0.1 Softmax [2 points]

1) [1 point] Prove that softmax is invariant to constant sifts in the input, *i.e.*, for any input vector $\mathbf{x}$ and a constant scalar c, the following holds:

$$\mathrm{softmax}(\mathbf{x}) = \mathrm{softmax}(\mathbf{x}+c) ,$$

where $\mathrm{softmax}(\mathbf{x})_i \triangleq \frac{e^{x_i}}{\sum_{i'} e^{x_{i'}}}$, and $\mathbf{x}+c$ means adding c to every dimension of $\mathbf{x}$.

2) [1 point] Let $\mathbf{z} = \mathbf{W}\mathbf{x} + \mathbf{c}$, where $\mathbf{W}$ and $\mathbf{c}$ are some matrix and vector, respectively. Let

$$J = \sum_i \log \mathrm{softmax}(\mathbf{z})_i \ .$$

Calculate the derivatives of J w.r.t. $\mathbf{W}$ and $\mathbf{c}$, respectively, *i.e.*, calculate $\frac{\partial J}{\partial \mathbf{W}}$ and $\frac{\partial J}{\partial \mathbf{c}}$.

1) $$\mathrm{softmax}(x+c)_i = \frac{e^{x_i+c}}{\Sigma_{i'} e^{x_{i'}+c}} = \frac{e^{x_i} \cdot e^c}{\Sigma_{i'} e^{x_{i'}} \cdot e^c} = \frac{\cancel{e^c} \cdot e^{x_i}}{\cancel{e^c} \cdot \sum_{i'} e^{x_{i'}}}$$

$$= \frac{e^{x_i}}{\Sigma_{i'} e^{x_{i'}}} = \mathrm{softmax}(x)$$

$$\therefore \mathrm{softmax}(x) = \mathrm{softmax}(x+c)$$

2) $$\frac{\partial J}{\partial W} = \frac{\partial J}{\partial Z_i} \cdot \frac{\partial Z}{\partial W}$$

$$\because J = \sum_i \log \mathrm{softmax}(Z)_i$$

$$\therefore \frac{\partial J}{\partial Z} = \sum_i \frac{1}{\mathrm{softmax}(Z)} \cdot (\mathrm{softmax}(Z))'$$

$$\because \mathrm{softmax}(x)_i = \frac{e^{x_i}}{\Sigma_j e^{x_j}}$$

$$\frac{\partial J}{\partial Z} = \sum_i \frac{\log \mathrm{softmax}(Z)_i}{\partial Z_i}$$

$$= \sum_i \frac{1}{\mathrm{softmax}(Z)_i} \cdot \left(\frac{e^{z_i}}{\Sigma_j e^{z_j}}\right)'$$

$$= \sum_i \frac{1}{\mathrm{softmax}(Z)_i} \cdot \frac{e^{z_i} \Sigma_j e^{z_j} - e^{z_i} \cdot e^{z_i}}{(\Sigma_j e^{z_j})^2}$$

$$= \frac{1}{\mathrm{softmax}(Z)_i} \cdot \underbrace{\frac{e^{z_i}}{\Sigma_j e^{z_j}}}_{\mathrm{softmax}(\,)} \cdot \frac{\Sigma_j e^{z_j} - e^{z_i}}{\Sigma_j e^{z_j}}$$

$$= \frac{\Sigma_j e^{z_j} - e^{z_i}}{\Sigma_j e^{z_j}}$$

$$= 1 - \mathrm{softmax}(Z)_i$$

$$\frac{\partial J}{\partial W} = \frac{\partial J}{\partial Z_j} \cdot \frac{\partial Z}{\partial W} = (1 - \mathrm{softmax}(Z)_i)\, X_i \quad i = (1 \cdots N)$$

$$\frac{\partial J}{\partial C} = \frac{\partial J}{\partial z_i} = (1 - \mathrm{Softmax}(z)_i) \qquad i = (1 \dots N)$$

0.2 Logistic Regression with Regularization [2 points]

1) [1 point] Let the data be $(\mathbf{x}_i, y_i)_{i=1}^{N}$, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{0, 1\}$. Logistic regression is a binary classification model, with the probability of y_i being 1 as:

$$p(y_i; \mathbf{x}_i, \boldsymbol{\theta}) = \sigma\left(\boldsymbol{\theta}^T \mathbf{x}_i\right) \triangleq \frac{1}{1 + e^{-\boldsymbol{\theta}^T \mathbf{x}_i}} ,$$

where $\boldsymbol{\theta}$ is the model parameter. Assume we impose an L_2 regularization term on the parameter, defined as:

$$\mathcal{R}(\boldsymbol{\theta}) = \frac{\lambda}{2} \boldsymbol{\theta}^T \boldsymbol{\theta}$$

with a positive constant λ. Write out the final objective function for this logistic regression with regularization model.

