Lecture5 习题作业

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。

解：进行二分类，标签为+1 和-1 时，上述说法正确。

Logistic Regression 算法在利用随机梯度下降法的权向量更新表达式为：$\boldsymbol{w}_{t+1} \leftarrow \boldsymbol{w}_t - \eta\theta(-y_n\boldsymbol{w}_t^T\mathbf{x}_n)(-y_n\mathbf{x}_n)$

感知器算法（PLA）的权向量更新表达式为：

$$\boldsymbol{w}_{t+1} \leftarrow \boldsymbol{w}_t + [\![\mathrm{sign}(\boldsymbol{w}_t^T\mathbf{x}_{n(t)}) \neq y_n]\!]\mathbf{x}_{n(t)}$$

当$\eta = 1$时, 逻辑斯蒂回归中的 Sigmoid 函数取值在 0 和 1 之间，而 PLA 的 BOOL 表达式取值不是 0 就是 1，所以，可以认为前者是“软性”的 PLA。

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

解：$L_{in} = \ln(1+exp(-y_n\boldsymbol{w}^T\mathbf{x}_n))$，

$$\frac{\partial L_{in}(\boldsymbol{w},\mathbf{x},y)}{\partial w_i} = \frac{\partial \ln(1+exp(-y_n\boldsymbol{w}^T\mathbf{x}_n))}{\partial(1+exp(-y_n\boldsymbol{w}^T\mathbf{x}_n))}\frac{\partial(1+\exp(-y_n\boldsymbol{w}^T\mathbf{x}_n))}{\partial(-y_n\boldsymbol{w}^T\mathbf{x}_n)}\frac{\partial(-y\boldsymbol{w}^T\mathbf{x})}{\partial w_i}$$

$$= \frac{1}{1+exp(-y_n\boldsymbol{w}^T\mathbf{x}_n)}\exp(-y_n\boldsymbol{w}^T\mathbf{x}_n)(-yx_i)$$

$$= \frac{\exp(-y_n\boldsymbol{w}^T\mathbf{x}_n)}{1+\exp(-y_n\boldsymbol{w}^T\mathbf{x}_n)}(-yx_i)$$

$$\nabla L_{in}(\boldsymbol{w},\mathbf{x},y) = \theta(-y\boldsymbol{w}^T\mathbf{x})(-y\mathbf{x})$$

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：$L_{in}(\boldsymbol{w})=(\theta(y\boldsymbol{w}^T\mathbf{x})-1)^2$

$$\frac{\partial L_{in}(\boldsymbol{w},\mathbf{x},y)}{\partial w_i}=2(\theta(y\boldsymbol{w}^T\mathbf{x})-1)\theta(y\boldsymbol{w}^T\mathbf{x})(1\text{-}\theta(y\boldsymbol{w}^T\mathbf{x}))yx_i$$

$$if\ \ (y\boldsymbol{w}^T\mathbf{x}))>0\quad \nabla L_{in}(\boldsymbol{w},\mathbf{x},\ y)=0$$

$$if\ \ (y\boldsymbol{w}^T\mathbf{x}))<0\quad \nabla L_{in}(\boldsymbol{w},\mathbf{x},y)=0$$

无论分类正确与否，梯度都为 0，影响学习性能。