假设训练样本集有N个样本$\{\vec{x}_1, \dots \vec{x}_n, \dots \vec{x}_N\}$，每个样本有d维特征，写成增广向量后是d+1维，$\vec{x}_n = (x_{n0}, x_{n1}, \dots x_{nd})^T$，所有的训练样本我们用X来表示成一个N*(d+1)维的矩阵：

$$\boldsymbol{X} = \begin{pmatrix} \vec{x}_1^T \\ \vdots \\ \vec{x}_n^T \\ \vdots \\ \vec{x}_N^T \end{pmatrix} = \begin{pmatrix} x_{10} & \cdots & x_{1d} \\ \vdots & \ddots & \vdots \\ x_{N0} & \cdots & x_{Nd} \end{pmatrix} \qquad (7)$$

所有训练样本标签对应的概率输出用N*K维矩阵表示，其中K是类别数，样本只能属于其中一个类别且概率取1，其他类别概率为0，假设如下表示的第一个样本属于类别1，第N个样本属于类别K：

$$\boldsymbol{Y} = \begin{pmatrix} \vec{y}_1 \\ \vdots \\ \vec{y}_n \\ \vdots \\ \vec{y}_N \end{pmatrix} = \begin{pmatrix} y_{11} & \cdots & y_{1K} \\ \vdots & \ddots & \vdots \\ y_{N1} & \cdots & y_{NK} \end{pmatrix} = \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix} \qquad (8)$$

经过式(1)、式(2)后，我们得到的样本类别的概率估计值为N*K维矩阵$\widehat{\boldsymbol{Y}}$：

$$\widehat{\boldsymbol{Y}} = \begin{pmatrix} \vec{\hat{y}}_1 \\ \vdots \\ \vec{\hat{y}}_n \\ \vdots \\ \vec{\hat{y}}_N \end{pmatrix} = \begin{pmatrix} \hat{y}_{11} & \cdots & \hat{y}_{1K} \\ \vdots & \ddots & \vdots \\ \hat{y}_{N1} & \cdots & \hat{y}_{NK} \end{pmatrix} \qquad (9)$$

根据式（6）得到E_{in}的梯度可以写为：

$$\nabla E_{in} = (\widehat{\boldsymbol{Y}} - \boldsymbol{Y})^T \boldsymbol{X} = \left(\vec{\hat{y}}_1 - \vec{y}_1, \dots \vec{\hat{y}}_n - \vec{y}_n, \dots \vec{\hat{y}}_N - \vec{y}_N\right) \begin{pmatrix} \vec{x}_1^T \\ \vdots \\ \vec{x}_n^T \\ \vdots \\ \vec{x}_N^T \end{pmatrix} = \begin{pmatrix} \sum_{n=1}^{N} (\hat{y}_{n1} - y_{n1}) \vec{x}_n^T \\ \vdots \\ \sum_{n=1}^{N} (\hat{y}_{nj} - y_{nj}) \vec{x}_n^T \\ \vdots \\ \sum_{n=1}^{N} (\hat{y}_{nK} - y_{nK}) \vec{x}_n^T \end{pmatrix} \qquad (10)$$

这相当于K*N维的矩阵与N*(d+1)维的矩阵做内积，得到K*(d+1)维的梯度，这里y_{nj}只会取0或者1。

假设类别对应的权系数向量用$\vec{w}$表示，加上常数项，它也是(d+1)维，一共K个类别，可以写成K*(d+1)维矩阵形式：