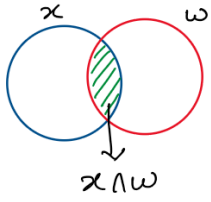
1. Prove Bayes' Theorem. Briefly explain why it is useful for machine learning problems, i.e., by converting posterior probability to likelihood and prior probability.

let "$x$" be the feature
"$\omega$" be the class we want to classify




Probability of having feature $x$ provided class $\omega$ have already classified

$$P(x|\omega) = \frac{P(x \cap \omega)}{P(\omega)} \quad -(1)$$

similarly,

$$P(\omega|x) = \frac{P(x \cap \omega)}{P(x)} \quad -(2)$$

from (1) and (2)

$$P(x \cap \omega) = P(x|\omega)\,P(\omega) = P(\omega|x)\,P(x)$$

$$\therefore \boxed{P(\omega|x) = \frac{P(x|\omega)P(\omega)}{P(x)}}$$

→ this is Bayes Theorem

where $P(x|\omega)$ = likelihood probability as we are calculating likelihood of having $x$ given $\omega$.
[we have all the information provided to calculate $P(x|\omega)$]

$P(\omega)$ = prior probability, it is probability of occuring $(\omega)$

$P(x)$ = evidence probability $\left(\sum_{i=1}^{i} P(x|\omega_i)\,P(\omega_i)\right)$
= It is just the multiplication of likelihood x prior for all classes.

This Bayes theorem is very useful in ML application as we have to predict the posterior probability (predicting class given feature $x$ already happened)

where we could calculate,
$P(x|\omega)$, $P(\omega)$ and $P(x)$ using the data available.

2. In Lecture 3-1, we gave the normal equation (i.e., closed-form solution) for linear regression using MSE as the cost function. **Prove that the closed-form solution for Ridge Regression** is $\boldsymbol{w}=(\lambda I+X_T\cdot X)_{-1}\cdot X_T\cdot \boldsymbol{y}$, where $I$ is the identity matrix, $X=(x_{(1)},x_{(2)},\ldots,x_{(m)})_T$ is the input data matrix, $x_{(i)}=(1,x_1,x_2,\ldots,x_n)$ is the $i$th data sample, and $y=(y_{(1)},y_{(2)},\ldots,y_m)$. Assume the hypothesis function $h_w(x)=w_0+w_1x_1+w_2x_2+\cdots+w_nx_n$, and $y_{(j)}$ is the measurement of $h_w(x)$ for the $j$th training sample. The cost function of the Ridge Regression is $E(\boldsymbol{w})=MSE(\boldsymbol{w})+\lambda_2\Sigma w_{i2mi=1}$. [Hint: please refer to the proof of the normal equation of linear regression. [ Note: Please use the following rectified definition of MSE when you prove: $MSE(w)=\Sigma(\boldsymbol{w}_T\cdot\boldsymbol{x}_{(i)}-y_{(i)})_{mi=12}$. ] .

Given:

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

where, $$MSE(\omega) = \sum_{i=1}^{m}(\omega^T.x^{(i)} - y^{(i)})^2$$

$\therefore$ $E(\omega)$ becomes

$$E(\omega) = \sum_{i=1}^{m}(\omega^T.x - y)^2 + \frac{\lambda}{2}\sum_{i=1}^{m}\omega_i^2$$

Representing above equation in matrix equivalent form.

$$E(\omega) = \frac{1}{2}(\omega^T.X - y)^T(\omega^T.X - y) + \frac{\lambda}{2}\omega^T\omega$$

$$= \frac{1}{2}\left[\{(\omega^T.x)^T - y^T\}(\omega^T x - y)\right] + \frac{\lambda}{2}\omega^T\omega$$

$$= \frac{1}{2}\left[(\omega^T x)^T.\omega^T.x - y^T.\omega^T x - (\omega^T x)^T y\right] - y^T.y + \frac{\lambda}{2}\omega^T\omega$$

$$= \frac{1}{2}\left[w \cdot x^T \cdot w^T \cdot x - y^T w^T x - w x^T \cdot y - y^T \cdot y\right] + \frac{\lambda}{2} w^T w$$

$$\frac{\partial E(w)}{\partial w} = \frac{1}{2}\left[2 x^T \cdot w \cdot x - y^T x - x^T \cdot y + 2\lambda w\right] \quad \text{—— (1)}$$

putting $y^T x = x^T y$ as

$$y = \begin{bmatrix} \\ \\ \end{bmatrix}_{m \times 1} \quad \text{and} \quad x = \begin{bmatrix} \\ \\ \end{bmatrix}_{m \times 1}$$

$$y^T = \begin{bmatrix} & & \end{bmatrix}_{1 \times m} \qquad x^T = \begin{bmatrix} & & \end{bmatrix}_{1 \times m}$$

$$y^T \cdot x = \begin{bmatrix} & & \end{bmatrix}_{1 \times m} \begin{bmatrix} \\ \\ \end{bmatrix}_{m \times 1} \qquad x^T \cdot y = \begin{bmatrix} & & \end{bmatrix}_{1 \times m} \begin{bmatrix} \\ \\ \end{bmatrix}_{m \times 1}$$

$$y^T x = \begin{bmatrix} \end{bmatrix}_{1 \times 1} \qquad x^T \cdot y = \begin{bmatrix} \end{bmatrix}_{1 \times 1}$$

