IML 第三次作业

习题 5.1

如果用线性函数，则每一层的输出都是上一层的线性模型，最终的输出也只是复杂的线性模型，无法表示复杂的非线性关系。

作业 5.2

当 x 很大时，$\exp(x)$ 的结果可能发生溢出而显示 NaN；假设 $x_j(1 \le j \le C)$ 的最大值为 x^*，可以如下处理

$$\frac{\exp(x_i)}{\sum_{j=1}^{C}\exp(x_j)} = \frac{\exp(x_i)\exp(-x^*)}{\exp(-x^*)\sum_{j=1}^{C}\exp(x_j)} = \frac{\exp(x_i - x^*)}{\sum_{j=1}^{C}\exp(x_j - x^*)}$$

这样可以避免数值上溢问题。

同样地，对于第二个式子，可以如下处理

$$\log\sum_{j=1}^{C}\exp(x_j) = \log\left[\exp(x^*)\sum_{j=1}^{C}\exp(x_j - x^*)\right] = x^* + \log\sum_{j=1}^{C}\exp(x_j - x^*)$$

作业 5.3

令 $f(x_i) = \dfrac{\exp(x_i)}{\sum_{j=1}^{C}\exp(x_j)}$， $g(x_i) = \log f(x_i)$。

首先

$$\frac{\partial f(x_i)}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial f(x_i)}{\partial x_1} & \frac{\partial f(x_i)}{\partial x_2} & \cdots & \frac{\partial f(x_i)}{\partial x_C} \end{bmatrix}$$

当 $i = k$ 时，有

$$\frac{\partial f(x_i)}{\partial x_k} = \frac{\exp(x_k)\sum_{j\neq k}^{C}\exp(x_j)}{\left[\sum_{j=1}^{C}\exp(x_j)\right]^2}$$

当 $i \neq k$ 时，有

$$\frac{\partial f(x_i)}{\partial x_k} = -\frac{\exp(x_i)\exp(x_k)}{\left[\sum_{j=1}^{C}\exp(x_j)\right]^2}$$

其次

$$\frac{\partial g(x_i)}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial g(x_i)}{\partial x_1} & \frac{\partial g(x_i)}{\partial x_2} & \cdots & \frac{\partial g(x_i)}{\partial x_C} \end{bmatrix}$$

由链式法则（视 log 以 e 为底）

$$\frac{\partial g(x_i)}{\partial x_k} = \frac{1}{f(x_i)} \cdot \frac{\partial f(x_i)}{\partial x_k}$$

当 $i = k$ 时，有

$$\frac{\partial g(x_i)}{\partial x_k} = \frac{\sum_{j \neq k}^{C} \exp(x_j)}{\sum_{j=1}^{C} \exp(x_j)}$$

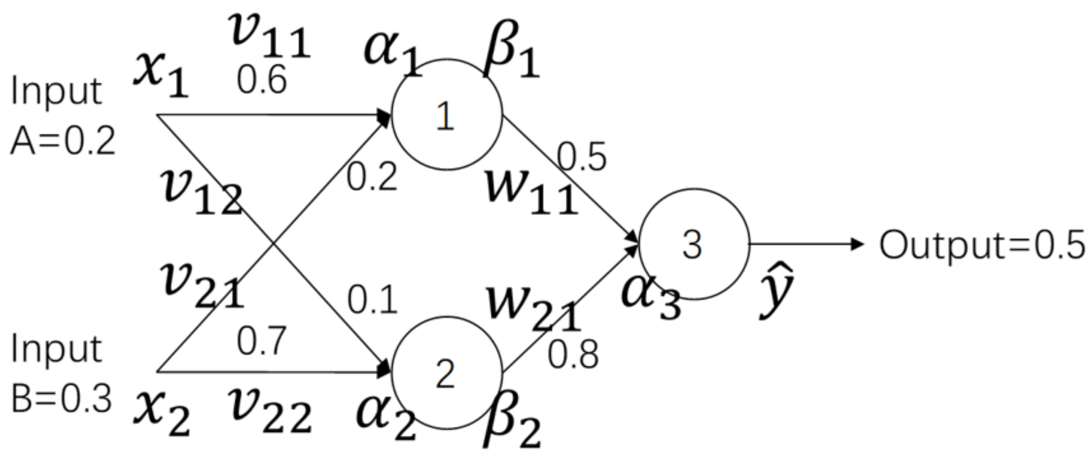
当 $i \neq k$ 时，有

$$\frac{\partial g(x_i)}{\partial x_k} = -\frac{\exp(x_i)}{\sum_{j=1}^{C} \exp(x_j)}$$

