Q1 Show that the following partial derivative for the logistic regression cost function holds:

$$\frac{dJ(\vec{\beta})}{d\beta_j} = \frac{1}{N}\sum_{i=1}^{N}\left(\sigma(\vec{\beta}^T\vec{x}_i) - y_i\right)x_{ij}$$

$\boxed{\text{Recall}}$: In logistic regression:

Sigmoid function: $\sigma(t_i) = \dfrac{1}{1+e^{-t_i}}$

logistic output: $P_i = \sigma(\underbrace{\vec{x}_i^T\vec{\beta}}_{})$

$$= t_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots$$

Cost function: $J(\vec{\beta}) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \underbrace{\log(P_i)}_{①} + (1-y_i)\underbrace{\log(1-P_i)}_{②}]$

$$P_i = \sigma(t_i) = \frac{1}{1+e^{-t_i}}$$

$\therefore$ $\log(P_i) = \log\frac{1}{1+e^{-t_i}} = -\log(1+e^{-t_i})$ — ①

$\log(1-P_i) = \log(1-\frac{1}{1+e^{-t_i}}) = \log(\frac{e^{-t_i}}{1+e^{-t_i}}) = \log(e^{-t_i}) - \log(1+e^{-t_i})$

$= -t_i - \log(1+e^{-t_i})$ — ②

plugin ① & ②

$$J(\vec{\beta}) = -\frac{1}{N}\sum_{i=1}^{N}[\cancel{-y_i\log(1+e^{-t_i})} + \underbrace{(1-y_i)(-t_i-\log(1+e^{-t_i}))}_{}]$$

$$= \underbrace{-t_i - \log(1+e^{-t_i})}_{} + y_i t_i + \cancel{y_i\log(1+e^{-t_i})}$$

$$-(\log e^{t_i} + \log(1+e^{-t_i}))$$
$$= -\log(e^{t_i}(1+e^{-t_i}))$$
$$= -\log(e^{t_i}+1)$$

$$J(\vec{\beta}) = -\frac{1}{N}\sum_{i=1}^{N}[y_i t_i - \log(1+e^{t_i})]$$

$$\frac{dJ(\vec{\beta})}{d\beta_j} = \frac{d}{d\beta_j}\left[\frac{1}{N}\sum_{i=1}^{N}\left[\underbrace{-y_i t_i}_{①} + \underbrace{\log(1+e^{t_i})}_{②}\right]\right]$$

$$① \quad \frac{d}{d\beta_j} -y_i t_i = \frac{d}{d\beta_j} -y_i \vec{\beta}^T\vec{x}_i = \frac{d}{d\beta_j} -y_i(\beta_0 + \beta_1 x_{i1} + \dots \beta_j x_{ij} \dots)$$

$$= -y_i x_{ij} \quad (\text{Non } \beta_j \text{ terms become zero})$$

$$② \quad \frac{d}{d\beta_j}\log(1+e^{t_i}) = \frac{d}{d\beta_j}\log(1+e^{\vec{\beta}^T\vec{x}_i})$$

$$= \frac{1}{1+e^{\vec{\beta}^T\vec{x}_i}} \cdot e^{\vec{\beta}^T\vec{x}_i} \cdot x_{ij} = \frac{1}{e^{-\vec{\beta}^T\vec{x}_i}+1} \cdot x_{ij}$$

plug in ① and ②

$$\frac{dJ(\vec{\beta})}{d\beta_j} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{1+e^{-\vec{\beta}^T\vec{x}_i}} x_{ij} - y_i x_{ij}\right)$$

$$\because \sigma(t) = \frac{1}{1+e^{-t}}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left(\sigma(\vec{\beta}^T\vec{x}_i) - y_i\right) x_{ij} \quad //$$

Q2a) What is the value of $\sum_{k=1}^{K} y_k^{(i)}$, as defined in Lecture 4 for softmax regression?

Ans: the value is 1

because each instance i can only belong to one class. If instance i belong to class j, then $y_j^{(i)} = 1$, $y_{k \neq j}^{(i)} = 0$

Q2b) Show that $\frac{1}{p_k^{(i)}} \frac{d p_k^{(i)}}{d t_{\bar{k}}} = \begin{cases} -p_{\bar{k}}^{(i)} & \text{if } k \neq \bar{k} \\ 1 - p_k^{(i)} = 1 - p_{\bar{k}}^{(i)} & \text{if } k = \bar{k} \end{cases}$

where $p_k^{(i)} = \frac{e^{t_k}}{\sum_{j=1}^{K} e^{t_j}}$, and $\bar{k}$ is an index where $1 \leq \bar{k} \leq K$

