Problem 1

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

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

My Solution:

First note that $J(\hat{\boldsymbol{\beta}}) = -\frac{1}{N}\sum_{i=1}^{N}[y_i log(p_i) + (1-y_i)log(1-p_i)]$. Then we also know that $p_i = \sigma(\boldsymbol{\beta^T x_i})$ thus if we let $t_i = \boldsymbol{\beta^T x_i}$ then $p_i = \sigma(t_i) = \frac{1}{1+e^{-t_i}}$. Briefly note that $t_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_N x_{iN}$ and thus t_i is a function of β_j. Here are two facts I will use to show my solution,

$$\frac{d}{dt_i}\sigma(t_i) = \frac{d}{dt_i}\left[\frac{1}{1+e^{-t_i}}\right] = \sigma(t_i)(1-\sigma(t_i)) \tag{1}$$

$$\frac{\partial t_i}{\partial \beta_j} = x_{ij} \tag{2}$$

Now I will replace every p_i with $\sigma(t_i)$ as they are equivalent expressions,

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

Note then,

$$\frac{\partial}{\partial \beta_j}\left[-\frac{1}{N}\sum_{i=1}^{N} y_i log(\sigma(t_i)) + (1-y_i)log(1-\sigma(t_i))\right] \tag{3}$$

$$\left[-\frac{1}{N}\sum_{i=1}^{N} y_i \frac{\partial}{\partial \beta_j} log(\sigma(t_i)) + (1-y_i)\frac{\partial}{\partial \beta_j} log(1-\sigma(t_i))\right] \tag{4}$$

Examining each partial independently I will use the fact that $\frac{d}{dx}log(f(x)) = \frac{f'(x)}{f(x)}$,

$$\frac{\partial}{\partial \beta_j} log(\sigma(t_i)) = \frac{1}{\sigma(t_i)} \cdot \frac{\partial}{\partial \beta_j}\sigma(t_i)$$

and,

$$\frac{\partial}{\partial \beta_j} log(1-\sigma(t_i)) = \frac{1}{1-\sigma(t_i)} \cdot \frac{\partial}{\partial \beta_j} 1 - \sigma(t_i) = -\frac{1}{1-\sigma(t_i)} \cdot \frac{\partial}{\partial \beta_j}\sigma(t_i)$$

Now we can rewrite (3) such that each term will have $\frac{\partial}{\partial \beta_j}\sigma(t_i)$ in common,

$$-\frac{1}{N}\sum_{i=1}^{N}\left[\frac{y_i}{\sigma(t_i)}-\frac{1-y_i}{1-\sigma(t_i)}\right]\cdot\frac{\partial}{\partial\beta_j}\sigma(t_i)$$

But note that we will apply the chain rule for $\frac{\partial}{\partial\beta_j}\sigma(t_i)$ given $t_i=\boldsymbol{\beta}^T\boldsymbol{x}_i=\sum_{i=1}^{N}\beta_N x_{iN}$ thus $\frac{\partial t_i}{\partial\beta_j}=x_{ij}$. Further, using the fact that $\frac{d}{dt}\sigma(t)=\frac{d}{dt}[\frac{1}{1+e^{-t_i}}]=\sigma(t)(1-\sigma(t))$,

$$-\frac{1}{N}\sum_{i=1}^{N}\left[\frac{y_i}{\sigma(t_i)}-\frac{1-y_i}{1-\sigma(t_i)}\right]\cdot(\sigma(t_i)(1-\sigma(t_i)))\cdot x_{ij}$$

$$-\frac{1}{N}\sum_{i=1}^{N}\left[\frac{y_i\cdot\sigma(t_i)(1-\sigma(t_i))}{\sigma(t_i)}-\frac{(1-y_i)\cdot\sigma(t_i)(1-\sigma(t_i))}{1-\sigma(t_i)}\right]\cdot x_{ij}$$

Note the above in red will cancel

$$-\frac{1}{N}\sum_{i=1}^{N}[y_i-y_i\sigma(t_i)-\sigma(t_i)+y_i\sigma(t_i)]\cdot x_{ij}$$

Therefore,

$$\frac{\partial}{\partial\beta_j}J(\boldsymbol{\beta})=\frac{1}{N}\sum_{i=1}^{N}(\sigma(t_i)-y_i)x_{ij}$$

But note $t_i=\boldsymbol{\beta}^T\boldsymbol{x}_i$ thus,

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

as desired.

Problem 2

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

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

Consider the following objective function, $J(B)=-\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}y_k^{(i)}log(p_k^{(i)})$. $y_k^{(i)}$ is the target probability that the ith instance belongs to class k. For the correct class $y_k^{(i)}$ is 1 and zero for all other classes implying that the sum is 1.

Therefore, $\sum_{k=1}^{K} y_k^{(i)} = 1$.

