Problem 1

Show that Eqn. (8) Lecture 3 holds, i.e., the partial derivative of $J(\boldsymbol{\beta})$ with respect to β_j is

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

Log Loss:

$$J(\beta) = \frac{-1}{N} \sum_{i=1}^{N} \left[y_i \log(p_i) + (1-y_i)\log(1-p_i) \right] \qquad (7)$$

$$\sigma(t) = \frac{1}{1+\exp(-t)}$$

$$\sigma(x_i^T\beta) = \frac{1}{1+\exp(-x_i^T\beta)} = p_i$$

$$\therefore \frac{\partial p_i}{\partial \beta_j} = \frac{\partial}{\partial \beta_j}(\sigma(x_i^T\beta))$$

$$= \sigma'(x_i^T\beta)\frac{\partial}{\partial \beta_j}(x_i^T\beta)$$

$$= \frac{\exp(-x_i^T\beta)\ x_{ij}}{(1+\exp(-x_i^T\beta))^2}$$

$$= \sigma(x_i^T\beta)(1-\sigma(x_i^T\beta))x_{ij}$$

$$\frac{\partial}{\partial \beta_j} J(\beta) = \frac{-1}{N} \sum_{i=1}^{N} \left[y_i \frac{1}{p_i}\frac{\partial p_i}{\partial \beta_j} + (1-y_i)\frac{1}{1-p_i}\left(-\frac{\partial p_i}{\partial \beta_j}\right) \right]$$

$$= \frac{-1}{N} \sum_{i=1}^{N} \left[\frac{y_i}{p_i} + \frac{1-y_i}{1-p_i} \right] \frac{\partial p_i}{\partial \beta_j}$$

now we can substitute p_i & $\frac{\partial p_i}{\partial \beta_j}$:

$$\frac{\partial}{\partial \beta_j} J(\beta) = \frac{-1}{N} \sum_{i=1}^{N} \left[\frac{y_i}{\cancel{\sigma}(x_i^T\beta)} + \frac{1-y_i}{\cancel{1-\sigma(x_i^T\beta)}} \right] \sigma(x_i^T\beta)\cancel{(1-\sigma(x_i^T\beta))}x_{ij}$$

$$= \frac{-1}{N} \sum^{N}$$

$$= \frac{-1}{N}\sum_{i=1}^{N}(y_i - \sigma(x_i^T\beta))\, x_{ij}$$

$$= \frac{1}{N}\sum_{i=1}^{N}(\sigma(\beta^T x_i) - y_j)\, x_{ij}$$

Problem 2

Consider the softmax regression. The number of data instances is N, and the number of classes is K.

(a) What is the value of $\sum_{k=1}^{K} y_k{}^{(i)}$, where $y_k^{(i)}$ is defined in Lecture 4.

(b) Let $t_k = s_k(\boldsymbol{x}_i)$. So p_k can be rewritten as:

$$p_k^{(i)} = \frac{\exp(t_k)}{\sum_{j=1}^{K}\exp(t_j)}$$

Let $\bar{k}$, $1 \le \bar{k} \le K$, be another index. Show that

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

(c) Show Eqn. (9) Lecture 4 holds, i.e.,

$$\nabla_{\boldsymbol{\beta}^{(\bar{k})}} J(B) = \frac{1}{N}\sum_{i=1}^{N}(p_{\bar{k}}^{(i)} - y_{\bar{k}}^{(i)})\boldsymbol{x}_i \tag{1}$$

(a) Lecture 4: $y_K^{(i)}$ is the target probability that the ith instance belongs to class K; either 0 or 1. Since an ith instance can only belong to 1 class, if $y_K^{(i)} = 1$ then $y_K^{(j)} = 0$ where $j \neq i$. This summation thus will always add up to 1, because each instance can belong only to one class. $\sum_{i=1}^{K} y_{(K)}^{(i)} = 1$

(b) case 1: $K = \bar{K}$

$$\frac{\partial p_K^{(i)}}{\partial t_K} = \frac{\exp(t_K)\sum_{j=1}^{K}\exp(t_j) - \exp(t_K)\exp(t_K)}{\sum_{j=1}^{K}\exp(t_j)}$$

$\partial t_K \qquad \sum_{j=1}^{K} \exp(t_j)$

$$\therefore \quad \frac{\partial p_K^{(i)}}{\partial t_K} = p_K^{(i)}(1 - p_K^{(i)})$$

Thus $\displaystyle \frac{1}{p_K^{(i)}} \frac{\partial p_K^{(i)}}{\partial t_K} = 1 - p_K^{(i)}$

case 2 : $K \neq \bar{K}$

$$\frac{\partial p_K^{(i)}}{\partial t_{\bar{K}}} = - \frac{\exp(t_K)\exp(t_{\bar{K}})}{\left(\sum_{j=1}^{K} \exp(t_j)\right)^2} = -p_K^{(i)} p_{\bar{K}}^{(i)}$$

$$\therefore \quad \frac{1}{p_K^{(i)}} \frac{\partial p_K^{(i)}}{\partial t_{\bar{K}}} = -p_{\bar{K}}^{(i)}$$

