Lecture1 作业

1，现在很多火车站都可以持身份证和车票“刷脸”进站，张同学拿着自己的身份证进站时机器却识别错误，从模式识别的角度看机器难以识别他的原因是：

（a）不存在模式；（b）没有训练样本；

（c）不同类的模式相似度高；（d）同一类的模式差异性大。

2，某研究人员用带有标签的数据库训练其目标识别算法，希望用在自动驾驶中对路况进行场景理解和行人检测，请问这个算法需要去解决哪种类型的学习问题：

（a）监督学习；（b）无监督学习；

（c）半监督学习；（d）增强学习。

3，如果一个好的学习算法通过训练样本集在解空间（假设空间）集合中找到一个最优解，使得其对所有训练样本都能够实现正确的模式识别，以下哪种说法正确：

（a）目标函数是已知的；

（b）所有训练样本都是线性可分的；

（c）训练样本集上全部分类正确，不能代表测试时也能正确；

（d）好的学习算法能够容许训练样本存在错误的标签。

4，某个电影网站每名网络用户都可以给电影打分，用户有 id 号，电影也有 id 号，网站已积累了 10 万个用户对所看电影喜爱程度的打分值，并将电影的特征如喜剧片、动作片、爱情片、…、明星甲、明星乙等与观众对这些特征的喜好匹配组合到一起设计了推荐系统，给每部电影一个百分制的评分结果。请用上述这些元素去构造一个学习问题。

解：

$S_1 = [0，100]$

$S_2 =$ 所有由用户 id 号和电影 id 号组成的向量：$[userid，movied]$

$S_3 =$ 某种数学模型，将电影的特征如喜剧片、动作片、爱情片、。。。、明星甲、明星乙等与观众对这些特征的喜好匹配组合到一起打分

$S_4 =$ 10 万个用户对他们所看过电影喜爱程度的一一评价结果：

$[(userid，movied)，rating]$

学习问题：$S_1 = Y，S_2 = X，S_3 = H，S_4 = D$

5，有一数据集共有 2000 张花卉图片，其中 1400 张是玫瑰，300 张是月季，300 张是蔷薇。某同学设计了一个玫瑰识别算法，从 2000 张图片中识别出 1000 张图片为玫瑰，但实际上其中只有 600 张是玫瑰，另外 300 张是月季、100 张是蔷薇。请计算分类正确率 (Accuracy)、分类错误率 (Error rate)、分类精度 (Precision)、召回率 (Recall)、F1 分数 (F1 Score)。如果该同学的算法把 2000 张图片都识别成玫瑰，请再次计算出上述指标。

解：（1）混淆矩阵为：

		预测结果	
		正 1000	负 1000
真 实结果	正 1400	TP 600	FN 800
	负 600	FP 400	TN 200

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} = \frac{600 + 200}{2000} = 40\%$$

$$Error = 1 - Accuracy = 1 - 40\% = 60\%$$

$$Precision = \frac{TP}{TP + FP} = \frac{600}{600 + 400} = 60\%$$

$$Recall = \frac{TP}{TP + FN} = \frac{600}{600 + 800} = 43\%$$

$$\text{F1 Score} = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * 0.6 * 0.43}{0.6 + 0.43} = 0.5$$

（2）混淆矩阵为：

		预测结果	
		正 2000	负 0
真 实结果	正 1400	TP 1400	FN 0
	负 600	FP 600	TN 0

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} = \frac{1400 + 0}{2000} = 70\%$$

$$Error = 1 - Accuracy = 1 - 70\% = 30\%$$

$$Precision = \frac{TP}{TP + FP} = \frac{1400}{1400 + 600} = 70\%$$

$$Recall = \frac{TP}{TP + FN} = \frac{1400}{1400 + 0} = 100\%$$

$$\text{F1 Score} = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * 0.7 * 1}{0.7 + 1} = 0.82$$

Lecture2 作业

1，假设训练样本集为$D=\{(\mathbf{x}_1,y_1)=((3,3)^T,1),(\mathbf{x}_2,y_2)=((4,3)^T,1),(\mathbf{x}_3,y_3)=((1,1)^T,-1)\}$, 使用感知器算法设计分类面，并判断测试样本$\mathbf{x}=(0,1)^T$属于哪个类别。

解：

样本增广后为：$\vec{x}_1=(1,3,3)^T$，$y_1=1$，$\vec{x}_2=(1,4,3)^T$，$y_2=1$，$\vec{x}_3=(1,1,1)^T$，$y_3=-1$

初始化权重：$\vec{w}^{(0)}=(0,0,0)^T$

$sign\left(\vec{w}^{(0)T}\vec{x}_1\right)=0\neq y_1$， $\therefore\ \vec{w}^{(1)}=\vec{w}^{(0)}+y_1\vec{x}_1=(1,3,3)^T$，

$sign\left(\vec{w}^{(1)T}\vec{x}_2\right)=1=y_2$， $\therefore\ \vec{w}^{(2)}=\vec{w}^{(1)}=(1,3,3)^T$

$sign\left(\vec{w}^{(2)T}\vec{x}_3\right)=1\neq y_3$， $\therefore\ \vec{w}^{(3)}=\vec{w}^{(2)}+y_3\vec{x}_3=(0,2,2)^T$

$sign\left(\vec{w}^{(3)T}\vec{x}_1\right)=1=y_1$， $\therefore\ \vec{w}^{(4)}=\vec{w}^{(3)}=(0,2,2)^T$

$sign\left(\vec{w}^{(4)T}\vec{x}_2\right)=1=y_2$， $\therefore\ \vec{w}^{(5)}=\vec{w}^{(4)}=(0,2,2)^T$

$sign\left(\vec{w}^{(5)T}\vec{x}_3\right)=1\neq y_3$， $\therefore\ \vec{w}^{(6)}=\vec{w}^{(5)}+y_3\vec{x}_3=(-1,1,1)^T$

$sign\left(\vec{w}^{(6)T}\vec{x}_1\right)=1=y_1$， $\therefore\ \vec{w}^{(7)}=\vec{w}^{(6)}=(-1,1,1)^T$

$sign\left(\vec{w}^{(7)T}\vec{x}_2\right)=1=y_2$， $\therefore\ \vec{w}^{(8)}=\vec{w}^{(7)}=(-1,1,1)^T$

$sign\left(\vec{w}^{(8)T}\vec{x}_3\right)=1\neq y_3$， $\therefore\ \vec{w}^{(9)}=\vec{w}^{(8)}+y_3\vec{x}_3=(-2,0,0)^T$

$sign\left(\vec{w}^{(9)T}\vec{x}_1\right)=-1\neq y_1$， $\therefore\ \vec{w}^{(10)}=\vec{w}^{(9)}+y_1\vec{x}_1=(-1,3,3)^T$

$sign\left(\vec{w}^{(10)T}\vec{x}_2\right)=1=y_2$， $\therefore\ \vec{w}^{(11)}=\vec{w}^{(10)}=(-1,3,3)^T$

$sign\left(\vec{w}^{(11)T}\vec{x}_3\right)=1\neq y_3$， $\therefore\ \vec{w}^{(12)}=\vec{w}^{(11)}+y_3\vec{x}_3=(-2,2,2)^T$

$sign\left(\vec{w}^{(12)T}\vec{x}_1\right) = 1 = y_1$， $\therefore\ \vec{w}^{(13)} = \vec{w}^{(12)} = (-2,2,2)^T$

$sign\left(\vec{w}^{(13)T}\vec{x}_2\right) = 1 = y_2$， $\therefore\ \vec{w}^{(14)} = \vec{w}^{(13)} = (-2,2,2)^T$

$sign\left(\vec{w}^{(14)T}\vec{x}_3\right) = 1 \neq y_3$， $\therefore\ \vec{w}^{(15)} = \vec{w}^{(14)} + y_3\vec{x}_3 = (-3,1,1)^T$

$sign\left(\vec{w}^{(15)T}\vec{x}_1\right) = 1 = y_1$， $\therefore\ \vec{w}^{(16)} = \vec{w}^{(15)} = (-3,1,1)^T$

$sign\left(\vec{w}^{(16)T}\vec{x}_2\right) = 1 = y_2$， $\therefore\ \vec{w}^{(17)} = \vec{w}^{(16)} = (-3,1,1)^T$

$sign\left(\vec{w}^{(17)T}\vec{x}_3\right) = -1 = y_3$， $\therefore\ \vec{w}^{(18)} = \vec{w}^{(17)} = (-3,1,1)^T$

$\therefore\ \vec{w} = (-3,1,1)^T$, 分类面为：$x_1 + x_2 - 3 = 0$

对测试样本进行增广，$\vec{\mathrm{x}} = (1,0,1)^T$，

$$sign\left(\vec{w}^T\vec{\mathrm{x}}\right) = sign((-3,1,1)(1,0,1)^T = -1, \quad \therefore\ \vec{x} \in -1\ 类$$

2，对于感知器算法（PLA），假设第 t 次迭代时，选择的是第 n 个样本：$\mathrm{sign}(\boldsymbol{w}^T\mathbf{x}_n) \neq y_n$，$\boldsymbol{w}_{t+1} \leftarrow \boldsymbol{w}_t + y_n\mathbf{x}_n$，下述那个式子正确？

（a）$\boldsymbol{w}_{t+1}^{\boldsymbol{T}}\mathbf{x}_n = y_n$

（b）$\mathrm{sign}\left(\boldsymbol{w}_{t+1}^{\boldsymbol{T}}\mathbf{x}_n\right) = y_n$

（c）$y_n\boldsymbol{w}_{t+1}^{\boldsymbol{T}}\mathbf{x}_n \geq y_n\boldsymbol{w}_t^{\boldsymbol{T}}\mathbf{x}_n$

（d）$y_n\boldsymbol{w}_{t+1}^{\boldsymbol{T}}\mathbf{x}_n < y_n\boldsymbol{w}_t^{\boldsymbol{T}}\mathbf{x}_n$

3，证明：针对线性可分训练样本集，PLA 算法中，当$\mathbf{W}_0 = \mathbf{0}$，在对分错样本进行了 T 次纠正后，下式成立：$\frac{\mathbf{w}_f^T}{\|\mathrm{w}_f\|}\frac{\mathbf{w}_T}{\|\mathbf{w}_T\|} \geq \sqrt{T} \cdot constant$

证明：由于

$$\begin{aligned}\boldsymbol{W}_f^T\boldsymbol{W}_{t+1} &= \boldsymbol{W}_f^T(\boldsymbol{W}_t + y_n(t)\boldsymbol{X}_n(t)) \\ &\geqslant \boldsymbol{W}_f^T\boldsymbol{W}_t + \min_n y_n(t)\boldsymbol{W}_f^T\boldsymbol{X}_n(t)\end{aligned}$$

且有$W_0=0$，故有$\boldsymbol{W}_f^T\boldsymbol{W}_T \geqslant T\cdot\min_n y_n\boldsymbol{W}_f^T\boldsymbol{X}_{\boldsymbol{n}}$；

又由于

$$\begin{aligned}\|\boldsymbol{W}_{t+1}\|^2 &= \|\boldsymbol{W}_t + y_n(t)\boldsymbol{X}_n(t)\|^2 \\ &= \|\boldsymbol{W}_t\|^2 + 2y_n(t)\boldsymbol{W}_t^T\boldsymbol{X}_n(t) + \|y_n(t)\boldsymbol{X}_n(t)\|^2 \\ &\leqslant \|\boldsymbol{W}_t\|^2 + 0 + \|y_n(t)\boldsymbol{X}_n(t)\|^2 \\ &\leqslant \|\boldsymbol{W}_t\|^2 + \max_n\|\boldsymbol{X}_n(t)\|^2\end{aligned}$$

故有$\|\boldsymbol{W}_T\| \leqslant \sqrt{T\cdot\max_n\|\boldsymbol{X}_n\|^2}$；

综上所述，有

$$\begin{aligned}\frac{\boldsymbol{W}_f^T\boldsymbol{W}_T}{\|\boldsymbol{W}_f\|\,\|\boldsymbol{W}_T\|} &\geqslant \frac{T\cdot\min_n y_n\boldsymbol{W}_f^T\boldsymbol{X}_{\boldsymbol{n}}}{\|\boldsymbol{W}_f\|\cdot\sqrt{T\cdot\max_n\|\boldsymbol{X}_n\|^2}} \\ &= \sqrt{T}\cdot constant\end{aligned}$$

4，针对线性可分训练样本集，PLA 算法中，假设对分错样本进行了 T 次纠正后得到的分类面不再出现错分状况，定义：$\boldsymbol{R^2 = \max_n\|x_n\|^2}$ ，$\rho = \min_n y_n\frac{\mathbf{w}_f^T}{\|\mathrm{w}_f\|}\mathbf{x}_n$，试证明：$\mathrm{T} \le \frac{\boldsymbol{R}^2}{\rho^2}$

证明：

$$\begin{aligned}\frac{\boldsymbol{W}_f^T\boldsymbol{W}_T}{\|\boldsymbol{W}_f\|\,\|\boldsymbol{W}_T\|} &\geqslant \frac{T\cdot\min_n y_n\boldsymbol{W}_f^T\boldsymbol{X}_{\boldsymbol{n}}}{\|\boldsymbol{W}_f\|\cdot\sqrt{T\cdot\max_n\|\boldsymbol{X}_n\|^2}} \\ &= \sqrt{T}\cdot\frac{\rho}{R}\end{aligned}$$

$$
\begin{aligned}
\sqrt{T} &\leqslant \frac{R}{\rho} \cdot \frac{\boldsymbol{W}_f^T \boldsymbol{W}_T}{\|\boldsymbol{W}_f\| \|\boldsymbol{W}_T\|} \\
&= \frac{R}{\rho} \cdot \cos\langle \boldsymbol{W}_f, \boldsymbol{W}_T \rangle \\
&\leqslant \frac{R}{\rho}
\end{aligned}
$$

因此有

$$
T \leqslant \frac{R^2}{\rho^2}
$$

5， 假设训练样本集为 $D = \{(\vec{x}_1, y_1) = ((0.2,0.7)^T, 1), (\vec{x}_2, y_2) = ((0.3,0.3)^T, 1), (\vec{x}_3, y_3) = ((0.4,0.5)^T, 1)$，$(\vec{x}_4, y_4) = ((0.6,0.5)^T, 1)$，$(\vec{x}_5, y_5) = ((0.1,0.4)^T, 1)$，$(\vec{x}_6, y_6) = ((0.4,0.6)^T, -1)$，$(\vec{x}_7, y_7) = ((0.6,0.2)^T, -1)$，$(\vec{x}_8, y_8) = ((0.7,0.4)^T, -1)$，$(\vec{x}_9, y_9) = ((0.8,0.6)^T, -1)$，$(\vec{x}_{10}, y_{10}) = ((0.7,0.5)^T, -1)\}$，用 Pocket 算法设计分类面。（可借助编程实现，迭代次数最多 10 次，需提交每次迭代的结果）

解：略

Lecture3 习题作业

1， 假设训练样本集为 $D=\{(\vec{x}_1,y_1)=((0.2,0.7)^T,1),(\vec{x}_2,y_2)=((0.3,0.3)^T,1),(\vec{x}_3,y_3)=((0.4,0.5)^T,1),(\vec{x}_4,y_4)=((0.6,0.5)^T,1),(\vec{x}_5,y_5)=((0.1,0.4)^T,1),(\vec{x}_6,y_6)=((0.4,0.6)^T,-1),(\vec{x}_7,y_7)=((0.6,0.2)^T,-1),(\vec{x}_8,y_8)=((0.7,0.4)^T,-1),(\vec{x}_9,y_9)=((0.8,0.6)^T,-1),(\vec{x}_{10},y_{10})=((0.7,0.5)^T,-1)\}$，使用线性回归算法（Linear Regression Algorithm），通过广义逆来求解，并设计这两类的分类函数，讨论结果。

解：令 $D=\{(\vec{x}_i,y_i)=((1,x_i^1,x_i^2),y_i)\},i=1\sim10$，故可写出

$$\boldsymbol{X}^T=\begin{bmatrix}1&1&1&1&1&1&1&1&1&1\\0.2&0.3&0.4&0.6&0.1&0.4&0.6&0.7&0.8&0.7\\0.7&0.3&0.5&0.5&0.4&0.6&0.2&0.4&0.6&0.5\end{bmatrix}$$

$$\boldsymbol{y}=(1,1,1,1,1,-1,-1,-1,-1,-1)$$

进而计算可得

$$\boldsymbol{X}^\dagger=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$$
$$=\begin{bmatrix}-0.16&0.7&0.11&-0.1&0.67&-0.13&0.63&0.04&-0.55&-0.20\\-0.53&-0.39&-0.16&0.25&-0.78&-0.14&0.20&0.43&0.67&0.45\\1.1&-0.88&0.14&0.17&-0.41&0.64&-1.33&-0.31&0.7&0.19\end{bmatrix}$$

于是有

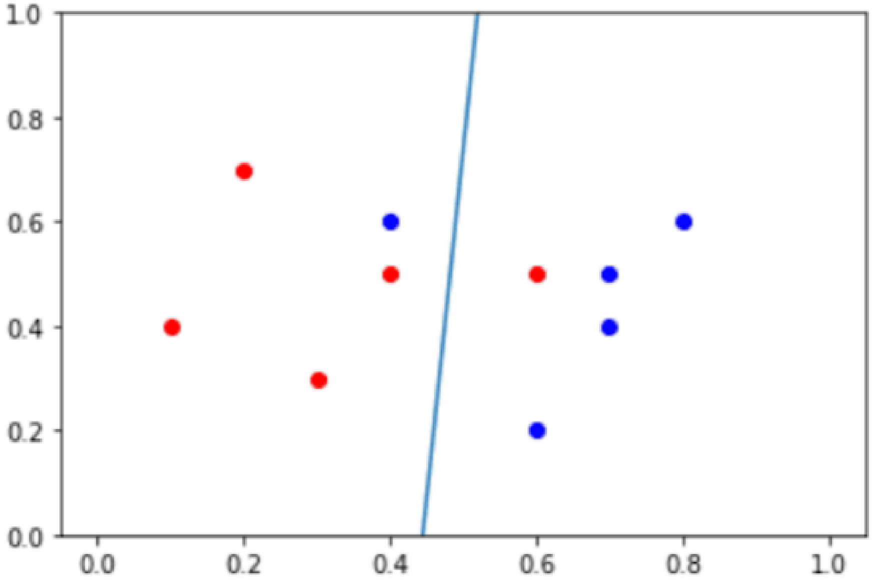
$$\begin{aligned}\boldsymbol{W}&=\boldsymbol{X}^\dagger\boldsymbol{y}\\&=(1.43,-3.22,0.24)^T\end{aligned}$$