Part 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 $\bar{k}$, $1 \leq \bar{k} \leq K$, be another index. Show that,

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

My Solution:

Case 1: Let $k = \bar{k}$

Then note that for $f(x) = \frac{g(x)}{h(x)} \Rightarrow f'(x) = \frac{g'(x)h(x) - h'(x)g(x)}{[h(x)]^2}$

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

So let $g(x) = e^{t_k}$ and $h(x) = \sum_{j=1}^{K} e^{t_j}$,

$$g'(x) = \frac{\partial}{\partial t_k} e^{t_k} = e^{t_k}$$

$$h'(x) = \frac{\partial}{\partial t_k} \sum_{j=1}^{K} e^{t_j} = \frac{\partial}{\partial t_k} (e^{t_1} + e^{t_2} + ... + e^{t_{k-1}} + e^{t_k} + e^{t_{k+1}} + ... + e^{t_K}) = e^{t_k}$$

$$\frac{\partial}{\partial t_k} p_k^{(i)} = \frac{e^{t_k} \sum_{j=1}^{K} e^{t_j} - e^{t_k} e^{t_k}}{[\sum_{j=1}^{K} e^{t_j}]^2}$$

$$\frac{\partial}{\partial t_k} p_k^{(i)} = \frac{e^{t_k} \sum_{j=1}^{K} e^{t_j}}{[\sum_{j=1}^{K} e^{t_j}]^2} - \frac{(e^{t_k})^2}{[\sum_{j=1}^{K} e^{t_j}]^2}$$

Note the definition of $p_k^{(i)} = \frac{e^{t_k}}{\sum_{j=1}^{K} e^{t_j}}$

Therefore, when $k = \bar{k}$ then $\frac{\partial p_k^{(i)}}{\partial t_k} = p_k^{(i)} - (p_k^{(i)})^2 = p_k^{(i)}(1 - p_k^{(i)})$

Case 2: Let $k \neq \bar{k}$

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

Let $g(x) = e^{t_k}$ and $h(x) = \sum_{j=1}^{K} e^{t_j}$

$$g'(x) = \frac{\partial}{\partial t_{\bar{k}}} e^{t_k} = 0$$

$$h'(x) = \frac{\partial}{\partial t_{\bar{k}}} (e^{t_1} + ... + e^{\bar{k}-1} + e^{\bar{k}} + e^{\bar{k}+1} + ... + e^{t_K}) = e^{t_{\bar{k}}}$$

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

Therefore, when $k \neq \bar{k}$ then $\frac{\partial}{\partial t_{\bar{k}}} p_k^{(i)} = -p_{\bar{k}}^{(i)} p_k^{(i)}$.
Giving,

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

If we bring out the reciprocal of $p_k^{(i)}$ for simplification we get,

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

as desired.

---

part c. Show $J(B) = -\frac{1}{N} \sum_{i=1}^N \sum_{k=1}^K y_k^{(i)} log(p_k^{(i)})$ holds,

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

My Solution:

Set-Up: Note that $J(B)$ is a function of $p_{\bar{k}}^{(i)}$ and $p_{\bar{k}}^{(i)}$ is a function of $t_{\bar{k}}$ in which $t_{\bar{k}} = x_i^T \beta^{\bar{k}}$ and is a function of $\beta^{\bar{k}}$. Therefore, we will use the chain rule and for reference, we will address each layer of $J(B)$ as follows,

$$\frac{\partial J(B)}{\partial p_k^{(i)}} \frac{\partial p_k^{(i)}}{\partial t_{\bar{k}}} \frac{\partial t_{\bar{k}}}{\partial \beta^{\bar{k}}}$$

Part a. gives that the $\sum_{k=1}^K y_k^{(i)} = 1$.
Part b. gives that

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

First note that $\frac{d}{dx} log(x) = \frac{1}{x}$,

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

Second note we have arrived at the fact established in b $\frac{1}{p_k^{(i)}}\frac{\partial p_k^{(i)}}{\partial t_{\bar{k}}}$,

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

Third, note we are considering $k = \bar{k}$ that is $y_{\bar{k}}^{(i)}$ is our target probability and for the correct class, $\bar{k}$, this is 1 and for all other classes the value is zero.

$$\frac{\partial J(B)}{\partial p_k^{(i)}}\frac{\partial p_k^{(i)}}{\partial t_{\bar{k}}} = -\frac{1}{N}\sum_{i=1}^{N}(\sum_{k\neq\bar{k}} y_k^{(i)}(-p_{\bar{k}}^{(i)}) + y_{\bar{k}}^{(i)}(1-p_{\bar{k}}^{(i)}))$$

Note above that $y_k^{(i)}(-p_{\bar{k}}^{(i)})$ will be zero since this is a one hot encoded vector. Also note for $k = \bar{k}$ we are just have $y_{\bar{k}}^{(i)}$.
Leaving us with the following expression (after distributing the negative sign),

$$\frac{\partial J(B)}{\partial p_k^{(i)}}\frac{\partial p_k^{(i)}}{\partial t_{\bar{k}}} = \frac{1}{N}\sum_{i=1}^{N}(p_{\bar{k}}^{(i)} - y_{\bar{k}}^{(i)})$$

Fourth, note that $t_{\bar{k}} = s_{\bar{k}}(\hat{x}_i) = \hat{x}_i^T\beta^{(\bar{k})}$ thus $\frac{\partial t_{\bar{k}}}{\partial \beta^{(\bar{k})}} = \boldsymbol{x}_i$

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

as desired.