$\boxed{\text{if } k \neq \bar{k}}$

$$\frac{d p_k^{(i)}}{d t_{\bar{k}}} = \frac{d}{d t_{\bar{k}}} \left(\frac{e^{t_k}}{\sum_{j=1}^{K} e^{t_j}} \right) = \frac{(e^{t_k})' \sum_{j=1}^{K} e^{t_j} - (\sum_{j=1}^{K} e^{t_j})' e^{t_k}}{\left[\sum_{j=1}^{K} e^{t_j} \right]^2}$$

$$= \frac{0 - e^{t_{\bar{k}}} \cdot e^{t_k}}{\left[\sum_{j=1}^{K} e^{t_j} \right]^2} = \frac{-e^{t_{\bar{k}}}}{\sum_{j=1}^{K} e^{t_j}} \cdot \frac{e^{t_k}}{\sum_{j=1}^{K} e^{t_j}} = (-p_{\bar{k}}^{(i)}) \cdot (p_k^{(i)})$$

$\therefore$ if $k \neq \bar{k}$, $\frac{d p_k^{(i)}}{d t_{\bar{k}}} = (-p_{\bar{k}}^{(i)}) \cdot (p_k^{(i)})$ —— ①

$\boxed{\text{if } k = \bar{k}}$

$$\frac{d p_k^{(i)}}{d t_{\bar{k}}} = \frac{d}{d t_{\bar{k}}} \left(\frac{e^{t_{\bar{k}}}}{\sum_{j=1}^{K} e^{t_j}} \right) = \frac{(e^{t_{\bar{k}}})' \sum_{j=1}^{K} e^{t_j} - (\sum_{j=1}^{K} e^{t_j})' e^{t_k}}{\left[\sum_{j=1}^{K} e^{t_j} \right]^2}$$

$$= \frac{e^{t_{\bar{k}}}\sum_{j=1}^{k} e^{t_j} - e^{t_{\bar{k}}} \cdot e^{t_{\bar{k}}}}{\left[\sum_{j=1}^{k} e^{t_j}\right]^2} = \underbrace{\frac{e^{t_{\bar{k}}}}{\sum_{j=1}^{k} e^{t_j}}}_{=(P_{\bar{k}}^{(i)})} \cdot \underbrace{\frac{\sum_{j=1}^{k} e^{t_j} - e^{t_{\bar{k}}}}{\sum_{j=1}^{k} e^{t_j}}}_{(1-P_{\bar{k}}^{(i)})}$$

$$\text{if } k=\bar{k}, \quad \frac{dP_k^{(i)}}{dt_{\bar{k}}} = (P_{\bar{k}}^{(i)})(1-P_{\bar{k}}^{(i)}) \quad \text{--- } 2$$

Summarize 1 & 2

$$\frac{dP_k^{(i)}}{dt_{\bar{k}}} = \begin{cases} (-P_{\bar{k}}^{(i)}) \cdot (P_k^{(i)}) & \text{if } k \neq \bar{k} \\ (P_{\bar{k}}^{(i)}) \cdot (1-P_{\bar{k}}^{(i)}) & \text{if } k = \bar{k} \end{cases}$$

$$\therefore \frac{1}{P_k^{(i)}} \cdot \frac{dP_k^{(i)}}{dt_{\bar{k}}} = \begin{cases} -P_{\bar{k}}^{(i)} & \text{if } k \neq \bar{k} \\ 1-P_{\bar{k}}^{(i)} = 1-P_k^{(i)} & \text{if } k = \bar{k} \end{cases}$$

Q2c) Softmax cost function : $J(\beta) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_k^i \log P_k^i$

show $\frac{dJ(\beta)}{d\beta^{(\bar{k})}} = \frac{1}{N}\sum_{i=1}^{N}(P_{\bar{k}}^{(i)} - y_{\bar{k}}^{(i)})x_i$

key terms to show

$$\frac{dJ(\beta)}{d\beta^{(\bar{k})}} = \frac{dJ(\beta)}{dP_k^{(i)}} \cdot \frac{dP_k^{(i)}}{dt_{\bar{k}}} \cdot \frac{dt_{\bar{k}}}{d\beta^{(\bar{k})}}$$

$= x_i$

$\because t_{\bar{k}} = x_i \beta^{(\bar{k})}$

from 2b $= \begin{cases} (-P_{\bar{k}}^{(i)}) \cdot (P_k^{(i)}) & \text{if } k \neq \bar{k} \\ (P_{\bar{k}}^{(i)}) \cdot (1-P_{\bar{k}}^{(i)}) & \text{if } k = \bar{k} \end{cases}$