因此这两类的分类函数为

$$h(\boldsymbol{x})=sign(\boldsymbol{W}^T\boldsymbol{x})$$

其中 $\boldsymbol{W}=(1.43,-3.22,0.24)^T$

并且将训练样本集$D=\{(\vec{x}_i,y_i)=((1,x_i^1,x_i^2),y_i)\},i=1\sim10$代入所得的分类函数$h(\boldsymbol{x})=sign(\boldsymbol{W}^T\boldsymbol{x})$可得该分类函数可大致正确分类训练样本。



2，根据向量或矩阵的计算性质，证明：

$$\|\mathbf{X}\boldsymbol{w}-Y\|^2=\boldsymbol{w}^T\mathbf{X}^T\mathbf{X}\boldsymbol{w}-2\boldsymbol{w}^T\mathbf{X}^T\mathrm{Y}+\mathrm{Y}^T\mathrm{Y}$$

解：

$$\|\mathbf{X}\boldsymbol{w}-Y\|^2=(\mathbf{X}\boldsymbol{w}-Y)^T(\mathbf{X}\boldsymbol{w}-Y)$$

$$=((\mathbf{X}\boldsymbol{w})^T-\mathrm{Y}^T)(\mathbf{X}\boldsymbol{w}-Y)$$

$$=(\boldsymbol{w}^T\mathbf{X}^T-\mathrm{Y}^T)(\mathbf{X}\boldsymbol{w}-Y)$$

$$=\boldsymbol{w}^T\mathbf{X}^T\mathbf{X}\boldsymbol{w}-\boldsymbol{w}^T\mathbf{X}^T\mathrm{Y}-\mathrm{Y}^T\mathbf{X}\boldsymbol{w}+\mathrm{Y}^T\mathrm{Y}$$

$$=\boldsymbol{w}^T\mathbf{X}^T\mathbf{X}\boldsymbol{w}-\boldsymbol{w}^T\mathbf{X}^T\mathrm{Y}-(\mathbf{X}\boldsymbol{w})^T\mathrm{Y}+\mathrm{Y}^T\mathrm{Y}$$

$$=\boldsymbol{w}^T\mathbf{X}^T\mathbf{X}\boldsymbol{w}-\boldsymbol{w}^T\mathbf{X}^T\mathrm{Y}-\boldsymbol{w}^T\mathbf{X}^T\mathrm{Y}+\mathrm{Y}^T\mathrm{Y}$$

$$=\boldsymbol{w}^T\mathbf{X}^T\mathbf{X}\boldsymbol{w}-2\boldsymbol{w}^T\mathbf{X}^T\mathrm{Y}+\mathrm{Y}^T\mathrm{Y}$$

3，总结梯度下降法、随机梯度下降法、Adagrad、RMSProp、动量法（Momentum）和 Adam 等方法权系数更新表达式。

解：对于任意的损失函数 L，假设任一单个样本 n 的梯度$\nabla L_n(\boldsymbol{w})$，$t$ 代表迭代次数

（1）梯度下降法：

$$\nabla L_{in}(\boldsymbol{w}) = \frac{1}{N}\sum_{n=1}^{N} \nabla L_n(\boldsymbol{w})$$

$$\boldsymbol{w}_{t+1} \leftarrow \boldsymbol{w}_t - \eta \nabla L_{in}(\boldsymbol{w}_t)$$

（2）随机梯度下降法：

$\nabla L_{in}(\boldsymbol{w}) = \frac{1}{B}\sum_{n=1}^{B} \nabla L_n(\boldsymbol{w})$，$B$ 代表批量大小，最小可以为 1

$$\boldsymbol{w}_{t+1} \leftarrow \boldsymbol{w}_t - \eta \nabla L_{in}(\boldsymbol{w}_t)$$

（3）Adagrad：

$$\nabla L_{in}(\boldsymbol{w}) = \frac{1}{B}\sum_{n=1}^{B} \nabla L_n(\boldsymbol{w})$$

$\sigma_t = \sqrt{\frac{1}{t+1}\sum_{t=0}^{t}(\nabla L_{in}(\boldsymbol{w}))^2} + \varepsilon$， ε代表极小量，防止σ_t为 0

$$\boldsymbol{w}_{t+1} \leftarrow \boldsymbol{w}_t - \frac{\eta}{\sigma_t} \nabla L_{in}(\boldsymbol{w}_t)$$

（4）RMSProp：

$$\nabla L_{in}(\boldsymbol{w}) = \frac{1}{B}\sum_{n=1}^{B} \nabla L_n(\boldsymbol{w})$$

$$\sigma_{t-1} = \sqrt{\frac{1}{t}\sum_{t=0}^{t-1}(\nabla L_{in}(\boldsymbol{w}))^2}$$

$$\sigma_t = \sqrt{\alpha(\sigma_{t-1})^2 + (1-\alpha)\left(\nabla L_{in}(\boldsymbol{w})\right)^2} + \varepsilon$$

$$\boldsymbol{w}_{t+1} \leftarrow \boldsymbol{w}_t - \frac{\eta}{\sigma_t} \nabla L_{in}(\boldsymbol{w}_t)$$

（5）动量法（Momentum）：

$\nabla L_{in}(\boldsymbol{w}) = \frac{1}{B}\sum_{n=1}^{B} \nabla L_n(\boldsymbol{w})$

$\boldsymbol{m}_{t+1} = \lambda \boldsymbol{m}_t - \eta \nabla L_{in}(\boldsymbol{w}_t)$，　　（$\boldsymbol{m_0} = 0$）

$\boldsymbol{w}_{t+1} \leftarrow \boldsymbol{w}_t + \boldsymbol{m}_{t+1}$

（6）Adam

$\boldsymbol{m_{t+1}} = \beta_1 \boldsymbol{m}_t - (1-\beta_1)\nabla L_{in}(\boldsymbol{w_t})$，　　（$\boldsymbol{m_0} = 0$）

$\boldsymbol{v_{t+1}} = \beta_2 \boldsymbol{v}_t - (1-\beta_2)\left(\nabla L_{in}(\boldsymbol{w})\right)^2$，　　（$\boldsymbol{v_0} = 0$）

$\widehat{\boldsymbol{m}}_{\boldsymbol{t+1}} = \boldsymbol{m}_{t+1}/(1-\beta_1^{t+1})$

$\widehat{\boldsymbol{v}}_{\boldsymbol{t+1}} = \boldsymbol{v}_{t+1}/(1-\beta_2^{t+1})$

$\boldsymbol{w}_{t+1} \leftarrow \boldsymbol{w}_t - \eta \widehat{\boldsymbol{m}}_{t+1}/(\sqrt{\widehat{\boldsymbol{v}}_{t+1}} + \varepsilon)$

Lecture4 习题作业

1，已知两类样本的数据如下：

$\omega_1: \{(5,37),(7,30),(10,35),(11.5,40),(14,38),(12,31)\}$

$\omega_2: \{(35,21.5),(39,21.7),(34,16),(37,17)\}$

试用 Fisher 判别函数法，求出最佳投影方向 W，及分类阈值 y_0

解：由题意知：

$$\mu_1 = \frac{1}{6}\sum_{n=1}^{6} X_n^{(1)} = (9.92\ 35.17)^T$$

$$\mu_{-1} = \frac{1}{4}\sum_{n=1}^{4} X_n^{(0)} = (36.25\ 19.05)^T$$

则可计算出类内离差阵：

$$\Sigma_1 = \sum_{n=1}^{6}(X_n^{(1)} - \mu_1)\cdot(X_n^{(1)} - \mu_1)^T = \begin{pmatrix} 56.21 & 16.58 \\ 16.58 & 78.83 \end{pmatrix}$$

$$\Sigma_{-1} = \sum_{n=1}^{4}(X_n^{(0)} - \mu_{-1})\cdot(X_n^{(0)} - \mu_{-1})^T = \begin{pmatrix} 14.75 & 9.55 \\ 9.55 & 26.53 \end{pmatrix}$$

$$S_w = \Sigma_1 + \Sigma_{-1} = \begin{pmatrix} 70.96 & 26.13 \\ 26.13 & 105.36 \end{pmatrix}$$

$$S_w^{-1} = \begin{pmatrix} 0.0155 & -0.0038 \\ -0.0038 & 0.0104 \end{pmatrix}$$

从而可计算出最佳投影方向：

$$W^* = S_w^{-1}(\mu_1 - \mu_{-1}) = (-0.4704, 0.2696)^T$$

$$y_0 = W^{*T}\frac{(\mu_1 + \mu_{-1})}{2} = -3.55$$

2，在 Fisher 判别中，用向量梯度的计算法则证明：$\boldsymbol{S}_B\boldsymbol{w} = \lambda\boldsymbol{S}_w\boldsymbol{w}$

证明： $L(\boldsymbol{w},\lambda) = \boldsymbol{w}^T\boldsymbol{S}_B\boldsymbol{w} + \lambda(K - \boldsymbol{w}^T\boldsymbol{S}_w\boldsymbol{w}) = \boldsymbol{w}^T(\boldsymbol{S}_B - \lambda\boldsymbol{S}_w)\boldsymbol{w} + \lambda K$

$$\nabla L_w(\boldsymbol{w},\lambda) = \frac{\partial L(\boldsymbol{w},\lambda)}{\partial \boldsymbol{w}} = \boldsymbol{0}^T$$

$$\frac{\partial L(\boldsymbol{w},\lambda)}{\partial \boldsymbol{w}} = \frac{\partial(\boldsymbol{w}^T(\boldsymbol{S}_B - \lambda\boldsymbol{S}_w)\boldsymbol{w})}{\partial \boldsymbol{w}} + 0$$

根据$\frac{\partial \boldsymbol{u}^T\boldsymbol{v}}{\partial \mathbf{x}} = \boldsymbol{u}^T\frac{\partial \boldsymbol{v}}{\partial \mathbf{x}} + \boldsymbol{v}^T\frac{\partial \boldsymbol{u}}{\partial \mathbf{x}}$，$\frac{\partial \boldsymbol{A}\mathbf{x}}{\partial \mathbf{x}} = \boldsymbol{A}$，以及$\frac{\partial \mathbf{x}}{\partial \mathbf{x}} = \boldsymbol{I}$，上式：

$$\frac{\partial(\boldsymbol{w}^T(\boldsymbol{S}_B - \lambda\boldsymbol{S}_w)\boldsymbol{w})}{\partial \boldsymbol{w}} = \boldsymbol{w}^T(\boldsymbol{S}_B - \lambda\boldsymbol{S}_w)\boldsymbol{I} + \left((\boldsymbol{S}_B - \lambda\boldsymbol{S}_w)\boldsymbol{w}\right)^T\boldsymbol{I}$$

$$= \boldsymbol{w}^T(\boldsymbol{S}_B - \lambda\boldsymbol{S}_w)\boldsymbol{I} + \boldsymbol{w}^T(\boldsymbol{S}_B - \lambda\boldsymbol{S}_w)^T\boldsymbol{I}$$

因为：$\boldsymbol{S}_B$和$\boldsymbol{S}_w$均为对称矩阵，

所以：$(\boldsymbol{S}_B - \lambda\boldsymbol{S}_w)^T = (\boldsymbol{S}_B - \lambda\boldsymbol{S}_w)$，

又：$(\boldsymbol{S}_B - \lambda\boldsymbol{S}_w)\boldsymbol{I} = (\boldsymbol{S}_B - \lambda\boldsymbol{S}_w)$

所以：
$$\frac{\partial L(\boldsymbol{w},\lambda)}{\partial \boldsymbol{w}} = 2\boldsymbol{w}^T(\boldsymbol{S}_B - \lambda\boldsymbol{S}_w) = \boldsymbol{0}^T$$

$$2(\boldsymbol{S}_B - \lambda\boldsymbol{S}_w)\boldsymbol{w} = \boldsymbol{0}$$

$$\boldsymbol{S}_B\boldsymbol{w} = \lambda\boldsymbol{S}_w\boldsymbol{w}$$

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：

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

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

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

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

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

无论分类正确与否，梯度都为 0，影响学习性能。

Lecture7-8 作业

1，假设两个样本 $\{(\mathbf{v}_1, y_1) = ((v_1, v_2)^T, 1), (\mathbf{v}_2, y_2) = ((-v_1, -v_2)^T, -1)\}$, 假设 H 是这两个样本的最大间隔分类面，写出其表达式。

解：两个样本关于原点对称，最大间隔分类面会垂直于两个样本的连线，且穿过原点，即样本连线的斜率与分类面（分类线）斜率的乘积为-1，而样本连线的斜率为$\frac{v_2}{v_1}$，所以，分类面（线）的斜率为：$-\frac{v_1}{v_2}$，且 b=0。

所以，最大间隔分类面为：

$$x_2 = -\frac{v_1}{v_2} x_1$$

即：$v_1 x_1 + v_2 x_2 = 0$

2，假设三个样本为 $D = \{(\mathbf{x}_1, y_1) = ((3,0)^T, 1), (\mathbf{x}_2, y_2) = ((0,4)^T, 1), (\mathbf{x}_3, y_3) = ((0,0)^T, -1)\}$,计算这三个样本到平面：$x_1 + x_2 = 1$的距离。

解：$d = \frac{|\boldsymbol{w}^T \mathbf{x} + b|}{\|\boldsymbol{w}\|}$

$$x_1 + x_2 = 1 \rightarrow x_1 + x_2 - 1 = 0$$

$$d_1 = \frac{|\boldsymbol{w}^T \mathbf{x}_1 + b|}{\|\boldsymbol{w}\|} = \frac{\left|(1,1)\binom{3}{0} - 1\right|}{\sqrt{(1^2 + 1^2)}} = \sqrt{2}$$

$$d_2 = \frac{|\boldsymbol{w}^T \mathbf{x}_2 + b|}{\|\boldsymbol{w}\|} = \frac{\left|(1,1)\binom{0}{4} - 1\right|}{\sqrt{(1^2 + 1^2)}} = \frac{3}{2}\sqrt{2}$$

$$d_3 = \frac{|\boldsymbol{w}^T \mathbf{x}_3 + b|}{\|\boldsymbol{w}\|} = \frac{\left|(1,1)\begin{pmatrix}0\\0\end{pmatrix} - 1\right|}{\sqrt{(1^2+1^2)}} = \frac{\sqrt{2}}{2}$$

3， 假设训练样本集为 $\mathrm{D} = \{(\mathbf{x}_1, y_1) = ((0,0)^T, -1), (\mathbf{x}_2, y_2) = ((2,2)^T, -1), (\mathbf{x}_3, y_3) = ((2,0)^T, 1), (\mathbf{x}_4, y_4) = ((3,0)^T, 1)\}$,使用 QP 求解器时，$\boldsymbol{a}_n^T$(n=1,2,3,4)分别为多少？

解：$\boldsymbol{a}_1^T = (-1,0,0)$，$\boldsymbol{a}_2^T = (-1,-2,-2)$，$\boldsymbol{a}_3^T = (1,2,0)$，$\boldsymbol{a}_4^T = (1,3,0)$

4， 假设训练样本集为： $\mathrm{D} = \{(\mathbf{x}_1, y_1) = ((1,1)^T, 1), (\mathbf{x}_2, y_2) = ((2,2)^T, 1), (\mathbf{x}_3, y_3) = ((2,0)^T, 1), (\mathbf{x}_4, y_4) = ((0,0)^T, -1), (\mathbf{x}_5, y_5) = ((1,0)^T, -1), (\mathbf{x}_6, y_6) = ((0,1)^T, -1)\}$, 请分别在$y_n(\boldsymbol{w}^T \mathbf{x}_n + b) \ge 1$和$y_n(\boldsymbol{w}^T \mathbf{x}_n + b) \ge 5$的条件下用 Primal SVM 方法来设计最优分类面g($\mathbf{x}$)，判断两种情况下的分类面是否一致，指出哪些是候选的支撑向量，并回答如何确认哪些是支撑向量。

解：（1）对于条件$y_n(\boldsymbol{w}^T \mathbf{x}_n + b) \ge 1$，可列出如下的式子

$$\min_{\boldsymbol{b},\boldsymbol{w}} \frac{1}{2} \boldsymbol{w}^T \boldsymbol{w}$$

$$s.t. \begin{cases} w_1 + w_2 + b \geqslant 1 \\ 2w_1 + 2w_2 + b \geqslant 1 \\ 2w_1 + b \geqslant 1 \\ -b \geqslant 1 \\ -w_1 - b \geqslant 1 \\ -w_2 - b \geqslant 1 \end{cases} \Longrightarrow \begin{cases} w_1 \geqslant 2 \\ w_2 \geqslant 2 \\ b \leqslant -3 \end{cases}$$

当且仅当$w_1 = 2, w_2 = 2, b = -3$，

$$\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}=\frac{1}{2}(w_1^2+w_2^2)\geqslant\frac{1}{2}(2^2+2^2)=4$$ 取得最小值。

可以验证 constraints 均满足。

故此时的最优分类面为

$$\boldsymbol{w}_1^T\boldsymbol{x}+b_1=0$$

其中$\boldsymbol{w}_1=[2\quad 2]^T, b_1=-3$。

可以验证，将$\boldsymbol{w}_1=[2\quad 2]^T, b_1=-3$代入上述 constraints 中有第 1、3、5、6 是严格等式，故候选支撑向量为$\boldsymbol{x}_1,\boldsymbol{x}_3,\boldsymbol{x}_5,\boldsymbol{x}_6$。

由 Dual SVM 知识可知，当求解 Dual SVM 问题时，在如下式子中

$$\alpha_n(1-y_n(\boldsymbol{w}^T\boldsymbol{x}_n+b))=0$$

$\boldsymbol{x}_1,\boldsymbol{x}_3,\boldsymbol{x}_5,\boldsymbol{x}_6$满足$\alpha_n>0, y_n(\boldsymbol{w}^T\boldsymbol{x}_n+b)=1$所对应的样本即为支撑向量。

（2）对于条件$y_n(\vec{w}^T\vec{x}_n+b)\geqslant 5$，可列出如下的式子

$$\min_{\boldsymbol{b},\boldsymbol{w}}\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}$$

$$s.t.\begin{cases} w_1+w_2+b\geqslant 5\\ 2w_1+2w_2+b\geqslant 5\\ 2w_1+b\geqslant 5\\ -b\geqslant 5\\ -w_1-b\geqslant 5\\ -w_2-b\geqslant 5\end{cases}\Longrightarrow\begin{cases} w_1\geqslant 10\\ w_2\geqslant 10\\ b\leqslant -15\end{cases}$$