$\therefore$ Equation (1) becomes,

$$= \frac{1}{2}\left[2 x^T w \cdot x - 2 x^T \cdot y + 2\lambda w\right]$$

$$= x^T w \cdot x - x^T \cdot y + \lambda w$$

$$\frac{\partial E(w)}{\partial (w)} = 0$$

$$x^T w \cdot x - x^T y + \lambda w = 0$$

$$x^T w \cdot x - \lambda w = x^T y$$

$$w\,(x^T \cdot x - \lambda I) = x^T y \quad \text{..... since } AI = A \text{ (matrix multiplication property)}$$

$$w = \frac{x^T \cdot y}{(x^T \cdot x - \lambda I)}$$

$$w = (x^T \cdot x - \lambda I)^{-1} \cdot x^T \cdot y$$

3. Recall the multi-class Softmax Regression model on page 16 of Lecture 3-3. Assume we have K different classes. The posterior probability is $\hat{p}_k = \delta(s_k(x))_k = \exp(s_k(x)) \Sigma \exp(s_j(x))_{Kj=1}$ for $k=1,2,\ldots,K$, where $s_k(x) = \theta_{kT} \cdot x$, and input $x$ is an $n$-dimension vector.

1) To learn this Softmax Regression model, how many parameters we need to estimate? What are these parameters?

We need to estimate $\theta_{n+1}$ parameters i.e, $\theta_0, \theta_1, \theta_2, \ldots, \theta_{n+1}$ where $\theta_0$ is bias parameter. These parameters are the numbers we use to adjust the features in order to predict the output.

for eq $y = \theta_0 + \theta_1 \omega_1 + \theta_2 \omega_2 + \ldots + \theta_n \omega_n$

$y$ ↑ output

where n = number of classes
∴ therefore there will be n+1 parameters.

2) Consider the cross-entropy cost function $J(\Theta)$ (see page 16 of Lecture 3-3) of $m$ training samples $\{(x_i, y_i)\}_{i=1,2,\ldots,m}$. Derive the gradient of $J(\Theta)$ regarding to $\theta_k$ as shown in page 17 of Lecture 3-3

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

We know,

$$p_K = \delta(s_K(x))_K = \frac{\exp(s_K(x))}{\sum_{j=1}^{K} \exp(s_j(x))}$$

Where $s_K(x) = \theta_{K.}^T x$

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} \sum_{K=1}^{K} y_K^{(i)} \log \left[ \frac{\exp(\theta_{K.}^T x)}{\sum_{j=1}^{K} \exp(\theta_j^T x)} \right]$$

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} \left[ \sum_{K=1}^{K} y_K^{(i)} \log [\exp(\theta_K^T \cdot x^i)] - \sum_{K=1}^{K} y_K^{(i)} \log \left[ \sum_{j=1}^{K} \exp(\theta_j^T x^i) \right] \right]$$

putting $y_K^{(i)} = 1$ as $y_K = 1$ .... if $i^{th}$ instance belongs to K
$0$ otherwise
} lecture slide 3-3 pg 16

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} \left[ \sum_{K=1}^{K} \log [\exp(\theta_K^T x^i)] - \sum_{K=1}^{K} \log \left[ \sum_{j=1}^{K} \exp(\theta_j^T x^i) \right] \right]$$

$= \theta_K^T x$ --- since $\ln(e^x) = x$

$$J(\theta) = \frac{-1}{m} \sum_{i=1}^{m} \left[ \sum_{k=1}^{K} \theta_k^T x^i - \sum_{k=1}^{K} \log \left[ \sum_{j=1}^{K} \exp(\theta_j^T x^{(i)}) \right] \right]$$

$$\frac{\partial J(\theta)}{\partial \theta_k} = \frac{-1}{m} \sum_{i=1}^{m} \left[ x^{(i)} - \frac{1}{\sum_{j=1}^{K} \exp(\theta_j^T x)} \cdot \exp(\theta_k^T x^i)\, x^{(i)} \right]$$

$$= \frac{-1}{m} \sum_{i=1}^{m} \left[ 1 - \frac{\exp(\theta_k^T x)}{\sum_{j=1}^{K} \exp(\theta_j^T x)} \right] x^{(i)}$$

$$= \frac{-1}{m} \sum_{i=1}^{m} \left[ 1 - \frac{\exp(s_k(x))}{\sum_{j=1}^{K} \exp(s_j(x))} \right] x^{(i)} \quad \cdots\cdots \text{ as } \theta_k^T x = s_k(x), \; \theta_j^T x = s_j(x)$$

$$= \frac{-1}{m} \sum_{i=1}^{m} \left[ 1 - p_k^{(j)} \right] x^{(i)} \quad \cdots \text{ as } p_k = \frac{\exp(s_k(x))}{\sum_{j=1}^{K} \exp(s_j(x))}$$

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