2) [1 point] If we use gradient descent to solve the model parameter. Derive the updating rule for $\boldsymbol{\theta}$. Your answer should contain the derivation, not just the final answer.

1) Loss function of logistic regression.

$$J(\theta) = \frac{1}{N}\left(-\log \prod_{i=1}^{N} P_\theta(y_i; x_i, \theta) + \frac{\lambda}{2}\theta^T\theta\right)$$

$$= \frac{1}{N}\left[\sum_{i=1}^{1} \log P_\theta(y_i; x_i, \theta) + \frac{\lambda}{2}\theta^T\theta\right]$$

$$\because P(y_i; x_i, \theta) = \frac{1}{1+e^{-\theta^T x_i}}$$

$$\therefore J(\theta) = \frac{1}{N}\sum_{i=1}^{1} \log\left(\frac{1}{1+e^{-\theta^T x_i}}\right) + \frac{\lambda}{2N}\theta^T\theta.$$

2) $$\frac{\partial J(\theta)}{\partial \theta_i} = \frac{1}{N}\left[\sum_{i=1}^{N} (\sigma(\theta^T x^i) - y_i) x_i\right] + \frac{\lambda}{N}\theta_i$$

$$\theta_j = \theta_j - \eta\left[\frac{1}{N}\sum_{i=1}^{N} (\sigma(\theta^T x^i) - y^i) x_j^i + \frac{\lambda}{N}\theta_j\right] \quad (j = 0, 1, 2 \dots)$$

0.3 Derivative of the Softmax Function [3 points]

1) [1 point] Define the loss function as

$$J(\mathbf{z}) = -\sum_{k=1}^{K} y_k \log \tilde{y}_k ,$$

where $\tilde{y}_k = \frac{e^{z_k}}{\sum_{k'} e^{z_{k'}}}$, and $(y_1, \cdots, y_K)$ is a known probability vector. Derive the $\frac{\partial J(\mathbf{z})}{\partial \mathbf{z}}$. Note $\mathbf{z} = (z_1, \cdots, z_K)$ is a vector so $\frac{\partial J(\mathbf{z})}{\partial \mathbf{z}}$ is in the form of a vector. Your answer should contain the derivation, not just the final answer.

2 [1 point] Assume the above softmax is the output layer of an FNN. Briefly explain how the derivative is used in the backpropagation algorithm.

3) [1 points] Let $\mathbf{z} = \mathbf{W}^T \mathbf{h} + \mathbf{b}$, where $\mathbf{W}$ is a matrix, $\mathbf{b}$ and $\mathbf{h}$ are vectors. Use the chain rule to calculate the gradient of $\mathbf{W}$ and $\mathbf{b}$, *i.e.*, $\frac{\partial J}{\partial \mathbf{W}}$ and $\frac{\partial J}{\partial \mathbf{b}}$, respectively.

1) $\frac{\partial J(z)}{\partial z} = -\sum_{k=1}^{k} y_k (\log \tilde{y}_k)'$

$= -\sum_{k=1}^{k} y_k \frac{1}{\tilde{y}_k} \cdot \tilde{y}_k'$

$= -\sum_{k=1}^{k} y_k \frac{1}{\tilde{y}_k} \cdot \frac{e^{z_k} \Sigma_k e^{z_{k'}} - e^{z_k} e^{z_k}}{(\Sigma_{k'} e^{z_{k'}})^2}$

$= - y_k \cdot \frac{1}{\tilde{y}_k} \cdot \tilde{y}_k \frac{\Sigma_{k'} e^{z_{k'}} - e^{z_k}}{\Sigma_{k'} e^{z_{k'}}}$

$= - y_k \cdot \frac{\Sigma_{k'} e^{z_{k'}} - e^{z_k}}{\Sigma_{k'} e^{z_{k'}}}$

$= - y_k \cdot (1 - \tilde{y}_k)$

$= y_k \tilde{y}_k - y_k$

2) use the result of question (1),

update the $\theta = \theta - \eta \frac{\partial J(z)}{\partial z} = \theta - \eta (y_k \tilde{y}_k - y_k)$

3) $\frac{\partial J}{\partial w} = \frac{\partial J}{\partial z_k} \cdot \frac{\partial z_k}{\partial w}$

$= (y_k \tilde{y}_k - y_k) h.$ $\quad k = (1 \cdots k)$

$\frac{\partial J}{\partial b} = \frac{\partial J}{\partial z_k} \cdot \frac{\partial z_k}{\partial b} = \frac{\partial J}{\partial z_k} \cdot 1 = y_k \tilde{y}_k - y_k \quad k = (1 \cdots k)$