当且仅当$w_1=10, w_2=10, b=-15$时有

$$\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}=\frac{1}{2}(w_1^2+w_2^2)\geqslant\frac{1}{2}(10^2+10^2)=100$$ 取得最小值，

可以验证 constraints 均满足。

故此时的最优分类面为

$$\boldsymbol{w}_2^T\boldsymbol{x}+b_2=0, which\ is\ exactly\ equivalent\ to\ \boldsymbol{w}_1^T\boldsymbol{x}+b_1=0$$

其中$\boldsymbol{w}_2=[10\quad 10]^T, b_2=-15$。

可以验证，将$\boldsymbol{w}_2=[10\quad 10]^T, b_2=-15$代入上述 constraints 中有第 1、3、5、6 是严格等式，故候选支撑向量为$\boldsymbol{x}_1,\boldsymbol{x}_3,\boldsymbol{x}_5,\boldsymbol{x}_6$。

由 Dual SVM 知识可知，当求解 Dual SVM 问题时，在如下式子中

$$\alpha_n(1-y_n(\boldsymbol{w}^T\boldsymbol{x}_n+b))=0$$

$\boldsymbol{x}_1,\boldsymbol{x}_3,\boldsymbol{x}_5,\boldsymbol{x}_6$满足$\alpha_n>0, y_n(\boldsymbol{w}^T\boldsymbol{x}_n+b)=5$所对应的样本即为支撑向量。

5，Hinge Loss 是支撑向量机的误差函数，因此，除了用二次规划求解最佳分类面外，也能用梯度下降法求解，（1）请推导梯度并写出算法流程；（2）假设初始增广权向量$\vec{w}=(0,0,0)^T$，用第 4 题训练样本集去设计分类面，指出哪些向量在边界上？假设它们都是支撑向量的话，请问最佳权系数向量是否是这些支撑向量的线性组合？

解：（1）已知样本集合$\{(\vec{x}_1,y_1),(\vec{x}_2,y_2),\ldots,(\vec{x}_N,y_N)\}$ $\{(\mathbf{x}_1,y_1),(\mathbf{x}_2,y_2),\ldots,(\mathbf{x}_N,y_N)\}$，每个样本的标签为$y_n\in\{+1，-1\}$，我们基于 Hinge Loss，对于每个样本定义其误差函数为：

$$err_{SVM}=\max(0,1-y_n(\boldsymbol{w}^T\mathbf{x}_n+b))$$

对其求梯度，得到：

当$1-y_n(\boldsymbol{w}^T\mathbf{x}_n+b)\geq 0$，$\frac{\partial L_{in}(\boldsymbol{w})}{\partial \boldsymbol{w}}=-y_n\mathbf{x}_n$

当$1-y_n(\boldsymbol{w}^T\mathbf{x}_n+b)<0$，$\frac{\partial L_{in}(\boldsymbol{w})}{\partial \boldsymbol{w}}=0$

利用随机梯度下降法得到新的$\boldsymbol{w}$

$$\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} - \eta \frac{\partial L_{in}\left(\boldsymbol{w}^{(t)}\right)}{\partial \boldsymbol{w}^{(t)}} = \boldsymbol{w}^{(t)} + \eta[\![1 - y_n(\boldsymbol{w}^T\mathbf{x}_n + b) \geq 0]\!]y_n\mathbf{x}_n$$

其中，$[\![\cdot]\!] = \begin{cases} 1, \ if\ condition\ is\ satisfied \\ 0, \ otherwise \end{cases}$

（2） 初始增广权向量$\boldsymbol{w}^{(0)} = (0,0,0)^T$

$\mathbf{x}_1 = (1,1,1)^T, \mathbf{x}_2 = (1,2,2)^T, \mathbf{x}_3 = (1,2,0)^T,$

$\mathbf{x}_4 = (1,0,0)^T, \mathbf{x}_5 = (1,1,0)^T, \mathbf{x}_6 = (1,0,1)^T$

$y_1 = 1, y_2 = 1, y_3 = 1, y_4 = -1, y_5 = -1, y_6 = -1$

取学习率$\eta = 1$

第一轮迭代

$$\max\left(0,1 - y_1\left(\boldsymbol{w}^{(0)^T}\mathbf{x}_1\right)\right) = \max(0,1) = 1$$

$$\frac{\partial L_{in}(\boldsymbol{w}^{(0)})}{\partial \boldsymbol{w}^{(0)}} = -y_1\mathbf{x}_1 = (-1,-1,-1)^T$$

$$\boldsymbol{w}^{(1)} = \boldsymbol{w}^{(0)} - \eta \frac{\partial L_{in}\left(\boldsymbol{w}^{(1)}\right)}{\partial \boldsymbol{w}^{(1)}} = \boldsymbol{w}^{(0)} + y_1\mathbf{x}_1 = (1,1,1)^T$$

第二轮迭代

$$\max\left(0,1 - y_2\left(\boldsymbol{w}^{(1)^T}\mathbf{x}_2\right)\right) = \max(0,-4) = 0$$

$$\boldsymbol{w}^{(2)} = \boldsymbol{w}^{(1)} = (1,1,1)^T$$

第三轮迭代

$$\max\left(0,1 - y_3\left(\boldsymbol{w}^{(2)^T}\mathbf{x}_3\right)\right) = \max(0,-2) = 0$$

$$\boldsymbol{w}\boldsymbol{w}^{(3)} = \boldsymbol{w}^{(2)} = (1,1,1)^T$$

第四轮迭代

$$\max\left(0,1-y_4\left({\boldsymbol{w}^{(3)}}^T\mathbf{x}_4\right)\right)=\max(0,2)=2$$

$$\frac{\partial L_{in}(\boldsymbol{w}^{(3)})}{\partial \boldsymbol{w}^{(3)}}=-y_4\mathbf{x}_4=(1,0,0)^T$$

$$\boldsymbol{w}^{(4)}=\boldsymbol{w}^{(3)}+y_4\mathbf{x}_4=(0,1,1)^T$$

第五轮迭代

$$\max\left(0,1-y_5\left({\boldsymbol{w}^{(4)}}^T\mathbf{x}_5\right)\right)=\max(0,2)=2$$

$$\frac{\partial L_{in}(\boldsymbol{w}^{(4)})}{\partial \boldsymbol{w}^{(4)}}=-y_5\mathbf{x}_5=(1,1,0)^T$$

$$\boldsymbol{w}^{(5)}=\boldsymbol{w}^{(4)}+y_5\mathbf{x}_5=(-1,0,1)^T$$

第六轮迭代

$$\max\left(0,1-y_6\left({\boldsymbol{w}^{(5)}}^T\mathbf{x}_6\right)\right)=\max(0,1)=1$$

$$\frac{\partial L_{in}(\boldsymbol{w}^{(5)})}{\partial \boldsymbol{w}^{(5)}}=-y_6\mathbf{x}_6=(1,0,1)^T$$

$$\boldsymbol{w}\boldsymbol{w}^{(6)}=\boldsymbol{w}^{(5)}+y_6\mathbf{x}_6=(-2,0,0)^T$$

第七轮迭代

$$\max\left(0,1-y_1\left({\boldsymbol{w}^{(6)}}^T\mathbf{x}_1\right)\right)=\max(0,3)=3$$

$$\frac{\partial L_{in}(\boldsymbol{w}^{(7)})}{\partial \boldsymbol{w}^{(7)}}=-y_1\mathbf{x}_1=(-1,-1,-1)^T$$

$$\boldsymbol{w}^{(7)}=\boldsymbol{w}\boldsymbol{w}^{(6)}+y_1\mathbf{x}_1=(-1,1,1)^T$$

第八轮迭代

$$\max\left(0,1-y_2\left({\boldsymbol{w}^{(7)}}^T\mathbf{x}_2\right)\right)=\max(0,-2)=0$$

$$\boldsymbol{w}^{(8)}=\boldsymbol{w}^{(7)}=(-1,1,1)^T$$

第九轮迭代

$$\max\left(0,1-y_3\left({\boldsymbol{w}^{(8)}}^T\mathbf{x}_3\right)\right)=\max(0,0)=0$$

$$\frac{\partial L_{in}(\boldsymbol{w}\boldsymbol{w}^{(8)})}{\partial \boldsymbol{w}^{(8)}} = -y_3\mathbf{x}_3 = (-1,-2,0)^T$$

$$\boldsymbol{w}^{(9)} = \boldsymbol{w}\boldsymbol{w}^{(8)} + y_3\mathbf{x}_3 = (0,3,1)^T$$

第十轮迭代

$$\max\left(0,1 - y_4\left(\boldsymbol{w}^{(9)^T}\mathbf{x}_4\right)\right) = \max(0,1) = 1$$

$$\frac{\partial L_{in}(\boldsymbol{w}\boldsymbol{w}^{(9)})}{\partial \overrightarrow{w}^{(9)}} = -y_4\mathbf{x}_4 = (1,0,0)^T$$

$$\boldsymbol{w}^{(10)} = \boldsymbol{w}^{(9)} + y_4\mathbf{x}_4 = (-1,3,1)^T$$

第十一轮迭代

$$\max\left(0,1 - y_5\left(\boldsymbol{w}^{(10)^T}\mathbf{x}_5\right)\right) = \max(0,1) = 1$$

$$\frac{\partial L_{in}(\boldsymbol{w}^{(4)})}{\partial \boldsymbol{w}^{(4)}} = -y_5\mathbf{x}_5 = (1,1,0)^T$$

$$\boldsymbol{w}^{(11)} = \boldsymbol{w}^{(10)} + y_5\mathbf{x}_5 = (-2,2,1)^T$$

第十二轮迭代

$$\max\left(0,1 - y_6\left(\boldsymbol{w}^{(11)^T}\mathbf{x}_6\right)\right) = \max(0,0) = 0$$

$$\frac{\partial L_{in}(\boldsymbol{w}^{(11)})}{\partial \boldsymbol{w}^{(11)}} = -y_6\mathbf{x}_6 = (1,0,1)^T$$

$$\boldsymbol{w}^{(12)} = \boldsymbol{w}^{(11)} + y_6\mathbf{x}_6 = (-3,2,0)^T$$

第十三轮迭代

$$\max\left(0,1 - y_1\left(\boldsymbol{w}^{(12)^T}\mathbf{x}_1\right)\right) = \max(0,2) = 2$$

$$\frac{\partial L_{in}(\boldsymbol{w}^{(12)})}{\partial \boldsymbol{w}^{(12)}} = -y_1\mathbf{x}_1 = (-1,-1,-1)^T$$

$$\boldsymbol{w}^{(7)} = \boldsymbol{w}^{(6)} + y_1\mathbf{x}_1 = (-2,3,1)^T$$

第十四轮迭代

对于$\mathbf{x}_2$、$\mathbf{x}_3$、$\mathbf{x}_4$满足$1-y_n\left(\boldsymbol{w}^{(13)^T}\mathbf{x}_n\right)<0$

$\max\left(0,1-y_5\left(\boldsymbol{w}^{(13)^T}\mathbf{x}_5\right)\right)=\max(0,2)=2$

$\boldsymbol{w}^{(14)}=\boldsymbol{w}^{(13)}+y_5\mathbf{x}_5=(-3,2,1)^T$

第十五轮迭代

对于$\mathbf{x}_6$满足$1-y_n\left(\boldsymbol{w}^{(13)^T}\mathbf{x}_n\right)<0$

$\max\left(0,1-y_1\left(\boldsymbol{w}^{(14)^T}\mathbf{x}_1\right)\right)=\max(0,1)=1$

$\boldsymbol{w}^{(15)}=\boldsymbol{w}^{(14)}+y_1\mathbf{x}_1=(-2,3,2)^T$

第十六轮迭代

对于$\mathbf{x}_2$、$\mathbf{x}_3$、$\mathbf{x}_4$满足$1-y_n\left(\boldsymbol{w}^{(15)^T}\mathbf{x}_n\right)<0$

$\max\left(0,1-y_5\left(\boldsymbol{w}^{(15)^T}\mathbf{x}_5\right)\right)=\max(0,2)=2$

$\boldsymbol{w}^{(16)}=\boldsymbol{w}^{(15)}+y_5\mathbf{x}_5=(-3,2,2)^T$

检验对任意$\mathbf{x}_n$满足$1-y_n\left(\boldsymbol{w}^{(15)^T}\mathbf{x}_n\right)<0$，迭代结束

得到分类面为$2x_1+2x_2-3=0$

$\vec{w}=(-3,2,2)^T$

将 $\mathbf{x}_1$、$\mathbf{x}_3$、$\mathbf{x}_5$、$\mathbf{x}_6$ 代入 $1-y_n(\mathbf{w}^T\mathbf{x}_n+b)$均为 0，说明这四个样本在边界上，均为候选的支撑向量。

为简单起见（不用求解对偶 SVM），按照本题题意候选的支撑向量均为支撑向量，则：$\boldsymbol{w}=7x_1+0x_3-5x_5-5x_6$，即最佳权系数向量为支撑向量的线性组合。

6，假如做了非线性变换后的两个训练样本为：$\{(\boldsymbol{Z}_1,+1)=(\boldsymbol{z},1),(\boldsymbol{Z}_2,-1)=(-\boldsymbol{z},-1)\}$，请写出用于设计硬间隔 SVM 时的拉格朗日函数$L(\boldsymbol{w},\mathrm{b},\alpha)$。

解：根据定义：

$$L(\boldsymbol{w},\mathrm{b},\alpha)=\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}+\alpha_1(1-y_1(\boldsymbol{w}^T\boldsymbol{Z}_1+b)+\alpha_2(1-y_2(\boldsymbol{w}^T\boldsymbol{Z}_2+b))$$

将两个样本代入，得到：

$$L(\boldsymbol{w},\mathrm{b},\alpha)=\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}+\alpha_1(1-(\mathbf{w}^T\boldsymbol{z}+b)+\alpha_2(1+(-\boldsymbol{w}^T\boldsymbol{z}+b))$$

7，对于一个单变量w，假设要在$w\geq 1$和$w\leq 3$这两个线性约束条件下，求$\frac{1}{2}w^2$的最小值，请写出其拉格朗日函数$L(w,\alpha)$以及这个最优问题的 KKT 条件。

解：由于是单变量，根据定义及约束条件：

$$L(w,\alpha)=\frac{1}{2}w^2+\alpha_1(1-w)+\alpha_2(w-3)$$

KKT 条件为：

$\alpha_1\geq 0$，$\alpha_2\geq 0$，

$w=\alpha_1-\alpha_2$，（通过$\frac{\partial L(w,\alpha)}{\partial w}=0$得到）

$\alpha_1(1-w)=0$，$\alpha_2(w-3)=0.$

8，假如做了非线性变换后的两个训练样本为：$\{(\boldsymbol{Z}_1,+1)=(\boldsymbol{z},1),(\boldsymbol{Z}_2,-1)=(-\boldsymbol{z},-1)\}$，在求解硬间隔 SVM 的对偶问题时，假定得到的最佳$\alpha_1>0$，最佳$\alpha_2>0$，请问最佳 b 为多少？

解：由于$\alpha_1 > 0$，$\alpha_2 > 0$，所以：$\boldsymbol{Z}_1$和$\vec{Z}_2$为支撑向量，根据定义：

$$b = y_1 - \boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{Z}_1 = y_2 - \boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{Z}_2 = 1 - \boldsymbol{w}^T\boldsymbol{z} = -1 + \boldsymbol{w}^T\boldsymbol{z}$$

得到：$\boldsymbol{w}^T\boldsymbol{z} = 1$，$b = 0$

9，假设有 5566 个样本用以训练对偶硬间隔 SVM 时得到 1126 个支撑向量，请问落在分类面边界上的样本数（也就是候选的支撑向量）有可能是：（a）0；（b）1024；（c）1234；（d）9999。

解：　因为：支撑向量数≤候选的支撑向量数≤样本总数

所以选择（c）

10，如果两个样本$\mathbf{x}$和$\mathbf{x}'$的内积$\mathbf{x}^T\mathbf{x}' = 10$，计算其$\phi_2$核函数$K_{\phi_2}(\mathbf{x}, \mathbf{x}')$等于多少？

解：因为：$K_{\phi_2}(\mathbf{x}, \mathbf{x}') = 1 + \mathbf{x}^T\mathbf{x}' + (\mathbf{x}^T\mathbf{x}')^2$

所以：　$K_{\phi_2}(\mathbf{x}, \mathbf{x}') = 1 + 10 + 100 = 111$

11，假设训练样本集为：$\mathrm{D} = \{(\mathbf{x}_1, y_1) = ((2,2)^T, 1), (\mathbf{x}_2, y_2) = ((-2,-2)^T, 1), (\mathbf{x}_3, y_3) = ((2,-2)^T, -1), (\mathbf{x}_4, y_4) = ((-2,2)^T, -1)\}$，请用 Dual SVM 来设计最优分类面 $\mathrm{g}(\mathbf{x})$，并指出哪些是支撑向量。

解：样本为非线性分布，所以，需要首先进行非线性变换：

令$\phi_2(\vec{x}) = \{1, x_1, x_2, x_1x_2, x_1^2, x_2^2\}$

则：$(\mathbf{x}_1, y_1) \rightarrow (\mathbf{z}_1, y_1)$: $\{(2,2)^T, 1\} \rightarrow \{(1,2,2,4,4,4)^T, 1\}$

$(\mathbf{x}_2, y_2) \rightarrow (\boldsymbol{z}_2, y_2)$: $\{(-2,-2)^T, 1\} \rightarrow \{(1,-2,-2,4,4,4)^T, 1\}$

$(\mathbf{x}_3, y_3) \rightarrow (\mathbf{z}_3, y_3)$: $\{(-2,2)^T, -1\} \rightarrow \{(1,-2,2,-4,4,4)^T, -1\}$

$(\mathbf{x}_4, y_4) \rightarrow (\mathbf{z}_4, y_4)$: $\{(2,-2)^T, -1\} \rightarrow \{(1,2,-2,-4,4,4)^T, -1\}$

令$\alpha_1 \geq 0$、$\alpha_2 \geq 0$、$\alpha_3 \geq 0$、$\alpha_4 \geq 0$

由 SVM 对偶模型得到：

$$\begin{cases} L(\boldsymbol{w}, b, \alpha) = \dfrac{1}{2}\sum_{n=1}^{4}\sum_{m=1}^{4} \alpha_n \alpha_m y_n y_m \mathbf{z}_n^T \mathbf{z}_m - \sum_{n=1}^{4} \alpha_n \\ \sum_{n=1}^{4} y_n \alpha_n = 0 \end{cases}$$