作业 5.4

$\text{ReLU}(x) = \max(0,x)$，其导数为

$$\text{ReLU}'(x) = \mathbb{I}(x > 0)$$



参数更新前，各部分输入输出结果为

$$\alpha_1 = v_{11}x_1 + v_{21}x_2 = 0.18$$
$$\beta_1 = \text{ReLU}\left(\alpha_1\right) = 0.18$$
$$\alpha_2 = v_{12}x_1 + v_{22}x_2 = 0.23$$
$$\beta_2 = \text{ReLU}\left(\alpha_2\right) = 0.23$$
$$\alpha_3 = w_{11}\beta_1 + w_{21}\beta_2 = 0.274$$
$$\hat{y} = \text{ReLU}\left(\alpha_3\right) = 0.274$$
$$E = \frac{1}{2}(y - \hat{y})^2 = 0.026$$

计算梯度项

$$
\begin{aligned}
\frac{\partial E}{\partial v_{11}} &= -\left(y-\hat{y}\right)\mathrm{ReLU}'\left(\alpha_3\right)w_{11}\,\mathrm{ReLU}'\left(\alpha_1\right)x_1 = -0.023 \\
\frac{\partial E}{\partial v_{12}} &= -\left(y-\hat{y}\right)\mathrm{ReLU}'\left(\alpha_3\right)w_{11}\,\mathrm{ReLU}'\left(\alpha_2\right)x_1 = -0.023 \\
\frac{\partial E}{\partial v_{21}} &= -\left(y-\hat{y}\right)\mathrm{ReLU}'\left(\alpha_3\right)w_{21}\,\mathrm{ReLU}'\left(\alpha_1\right)x_2 = -0.054 \\
\frac{\partial E}{\partial v_{22}} &= -\left(y-\hat{y}\right)\mathrm{ReLU}'\left(\alpha_3\right)w_{21}\,\mathrm{ReLU}'\left(\alpha_2\right)x_2 = -0.054
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial E}{\partial w_{11}} &= -(y-\hat{y})\,\mathrm{ReLU}'\left(\alpha_3\right)\beta_1 = -0.041 \\
\frac{\partial E}{\partial w_{21}} &= -(y-\hat{y})\,\mathrm{ReLU}'\left(\alpha_3\right)\beta_2 = -0.052
\end{aligned}
$$

因为学习率 $\eta = 1$, 更新参数:

$$
\begin{aligned}
v_{11} &\leftarrow v_{11} - \frac{\partial E}{\partial v_{11}} = 0.623 \\
v_{12} &\leftarrow v_{12} - \frac{\partial E}{\partial v_{12}} = 0.123 \\
v_{21} &\leftarrow v_{21} - \frac{\partial E}{\partial v_{21}} = 0.254 \\
v_{22} &\leftarrow v_{22} - \frac{\partial E}{\partial v_{22}} = 0.754 \\
w_{11} &\leftarrow w_{11} - \frac{\partial E}{\partial w_{11}} = 0.541 \\
w_{21} &\leftarrow w_{21} - \frac{\partial E}{\partial w_{21}} = 0.852
\end{aligned}
$$

更新参数后

$$\alpha_1 = v_{11}x_1 + v_{21}x_2 = 0.201$$

$$\beta_1 = \mathrm{ReLU}\left(\alpha_1\right) = 0.201$$

$$\alpha_2 = v_{12}x_1 + v_{22}x_2 = 0.251$$

$$\beta_2 = \mathrm{ReLU}\left(\alpha_2\right) = 0.251$$

$$\alpha_3 = w_{11}\beta_1 + w_{21}\beta_2 = 0.221$$

$$\hat{y} = \mathrm{ReLU}\left(\alpha_3\right) = 0.323$$

$$E = \frac{1}{2}(y-\hat{y})^2 = 0.016$$

可见参数更新使平方损失值下降了。

习题 6.4

线性判别分析可以解决多分类问题，而线性核支持向量机只能解决二分类问题。并且线性判别分析只有处理线性可分样本时才工作地比较好。所以二者在处理二分类问题时且两类样本线性可分时等价。

