1. **请写出softmax的解析式。请证明 logistic regression本质上是softmax的一种特殊形式。**
   假设有K个类别，那么softmax regression的输出是：

$$y_i = \frac{\exp(x_i)}{\sum\limits_{j=1}^{K} \exp(x_j)}$$

   如果K=2，那么可以将两个类别分别记为0和1，那么有：

$$\begin{aligned} y_0 &= \frac{\exp(x_0)}{\exp(x_0) + \exp(x_1)} \\ y_1 &= \frac{\exp(x_1)}{\exp(x_0) + \exp(x_1)} \end{aligned}$$

   令 $z = x_0 - x_1$，可以得到：

$$\begin{aligned} y_0 &= \frac{1}{1 + \exp(-z)} \\ y_1 &= 1 - y_0 \end{aligned}$$

   这就是logistic regression的输出形式。因此，当K=2时，softmax regression就等价于logistic regression。
2. **请推导discriminative model的 log p(y|X) 对参数theta的梯度的形式，解释为什么可以称为learning from errors。见Page 21。**

$$\begin{aligned} Z(\theta) &= \sum_k \exp\left(f_\theta^{(k)}(X)\right) \\ p(y|X) &= \frac{\exp(f_\theta^{(k)}(X))}{Z(\theta)} \\ \frac{\partial}{\partial\theta}\log p(y|X) &= \frac{\partial}{\partial\theta}\left(f_\theta^{(k)}(X) - \log\left(Z(\theta)\right)\right) \\ &= \frac{\partial}{\partial\theta} f_\theta^{(k)}(X) - \frac{1}{Z(\theta)}\frac{\partial}{\partial\theta} Z(\theta) \\ &= \frac{\partial}{\partial\theta} f_\theta^{(k)}(X) - \frac{1}{Z(\theta)}\frac{\partial}{\partial\theta}\left(\sum_k \exp\left(f_\theta^{(k)}(X)\right)\right) \\ &= \frac{\partial}{\partial\theta} f_\theta^{(k)}(X) - \sum_{k'} \frac{1}{Z(\theta)}\frac{\partial}{\partial\theta}\exp\left(f_\theta^{(k')}(X)\right) \\ &= \frac{\partial}{\partial\theta} f_\theta^{(k)}(X) - \sum_{k'} \frac{\exp\left(f_\theta^{(k')}(X)\right)}{Z(\theta)}\frac{\partial}{\partial\theta} f_\theta^{(k')}(X) \\ &= \frac{\partial}{\partial\theta} f_\theta^{(k)}(X) - \sum_{k'} p_{k'}\frac{\partial}{\partial\theta} f_\theta^{(k')}(X) \\ &= \sum_{k'} \left(1(k = k') - p_{k'}\right)\frac{\partial}{\partial\theta} f_\theta^{(k')}(X) \\ &= \frac{\partial}{\partial\theta} f_\theta(X)^\top (Y - p) \\ &= \frac{\partial}{\partial\theta} f_\theta(X)^\top (Y - \mathrm{E}_\theta(Y|X)) \end{aligned}$$

   其中 Y 是 one-hot 向量，$k' = k$ 时为 1 其他为 0。
   梯度结果中 $Y - p$ 表示了真实结果和模型输出之间的差，因此是 learning from errors。