求$L(\boldsymbol{w}, b, \alpha)$对$\alpha$的梯度：$\frac{\partial L}{\partial \alpha_n} = \sum_{m=1}^{4} \alpha_m y_n y_m \mathbf{z}_n^T \mathbf{z}_m - 1$

且：$\alpha_1 + \alpha_2 - \alpha_3 - \alpha_4 = 0$

代入训练样本，

$$\frac{\partial L}{\partial \alpha_1} = 57\alpha_1 + 41\alpha_2 - 17\alpha_3 - 17\alpha_4 - 1 = 0 \rightarrow 40\alpha_1 + 24\alpha_2 - 1 = 0$$
$$\frac{\partial L}{\partial \alpha_2} = 57\alpha_2 + 41\alpha_1 - 17\alpha_3 - 17\alpha_4 - 1 = 0 \rightarrow 40\alpha_2 + 24\alpha_1 - 1 = 0$$
$$\frac{\partial L}{\partial \alpha_3} = 57\alpha_3 - 17\alpha_1 - 17\alpha_2 + 41\alpha_4 - 1 = 0 \rightarrow 40\alpha_3 + 24\alpha_4 - 1 = 0$$
$$\frac{\partial L}{\partial \alpha_4} = 57\alpha_4 - 17\alpha_1 - 17\alpha_2 + 41\alpha_3 - 1 = 0 \rightarrow 40\alpha_4 + 24\alpha_3 - 1 = 0$$

求解得到：$\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \frac{1}{64}$

$$\therefore\ \boldsymbol{w} = \sum_{n=1}^{4} \alpha_n y_n \mathbf{z}_n = \frac{1}{64}(\mathbf{z}_1 + \mathbf{z}_2 - \mathbf{z}_3 - \mathbf{z}_4) = (0,0,0,\frac{1}{4},0,0)^T$$

$$b = y_1 - \boldsymbol{w}^T \mathbf{z}_1 = 1 - \left(0,0,0,\frac{1}{4},0,0\right)(1,2,2,4,4,4)^T = 0$$

$$\therefore\ g_{SVM} = sign(\boldsymbol{w}^T \phi_2(\mathbf{x}) + b) = sign(\frac{1}{4}x_1 x_2)$$

且四个样本均为支撑向量。

Lecture 9 习题作业

1，假设有如下训练样本：$\mathbf{x}_1=(0,0)^T$属于第一类，$\mathbf{x}_2=(1,1)^T$属于第二类，$\mathbf{x}_3=(-1,1)^T$属于第三类，请用多类分类中的 OVO（One-versus-one）策略，设计上述三类别的两两分类器，并分析测试样本$\mathbf{x}=(1,-2)^T$属于哪个类别。

解：利用 OVO 策略，对三个类别两两求分类面：

（1）用感知器算法求第一类和第二类之间的分类面

样本增广后为：$\mathbf{x}_1=(1,0,0)^T$，$y_1=1$，$\mathbf{x}_2=(1,1,1)^T$，$y_2=-1$，

初始化权重：$\boldsymbol{w}_{[1,2]}^{(0)}=(0,0,0)^T$

$sign\left(\boldsymbol{w}_{[1,2]}^{(0)T}\mathbf{x}_1\right)=0\neq y_1$， $\therefore\ \boldsymbol{w}_{[1,2]}^{(1)}=\boldsymbol{w}_{[1,2]}^{(0)}+y_1\mathbf{x}_1=(1,0,0)^T$，

$sign\left(\boldsymbol{w}_{[1,2]}^{(1)T}\mathbf{x}_2\right)=1\neq y_2$， $\therefore\ \boldsymbol{w}_{[1,2]}^{(2)}=\boldsymbol{w}_{[1,2]}^{(1)}+y_2\mathbf{x}_2=(0,-1,-1)^T$

$sign\left(\boldsymbol{w}_{[1,2]}^{(2)T}\mathbf{x}_1\right)=0\neq y_1$， $\therefore\ \boldsymbol{w}_{[1,2]}^{(3)}=\boldsymbol{w}_{[1,2]}^{(2)}+y_1\mathbf{x}_1=(1,-1,-1)^T$

$sign\left(\boldsymbol{w}_{[1,2]}^{(3)T}\mathbf{x}_2\right)=-1=y_2,且\ sign\left(\boldsymbol{w}_{[1,2]}^{(3)T}\mathbf{x}_1\right)=1=y_1$

$\therefore\ \boldsymbol{w}_{[1,2]}=(1,-1,-1)^T$，分类面为：$1-x_1-x_2=0$

（2）用感知器算法求第一类和第三类之间的分类面

样本增广后为：$\mathbf{x}_1=(1,0,0)^T$，$y_1=1$，$\mathbf{x}_3=(1,-1,1)^T$，$y_3=-1$，

初始化权重：$\boldsymbol{w}_{[1,3]}^{(0)}=(0,0,0)^T$

$sign\left(\boldsymbol{w}_{[1,3]}^{(0)T}\mathbf{x}_1\right)=0\neq y_1$， $\therefore\ \boldsymbol{w}_{[1,3]}^{(1)}=\boldsymbol{w}_{[1,3]}^{(0)}+y_1\mathbf{x}_1=(1,0,0)^T$，

$sign\left(\boldsymbol{w}_{[1,3]}^{(1)T}\mathbf{x}_3\right)=1\neq y_3$， $\therefore\ \boldsymbol{w}_{[1,3]}^{(2)}=\boldsymbol{w}_{[1,3]}^{(1)}+y_3\mathbf{x}_3=(0,1,-1)^T$

$sign\left(\boldsymbol{w}_{[1,3]}^{(2)T}\mathbf{x}_1\right)=0\neq y_1$， $\therefore\ \boldsymbol{w}_{[1,3]}^{(3)}=\boldsymbol{w}_{[1,3]}^{(2)}+y_1\mathbf{x}_1=(1,1,-1)^T$

$sign\left(\boldsymbol{w}_{[1,3]}^{(3)T}\mathbf{x}_3\right)=-1=y_3,且\ sign\left(\boldsymbol{w}_{[1,3]}^{(3)T}\mathbf{x}_1\right)=1=y_1$

$\therefore\ \boldsymbol{w}_{[1,3]}=(1,1,-1)^T$，分类面为：$1+x_1-x_2=0$

（3）用感知器算法求第二类和第三类之间的分类面

样本增广后为：$\mathbf{x}_2=(1,1,1)^T$，$y_2=1$，$\mathbf{x}_3=(1,-1,1)^T$，$y_3=-1$，初始化权重：$\boldsymbol{w}_{[2,3]}^{(0)}=(0,0,0)^T$

$sign\left(\boldsymbol{w}_{[2,3]}^{(0)T}\mathbf{x}\mathbf{x}_2\right)=0\neq y_2$，$\therefore\ \boldsymbol{w}_{[2,3]}^{(1)}=\boldsymbol{w}_{[2,3]}^{(0)}+y_2\mathbf{x}_2=(1,1,1)^T$，

$sign\left(\boldsymbol{w}_{[2,3]}^{(1)T}\mathbf{x}_3\right)=1\neq y_3$，$\therefore\ \boldsymbol{w}_{[2,3]}^{(2)}=\boldsymbol{w}_{[2,3]}^{(1)}+y_3\mathbf{x}_3=(0,2,0)^T$

$sign\left(\boldsymbol{w}_{[2,3]}^{(2)T}\mathbf{x}_2\right)=1=y_2$, 且 $sign\left(\boldsymbol{w}_{[2,3]}^{(2)T}\mathbf{x}_3\right)=-1=y_3$

$\therefore\ \boldsymbol{w}_{[2,3]}=(0,2,0)^T$，分类面为：$x_1=0$

对测试样本进行增广，$\mathbf{x}=(1,1,-2)^T$，分别代入上述三个分类面：

第一类和第二类：

$sign\left(\boldsymbol{w}_{[1,2]}^T\mathbf{x}\right)=sign((1,-1,-1)(1,1,-2)^T=1,\quad\therefore\ \mathbf{x}\in$ *第一类*

第一类和第三类：

$sign\left(\boldsymbol{w}_{[1,3]}^T\mathbf{x}\right)=sign((1,1,-1)(1,1,-2)^T=1,\quad\therefore\ \mathbf{x}\in$ *第一类*

第二类和第三类：

$sign\left(\boldsymbol{w}_{[2,3]}^T\mathbf{x}\right)=sign((0,2,0)(1,1,-2)^T=1,\quad\therefore\ \mathbf{x}\in$ *第二类*

最终的投票结果是测试样本属于第一类。

2，现有四个样本，假设样本（3，0）和（3，6）属于第一类，样本（0，3）属于第二类，样本（-3，0）属于第三类，请用 Softmax 算法设计出这三个类别的分类器（假设这三个类别的初始权向量均为零向量，迭代步长取 1，需要写出计算过程）。

解：

（1）梯度的计算

假设输入样本$\mathbf{x}$属于K个类别$Y=\{1,2,\dots k,\dots K\}$中的某个类别k时，在Softmax中，我们按照式（1）计算其内积，按照式（2）计算其属于类别j的概率：

$$s_j=\boldsymbol{w}_j^T\mathbf{x} \qquad (1)$$

$$\hat{y}_j=\frac{e^{s_j}}{\sum_k e^{s_k}} \qquad (2)$$

经过Softmax函数后，得到的输出为K个类别的概率列向量：$\hat{\boldsymbol{y}}=(\hat{y}_1,\dots\hat{y}_j,\dots\hat{y}_K)^T$，假设理想的各个类别标签对应的概率为列向量：$\boldsymbol{y}=\{y_1,\dots y_j,\dots y_K\}$，且该列向量的一个元素为1，其他均为0，代表样本属于这个类别。我们选择用交叉熵作为误差函数其为表达式：

$$L_{in}(\boldsymbol{w}_k)=-\sum_{k=1}^{K}y_k ln\hat{y}_k=-ln\hat{y}_k \qquad (3)$$

我们可以计算L_{in}对于$\boldsymbol{w}_j(j=1,2,\dots,K)$的梯度：

$$\frac{\partial L_{in}}{\partial \boldsymbol{w}_j}=\frac{\partial E_{in}}{\partial \hat{y}_k}\frac{\partial \hat{y}_k}{\partial s_j}\frac{\partial s_j}{\partial \boldsymbol{w}_j}=-\frac{1}{\hat{y}_k}\frac{\partial \hat{y}_k}{\partial s_j}\mathbf{x}^T \qquad (4)$$

我们再来计算$\frac{\partial \hat{y}_k}{\partial s_j}$：

$$\frac{\partial \hat{y}_k}{\partial s_j}=\frac{\partial}{\partial s_j}\left(\frac{e^{s_k}}{\sum_k e^{s_k}}\right)=\frac{(e^{s_k})'\sum_k e^{s_k}-(\sum_k e^{s_k})'e^{s_k}}{(\sum_k e^{s_k})^2}=$$

$$\begin{cases}\frac{e^{s_j}\sum_k e^{s_k}-e^{s_j}e^{s_k}}{(\sum_k e^{s_k})^2}=\frac{e^{s_j}}{\sum_k e^{s_k}}-\frac{e^{s_j}}{\sum_k e^{s_k}}\frac{e^{s_k}}{\sum_k e^{s_k}}=\hat{y}_j(1-\hat{y}_k) & j=k\\ \frac{0\sum_k e^{s_k}-e^{s_j}e^{s_k}}{(\sum_k e^{s_k})^2}=0-\frac{e^{s_j}}{\sum_k e^{s_k}}\frac{e^{s_k}}{\sum_k e^{s_k}}=-\hat{y}_k\hat{y}_j & j\neq k\end{cases} \qquad (5)$$

将式（5）代入到式（4），我们得到E_{in}对于$\vec{w}_j$的梯度：

$$\frac{\partial L_{in}}{\partial \boldsymbol{w}_j}=\frac{\partial L_{in}}{\partial \hat{y}_k}\frac{\partial \hat{y}_k}{\partial s_j}\frac{\partial s_j}{\partial \boldsymbol{w}_j}=-\frac{1}{\hat{y}_k}\frac{\partial \hat{y}_k}{\partial s_j}\mathbf{x}^T=\begin{cases}(\hat{y}_j-1)\mathbf{x}^T & j=k\\ \hat{y}_j\mathbf{x}^T & j\neq k\end{cases} \qquad (6)$$

针对N个训练样本，将上述推导及求解过程写成矩阵或向量形式如下：

假设训练样本集有N个样本$\{\mathbf{x}_1, \ldots \mathbf{x}_n, \ldots \mathbf{x}_N\}$，每个样本有d维特征，写成增广向量后是d+1维，$\mathbf{x}_n = (x_{n0}, x_{n1}, \ldots x_{nd})^T$，所有的训练样本我们用$\mathbf{X}$来表示成一个N*(d+1)维的矩阵：

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \\ \vdots \\ \mathbf{x}_N^T \end{pmatrix} = \begin{pmatrix} x_{10} & \cdots & x_{1d} \\ \vdots & \ddots & \vdots \\ x_{N0} & \cdots & x_{Nd} \end{pmatrix} \tag{7}$$

所有训练样本标签对应的概率输出用N*K维矩阵表示，其中K是类别数，样本只能属于其中一个类别且概率取1，其他类别概率为0，假设如下表示的第一个样本属于类别1，第N个样本属于类别K：

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_n \\ \vdots \\ \boldsymbol{y}_N \end{pmatrix} = \begin{pmatrix} y_{11} & \cdots & y_{1K} \\ \vdots & \ddots & \vdots \\ y_{N1} & \cdots & y_{NK} \end{pmatrix} = \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix} \tag{8}$$

经过式(1)、式(2)后，我们得到的样本类别的概率估计值为N*K维矩阵$\widehat{\boldsymbol{Y}}$：

$$\widehat{\boldsymbol{Y}} = \begin{pmatrix} \widehat{\boldsymbol{y}}_1 \\ \vdots \\ \widehat{\boldsymbol{y}}_n \\ \vdots \\ \widehat{\boldsymbol{y}}_N \end{pmatrix} = \begin{pmatrix} \hat{y}_{11} & \cdots & \hat{y}_{1K} \\ \vdots & \ddots & \vdots \\ \hat{y}_{N1} & \cdots & \hat{y}_{NK} \end{pmatrix} \tag{9}$$

根据式（6）得到L_{in}的梯度可以写为：

$$\nabla L_{in} = (\widehat{\boldsymbol{Y}} - \boldsymbol{Y})^T \mathbf{X} = \left(\widehat{\boldsymbol{y}}_1 - \boldsymbol{y}_1, \ldots \widehat{\boldsymbol{y}}_n - \boldsymbol{y}_n, \ldots \widehat{\boldsymbol{y}}_N - \boldsymbol{y}_N\right) \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \\ \vdots \\ \mathbf{x}_N^T \end{pmatrix} = \begin{pmatrix} \sum_{n=1}^{N} (\hat{y}_{n1} - y_{n1}) \mathbf{x}_\mathrm{n}^\mathrm{T} \\ \vdots \\ \sum_{n=1}^{N} (\hat{y}_{nj} - y_{nj}) \mathbf{x}_n^T \\ \vdots \\ \sum_{n=1}^{N} (\hat{y}_{nK} - y_{nK}) \mathbf{x}_n^T \end{pmatrix}$$

10)

这相当于K*N维的矩阵与N*(d+1)维的矩阵做内积，得到K*(d+1)维的梯度，这里y_{nj}只会取0或者1。

假设类别对应的权系数向量用$\boldsymbol{w}$表示，加上常数项，它也是(d+1)维，一共K个类别，可以写成(d+1)*K维矩阵形式：

$$\mathbf{W}=(\boldsymbol{w_1}\quad\cdots\quad\boldsymbol{w_j}\quad\cdots\quad\boldsymbol{w_K})=\begin{pmatrix}w_{01} & \cdots & w_{0K}\\ \vdots & \ddots & \vdots\\ w_{d1} & \cdots & w_{dK}\end{pmatrix} \tag{11}$$

假设学习率为η，迭代次数用上标t表示，利用梯度下降法得到权重的更新式：

$$\boldsymbol{w}^{\boldsymbol{T}^{(t+1)}}=\boldsymbol{w}^{\boldsymbol{T}^{(t)}}-\eta\nabla L_{in}=\begin{pmatrix}\boldsymbol{w}_1^{(t)}-\eta\sum_{n=1}^{N}\left(\hat{y}_{n1}-y_{n1}\right)\mathbf{x}_{\mathrm{n}}^{\mathrm{T}}\\ \vdots\\ \boldsymbol{w}_j^{(t)}-\eta\sum_{n=1}^{N}\left(\hat{y}_{nj}-y_{nj}\right)\mathbf{x}_{\mathrm{n}}^{\mathrm{T}}\\ \vdots\\ \boldsymbol{w}_K^{(t)}-\eta\sum_{n=1}^{N}\left(\hat{y}_{nK}-y_{nK}\right)\mathbf{x}_{\mathrm{n}}^{\mathrm{T}}\end{pmatrix} \tag{12}$$

根据更新后的权重，我们可以重新计算每个样本在每个类别权系数向量下的内积S，同样，我们也可以把S写成矩阵形式，它是N*K维矩阵：

$$\boldsymbol{S}=\mathbf{X}\boldsymbol{W}^{(\mathrm{t+1})}=\begin{pmatrix}\mathbf{x}_1^T\\ \vdots\\ \mathbf{x}_n^T\\ \vdots\\ \mathbf{x}_N^T\end{pmatrix}\left(\boldsymbol{w}_1{}^{(t+1)},\dots,\boldsymbol{w}_j{}^{(t+1)},\dots\boldsymbol{w}_K{}^{(t+1)}\right)=$$

$$\begin{pmatrix}(\boldsymbol{w}_1{}^{(t+1)})^T\mathbf{x}_1 & \cdots & (\boldsymbol{w}_K{}^{(t+1)})^T\mathbf{x}_1\\ \vdots & \ddots & \vdots\\ (\boldsymbol{w}_1{}^{(t+1)})^T\mathbf{x}_N & \cdots & (\boldsymbol{w}_K{}^{(t+1)})^T\mathbf{x}_N\end{pmatrix}=\begin{pmatrix}s_{11} & \cdots & s_{1j} & \cdots & s_{1K}\\ \vdots & \cdots & \cdots & \cdots & \vdots\\ s_{n1} & \cdots & s_{nj} & \cdots & s_{nK}\\ \vdots & \cdots & \cdots & \cdots & \vdots\\ s_{N1} & \cdots & s_{Nj} & \cdots & s_{NK}\end{pmatrix} \tag{13}$$

