1，有人说当批量大小为 1 时基于随机梯度下降法（Stochastic Gradient Descent，SGD）的逻辑斯蒂回归（Logistic Regression）算法可以被看作为“软性”的感知器算法（PLA），你认同这个说法吗？请给出你的理由。

答：同意①采用批量为1的SGD优化逻辑斯蒂回归，其权重更新方式与PLA权重更新方式类似②PLA与逻辑斯蒂回归均适用于二分类问题③逻辑斯蒂回归3所以被称为"软性"，是因为它输为[0,1]间的连续值，表征属于某类别的概率

2，在 Logistic regression 中当标签 y={+1,-1}时常用交叉熵作为损失函数：$L_{in}(\boldsymbol{w}) = \frac{1}{N}\sum_1^N \ln\left(1+\exp\left(-y_n \boldsymbol{w}^T \mathbf{x}_n\right)\right)$，请推导出该函数的梯度表达式

解：$\frac{\partial L_{in}(w)}{\partial w} = \frac{1}{N}\sum_{n=1}^{N} \frac{\exp(-y_n w^T x_n)}{1+\exp(-y_n w^T x_n)} \cdot (-y_n x_n) = \frac{1}{N}\sum_{n=0}^{N} \theta(-y_n w^T x_n)(-y_n x_n)$

3，为什么在 Logistic Regression 中不用$L_{in}(\boldsymbol{w}) = (\theta(y\boldsymbol{w}^T\mathbf{x}) - 1)^2$作为损失函数，这里假设$\theta(.)$是 *Sigmoid* 函数，标签 y={+1, −1}。

解：使用MSE作损失函数时，其梯度

$$\frac{\partial L_{in}(w)}{\partial w} = 2(\theta(yw^T x)-1)\cdot \frac{\partial \theta(yw^T x)}{\partial w}$$

而其中 $\frac{\partial \theta(yw^T x)}{\partial w}$ 项在 $yw^T x \to -\infty$ 与 $yw^T x \to +\infty$ 时均趋于0，使梯度极平稳，非常不利于使用梯度下降计算