Problem 1.

Derive the gradient in softmax regression $\frac{\partial J}{\partial w_{k,i}}, \frac{\partial J}{\partial b_k}$

1° Consider the loss for one sample. Superscript (m) is omitted for simplicity

$$J = - \sum_{k'} t_{k'} \log y_{k'} \quad \text{where} \quad y_k = \frac{\exp(z_k)}{\sum_{k''} \exp(z_{k''})} \quad \text{and} \quad z_k = w_k^T x + b$$

$$= - \sum_{k'} t_{k'} \left[\log \exp(z_{k'}) - \log \sum_{k''} \exp(z_{k''}) \right]$$

$$= - \sum_{k'} t_{k'} \left[z_{k'} - \log \sum_{k''} \exp(z_{k''}) \right]$$

$$= - \sum_{k'} t_{k'} z_{k'} - \log \sum_{k''} \exp(z_{k''})$$

[because one t^k is one] $\sum_{k'}$ for the second term is not needed

$$\frac{\partial J}{\partial z_k} = \frac{\partial}{\partial z_k} \left[t_k z_k - \log \sum_{k''} \exp(z_{k''}) \right]$$

$$= -t_k + \frac{1}{\sum_{k''} \exp(z_{k''})} \cdot \exp(z_k).$$

$$= -t_k + y_k$$

Thus $$\frac{\partial J}{\partial w_{k,j}} = \frac{\partial J}{\partial z_k} \cdot \frac{\partial z_k}{\partial w_{kj}} = (y_k - t_k)\, x_j$$

$$\frac{\partial J}{\partial b_k} = \frac{\partial J}{\partial z_k} \cdot \frac{\partial z_k}{\partial b_k} = y_k - t_k$$

If you consider the total loss of m samples

$$\frac{\partial J_{total}}{\partial w_{kj}} = \sum_{m=1}^{M} \left(y_k^{(m)} - t_k^{(m)} \right) x_j^{(m)} \qquad \frac{\partial J_{total}}{\partial b_k} = \sum_{m=1}^{M} \left(y_k^{(m)} - t_k^{(m)} \right)$$

Problem 2.

Read the section "**Logistic regression vs. softmax**" in the lecture note. It shows that a two-way softmax can be reduced to logistic regression.

Please show that logistic regression can also be reduced to 2-way softmax, i.e., for any parameter of the logistic regression model, there exists some parameter of the softmax regression model that does the same thing.

2° Consider logistic regression:

$$y = \sigma(w^T x + b) = \frac{1}{1 + \exp(-(w^T x + b))} = \frac{\exp(w^T x + b)}{\exp(w^T x + b) + 1}$$

$$= \frac{\exp(w^T x + b)}{\exp(w^T x + b) + \exp(0^T x + 0)}$$

Thus, it is equivalent to a two-way softmax

with weights $\begin{bmatrix} - & w^T & - \\ - & 0^T & - \end{bmatrix}$ bias $\begin{bmatrix} b \\ 0 \end{bmatrix}$

Problem 3.

Consider a k-way classification. The predicted probability of a sample is $\boldsymbol{y} \in \mathbb{R}^K$, where y_k is the predicted probability of the kth category. Suppose correctly predicting a sample of category k leads to a utility of u_k. Incorrect predictions do not have utilities or losses.

Give the decision rule, i.e., a mapping from $\boldsymbol{y}$ to $\hat{t}$, that maximizes the total expected utility.

3°
$$\mathbb{E}_{t\sim y}[u] = \sum_k y_k u_k \cdot \mathbb{1}\{\hat{t}=k\}$$

To maximize the utility

$$\hat{t} = \arg\max_k \; y_k u_k$$

END OF W7

First (soft) deadline: Nov 9
Second (hard) deadline: Nov 16 before exam

Reference solutions will be released on Nov 11 for students to better prepare for the mid-term.

Students may refer to the provided solutions, but must submit their own written/typed solutions before Nov 16 to get marks. Copy and paste the provided solutions will be considered as plagiarism.