利用Softmax可以得到：

$$\widehat{\boldsymbol{Y}}=\begin{pmatrix}\widehat{\boldsymbol{y}}_1\\ \vdots\\ \widehat{\boldsymbol{y}}_n\\ \vdots\\ \widehat{\boldsymbol{y}}_N\end{pmatrix}=\begin{pmatrix}\hat{y}_{11} & \cdots & \hat{y}_{1K}\\ \vdots & \ddots & \vdots\\ \hat{y}_{N1} & \cdots & \hat{y}_{NK}\end{pmatrix} \tag{14}$$

因为对于一个样本的误差函数为式（3），所以，对于所有样本其误差函数（损失函数）为：

$$L_{in}=\frac{1}{N}\sum_{n=1}^{N}(-ln\hat{y}_{nk}) \tag{15}$$

（2）习题的求解

首 先， 将 样 本 变 为 增 广 向 量： $\mathbf{x}_1=(1,3,0)^T,\mathbf{x}_2=(1,3,6)^T,\mathbf{x}_3=(1,0,3)^T,\mathbf{x}_4=$

$(1,-3,0)^T$，得到：

$$\mathbf{X}=\begin{pmatrix}1&3&0\\1&3&6\\1&0&3\\1&-3&0\end{pmatrix}$$

四个样本对应的理想概率值为 $\boldsymbol{y}_1=(1,0,0)^T$，$\boldsymbol{y}_2=(1,0,0)^T$，$\boldsymbol{y}_3=(0,1,0)^T$，$\boldsymbol{y}_4=(0,0,1)^T$，即：

$$\boldsymbol{Y}=\begin{pmatrix}1&0&0\\1&0&0\\0&1&0\\0&0&1\end{pmatrix}$$

假设三个类别的初始权向量为：$\boldsymbol{w}_1^{(0)}=(0,0,0)^T$，$\boldsymbol{w}_2^{(0)}=(0,0,0)^T$，$\boldsymbol{w}_3^{(0)}=(0,0,0)^T$，即：

$$\boldsymbol{W}^{(0)}=\begin{pmatrix}0&0&0\\0&0&0\\0&0&0\end{pmatrix}$$

令$\eta=1$。

第一次迭代：t=0，将$\mathbf{x}_n$，$(n=1,2,3,4)$，$\boldsymbol{w}_k^{(0)},(k=1,2,3)$代入到式（13），得到：

$$\boldsymbol{S}=\mathbf{X}\boldsymbol{W}^{(0)}=\begin{pmatrix}1&3&0\\1&3&6\\1&0&3\\1&-3&0\end{pmatrix}\begin{pmatrix}0&0&0\\0&0&0\\0&0&0\end{pmatrix}=\begin{pmatrix}0&0&0\\0&0&0\\0&0&0\\0&0&0\end{pmatrix}$$

利用式(2)和式(14)得到：

$$\widehat{\boldsymbol{Y}}=\begin{pmatrix}\widehat{\boldsymbol{y}}_1\\\widehat{\boldsymbol{y}}_2\\\widehat{\boldsymbol{y}}_3\\\widehat{\boldsymbol{y}}_4\end{pmatrix}=\begin{pmatrix}\frac{1}{3}&\frac{1}{3}&\frac{1}{3}\\\frac{1}{3}&\frac{1}{3}&\frac{1}{3}\\\frac{1}{3}&\frac{1}{3}&\frac{1}{3}\\\frac{1}{3}&\frac{1}{3}&\frac{1}{3}\end{pmatrix}$$

显然所有样本都没有正确分类，按照式(15)，每一个样本任意选择一个类别获得其概率，计算$L_{in}=\frac{1}{4}\sum_{n=1}^{4}(-ln\frac{1}{3})=1.099$

所以，我们按照式（10）求得梯度：

$$\nabla L_{in} = (\widehat{\boldsymbol{Y}} - \boldsymbol{Y})^T \mathbf{X} = \begin{pmatrix} \frac{1}{3}-1 & \frac{1}{3}-1 & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3}-1 & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3}-1 \end{pmatrix} \begin{pmatrix} 1 & 3 & 0 \\ 1 & 3 & 6 \\ 1 & 0 & 3 \\ 1 & -3 & 0 \end{pmatrix} = \begin{pmatrix} -\frac{2}{3} & -5 & -3 \\ \frac{1}{3} & 1 & 0 \\ \frac{1}{3} & 4 & 3 \end{pmatrix}$$

用梯度下降法式(12)进行权系数向量更新:

$$\boldsymbol{w}^{\boldsymbol{T}^{(1)}} = \boldsymbol{w}^{\boldsymbol{T}^{(0)}} - \eta \nabla L_{in} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} - \begin{pmatrix} -\frac{2}{3} & -5 & -3 \\ \frac{1}{3} & 1 & 0 \\ \frac{1}{3} & 4 & 3 \end{pmatrix} = \begin{pmatrix} \frac{2}{3} & 5 & 3 \\ -\frac{1}{3} & -1 & 0 \\ -\frac{1}{3} & -4 & -3 \end{pmatrix}$$

根据式(13)得到S矩阵:

$$\boldsymbol{S} = \mathbf{X}\boldsymbol{W}^{(1)} = \begin{pmatrix} 1 & 3 & 0 \\ 1 & 3 & 6 \\ 1 & 0 & 3 \\ 1 & -3 & 0 \end{pmatrix} \begin{pmatrix} \frac{2}{3} & -\frac{1}{3} & -\frac{1}{3} \\ 5 & -1 & -4 \\ 3 & 0 & -3 \end{pmatrix} = \begin{pmatrix} 15.67 & -3.33 & -12.33 \\ 33.67 & -3.33 & -30.33 \\ 9.67 & -0.33 & -9.33 \\ -14.33 & 2.67 & 11.67 \end{pmatrix}$$

利用Softmax得到:

$$\widehat{\boldsymbol{Y}} = \begin{pmatrix} \widehat{\boldsymbol{y}}_1 \\ \widehat{\boldsymbol{y}}_2 \\ \widehat{\boldsymbol{y}}_3 \\ \widehat{\boldsymbol{y}}_4 \end{pmatrix} = \begin{pmatrix} 1.00 & 0.00 & 0.00 \\ 1.00 & 0.00 & 0.00 \\ 1.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 1.00 \end{pmatrix}$$

第三个样本错分，计算$L_{in} = (-ln1 - ln1 - ln0 - ln1)/4 = \infty$

第二次迭代:

我们按照式（10）求得梯度:

$$\nabla L_{in} = (\widehat{\boldsymbol{Y}} - \boldsymbol{Y})^T \mathbf{X} = \begin{pmatrix} 1-1 & 1-1 & 1 & 0 \\ 0 & 0 & 0-1 & 0 \\ 0 & 0 & 0 & 1-1 \end{pmatrix} \begin{pmatrix} 1 & 3 & 0 \\ 1 & 3 & 6 \\ 1 & 0 & 3 \\ 1 & -3 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 3 \\ -1 & 0 & -3 \\ 0 & 0 & 0 \end{pmatrix}$$

用梯度下降法式(12)进行权系数向量更新:

$$\boldsymbol{w}^{\boldsymbol{T}^{(2)}} = \boldsymbol{w}^{\boldsymbol{T}^{(1)}} - \eta \nabla L_{in} = \begin{pmatrix} \frac{2}{3} & 5 & 3 \\ -\frac{1}{3} & -1 & 0 \\ -\frac{1}{3} & -4 & -3 \end{pmatrix} - \begin{pmatrix} 1 & 0 & 3 \\ -1 & 0 & -3 \\ 0 & 0 & 0 \end{pmatrix}$$

$$= \begin{pmatrix} -0.33 & 5 & 0 \\ 0.67 & -1 & 3 \\ -0.33 & -4 & -3 \end{pmatrix}$$

根据式(13)得到S矩阵:

$$\boldsymbol{S} = \mathbf{X}\boldsymbol{W}^{(2)} = \begin{pmatrix} 1 & 3 & 0 \\ 1 & 3 & 6 \\ 1 & 0 & 3 \\ 1 & -3 & 0 \end{pmatrix} \begin{pmatrix} -0.33 & 0.67 & -0.33 \\ 5 & -1 & -4 \\ 0 & 3 & -3 \end{pmatrix} = \begin{pmatrix} 14.67 & -2.33 & -12.33 \\ 14.67 & 15.67 & -30.33 \\ -0.33 & 9.67 & -9.33 \\ -15.33 & 3.67 & 11.27 \end{pmatrix}$$

利用Softmax得到:

$$\widehat{\boldsymbol{Y}} = \begin{pmatrix} \widehat{\boldsymbol{y}}_1 \\ \widehat{\boldsymbol{y}}_2 \\ \widehat{\boldsymbol{y}}_3 \\ \widehat{\boldsymbol{y}}_4 \end{pmatrix} = \begin{pmatrix} 1.00 & 0.00 & 0.00 \\ 0.27 & 0.73 & 0.00 \\ 0.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 1.00 \end{pmatrix}$$

第二个样本错分，计算$L_{in} = (-ln1 - ln0.27 - ln1 - ln1)/4 = 0.33$

第三次迭代:

我们按照式（10）求得梯度:

$$\nabla L_{in} = (\widehat{\boldsymbol{Y}} - \boldsymbol{Y})^T \mathbf{X} = \begin{pmatrix} 1-1 & 0.27-1 & 0 & 0 \\ 0 & 0.73 & 1-1 & 0 \\ 0 & 0 & 0 & 1-1 \end{pmatrix} \begin{pmatrix} 1 & 3 & 0 \\ 1 & 3 & 6 \\ 1 & 0 & 3 \\ 1 & -3 & 0 \end{pmatrix}$$

$$= \begin{pmatrix} -0.73 & -2.19 & -4.38 \\ 0.73 & 2.19 & 4.38 \\ 0 & 0 & 0 \end{pmatrix}$$

用梯度下降法式(12)进行权系数向量更新:

$$\boldsymbol{w}^{\boldsymbol{T}^{(3)}} = \boldsymbol{w}^{\boldsymbol{T}^{(2)}} - \eta \nabla L_{in} = \begin{pmatrix} -0.33 & 5 & 0 \\ 0.67 & -1 & 3 \\ -0.33 & -4 & -3 \end{pmatrix} - \begin{pmatrix} -0.73 & -2.19 & -4.38 \\ 0.73 & 2.19 & 4.38 \\ 0 & 0 & 0 \end{pmatrix}$$

$$= \begin{pmatrix} 0.40 & 7.19 & 4.38 \\ -0.06 & -3.19 & -1.38 \\ -0.33 & -4 & -3 \end{pmatrix}$$

根据式(13)得到S矩阵：

$\boldsymbol{S}=\mathbf{X}\boldsymbol{W}^{(3)}=$

$$\begin{pmatrix}1&3&0\\1&3&6\\1&0&3\\1&-3&0\end{pmatrix}\begin{pmatrix}0.40&-0.06&-0.33\\7.19&-3.19&-4\\4.38&-1.38&-3\end{pmatrix}=\begin{pmatrix}21.97&-9.63&-12.33\\48.25&-17.91&-30.33\\13.54&-4.20&-9.33\\-21.17&9.51&11.67\end{pmatrix}$$

利用Softmax得到：

$$\widehat{\boldsymbol{Y}}=\begin{pmatrix}\widehat{\boldsymbol{y}}_1\\\widehat{\boldsymbol{y}}_2\\\widehat{\boldsymbol{y}}_3\\\widehat{\boldsymbol{y}}_4\end{pmatrix}=\begin{pmatrix}1.00&0.00&0.00\\1.00&0.00&0.00\\1.00&0.00&0.00\\0.00&0.11&0.89\end{pmatrix}$$

第三个样本错分，计算$L_{in}=(-ln1-ln1-ln0-ln0.89)/4=\infty$

第四次迭代：

我们按照式（10）求得梯度：

$$\nabla L_{in}=(\widehat{\boldsymbol{Y}}-\boldsymbol{Y})^T\mathbf{X}=\begin{pmatrix}1-1&1-1&1&0\\0&0&0-1&0.11\\0&0&0&0.89-1\end{pmatrix}\begin{pmatrix}1&3&0\\1&3&6\\1&0&3\\1&-3&0\end{pmatrix}$$

$$=\begin{pmatrix}1&0&3\\-0.89&-0.33&-3\\-0.11&0.33&0\end{pmatrix}$$

用梯度下降法式(12)进行权系数向量更新：

$$\boldsymbol{w}^{T^{(4)}}=\boldsymbol{w}^{T^{(3)}}-\eta\nabla L_{in}=\begin{pmatrix}0.40&7.19&4.38\\-0.06&-3.19&-1.38\\-0.33&-4&-3\end{pmatrix}-\begin{pmatrix}1&0&3\\-0.89&-0.33&-3\\-0.11&0.33&0\end{pmatrix}$$

$$=\begin{pmatrix}-0.60&7.19&1.38\\0.83&-2.86&1.62\\-0.22&-4.33&-3\end{pmatrix}$$

根据式(13)得到S矩阵：

$\boldsymbol{S}=\mathbf{X}\boldsymbol{W}^{(4)}=$

$$\begin{pmatrix}1&3&0\\1&3&6\\1&0&3\\1&-3&0\end{pmatrix}\begin{pmatrix}-0.60&0.83&-0.22\\7.19&-2.86&-4.33\\1.38&1.62&-3\end{pmatrix}=\begin{pmatrix}20.97&-7.75&-13.21\\29.25&1.97&-31.21\\3.54&5.69&-9.22\\-22.17&9.41&12.77\end{pmatrix}$$

利用Softmax得到：

$$\widehat{\boldsymbol{Y}}=\begin{pmatrix}\widehat{\boldsymbol{y}}_1\\\widehat{\boldsymbol{y}}_2\\\widehat{\boldsymbol{y}}_3\\\widehat{\boldsymbol{y}}_4\end{pmatrix}=\begin{pmatrix}1.00 & 0.00 & 0.00\\1.00 & 0.00 & 0.00\\0.10 & 0.90 & 0.00\\0.00 & 0.02 & 0.98\end{pmatrix}$$

所有样本均正确分类，计算$L_{in}=(-ln1-ln1-ln0.90-ln0.98)/4=0.03$

此时求得的权系数向量矩阵为：

$$\boldsymbol{W}^{(4)}=\begin{pmatrix}-0.60 & 0.83 & -0.22\\7.19 & -2.86 & -4.33\\1.38 & 1.62 & -3\end{pmatrix}$$

即：

$\boldsymbol{w}_1=(-0.60,7.19,1.38)^T$

$\boldsymbol{w}_2=(0.83,-2.86,1.62)^T$

$\boldsymbol{w}_3=(-0.22,-4.33,-3)^T$

不习惯看矩阵的，可以看如下求解过程：

第一次迭代：将$\mathbf{x}_n$，$(n=1,2,3,4)$，$\boldsymbol{w}_k^{(0)},(k=1,2,3)$代入到式（1），对每一个样本均得到$s_1=s_2=s_3=0$，代入式（2）得到：$\widehat{\boldsymbol{Y}}=(\hat{y}_1,\hat{y}_2,\hat{y}_3)^T=(\frac{1}{3},\frac{1}{3},\frac{1}{3})^T$，显然这个样本没有正确分类，所以，我们按照式（6）求得梯度去计算新的$\boldsymbol{w}_k$，我们以计算$\boldsymbol{w}_1$为例，先用式（6）计算梯度：

$$\frac{\partial L_{in}}{\partial \boldsymbol{w}_1}=\sum_{n=1}^{4}\frac{\partial L_{in}(\mathbf{x}_n)}{\partial \boldsymbol{w}_1}=(\hat{y}_1-1)\mathbf{x}_1^T+(\hat{y}_1-1)\mathbf{x}_2^T+\hat{y}_2\mathbf{x}_3^T+\hat{y}_3\mathbf{x}_4^T$$

$$=\left(\frac{1}{3}-1\right)\mathbf{x}_1^T+\left(\frac{1}{3}-1\right)\mathbf{x}_2^T+\frac{1}{3}\mathbf{x}_3^T+\frac{1}{3}\mathbf{x}_4^T=(-\frac{2}{3},-5,-3)^T$$

同理，我们可以得到：$\frac{\partial L_{in}}{\partial \boldsymbol{w}_2}=(\frac{1}{3},1,0)^T$，$\frac{\partial L_{in}}{\partial \boldsymbol{w}_3}=(\frac{1}{3},4,3)^T$

用梯度下降法对$\boldsymbol{w}_k$进行更新：

$$\boldsymbol{w}_1^{(1)} = \boldsymbol{w}_1^{(0)} - \frac{\partial L_{in}}{\partial \boldsymbol{w}_1} = (0,0,0)^T - \left(-\frac{2}{3}, -5, -3\right)^T = (\frac{2}{3}, 5,3)^T$$

$$\boldsymbol{w}_2^{(1)} = \boldsymbol{w}_2^{(0)} - \frac{\partial L_{in}}{\partial \boldsymbol{w}_2} = (0,0,0)^T - \left(\frac{1}{3}, 1,0\right)^T = (-\frac{1}{3}, -1,0)^T$$

$$\boldsymbol{w}_3^{(1)} = \boldsymbol{w}_3^{(0)} - \frac{\partial L_{in}}{\partial \boldsymbol{w}_3} = (0,0,0)^T - \left(\frac{1}{3}, 4,3\right)^T = (-\frac{1}{3}, -4, -3)^T$$

根据$\boldsymbol{w}_1^{(1)}$，$\boldsymbol{w}_2^{(1)}$和$\boldsymbol{w}_3^{(1)}$，我们用式（1）得到：

对于$\mathbf{x}_1$，我们有：$s_1 = \boldsymbol{w}_1^T \mathbf{x}_1 = \left(\frac{2}{3}, 5,3\right)\begin{pmatrix}1\\3\\0\end{pmatrix} = 15.67$，$s_2 = \boldsymbol{w}_2^T \mathbf{x}_1 = \left(-\frac{1}{3}, -1,0\right)\begin{pmatrix}1\\3\\0\end{pmatrix} = -3.33$，$s_3 = \boldsymbol{w}_3^T \mathbf{x}_1 = \left(-\frac{1}{3}, -4, -3\right)\begin{pmatrix}1\\3\\0\end{pmatrix} = -12.33$