习题 6.6

SVM 训练出来的模型只和稀疏的支持向量有关，但是若噪声出现在这些支持向量中，就会参与到优化问题的最大化中，即对 SVM 的最终模型造成很大的影响。

习题 6.9

书中式 (6.29) 如下：

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m} \ell_{0/1}\left(y_i\left(\boldsymbol{w}^\top \boldsymbol{x}_i + b\right) - 1\right)$$

用对率损失 $\ell_{log}(z) = \log(1+\exp(-z))$ 替代 $\ell_{0/1}$，令 $z = y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)$，则式 (6.29) 变为

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m} \log(1+\exp(-z))$$

这是一个带 L_2 正则项的正则化问题，根据表示定理，其解总可以写成

$$\boldsymbol{w}_* = \sum_{i=1}^{m} \alpha_i \phi(\boldsymbol{x}_i)$$

按书上对对偶问题的推导，其对偶问题为

$$\min_{\boldsymbol{\alpha}} \sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i\alpha_j\kappa\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) + C\sum_{i=1}^{m} \log\left[1+\exp\left(-y_i\sum_{j=1}^{m}\alpha_j\kappa\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)\right)\right]$$

上面这个问题没有约束，可用 GD 算法求解后得到 $\boldsymbol{w}_*$,b_*。

作业 6.4

令 $\boldsymbol{\alpha}^* = \begin{pmatrix} \boldsymbol{\alpha} \\ \hat{\boldsymbol{\alpha}} \end{pmatrix}$, 则

$$\begin{aligned}
\sum_{i=1}^{m}\sum_{j=1}^{m}\left(\alpha_i - \hat{\alpha}_i\right)\left(\alpha_j - \hat{\alpha}_j\right)\kappa\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) &= \sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i\alpha_j k_{ij} - \widehat{\alpha}_i\alpha_j k_{ij} - \alpha_i\widehat{\alpha}_j k_{ij} + \widehat{\alpha}_i\widehat{\alpha}_j k_{ij} \\
&= \boldsymbol{\alpha}^{*\top}\begin{bmatrix} k & -k \\ -k & k \end{bmatrix}\boldsymbol{\alpha}^*
\end{aligned}$$

令 $\boldsymbol{v}=\begin{pmatrix} -\boldsymbol{y}-\epsilon \\ \boldsymbol{y}-\epsilon \end{pmatrix}$，则有

$$\sum_{i=1}^{m}\left(y_i\left(\hat{\alpha}_i-\alpha_i\right)-\epsilon\left(\hat{\alpha}_i+\alpha_i\right)\right)=\boldsymbol{\alpha}^{*\top}\boldsymbol{v}$$

因此原式可化为

$$\max_{\boldsymbol{\alpha}^*} g\left(\boldsymbol{\alpha}^*\right)=\boldsymbol{\alpha}^{*\top}\boldsymbol{v}-\frac{1}{2}\boldsymbol{\alpha}^{*\top}\boldsymbol{K}\boldsymbol{\alpha}^*$$
$$\text{s.t.}\quad C\succcurlyeq\boldsymbol{\alpha}^*\succcurlyeq 0,\boldsymbol{\alpha}^{*\top}\boldsymbol{v}=0$$

这里 $\boldsymbol{K}=\begin{bmatrix} k & -k \\ -k & k \end{bmatrix}$。

作业 6.5

简单起见，不妨令两个变量为 $\boldsymbol{x}$,$\boldsymbol{y}$，并且他们都是 n 维的

$$\begin{aligned}\kappa(\boldsymbol{x},\boldsymbol{y})&=\left(\sum_{i=1}^{n}x_iy_i\right)^2\\&=\sum_{i=1}^{n}\left(x_i^2\right)\left(y_i^2\right)+\sum_{i=2}^{n}\sum_{j=1}^{i-1}\left(\sqrt{2}x_ix_j\right)\left(\sqrt{2}y_iy_j\right)\end{aligned}$$

所以

$$\phi(\boldsymbol{x})=\left(x_n^2,\ldots,x_1^2,\sqrt{2}x_nx_{n-1},\ldots,\sqrt{2}x_nx_1,\sqrt{2}x_{n-1}x_{n-2},\ldots,\sqrt{2}x_2x_1\right)$$