(c) given $\displaystyle J(B) = \frac{-1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} y_k^{(i)} \log(p_k^{(i)})$

& $\displaystyle p_K^{(i)} = \frac{\exp(X_i^T \beta^{(K)})}{\sum_{j=1}^{K} \exp(X_i^T \beta^{(j)})}$

$$\frac{\partial J(B)}{\partial \beta^{(\bar{K})}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} y_k^{(i)} \frac{\partial}{\partial \beta^{(\bar{K})}} \log(p_K^{(i)})$$

$$\frac{\partial}{\partial \beta^{(\bar{K})}} \log(p_K^{(i)}) = \frac{1}{p_K^{(i)}} \frac{\partial p_K^{(i)}}{\partial \beta^{(\bar{K})}}$$

from softmax properties:

$$\frac{\partial p_K^{(i)}}{\partial t_{\bar{K}}} = \begin{cases} p_K^{(i)}(1 - p_K^{(i)}) & \text{if } K = \bar{K} \\ -p_K^{(i)} p_{\bar{K}}^{(i)} & K \neq \bar{K} \end{cases}$$

$$\frac{\partial t_{\bar{K}}}{\partial \beta^{(K)}} = x_i$$

$$\therefore \quad \frac{\partial p_K^{(i)}}{\partial \beta^{(\bar{K})}} = \frac{\partial p_K^{(i)}}{\partial t_{\bar{K}}} \cdot \frac{\partial t_{\bar{K}}}{\partial \beta^{(\bar{K})}} = \frac{\partial p_K^{(i)}}{\partial t_{\bar{K}}} \cdot x_i$$

$$\partial\beta^{(k)} \quad \partial t_{\bar{k}} \quad \overline{\partial\beta^{(\bar{k})}} \quad \overline{\partial t_{\bar{k}}}$$

now we substitute

$$\frac{\partial J(B)}{\partial\beta^{(\bar{k})}} = -\frac{1}{N}\sum_{i=1}^{N}\left(y_{\bar{k}}^{(i)} - p_{\bar{k}}^{(i)}\right)X_i$$

$$\therefore \nabla_{\beta^{(\bar{k})}} J(B) = \frac{1}{N}\sum_{i=1}^{N}\left(p_{\bar{k}}^{(i)} - y_{\bar{k}}^{(i)}\right)X_i$$

Problem 3 (MATH 5388 Only)

Show the cost function of softmax regression (Eqn. 8 Lecture 4) is equivalent to that of the logistic regression (Eqn. 7 lecture 3), if there are only 2 classes: 1 and 0. Hint: suppose the coefficient for the first class 1 is $\boldsymbol{\beta}^{(1)}$ and the coefficient for the second class 0 is $\boldsymbol{\beta}^{(2)}$. Consider consolidating the coefficients to a single parameter $\boldsymbol{\beta} = \boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(2)}$.

softmax regression for k classes:

$$J(\beta) = \frac{-1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_k^{(i)} \log(p_k^{(i)})$$

we are told only 2 classes

$$\therefore K = 2$$

$$J(\beta) = \frac{-1}{N}\sum_{i=1}^{N}\left[y_i \log(p_i) + (1-y_i)\log(1-p_i)\right]$$

$$p_1^{(i)} = \frac{\exp(\beta^{(1)T} x_i)}{\exp(\beta^{(1)T} x_i) + \exp(\beta^{(2)T} x_i)}$$

$$p_2^{(i)} = \frac{\exp(\beta^{(2)T} x_i)}{\exp(\beta^{(1)T} x_i) + \exp(\beta^{(2)T} x_i)} = 1 - p_1^{(i)}$$

hint : $\beta = \beta^{(1)} - \beta^{(2)}$

$$\therefore \quad p_1^{(i)} = \frac{\exp(\beta^T x_i)}{\exp(\beta^T x_i)+1}$$

$$p_2^{(i)} = \frac{1}{\exp(\beta^T x_i)+1} = 1 - p_1^{(i)}$$

$$J(\beta) = \frac{-1}{N} \sum_{i=1}^{N} \left[y_1^{(i)} \log(p_1^{(i)}) + y_2^{(i)} \log(p_2^{(i)}) \right]$$

$$= \frac{-1}{N} \sum_{i=1}^{N} \left[y_i \log\left(\frac{\exp(\beta^T x_i)}{\exp(\beta^T x_i)+1}\right) + (1-y_i)\log\left(\frac{1}{\exp(\beta^T x_i)+1}\right)\right]$$

same as that of logistic regression