利用式（2），我们可以得到：$\hat{y}_1 = \frac{e^{s_1}}{e^{s_1}+e^{s_2}+e^{s_3}} = 1.00$，$\hat{y}_2 = \frac{e^{s_2}}{e^{s_1}+e^{s_2}+e^{s_3}} = 0.00$，$\hat{y}_3 = \frac{e^{s_3}}{e^{s_1}+e^{s_2}+e^{s_3}} = 0.00$，即，$\hat{\boldsymbol{y}}_1 = (1.00,0.00,0.00)^T$，对照$\boldsymbol{y}_1 = (1,0,0)^T$，此时对于样本$\mathbf{x}_1$分类是正确的。

同理：对于$\mathbf{x}_2$，我们有$s_1 = 33.67$，$s_2 = -3.33$，$s_3 = -30.33$，对应的我们可以计算出$\hat{\boldsymbol{y}}_2 = (1.00,0.00,0.00)^T$，对照$\boldsymbol{y}_2 = (1,0,0)^T$，此时对于样本$\mathbf{x}_2$分类是正确的。

对于$\mathbf{x}_3$，我们有$s_1 = 9.67$，$s_2 = -0.33$，$s_3 = -9.33$，对应的我们可以计算出$\hat{\boldsymbol{y}}_3 = (1.00,0.00,0.00)^T$，对照$\boldsymbol{y}_3 = (0,1,0)^T$，此时对于样本$\mathbf{x}_3$分类是错误的。

对于$\mathbf{x}_4$，我们有$s_1 = -14.33$，$s_2 = 2.67$，$s_3 = 11.67$，对应的我们可以计算出$\hat{\boldsymbol{y}}_4 = (0.00,0.00,1.00)^T$，对照$\boldsymbol{y}_4 = (0,0,1)^T$，此时对于样本$\mathbf{x}_4$分类是正确的。

第三个样本错分，计算$L_{in} = (-ln1 - ln1 - ln0 - ln1)/4 = \infty$

第二次迭代：我们需要按照式（6）重新计算梯度去得到新的$\boldsymbol{w}_k$，仍以计算$\boldsymbol{w}_1$为

例，先用式（6）计算梯度：

$$\frac{\partial L_{in}}{\partial \boldsymbol{w}_1}=\sum_{n=1}^{4}\frac{\partial L_{in}(\mathbf{x}_n)}{\partial \boldsymbol{w}_1}=(\hat{y}_1-1)\mathbf{x}_1^T+(\hat{y}_1-1)\mathbf{x}_2^T+\hat{y}_2\mathbf{x}_3^T+\hat{y}_3\mathbf{x}_4^T$$

$$=(1-1)\mathbf{x}_1^T+(1-1)\mathbf{x}_2^T+1\mathbf{x}_3^T+0\mathbf{x}_4^T=(1,0,3)^T$$

同理，我们可以得到：$\frac{\partial L_{in}}{\partial \boldsymbol{w}_2}=0\mathbf{x}_1^T+0\mathbf{x}_2^T+(0-1)\mathbf{x}_3^T+0\mathbf{x}_4^T=(-1,0,-3)^T$，$\frac{\partial L_{in}}{\partial \boldsymbol{w}_3}=0\mathbf{x}_1^T+0\mathbf{x}_2^T+0\mathbf{x}_3^T+(1-1)\mathbf{x}_4^T=(0,0,0)^T$

用梯度下降法对$\boldsymbol{w}_k$进行更新：

$$\boldsymbol{w}_1^{(2)}=\boldsymbol{w}_1^{(1)}-\frac{\partial L_{in}}{\partial \boldsymbol{w}_1}=(0.67,5,3)^T-(1,0,3)^T=(-0.33,5,0)^T$$

$$\boldsymbol{w}_2^{(2)}=\boldsymbol{w}_2^{(1)}-\frac{\partial L_{in}}{\partial \boldsymbol{w}_2}=(-0.33,-1,0)^T-(-1,0,-3)^T=(0.67,-1,3)^T$$

$$\boldsymbol{w}_3^{(2)}=\boldsymbol{w}_3^{(1)}-\frac{\partial L_{in}}{\partial \boldsymbol{w}_3}=(-0.33,-4,-3)^T-(0,0,0)^T=(-0.33,-4,-3)^T$$

根据$\boldsymbol{w}_1^{(2)}$，$\boldsymbol{w}_2^{(2)}$和$\boldsymbol{w}_3^{(2)}$，我们用式（1）得到：

对于$\mathbf{x}_1$，我们有：$s_1=\vec{w}_1^T\mathbf{x}_1=(-0.33,5,0)\begin{pmatrix}1\\3\\0\end{pmatrix}=14.67$，$s_2=\boldsymbol{w}_2^T\mathbf{x}_1=(0.67,-1,3)\begin{pmatrix}1\\3\\0\end{pmatrix}=-2.33$，$s_3=\boldsymbol{w}_3^T\mathbf{x}_1=(-0.33,-4,-3)\begin{pmatrix}1\\3\\0\end{pmatrix}=-12.33$

利用式（2），我们可以得到：$\hat{y}_1=\frac{e^{s_1}}{e^{s_1}+e^{s_2}+e^{s_3}}=1.00$，$\hat{y}_2=\frac{e^{s_2}}{e^{s_1}+e^{s_2}+e^{s_3}}=0.00$，$\hat{y}_3=\frac{e^{s_3}}{e^{s_1}+e^{s_2}+e^{s_3}}=0.00$，即，$\hat{\boldsymbol{y}}_1=(1.00,0.00,0.00)^T$，对照$\boldsymbol{y}_1=(1,0,0)^T$，此时对于样本$\mathbf{x}_1$分类是正确的。

同理：对于$\mathbf{x}_2$，我们有$s_1=14.67$，$s_2=15.67$，$s_3=-30.33$，对应的我们可以计算出$\hat{\boldsymbol{y}}_2=(0.27,0.73,0.00)^T$，对照$\boldsymbol{y}_2=(1,0,0)^T$，此时对于样本$\mathbf{x}_2$分类是错误的。

对于$\mathbf{x}_3$，我们有$s_1=-0.33$，$s_2=9.67$，$s_3=-9.33$，对应的我们可以计算

出$\hat{\boldsymbol{y}}_3 = (0.00,1.00,0.00)^T$，对照$\boldsymbol{y}_3 = (0,1,0)^T$，此时对于样本$\mathbf{x}_3$分类是正确的。

对于$\mathbf{x}_4$，我们有$s_1 = -15.33$，$s_2 = 3.67$，$s_3 = 11.27$，对应的我们可以计算出$\hat{\boldsymbol{y}}_4 = (0.00,0.00,1.00)^T$，对照$\boldsymbol{y}_4 = (0,0,1)^T$，此时对于样本$\mathbf{x}_4$分类是正确的。

第二个样本错分，计算$L_{in} = (-ln1 - ln0.27 - ln1 - ln1)/4 = 0.33$

第三次迭代：我们需要按照式（6）重新计算梯度去得到新的$\boldsymbol{w}_k$，仍以计算$\boldsymbol{w}_1$为例，先用式（6）计算梯度：

$$\frac{\partial L_{in}}{\partial \boldsymbol{w}_1} = \sum_{n=1}^{4} \frac{\partial L_{in}(\mathbf{x}_n)}{\partial \boldsymbol{w}_1} = (\hat{y}_1 - 1)\mathbf{x}_1^T + (\hat{y}_1 - 1)\mathbf{x}_2^T + \hat{y}_2\mathbf{x}_3^T + \hat{y}_3\mathbf{x}_4^T$$

$$= (1-1)\mathbf{x}_1^T + (0.27-1)\mathbf{x}_2^T + 0\mathbf{x}_3^T + 0\mathbf{x}_4^T$$

$$= (-0.73, -2.19, -4.38)^T$$

同理，我们可以得到：$\frac{\partial L_{in}}{\partial \boldsymbol{w}_2} = 0\mathbf{x}_1^T + 0.73\mathbf{x}_2^T + (1-1)\mathbf{x}_3^T + 0\mathbf{x}_4^T = (0.73,2.19,4.38)^T$，$\frac{\partial L_{in}}{\partial \boldsymbol{w}_3} = 0\mathbf{x}_1^T + 0\mathbf{x}_2^T + 0\mathbf{x}_3^T + (1-1)\mathbf{x}_4^T = (0,0,0)^T$

用梯度下降法对$\boldsymbol{w}_k$进行更新：

$$\boldsymbol{w}_1^{(3)} = \boldsymbol{w}_1^{(2)} - \frac{\partial L_{in}}{\partial \boldsymbol{w}_1} = (-0.33,5,0)^T - (-0.73, -2.19, -4.38)^T$$

$$= (0.40, 7.19, 4.38)^T$$

$$\boldsymbol{w}_2^{(3)} = \boldsymbol{w}_2^{(2)} - \frac{\partial L_{in}}{\partial \boldsymbol{w}_2} = (0.67, -1,3)^T - (0.73,2.19,4.38)^T$$

$$= (-0.06, -3.19, -1.38)^T$$

$$\boldsymbol{w}_3^{(3)} = \boldsymbol{w}_3^{(2)} - \frac{\partial L_{in}}{\partial \boldsymbol{w}_3} = (-0.33, -4, -3)^T - (0,0,0)^T = (-0.33, -4, -3)^T$$

根据$\boldsymbol{w}_1^{(3)}$，$\boldsymbol{w}_2^{(3)}$和$\boldsymbol{w}_3^{(3)}$，我们用式（1）得到：

对于$\mathbf{x}_1$，我们有：$s_1 = \boldsymbol{w}_1^T\mathbf{x}_1 = (0.40, 7.19, 4.38)\begin{pmatrix}1\\3\\0\end{pmatrix} = 21.97$，$s_2 = \boldsymbol{w}_2^T\mathbf{x}_1 =$

$(-0.06,-3.19,-1.38)\begin{pmatrix}1\\3\\0\end{pmatrix}=-9.63$ ， $s_3=\boldsymbol{w}_3^T\mathbf{x}_1=(-0.33,-4,-3)\begin{pmatrix}1\\3\\0\end{pmatrix}=-12.33$

利用式（2），我们可以得到：$\hat{y}_1=\frac{e^{s_1}}{e^{s_1}+e^{s_2}+e^{s_3}}=1.0000$，$\hat{y}_2=\frac{e^{s_2}}{e^{s_1}+e^{s_2}+e^{s_3}}=0.0000$，$\hat{y}_3=\frac{e^{s_3}}{e^{s_1}+e^{s_2}+e^{s_3}}=0.0000$，即，$\hat{\boldsymbol{y}}_1=(1.00,0.00,0.00)^T$，对照$\boldsymbol{y}_1=(1,0,0)^T$，此时对于样本$\mathbf{x}_1$分类是正确的。

同理：对于$\mathbf{x}_2$，我们有$s_1=48.25$，$s_2=-17.91$，$s_3=-30.33$，对应的我们可以计算出$\hat{\boldsymbol{y}}_2=(1.00,0.00,0.00)^T$，对照$\boldsymbol{y}_2=(1,0,0)^T$，此时对于样本$\mathbf{x}_2$分类是正确的。

对于$\mathbf{x}_3$，我们有$s_1=13.54$，$s_2=-4.20$，$s_3=-9.33$，对应的我们可以计算出$\hat{\boldsymbol{y}}_3=(1.00,0.00,0.00)^T$，对照$\boldsymbol{y}_3=(0,1,0)^T$，此时对于样本$\mathbf{x}_3$分类是错误的。

对于$\mathbf{x}_4$，我们有$s_1=-21.17$，$s_2=9.51$，$s_3=11.67$，对应的我们可以计算出$\hat{\boldsymbol{y}}_4=(0.0000,0.11,0.89)^T$，对照$\boldsymbol{y}_4=(0,0,1)^T$，此时对于样本$\mathbf{x}_4$分类是正确的。

第三个样本错分，计算$L_{in}=(-ln1-ln1-ln0-ln0.89)/4=\infty$

第四次迭代：我们需要按照式（6）重新计算梯度去得到新的$\boldsymbol{w}_k$，仍以计算$\boldsymbol{w}_1$为例，先用式（6）计算梯度：

$$\frac{\partial L_{in}}{\partial \boldsymbol{w}_1}=\sum_{n=1}^{4}\frac{\partial L_{in}(\mathbf{x}_n)}{\partial \boldsymbol{w}_1}=(\hat{y}_1-1)\mathbf{x}_1^T+(\hat{y}_1-1)\mathbf{x}_2^T+\hat{y}_2\mathbf{x}_3^T+\hat{y}_3\mathbf{x}_4^T$$

$$=(1-1)\mathbf{x}_1^T+(1-1)\mathbf{x}_2^T+1\mathbf{x}_3^T+0\mathbf{x}_4^T=(1,0,3)^T$$

同理，我们可以得到：

$\frac{\partial L_{in}}{\partial \boldsymbol{w}_2} = 0\mathbf{x}_1^T + 0\mathbf{x}_2^T + (0-1)\mathbf{x}_3^T + 0.11\mathbf{x}_4^T = (-0.89, -0.33, -3)^T$,

$$\frac{\partial L_{in}}{\partial \boldsymbol{w}_3} = 0\mathbf{x}_1^T + 0\mathbf{x}_2^T + 0\mathbf{x}_3^T + (0.89-1)\mathbf{x}_4^T = (-0.11, 0.33, 0)^T$$

用梯度下降法对$\boldsymbol{w}_k$进行更新:

$$\boldsymbol{w}_1^{(4)} = \boldsymbol{w}_1^{(3)} - \frac{\partial L_{in}}{\partial \boldsymbol{w}_1} = (0.40, 7.19, 4.38)^T - (1,0,3)^T = (-0.60, 7.19, 1.38)^T$$

$$\boldsymbol{w}_2^{(4)} = \boldsymbol{w}_2^{(3)} - \frac{\partial L_{in}}{\partial \boldsymbol{w}_2} = (-0.06, -3.19, -1.38)^T - (-0.89, -0.33, -3)^T$$
$$= (0.83, -2.86, 1.62)^T$$

$$\boldsymbol{w}_3^{(4)} = \boldsymbol{w}_3^{(3)} - \frac{\partial L_{in}}{\partial \boldsymbol{w}_3} = (-0.33, -4, -3)^T - (-0.11, 0.33, 0)^T$$
$$= (-0.22, -4.33, -3)^T$$

根据$\boldsymbol{w}_1^{(4)}$,$\boldsymbol{w}_2^{(4)}$和$\boldsymbol{w}_3^{(4)}$,我们用式(1)得到:

对于$\mathbf{x}_1$,我们有:$s_1 = \boldsymbol{w}_1^T\mathbf{x}_1 = (-0.60, 7.19, 1.38)\begin{pmatrix}1\\3\\0\end{pmatrix} = 20.97$,$s_2 = \boldsymbol{w}_2^T\mathbf{x}_1 = (0.83, -2.86, 1.62)\begin{pmatrix}1\\3\\0\end{pmatrix} = -7.75$,$s_3 = \boldsymbol{w}_3^T\mathbf{x}_1 = (-0.18, -4.45, -3)\begin{pmatrix}1\\3\\0\end{pmatrix} = -13.21$

利用式(2),我们可以得到:$\hat{y}_1 = \frac{e^{s_1}}{e^{s_1}+e^{s_2}+e^{s_3}} = 1.00$,$\hat{y}_2 = \frac{e^{s_2}}{e^{s_1}+e^{s_2}+e^{s_3}} = 0.00$,$\hat{y}_3 = \frac{e^{s_3}}{e^{s_1}+e^{s_2}+e^{s_3}} = 0.00$,即,$\hat{\boldsymbol{y}}_1 = (1.00, 0.00, 0.00)^T$,对照$\boldsymbol{y}_1 = (1,0,0)^T$,此时对于样本$\mathbf{x}_1$分类是正确的。

同理:对于$\mathbf{x}_2$,我们有$s_1 = 29.25$,$s_2 = 1.97$,$s_3 = -31.21$,对应的我们可以计算出$\hat{\boldsymbol{y}}_2 = (1.00, 0.00, 0.00)^T$,对照$\boldsymbol{y}_2 = (1,0,0)^T$,此时对于样本$\mathbf{x}_2$分类是正确的。

对于$\mathbf{x}_3$,我们有$s_1 = 3.54$,$s_2 = 5.69$,$s_3 = -9.22$,对应的我们可以计算出$\hat{\boldsymbol{y}}_3 = (0.10, 0.90, 0.00)^T$,对照$\boldsymbol{y}_3 = (0,1,0)^T$,此时对于样本$\mathbf{x}_3$分类是正确的。

对于$\mathbf{x}_4$，我们有$s_1 = -22.17$，$s_2 = 9.41$，$s_3 = 12.77$，对应的我们可以计算出$\hat{\boldsymbol{y}}_4 = (0.00,0.02,0.98)^T$，对照$\boldsymbol{y}_4 = (0,0,1)^T$，此时对于样本$\mathbf{x}_4$分类是正确的。

计算$L_{in} = (-ln1 - ln1 - ln0.90 - ln0.98)/4 = 0.03$

于是我们最终得到的是：

$\boldsymbol{w}_1 = (-0.60, 7.19, 1.38)^T$

$\boldsymbol{w}_2 = (0.83, -2.86, 1.62)^T$

$\boldsymbol{w}_3 = (-0.22, -4.33, -3)^T$

Lecture10-11 作业

1，假设 $g_0(\vec{x})=1$，以下哪一组（$\alpha_0,\alpha_1,\alpha_2$）允许 $G(\vec{x})=sign(\sum_{t=0}^{2}\alpha_t g_t(\vec{x}))$实现$OR(g_1,g_2)$的功能。（a）（-3，+1，+1）；（b）（-1，+1，+1）；（c）（+1，+1，+1）；（d）（+3，+1，+1）。

解：$OR(g_1,g_2)$的关系意味着只要有一个$g_i=1$，输出即为 1，当两个都为“-1”时，输出才为-1。

根据题目条件:

(a)$G(\vec{x})=sign(-3+g_1(\vec{x})+g_2(\vec{x}))$，当$g_1(\vec{x})=g_2(\vec{x})=1$时，$G(\vec{x})=-1$，不满足定义;

(b) $G(\vec{x})=sign(-1+g_1(\vec{x})+g_2(\vec{x}))$，当$g_1(\vec{x})=-1$，$g_2(\vec{x})=1$，*或者*$g_1(\vec{x})=1$，$g_2(\vec{x})=-1$时，$G(\vec{x})=-1$，不满足定义;

