k; $=0$, otherwise]

~~expanding inner summation,~~

$$\cancel{= -\frac{1}{m}\sum_{i=1}^{m}\left[\underbrace{y_1^i \log(\hat{p}_1^i)}_{0} + \underbrace{y_2^i \log(\hat{p}_2^i)}_{0} + \dots + y_k^i \log(\hat{p}_k^i) + \dots\right]}$$

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

or, $$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{k} y_j^i \log(\hat{p}_j)^i$$

$$\Rightarrow \frac{\partial J(\theta)}{\partial \theta} = -\frac{1}{m}\sum_{i=1}^{m}\sum_{k} y_k \frac{\partial[\log p_k]}{\partial \theta}$$

$$= -\frac{1}{m}\sum_{i=1}^{m}\sum_{k} y_k^i \cdot \frac{1}{p_k^i} \cdot \frac{\partial p_k^i}{\partial \theta}.$$ [Chainrule]

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

Now, $$\frac{\partial J(\theta)}{\partial \theta} = -\frac{1}{m}\sum_{i=1}^{m}\left\{-y_k^i(1-p_k^i) + \sum y_k(p_j^i)\right\}x^i$$

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

$$= \frac{1}{m}\sum_{i=1}^{m}\left\{p_k^i\Big(\underbrace{\sum y_k}_{1}\Big) - y_k^i + \underbrace{y_k^i p_j^i}_{0}\right\}x^i$$

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