解：

（1）梯度的计算

假设输入样本$\vec{x}$ 属于K个类别$Y=\{1,2,\ldots k,\ldots K\}$中的某个类别k时，在Softmax中，我们按照式（1）计算其内积，按照式（2）计算其属于类别j的概率：

$$s_j=\vec{w}_j^T\vec{x} \qquad (1)$$

$$\hat{y}_j=\frac{e^{s_j}}{\sum_k e^{s_k}} \qquad (2)$$

经过Softmax函数后，得到的输出为K个类别的概率列向量：$\hat{Y}=(\hat{y}_1,\ldots\hat{y}_j,\ldots\hat{y}_K)^T$，假设理想的各个类别标签对应的概率为列向量：$Y=\{y_1,\ldots y_j,\ldots y_K\}$，且该列向量的一个元素为1，其他均为0，代表样本属于这个类别。我们选择用交叉熵作为误差函数其为表达式：

$$E_{in}(\vec{w}_k)=-\sum_{k=1}^{K}y_k ln\hat{y}_k=-ln\hat{y}_k \qquad (3)$$

我们可以计算E_{in}对于$\vec{w}_j(j=1,2,\ldots,K)$的梯度：

$$\frac{\partial E_{in}}{\partial\vec{w}_j}=\frac{\partial E_{in}}{\partial\hat{y}_k}\frac{\partial\hat{y}_k}{\partial s_j}\frac{\partial s_j}{\partial\vec{w}_j}=-\frac{1}{\hat{y}_k}\frac{\partial\hat{y}_k}{\partial s_j}\vec{x} \qquad (4)$$

我们再来计算$\frac{\partial\hat{y}_k}{\partial s_j}$：

$$\frac{\partial\hat{y}_k}{\partial s_j}=\frac{\partial}{\partial s_j}\left(\frac{e^{s_k}}{\sum_k e^{s_k}}\right)=\frac{(e^{s_k})'\sum_k e^{s_k}-(\sum_k e^{s_k})'e^{s_k}}{(\sum_k e^{s_k})^2}=$$

$$\begin{cases}\frac{e^{s_j}\sum_k e^{s_k}-e^{s_j}e^{s_k}}{(\sum_k e^{s_k})^2}=\frac{e^{s_j}}{\sum_k e^{s_k}}-\frac{e^{s_j}}{\sum_k e^{s_k}}\frac{e^{s_k}}{\sum_k e^{s_k}}=\hat{y}_j(1-\hat{y}_k) & j=k\\ \frac{0\sum_k e^{s_k}-e^{s_j}e^{s_k}}{(\sum_k e^{s_k})^2}=0-\frac{e^{s_j}}{\sum_k e^{s_k}}\frac{e^{s_k}}{\sum_k e^{s_k}}=-\hat{y}_k\hat{y}_j & j\neq k\end{cases} \qquad (5)$$

将式（5）代入到式（4），我们得到E_{in}对于$\vec{w}_j$的梯度：

$$\frac{\partial E_{in}}{\partial\vec{w}_j}=\frac{\partial E_{in}}{\partial\hat{y}_k}\frac{\partial\hat{y}_k}{\partial s_j}\frac{\partial s_j}{\partial\vec{w}_j}=-\frac{1}{\hat{y}_k}\frac{\partial\hat{y}_k}{\partial s_j}\vec{x}=\begin{cases}(\hat{y}_j-1)\vec{x} & j=k\\ \hat{y}_j\vec{x} & j\neq k\end{cases} \qquad (6)$$

针对N个训练样本，将上述推导及求解过程写成矩阵或向量形式如下：