(c) $G(\vec{x})=sign(1+g_1(\vec{x})+g_2(\vec{x}))$，$g_1(\vec{x})$*与*$g_2(\vec{x})$取任意的+1 和-1 时，均能满足$OR(g_1,g_2)$的定义;

(d) $G(\vec{x})=sign(3+g_1(\vec{x})+g_2(\vec{x}))$，当$g_1(\vec{x})=g_2(\vec{x})=-1$时，$G(\vec{x})=1$，不满足定义。

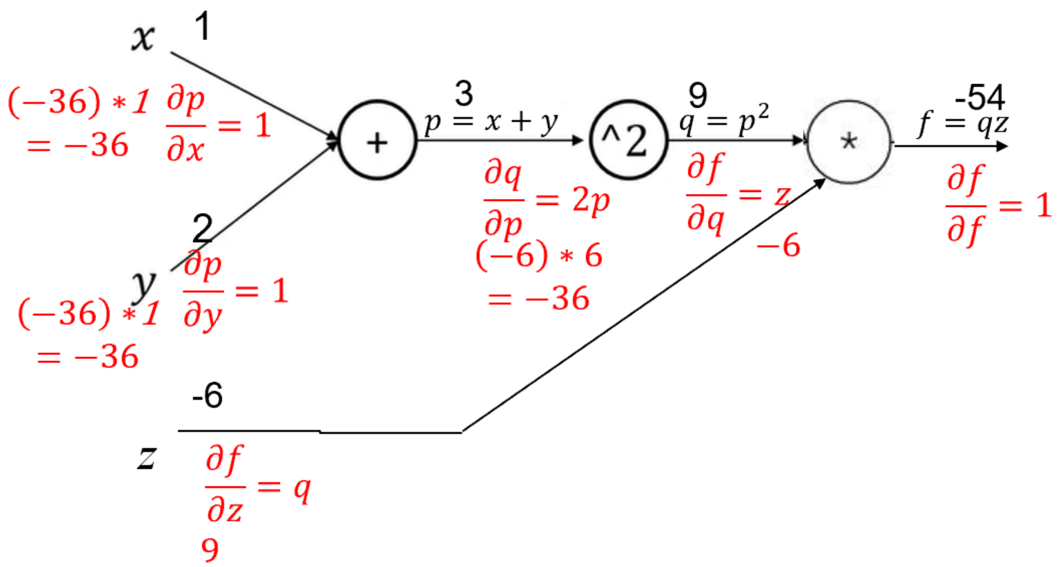
所以，只有(c)满足定义。

2，在 3-5-1 的神经网络中，网络参数有多少？

解：在第一层 3-5 中的参数为:（3+1（常数项））*5=20；在第二层 5-1 中的参数为（5+1（常数项））*1=6，所以，网络参数一共为 26。

3，画出$(x+y)^2 z$的计算图，当 x=1,y=2,z=-6 时，写出前向传播的数值和反向传播的梯度值。

解：

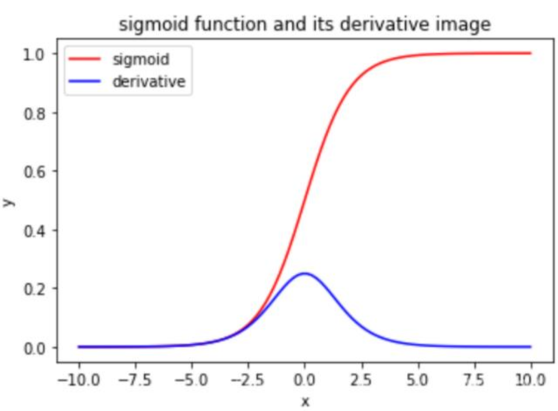


4，计算 Sigmoid 函数、双曲正切函数和 ReLU 函数的导数函数，分析这三个函数作为激活函数时的优缺点。

解：（1）对于 Sigmoid 函数，$\theta(x) = \frac{1}{1+e^{-x}}$

$$\frac{\partial \theta(x)}{\partial x} = \frac{e^{-x}}{(1+e^{-x})^2} = \frac{1+e^{-x}-1}{1+e^{-x}} \frac{1}{1+e^{-x}} = (1-\theta(x))\theta(x)$$

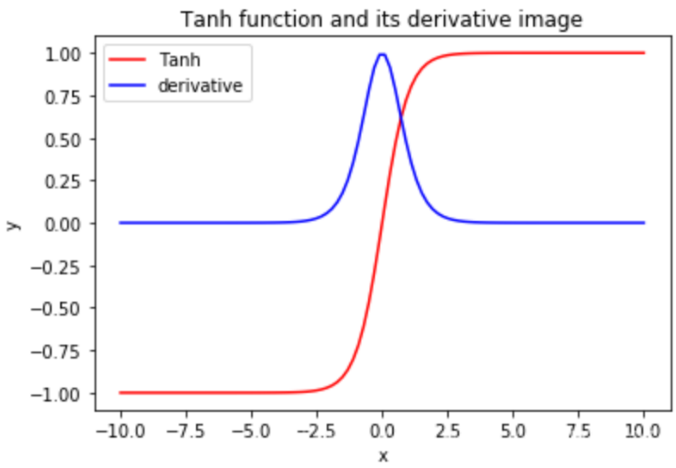
Sigmoid 函数作为激活函数的优点是连续、单调、可导且具有非线性特点。但其缺点是非中心对称，输出不是 0 均值的，同时它的导数函数曲线见右图，x 变化很小的范围内，导数才有值且不大，在神经



网络应用中，训练过程使用的是反向传播算法，通过链式法则回传的梯度不断相乘，因此，Sigmoid 函数作为激活函数时会导致梯度消失问题，尤其在网络比较深时会达不到训练效果。

（2）对于双曲正切函数，$tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$

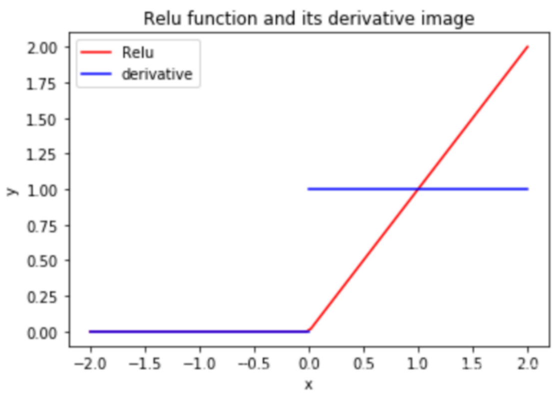
$$\frac{\partial \tanh(x)}{\partial x} = \frac{(e^x+e^{-x})(e^x+e^{-x})-(e^x-e^{-x})(e^x-e^{-x})}{(e^x+e^{-x})^2} = 1 - tanh^2(x)$$



双曲正切函数作为激活函数的优点是中心对称、单调、连续、可导且具有非线性特点。但其缺点是它的导数函数曲线见右图，x 变化更小的范围内，导数才有值且不大于 1，在神经网络应用中，训练过程使用的是反向传播算法，通过链式法则回传的梯度不断相乘，因此，双曲正切函数作为激活函数时同样会导致梯度消失问题，尤其在网络比较深时会达不到训练效果。

（3）对于 ReLU 函数，$ReLU(x) = \max(0, x)$

$$\frac{\partial \mathrm{ReLU}(x)}{\partial x} = \begin{cases} 1 & x > 0 \\ 0 & x \leq 0 \end{cases}$$



ReLU 函数作为激活函数的优点是梯度不会饱和，解决了梯度消失问题，没有指数计算，计算复杂度低。它的缺点是非中心对称，负数部分的梯度为

0，导致相应参数不会更新。

5，对于一幅 300*300 大小的彩色（RGB）图像，（1）如果输入端与有 100 个神经元的第一层隐含层用全链接方式（Fully Connected neural Network）连接时，请问这一层会包含多少参数？（2）如果用 100 个 5*5*3 大小的滤波器作卷积操作，那么这一层的参数为多少？如果滤波器移动步长（stride=1）为 1，经过卷积计算后的输出端神经元个数有多少？

解：（1）因为是 RGB 图像，所以共有 3 个通道，全连接情况下包含的参数是：300*300（图像大小）*3（颜色通道数）*100（第一层神经元个数）+1*100（每个神经元都要与输入层的常数项连接）=27000100；

（2）卷积操作时，第一层的参数是由卷积滤波器大小和常数项共同确定的，因此其包含的参数为：（5*5（滤波器大小）*3（颜色通道数）+1（常数项））*100（滤波器个数）=7600；（即：F=5，K=100，D1=3，参数量=（F*F*D1）*K+K=5*5*3*100+100=7600）

因为卷积核为 5*5*3，且填充值为 0，卷积后第一层神经元的个数为（（300（图像长或宽）-5（滤波器大小）/1（移动步长））+1）^2*100（滤波器个数）=8761600（相当于 W1=300，H1=300，D1=3，F=5，K=100，S=1，P=0，W2=（W1-F+2P）/S+1=（300-5+0）/1+1=296；H2=（H1-F+2P）/S+1=（300-5+0）/1+1=296；D2=K=100，神经元个数为：W2*H2*D2=296*296*100=8761600）

6，某一个卷积神经网络结构如下：

（i）输入层 Input 的 RGB 图像大小是 227*227*3。

（ii）第 1 层卷积层 Conv-1 是通过对输入图像用 96 个 11*11*3 大小的滤波器通过步长(stride)为 4，不做边缘填充（padding）得到的。

（iii）接下来是池化层 MaxPool-1，它用 3*3 尺寸、步长为 2 对 Conv-1 做 Max Pooling 操作。

（iv）然后我们对图像进行边缘填充，填充值为 2（如原来图像大小为 7*7 时，做填充值为 2 的填充后，图像大小变为 11*11），用 256 个 5*5 大小的滤波器按步长为 1，做第二次卷积操作，得到 Conv-2 层。

（v）再接一个池化层 MaxPool-2，它用 3*3 尺寸、步长为 2 做一次 Max Pooling 操作。

（vi）MaxPool-2 层输出去接一个有 4096 个神经元的全连接层 FC-1。

（vii）再接一个全连接层 FC-2 实现对 1000 个类别的分类。

请计算：（1）输入层到 Conv-1 层的参数量有多少？（2）经过池化层 MaxPool-1 后的神经元是多少？（3）经过第二次卷积操作后的图像大小为多少？（4）MaxPool-2 层到 FC-1 层的参数量是多少？（5）FC-1 层到 FC-2 层的参数量是多少？

解：输入图像大小为 227*227*3（即：W1=227，H1=227，D1=3），第一层卷积核为 11*11（即：F=11），共 96 个滤波器（即：K=96），步长为 4（即：S=4），边缘填充为 0（即：P=0），则卷积以后的图像边

长为：((227-11+2*0）/4）+1=55，大小为 55*55（即 W2=H2=55），与 96 个滤波器构成特征图，所以卷积层 Conv-1 的神经元个数为 55*55*96=290400，（1）输入层到 Conv-1 层的参数量为 F*F*D1*K+1*K，即：11*11*3*96+1*96=34944；对 55*55 大小的图像做第一次 Maxpooling，这时候通道数 96 保持不变，因为它用 3*3 大小的尺寸以步长为 2 做 Maxpooling，则得到的图像边长为（(55-3）/2）+1=27，图像大小为 27*27，（2）经过池化层 MaxPool-1 后的神经元为 27*27*96；再做第二次卷积，此时是对 27*27 大小的图像，用 5*5 大小的滤波器按步长为 1，边缘填充值为 2 做卷积，滤波器个数为 256 个，所以，卷积后图像的边长为（(27-5+2*2）/1）+1=27，大小为 27*27，与 256 个滤波器构成特征图，所以卷积层 Conv-2 的神经元个数为 27*27*256=186624，（3）经过第二次卷积操作后的图像大小为 27*27，MaxPool-1 层到 Conv-2 层的参数量为：5*5*96（池化层的通道数）*256（滤波器个数）+256（常数项）=614656；再经过池化层 MaxPool-2，3*3 尺寸、步长为 2，则得到的图像边长为（(27-3）/2）+1=13，图像大小为 13*13，上一层的通道数是 256，所以，MaxPool-2 层的神经元个数为 13*13*256=43264；MaxPool-2 层输出去接一个有 4096 个神经元的全连接层 FC-1，所以，（4）MaxPool-2 层到 FC-1 层的参数量是 13*13*256*4096+4096（常数项）=177213440；最后的输出层要对 1000 个类别进行分类，即 FC-2 层的神经元个数是 1000 个，而输入是 4096 个神经元，所以，（5）FC-1 层到 FC-2 层的参数量是 4096*1000+1000=4097000

7，有训练样本集为：$\mathrm{D}=\{(\vec{x}_1,y_1)=((1,1)^T,1),(\vec{x}_2,y_2)=((-1,-1)^T,1),(\vec{x}_3,y_3)=((-1,1)^T,-1),(\vec{x}_4,y_4)=((1,-1)^T,-1)\}$，假设某神经网络结构为第一层有两个神经元，第二层有三个神经元，第三层有一个神经元，前两层每个神经元的激活函数为ReLU（即$x_d^{(l)}=\max(0,s_d^{(l)})$，这里$s_d^{(l)}$代表第l层第d个神经元的输入，$x_d^{(l)}$代表该神经元的输出），第三层为线性输出，即$\hat{y}=\mathrm{s}_1^{(3)}$。误差函数为：$E_{in}=\frac{1}{N}\sum_n(y_n-\hat{y}_n)^2$，学习率为0.01。假设初始权系数矩阵定义如下：

$$\mathbf{w}_0^{(1)}=\begin{pmatrix}1&1\\1&1\\1&1\end{pmatrix},\ \mathbf{w}_0^{(2)}=\begin{pmatrix}1&1&1\\1&1&1\\1&1&1\end{pmatrix},\ \mathbf{w}_0^{(3)}=\begin{pmatrix}1\\1\\1\\1\end{pmatrix}$$

其中w的下标0代表迭代次数为0（即初始状态），上标数字分别代表第1、2、3层。要求将上述训练样本集的样本用反向传播法按顺序进行一轮训练，写出每一次迭代时各层的权系数矩阵，即：t=1时，进入样本$\vec{x}_1$，得到$\mathbf{w}_1^{(1)}$、$\mathbf{w}_1^{(2)}$和$\mathbf{w}_1^{(3)}$；t=2时，进入样本$\vec{x}_2$，得到$\mathbf{w}_2^{(1)}$、$\mathbf{w}_2^{(2)}$和$\mathbf{w}_2^{(3)}$；t=3时，进入样本$\vec{x}_3$，得到$\mathbf{w}_3^{(1)}$、$\mathbf{w}_3^{(2)}$和$\mathbf{w}_3^{(3)}$；t=4时，进入样本$\vec{x}_4$，得到$\mathbf{w}_4^{(1)}$、$\mathbf{w}_4^{(2)}$和$\mathbf{w}_4^{(3)}$

解：

（1）算法步骤描述：

假设训练样本集有 N 个样本$\{\vec{x}_1,\dots\vec{x}_n,\dots\vec{x}_N\}$，每个样本有 d 维特征，写成增广向量后是 d+1 维，$\vec{x}_n=(1,x_{n1},\dots x_{nd})^T$，将神经网络的输入层当第 0 层，所以写为：$\vec{x}_n^{(0)}=(1,x_{n1}^{(0)},\dots x_{nd}^{(0)})^T$，当 d=2 时，$\vec{x}_n^{(0)}=(1,x_{n1}^{(0)},x_{n2}^{(0)})^T$

假设第一层有两个神经元，第二层有三个神经元，第三层有一个神经元。

第一层、第二层和第三层的权系数矩阵分别为：

$$\mathbf{w}^{(1)}=\begin{pmatrix} w_{01}^{(1)} & w_{02}^{(1)} \\ w_{11}^{(1)} & w_{12}^{(1)} \\ w_{21}^{(1)} & w_{22}^{(1)} \end{pmatrix},\quad \mathbf{w}^{(2)}=\begin{pmatrix} w_{01}^{(2)} & w_{02}^{(2)} & w_{03}^{(2)} \\ w_{11}^{(2)} & w_{12}^{(2)} & w_{13}^{(2)} \\ w_{21}^{(2)} & w_{22}^{(2)} & w_{23}^{(2)} \end{pmatrix},\quad \mathbf{w}^{(3)}=\begin{pmatrix} w_{01}^{(3)} \\ w_{11}^{(3)} \\ w_{21}^{(3)} \\ w_{31}^{(3)} \end{pmatrix}$$

则第一层神经元的输入为:

$$\begin{pmatrix} s_1^{(1)} \\ s_2^{(1)} \end{pmatrix}=(\mathbf{w}^{(1)})^T\vec{x}_n^{(0)}$$

假设第一层神经元的激活函数为ReLU，即：$x^{(1)}=\max(0,s^{(1)})$，则:

$$\begin{pmatrix} x_1^{(1)} \\ x_2^{(1)} \end{pmatrix}=\begin{pmatrix} \max(0,s_1^{(1)}) \\ \max(0,s_2^{(1)}) \end{pmatrix}$$

第二层神经元的输入为:

$$\begin{pmatrix} s_1^{(2)} \\ s_2^{(2)} \\ s_3^{(2)} \end{pmatrix}=(\mathbf{w}^{(2)})^T\begin{pmatrix} 1 \\ x_1^{(1)} \\ x_2^{(1)} \end{pmatrix}$$

假设第二层神经元的激活函数为ReLU，即：$x^{(2)}=\max(0,s^{(2)})$，则:

$$\begin{pmatrix} x_1^{(2)} \\ x_2^{(2)} \\ x_3^{(2)} \end{pmatrix}=\begin{pmatrix} \max(0,s_1^{(2)}) \\ \max(0,s_2^{(2)}) \\ \max(0,s_3^{(2)}) \end{pmatrix}$$

则第三层的输入为:

$$\mathrm{s}_1^{(3)}=(\mathbf{w}^{(3)})^T\begin{pmatrix} 1 \\ x_1^{(2)} \\ x_2^{(2)} \\ x_3^{(2)} \end{pmatrix}$$

