1， 假设你用一个 2 层 hidden layer 的神经网络来解决 K-class 的分类问题。具体定义如下：

$$
\begin{aligned}
\mathbf{z}^{[1]} &= W^{[1]}\mathbf{x} + \mathbf{b}^{[1]} \\
\mathbf{a}^{[1]} &= \mathrm{LeakyReLU}(\mathbf{z}^{[1]}, \alpha = 0.01) \\
\mathbf{z}^{[2]} &= W^{[2]}\mathbf{a}^{[1]} + \mathbf{b}^{[2]} \\
\hat{\mathbf{y}} &= \mathrm{softmax}(\mathbf{z}^{[2]}) \\
L &= -\sum_{i=1}^{K} \mathbf{y}_i \log(\hat{\mathbf{y}}_i)
\end{aligned}
$$

其中输入 $X \in R^{D_x \times 1}$, one-hot encoded 标签$y \in \{0,1\}^K$，$z^{[1]} \in R^{D_a \times 1}$。

softmax 定义如下：

$$\hat{\mathbf{y}} = [\frac{\exp(\mathbf{z}_1^{[2]})}{Z}, \ldots, \frac{\exp(\mathbf{z}_K^{[2]})}{Z}]$$

$$Z = \sum_{j=1}^{K} \exp(\mathbf{z}_j^{[2]})$$

i. 请问$W^{[2]}$和$b^{[2]}$的维度是多少（dimensionality）？如果我们一次给一个 batch 的输入数据 $X \in R^{D_x \times m}$（m 个样本）,请问最后一层输出的结果的维度是多少？

(1)

$$W^{[2]} \in R^{K \times D_a} \quad b^{[2]} \in R^{K \times 1}$$

$$\text{输出维度} K \times m$$

ii. 请计算$\partial \hat{y}_k / \partial z_k^{[2]}$？请尽量用$\hat{y}$里的元素表示最终结果。

(2)

$$
\begin{aligned}
\frac{\partial \hat{y}_k}{\partial z_k^{[2]}} &= \frac{\partial(\frac{e^{z_k^{[2]}}}{\sum_{j=1}^{K} e^{z_j^{[2]}}})}{\partial z_k^{[2]}} \\
&= \frac{\partial(1-\sum_{j\neq k} e^{z_j^{[2]}} \times \frac{1}{\sum_{j=1}^{K} e^{z_j^{[2]}}})}{\partial z_k^{[2]}} \\
&= -\sum_{j\neq k} e^{z_j^{[2]}} \frac{\partial(\frac{1}{\sum_{j=1}^{K} e^{z_j^{[2]}}})}{\partial z_k^{[2]}} \\
&= \sum_{j\neq k} e^{z_j^{[2]}} \frac{e^{z_k^{[2]}}}{Z^2} \\
&= \frac{(Z-e^{z_k^{[2]}})e^{z_k^{[2]}}}{Z^2} \\
&= (1-\hat{y}_k)\hat{y}_k
\end{aligned}
$$

iii. 如果$i \neq k$，请给出$\partial \hat{y}_k / \partial z_i^{[2]}$的计算结果，请尽量用$\hat{y}$里的元素

(3)

$$
\begin{aligned}
\frac{\partial \hat{y}_k}{\partial z_i^{[2]}} &= \frac{\partial(\frac{e^{z_k^{[2]}}}{\sum_{j=1}^{K} e^{z_j^{[2]}}})}{\partial z_i^{[2]}} \\
&= e^{z_k^{[2]}} \frac{\partial(\frac{1}{\sum_{j=1}^{K} e^{z_j^{[2]}}})}{\partial z_i^{[2]}} \\
&= -e^{z_k^{[2]}} \frac{e^{z_i^{[2]}}}{Z^2} \\
&= -\hat{y}_i \hat{y}_k
\end{aligned}
$$

iv. 假设 y 向量里的第 k 个元素值是 1，其它都是 0。请分情况计算$\partial L / \partial z_i^{[2]}$？请尽量用$\hat{y}$里的元素表示最终结果。

(4)

$$
\begin{aligned}
\frac{\partial L}{\partial z_i^{[2]}} &= -\frac{\partial(y_k \log(\hat{y}_k))}{\partial z_i^{[2]}} \\
&= -\frac{\partial \log(\hat{y}_k)}{\partial z_i^{[2]}}
\end{aligned}
$$

若$k = i$：

$$
\begin{aligned}
\frac{\partial L}{\partial z_i^{[2]}} &= -\frac{\partial \log(\hat{y}_i)}{\partial z_i^{[2]}} \\
&= -\frac{\partial (z_i^{[2]} - \log(Z))}{\partial z_i^{[2]}} \\
&= \hat{y}_i - 1
\end{aligned}
$$

若$k \neq i$：

$$
\begin{aligned}
\frac{\partial L}{\partial z^{[2]}} &= -\frac{\partial \log(\hat{y}_k)}{\partial z_i^{[2]}} \\
&= -\frac{\partial (z_k^{[2]} - \log(Z))}{\partial z_i^{[2]}} \\
&= \hat{y}_i
\end{aligned}
$$