逻辑斯蒂回归习题解答

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：

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

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

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

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

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

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

3， 什么情况下朴素贝叶斯模型预测+1 类的概率可写成：$P(y=1|\mathbf{x}_n)=\frac{1}{1+\exp(-u)}$的形式（其中，$u=\boldsymbol{w}^T\mathbf{x}_n+w_0$）？与逻辑斯蒂回归相比较，两者在模型的形式上相似，差异体现在哪里呢？

解：在朴素贝叶斯模型中，+1 类的概率为：

$$P(y=1|\mathbf{x}_n)=\frac{p(\mathbf{x}_n|y=1)P(y=1)}{p(\mathbf{x}_n|y=-1)P(y=-1)+p(\mathbf{x}_n|y=1)P(y=1)}$$

$$=\frac{1}{1+\dfrac{p(\mathbf{x}_n|y=-1)P(y=-1)}{p(\mathbf{x}_n|y=1)P(y=1)}}$$

令：$u=log\frac{p(\mathbf{x}_n|y=1)P(y=1)}{p(\mathbf{x}_n|y=-1)P(y=-1)}$

则：$P(y=1|\mathbf{x}_n)=\frac{1}{1+\exp(-u)}$

当：$p(\mathbf{x}_n|y=1,\,\boldsymbol{w})=\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma})$，$p(\mathbf{x}_n|y=-1,\,\boldsymbol{w})=\mathcal{N}(\boldsymbol{\mu}_{-1},\boldsymbol{\Sigma})$

即：$p(\mathbf{x}_n|y=i,\,\boldsymbol{w})=\frac{1}{(2\boldsymbol{\pi})^{d/2}|\boldsymbol{\Sigma}|^{1/2}}\exp(-\frac{1}{2}(\mathbf{x}_n-\boldsymbol{\mu_i})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu_i}))$

因此：$u=log\frac{\frac{1}{(2\boldsymbol{\pi})^{\frac{d}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\mathbf{x}_n-\boldsymbol{\mu_1})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu_1})\right)P(y=1)}{\frac{1}{(2\boldsymbol{\pi})^{\frac{d}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\mathbf{x}_n-\boldsymbol{\mu_{-1}})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu_{-1}})\right)P(y=-1)}$

$$=-\frac{1}{2}(\mathbf{x}_n-\boldsymbol{\mu_1})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu_1})+logP(y=1)$$

$$+\frac{1}{2}(\mathbf{x}_n-\boldsymbol{\mu_{-1}})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu_{-1}})-logP(y=-1)$$

$$=-\frac{1}{2}(\mathbf{x}_n-\boldsymbol{\mu_1})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu_1})+\frac{1}{2}(\mathbf{x}_n-\boldsymbol{\mu_{-1}})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu_{-1}})$$

$$+log\frac{P(y=1)}{P(y=-1)}$$

$$= -\frac{1}{2}(\mathbf{x}_n^T - \boldsymbol{\mu}_1^T)\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x}_n^T - \boldsymbol{\mu}_{-1}^T)\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_{-1})$$

$$+ \log\frac{P(y=1)}{P(y=-1)}$$

$$= -\frac{1}{2}(\mathbf{x}_n^T\boldsymbol{\Sigma}^{-1} - \boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1})(\mathbf{x}_n - \boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x}_n^T\boldsymbol{\Sigma}^{-1} - \boldsymbol{\mu}_{-1}^T\boldsymbol{\Sigma}^{-1})(\mathbf{x}_n - \boldsymbol{\mu}_{-1})$$

$$+ \log\frac{P(y=1)}{P(y=-1)}$$

$$= -\frac{1}{2}(\cancel{\mathbf{x}_n^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_n} - \mathbf{x}_n^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_n + \boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1) + \frac{1}{2}(\cancel{\mathbf{x}_n^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_n}$$

$$- \mathbf{x}_n^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{-1} - \boldsymbol{\mu}_{-1}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_n + \boldsymbol{\mu}_{-1}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{-1})$$

$$+ \log\frac{P(y=1)}{P(y=-1)}$$

$$= \frac{1}{2}\mathbf{x}_n^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_{-1}) + \frac{1}{2}(\boldsymbol{\mu}_1^T - \boldsymbol{\mu}_{-1}^T)\boldsymbol{\Sigma}^{-1}\mathbf{x}_n - \frac{1}{2}\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1$$

$$+ \frac{1}{2}\boldsymbol{\mu}_{-1}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{-1} + \log\frac{P(y=1)}{P(y=-1)}$$

$$= \frac{1}{2}(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_{-1}))^T\mathbf{x}_n + \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_{-1})^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_n - \frac{1}{2}\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1$$

$$+ \frac{1}{2}\boldsymbol{\mu}_{-1}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{-1} + \log\frac{P(y=1)}{P(y=-1)}$$

由于：$(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_{-1}))^T = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_{-1})^T(\boldsymbol{\Sigma}^{-1})^T$

又由于：$\boldsymbol{\Sigma}^{-1}$是对称阵，所以$(\boldsymbol{\Sigma}^{-1})^T = \boldsymbol{\Sigma}^{-1}$

则：

$$u = \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_{-1})^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_n + \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_{-1})^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_n - \frac{1}{2}\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1$$

$$+ \frac{1}{2}\boldsymbol{\mu}_{-1}^T\boldsymbol{\Sigma}^{-1}\mu_{-1} + \log\frac{P(y=1)}{P(y=-1)}$$

$$= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_{-1})^T \boldsymbol{\Sigma}^{-1} \mathbf{x}_n - \frac{1}{2} \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_{\mathbf{1}} + \frac{1}{2} \boldsymbol{\mu}_{-1}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_{\mathbf{-1}} + log \frac{P(y=1)}{P(y=-1)}$$

假设：$\boldsymbol{w} = \boldsymbol{\Sigma}^{\mathbf{-1}} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_{-1})$，

$$w_0 = -\frac{1}{2} \boldsymbol{\mu}_{\mathbf{1}}^{\boldsymbol{T}} \boldsymbol{\Sigma}^{\mathbf{-1}} \boldsymbol{\mu}_{\mathbf{1}} + \frac{1}{2} \boldsymbol{\mu}_{\mathbf{-1}}^{\boldsymbol{T}} \boldsymbol{\Sigma}^{\mathbf{-1}} \boldsymbol{\mu}_{\mathbf{-1}} + log \frac{P(y=1)}{P(y=-1)}$$

则：$u = \boldsymbol{w}^T \mathbf{x}_n + w_0$，且：$P(y = 1|\mathbf{x}_n) = \frac{1}{1+exp(-u)}$

对于逻辑斯蒂回归：

$$P(y = 1|\mathbf{x}_n) = h(\mathbf{x}_n) = \theta(\boldsymbol{w}^T \mathbf{x}_n + w_0) = \frac{1}{1 + exp(-(\boldsymbol{w}^T \mathbf{x}_n + w_0))}$$

$$= \frac{1}{1+exp(-u)}$$

可见，两者形式上是相似的。但在实际应用中却完全不同，对于朴素贝叶斯模型来说，需要学习$p(\mathbf{x}_n|y = i, \boldsymbol{w})$和$P(y = i)$，或者说需要学习$P(\mathbf{x}_n, y)$联合分布，这属于生成式模型范畴。但是，对于逻辑斯蒂回归而言，这只需要直接建模$P(y|\mathbf{x}_n)$，或者直接寻找输入样本和输出类别间的映射关系，属于判别式模型。