Goal: Solve $\frac{dJ(\beta)}{dP_{\bar{K}}^{(i)}} \cdot \frac{dP_{K}^{(i)}}{dt_{\bar{K}}}$ and show that

$$\frac{d}{d\beta^{(\bar{K})}} \sum_{K=1}^{K} -y_K^i \log P_K^i = (P_{\bar{K}}^i - y_{\bar{K}}^i)$$

$$\frac{d}{d\beta^{(\bar{K})}} \sum_{K=1}^{K} -y_K^i \log P_K^i = -\sum_{K=1}^{K} y_K^i \frac{d}{d\beta^{\bar{K}}} \log P_K^i = -\sum_{K=1}^{K} y_K^i \cdot \underbrace{\frac{1}{P_K^i} \cdot \frac{dP_K^i}{d\beta^{\bar{K}}}}_{\text{use reference result of Q2b}}$$

$$= -y_{\bar{K}}^i \cdot \frac{(P_{\bar{K}}^i)(1-P_{\bar{K}}^i)}{P_{\bar{K}}^i} - \sum_{K \neq \bar{K}} y_K^i \cdot \frac{(-P_{\bar{K}}^i)(P_K^i)}{P_K^i}$$

$$= -y_{\bar{K}}^i (1-P_{\bar{K}}^i) + \sum_{K \neq \bar{K}} y_K^i (P_{\bar{K}}^i)$$

$$= -y_{\bar{K}}^i + y_{\bar{K}}^i P_{\bar{K}}^i + \sum_{K \neq \bar{K}} y_K^i (P_{\bar{K}}^i)$$

$$= -y_{\bar{K}}^i + P_{\bar{K}}^i \underbrace{\sum_{K=1}^{K} y_K^i}_{=1 \text{ as answered in 2a}}$$

$$= P_{\bar{K}}^i - y_{\bar{K}}^i$$

→ Now, multiply $\frac{dt_{\bar{K}}}{d\beta^{\bar{K}}}$'s result X_i

we get:

$$\frac{d}{d\beta^{\bar{K}}} \sum_{K=1}^{K} -y_K^i \log P_K^i = (P_{\bar{K}}^{(i)} - y_{\bar{K}}^{(i)}) X_i$$

put the summation and $\frac{1}{N}$ back and pull out negative sign:

$$\frac{d}{d\beta^{\bar{K}}} -\frac{1}{N} \sum_{i=1}^{N} \sum_{K=1}^{K} y_K^i \log P_K^i = \frac{1}{N} \sum_{i=1}^{N} (P_{\bar{K}}^{(i)} - y_{\bar{K}}^{(i)}) X_i$$

//

Q3. Show that if there are only 2 classes, cost function of softmax = cost function of logistic regression

$\boxed{\text{Recall}}$: Logistic probability output $= P = \dfrac{1}{1+e^{-X^T\beta}}$ — ①

Softmax score of each class k $= S_k(x) = x^T\beta^{(k)}$

Softmax log-odd $= P_k = \sigma_k(S(x)) = \dfrac{e^{S_k(x)}}{\sum_{j=1}^{K} e^{S_j(x)}}$

For softmax, if there're only 2 classes:
with class 1 = 1, class 2 = 0
Then k=2, with Score = $\underbrace{X^T\beta^{(1)}}_{\text{class } 1=1}$ or $\underbrace{X^T\beta^{(2)}}_{\text{class } 2=0}$

Then $P_1 = \dfrac{e^{X^T\beta^{(1)}}}{e^{X^T\beta^{(1)}} + e^{X^T\beta^{(2)}}}$ $P_2 = \dfrac{e^{X^T\beta^{(2)}}}{e^{X^T\beta^{(1)}} + e^{X^T\beta^{(2)}}}$

$$P_1 = \frac{1}{1+\frac{e^{X^T\beta^{(2)}}}{e^{X^T\beta^{(1)}}}} = \frac{1}{1+e^{X^T(\beta^{(2)}-\beta^{(1)})}} = \frac{1}{1+e^{-X^T(\beta^{(1)}-\beta^{(2)})}} \quad \text{— ②}$$

Compare ① & ②, with $\beta = \beta^{(1)} - \beta^{(2)}$,
the probability given by softmax & logistic regression are the same.

Softmax cost function: $J(\beta) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_k^i \log P_k^i$ ☆

$\Downarrow$

if k=2, $= y_1^i \log P_1^i + y_2^i \log P_2^i$

if k=2, $y_2^i = (1-y_1^i)$
$P_2^i = (1-P_1^i)$ $\Rightarrow$

$\Downarrow$

$= y_1^i \log P_1^i + (1-y_1^i)\log P_1^i)$ ☆

$\therefore J(\beta) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(P_i) + (1-y_i)\log(1-P_i)]$
= Logistic cost function //