因为第三层是线性操作，即输出$\hat{y}=\mathrm{s}_1^{(3)}$

对于输入样本$\vec{x}_n$，假设其标签为y_n，采用平方误差函数。即：$e_n=(y_n-\hat{y}_n)^2$

$$\delta_1^{(3)}=-2(y_n-\mathrm{s}_1^{(3)})$$

运用反向传播法，于是：$\delta_j^{(2)}=\sum_k(\delta_k^{(3)})(w_{jk}^{(3)})(x_j^{(2)})'$

对于ReLU来说，其导数为：$(x_j^{(L)})'=[\![s_j^{(L-1)}\geq 0]\!]$

所以：$\delta_j^{(2)} = \sum_k(\delta_k^{(3)})(w_{jk}^{(3)})[\![s_j^{(2)} \ge 0]\!] = \delta_1^{(3)} w_{j1}^{(3)} [\![s_j^{(2)} \ge 0]\!]$

即：$\vec{\delta}^{(2)} = \begin{pmatrix} \delta_1^{(2)} \\ \delta_2^{(2)} \\ \delta_3^{(2)} \end{pmatrix} = \begin{pmatrix} \delta_1^{(3)} w_{11}^{(3)} [\![s_1^{(2)} \ge 0]\!] \\ \delta_1^{(3)} w_{21}^{(3)} [\![s_2^{(2)} \ge 0]\!] \\ \delta_1^{(3)} w_{31}^{(3)} [\![s_3^{(2)} \ge 0]\!] \end{pmatrix}$

继续运用反向传播法，于是：$\delta_j^{(1)} = \sum_k(\delta_k^{(2)})(w_{jk}^{(2)})(x_j^{(1)})'$，所以：

$$\delta_j^{(1)} = \sum\nolimits_k(\delta_k^{(2)})(w_{jk}^{(2)})[\![s_j^{(1)} \ge 0]\!] = (\delta_1^{(2)} w_{j1}^{(2)} + \delta_2^{(2)} w_{j2}^{(2)} + \delta_3^{(2)} w_{j3}^{(2)})[\![s_j^{(1)} \ge 0]\!]$$

由此可以得到：

$$\vec{\delta}^{(1)} = \begin{pmatrix} \delta_1^{(1)} \\ \delta_2^{(1)} \end{pmatrix} = \begin{pmatrix} [\![s_1^{(1)} \ge 0]\!] & 0 \\ 0 & [\![s_2^{(1)} \ge 0]\!] \end{pmatrix} \begin{pmatrix} w_{11}^{(2)} & w_{12}^{(2)} & w_{13}^{(2)} \\ w_{21}^{(2)} & w_{22}^{(2)} & w_{23}^{(2)} \end{pmatrix} \begin{pmatrix} \delta_1^{(2)} \\ \delta_2^{(2)} \\ \delta_3^{(2)} \end{pmatrix}$$

假定t表示迭代次数，η为学习步长，利用梯度下降法进行权系数更新：

$$\mathbf{w}_{t+1}^{(1)} = \mathbf{w}_t^{(1)} - \eta \vec{x}_n^{(0)} (\overrightarrow{\delta}^{(1)})^T = \mathbf{w}_t^{(1)} - \eta \begin{pmatrix} 1 \\ x_{n1}^{(0)} \\ x_{n2}^{(0)} \end{pmatrix} \left(\delta_1^{(1)}, \delta_2^{(1)}\right)$$

$$= \begin{pmatrix} w_{01}^{(1)} & w_{02}^{(1)} \\ w_{11}^{(1)} & w_{12}^{(1)} \\ w_{21}^{(1)} & w_{22}^{(1)} \end{pmatrix} - \eta \begin{pmatrix} \delta_1^{(1)} & \delta_2^{(1)} \\ x_{n1}^{(0)} \delta_1^{(1)} & x_{n1}^{(0)} \delta_2^{(1)} \\ x_{n2}^{(0)} \delta_1^{(1)} & x_{n2}^{(0)} \delta_2^{(1)} \end{pmatrix}$$

$$\mathbf{w}_{t+1}^{(2)} = \mathbf{w}_t^{(2)} - \eta \vec{x}_n^{(1)} (\overrightarrow{\delta}^{(2)})^T = \mathbf{w}_t^{(2)} - \eta \begin{pmatrix} 1 \\ x_1^{(1)} \\ x_2^{(1)} \end{pmatrix} \left(\delta_1^{(2)}, \delta_2^{(2)}, \delta_3^{(2)}\right)$$

$$= \begin{pmatrix} w_{01}^{(2)} & w_{02}^{(2)} & w_{03}^{(2)} \\ w_{11}^{(2)} & w_{12}^{(2)} & w_{13}^{(2)} \\ w_{21}^{(2)} & w_{22}^{(2)} & w_{23}^{(2)} \end{pmatrix} - \eta \begin{pmatrix} \delta_1^{(2)} & \delta_2^{(2)} & \delta_3^{(2)} \\ x_1^{(1)} \delta_1^{(2)} & x_1^{(1)} \delta_2^{(2)} & x_1^{(1)} \delta_3^{(2)} \\ x_2^{(1)} \delta_1^{(2)} & x_2^{(1)} \delta_2^{(2)} & x_2^{(1)} \delta_3^{(2)} \end{pmatrix}$$

$$\mathbf{w}_{t+1}^{(3)} = \mathbf{w}_t^{(3)} - \eta \vec{x}_n^{(2)} (\overrightarrow{\delta}^{(3)})^T = \mathbf{w}_t^{(3)} - \eta \begin{pmatrix} 1 \\ x_1^{(2)} \\ x_2^{(2)} \\ x_3^{(2)} \end{pmatrix} \delta_1^{(3)} = \begin{pmatrix} w_{01}^{(3)} \\ w_{11}^{(3)} \\ w_{21}^{(3)} \\ w_{31}^{(3)} \end{pmatrix} - \eta \begin{pmatrix} \delta_1^{(3)} \\ x_1^{(2)} \delta_1^{(3)} \\ x_2^{(2)} \delta_1^{(3)} \\ x_3^{(2)} \delta_1^{(3)} \end{pmatrix}$$

反复迭代至T次。

（2）代入习题数据的解答流程：

t=0

$$\mathbf{w}_0^{(1)}=\begin{pmatrix}1&1\\1&1\\1&1\end{pmatrix},\ \mathbf{w}_0^{(2)}=\begin{pmatrix}1&1&1\\1&1&1\\1&1&1\end{pmatrix},\ \mathbf{w}_0^{(3)}=\begin{pmatrix}1\\1\\1\\1\end{pmatrix}$$

t=1时，对于第一个样本$\vec{x}_1=(1,1)^T$，则第一层神经元的输入为：

$$\begin{pmatrix}s_1^{(1)}\\s_2^{(1)}\end{pmatrix}=(\mathbf{w}^{(1)})^T\vec{x}_n^{(0)}=\begin{pmatrix}1&1&1\\1&1&1\end{pmatrix}\begin{pmatrix}1\\1\\1\end{pmatrix}=\begin{pmatrix}3\\3\end{pmatrix}$$

则：$\begin{pmatrix}x_1^{(1)}\\x_2^{(1)}\end{pmatrix}=\begin{pmatrix}\max(0,s_1^{(1)})\\\max(0,s_2^{(1)})\end{pmatrix}=\begin{pmatrix}3\\3\end{pmatrix}$

第二层神经元的输入为：

$$\begin{pmatrix}s_1^{(2)}\\s_2^{(2)}\\s_3^{(2)}\end{pmatrix}=(\mathbf{w}^{(2)})^T\begin{pmatrix}1\\x_1^{(1)}\\x_2^{(1)}\end{pmatrix}=\begin{pmatrix}1&1&1\\1&1&1\\1&1&1\end{pmatrix}\begin{pmatrix}1\\3\\3\end{pmatrix}=\begin{pmatrix}7\\7\\7\end{pmatrix}$$

则：$\begin{pmatrix}x_1^{(2)}\\x_2^{(2)}\\x_3^{(2)}\end{pmatrix}=\begin{pmatrix}7\\7\\7\end{pmatrix}$

则第三层的输入为：

$$\mathrm{s}_1^{(3)}=(\mathbf{w}^{(3)})^T\begin{pmatrix}1\\x_1^{(2)}\\x_2^{(2)}\\x_3^{(2)}\end{pmatrix}=\begin{pmatrix}1&1&1&1\end{pmatrix}\begin{pmatrix}1\\7\\7\\7\end{pmatrix}=22$$

即输出$\hat{y}=\mathrm{s}_1^{(3)}=22$

对于样本$\vec{x}_1$，其标签为1，采用平方误差函数：$e_n=(y_n-\hat{y}_n)^2$，则：

$$\delta_1^{(3)}=-2\left(y_n-\mathrm{s}_1^{(3)}\right)=-2(1-22)=42$$

运用反向传播法，于是：

$$\delta_j^{(2)}=\sum\nolimits_k(\delta_k^{(3)})(w_{jk}^{(3)})\left[\!\left[s_j^{(2)}\geq 0\right]\!\right]=\delta_1^{(3)}w_{j1}^{(3)}\left[\!\left[s_j^{(2)}\geq 0\right]\!\right]$$

即：$\vec{\delta}^{(2)}=\begin{pmatrix}\delta_1^{(2)}\\ \delta_2^{(2)}\\ \delta_3^{(2)}\end{pmatrix}=\begin{pmatrix}\delta_1^{(3)}w_{11}^{(3)}[\![s_1^{(2)}\geq 0]\!]\\ \delta_1^{(3)}w_{21}^{(3)}[\![s_2^{(2)}\geq 0]\!]\\ \delta_1^{(3)}w_{31}^{(3)}[\![s_3^{(2)}\geq 0]\!]\end{pmatrix}=\begin{pmatrix}42*1*1\\ 42*1*1\\ 42*1*1\end{pmatrix}=\begin{pmatrix}42\\ 42\\ 42\end{pmatrix}$

继续运用反向传播法，于是：$\delta_j^{(1)}=\sum_k(\delta_k^{(2)})(w_{jk}^{(2)})(x_j^{(1)})'$，所以：

$$\delta_j^{(1)}=\sum_k(\delta_k^{(2)})(w_{jk}^{(2)})[\![s_j^{(1)}\geq 0]\!]=(\delta_1^{(2)}w_{j1}^{(2)}+\delta_2^{(2)}w_{j2}^{(2)}+\delta_3^{(2)}w_{j3}^{(2)})[\![s_j^{(1)}\geq 0]\!]$$

由此可以得到：

$$\vec{\delta}^{(1)}=\begin{pmatrix}\delta_1^{(1)}\\ \delta_2^{(1)}\end{pmatrix}=\begin{pmatrix}[\![s_1^{(1)}\geq 0]\!] & 0\\ 0 & [\![s_2^{(1)}\geq 0]\!]\end{pmatrix}\begin{pmatrix}w_{11}^{(2)} & w_{12}^{(2)} & w_{13}^{(2)}\\ w_{21}^{(2)} & w_{22}^{(2)} & w_{23}^{(2)}\end{pmatrix}\begin{pmatrix}\delta_1^{(2)}\\ \delta_2^{(2)}\\ \delta_3^{(2)}\end{pmatrix}$$

$$=\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\begin{pmatrix}1 & 1 & 1\\ 1 & 1 & 1\end{pmatrix}\begin{pmatrix}42\\ 42\\ 42\end{pmatrix}=\begin{pmatrix}126\\ 126\end{pmatrix}$$

令$\eta=0.01$，利用梯度下降法进行权系数更新：

$$\mathbf{w}_1^{(1)}=\mathbf{w}_0^{(1)}-\eta\vec{x}_n^{(0)}(\overrightarrow{\delta}^{(1)})^T=\mathbf{w}_t^{(1)}-\eta\begin{pmatrix}1\\ x_{n1}^{(0)}\\ x_{n2}^{(0)}\end{pmatrix}\left(\delta_1^{(1)},\delta_2^{(1)}\right)$$

$$=\begin{pmatrix}w_{01}^{(1)} & w_{02}^{(1)}\\ w_{11}^{(1)} & w_{12}^{(1)}\\ w_{21}^{(1)} & w_{22}^{(1)}\end{pmatrix}-\eta\begin{pmatrix}\delta_1^{(1)} & \delta_2^{(1)}\\ x_{n1}^{(0)}\delta_1^{(1)} & x_{n1}^{(0)}\delta_2^{(1)}\\ x_{n2}^{(0)}\delta_1^{(1)} & x_{n2}^{(0)}\delta_2^{(1)}\end{pmatrix}$$

$$=\begin{pmatrix}1 & 1\\ 1 & 1\\ 1 & 1\end{pmatrix}-0.01\begin{pmatrix}126 & 126\\ 1*126 & 1*126\\ 1*126 & 1*126\end{pmatrix}=\begin{pmatrix}-0.26 & -0.26\\ -0.26 & -0.26\\ -0.26 & -0.26\end{pmatrix}$$

$$\mathbf{w}_1^{(2)}=\mathbf{w}_0^{(2)}-\eta\vec{x}_n^{(1)}(\overrightarrow{\delta}^{(2)})^T=\mathbf{w}_0^{(2)}-\eta\begin{pmatrix}1\\ x_1^{(1)}\\ x_2^{(1)}\end{pmatrix}\left(\delta_1^{(2)},\delta_2^{(2)},\delta_3^{(2)}\right)$$

$$=\begin{pmatrix}w_{01}^{(2)} & w_{02}^{(2)} & w_{03}^{(2)}\\ w_{11}^{(2)} & w_{12}^{(2)} & w_{13}^{(2)}\\ w_{21}^{(2)} & w_{22}^{(2)} & w_{23}^{(2)}\end{pmatrix}-\eta\begin{pmatrix}\delta_1^{(2)} & \delta_2^{(2)} & \delta_3^{(2)}\\ x_1^{(1)}\delta_1^{(2)} & x_1^{(1)}\delta_2^{(2)} & x_1^{(1)}\delta_3^{(2)}\\ x_2^{(1)}\delta_1^{(2)} & x_2^{(1)}\delta_2^{(2)} & x_2^{(1)}\delta_3^{(2)}\end{pmatrix}$$

$$=\begin{pmatrix}1 & 1 & 1\\ 1 & 1 & 1\\ 1 & 1 & 1\end{pmatrix}-0.01\begin{pmatrix}42 & 42 & 42\\ 3*42 & 3*42 & 3*42\\ 3*42 & 3*42 & 3*42\end{pmatrix}$$

$$=\begin{pmatrix}0.58 & 0.58 & 0.58\\ -0.26 & -0.26 & -0.26\\ -0.26 & -0.26 & -0.26\end{pmatrix}$$

$$\mathbf{w}_1^{(3)}=\mathbf{w}_0^{(3)}-\eta\vec{x}_n^{(2)}\overrightarrow{(\delta^{(3)})}^T=\mathbf{w}_0^{(3)}-\eta\begin{pmatrix}1\\x_1^{(2)}\\x_2^{(2)}\\x_3^{(2)}\end{pmatrix}\delta_1^{(3)}=\begin{pmatrix}1\\1\\1\\1\end{pmatrix}-0.01\begin{pmatrix}42\\7*42\\7*42\\7*42\end{pmatrix}$$

$$=\begin{pmatrix}0.58\\-1.94\\-1.94\\-1.94\end{pmatrix}$$

t=2，对于第二个样本$\vec{x}_2=(-1,-1)^T$，则第一层神经元的输入为：

$$\begin{pmatrix}s_1^{(1)}\\s_2^{(1)}\end{pmatrix}=(\mathbf{w}^{(1)})^T\vec{x}_n^{(0)}=\begin{pmatrix}-0.26&-0.26&-0.26\\-0.26&-0.26&-0.26\end{pmatrix}\begin{pmatrix}1\\-1\\-1\end{pmatrix}=\begin{pmatrix}0.26\\0.26\end{pmatrix}$$

则：$\begin{pmatrix}x_1^{(1)}\\x_2^{(1)}\end{pmatrix}=\begin{pmatrix}\max(0,s_1^{(1)})\\\max(0,s_2^{(1)})\end{pmatrix}=\begin{pmatrix}0.26\\0.26\end{pmatrix}$

第二层神经元的输入为：

$$\begin{pmatrix}s_1^{(2)}\\s_2^{(2)}\\s_3^{(2)}\end{pmatrix}=(\mathbf{w}^{(2)})^T\begin{pmatrix}1\\x_1^{(1)}\\x_2^{(1)}\end{pmatrix}=\begin{pmatrix}0.58&-0.26&-0.26\\0.58&-0.26&-0.26\\0.58&-0.26&-0.26\end{pmatrix}\begin{pmatrix}1\\0.26\\0.26\end{pmatrix}=\begin{pmatrix}0.44\\0.44\\0.44\end{pmatrix}$$

则：$\begin{pmatrix}x_1^{(2)}\\x_2^{(2)}\\x_3^{(2)}\end{pmatrix}=\begin{pmatrix}0.44\\0.44\\0.44\end{pmatrix}$

则第三层的输入为：

$$\mathrm{s}_1^{(3)}=(\mathbf{w}^{(3)})^T\begin{pmatrix}1\\x_1^{(2)}\\x_2^{(2)}\\x_3^{(2)}\end{pmatrix}=(0.58\quad -1.94\quad -1.94\quad -1.94)\begin{pmatrix}1\\0.44\\0.44\\0.44\end{pmatrix}=-1.98$$

即输出$\hat{y}=\mathrm{s}_1^{(3)}=-1.98$

对于样本$\vec{x}_2$，其标签为1，采用平方误差函数：$e_n=(y_n-\hat{y}_n)^2$，则：

$$\delta_1^{(3)}=-2\left(y_n-\mathrm{s}_1^{(3)}\right)=-2\left(1-(-1.98)\right)=-5.96$$

运用反向传播法，于是：

$$\delta_j^{(2)} = \sum_k (\delta_k^{(3)})(w_{jk}^{(3)}) [\![s_j^{(2)} \geq 0]\!] = \delta_1^{(3)} w_{j1}^{(3)} [\![s_j^{(2)} \geq 0]\!]$$

即：$\vec{\delta}^{(2)} = \begin{pmatrix} \delta_1^{(2)} \\ \delta_2^{(2)} \\ \delta_3^{(2)} \end{pmatrix} = \begin{pmatrix} \delta_1^{(3)} w_{11}^{(3)} [\![s_1^{(2)} \geq 0]\!] \\ \delta_1^{(3)} w_{21}^{(3)} [\![s_2^{(2)} \geq 0]\!] \\ \delta_1^{(3)} w_{31}^{(3)} [\![s_3^{(2)} \geq 0]\!] \end{pmatrix} = \begin{pmatrix} -5.96 * (-1.94) * 1 \\ -5.96 * (-1.94) * 1 \\ -5.96 * (-1.94) * 1 \end{pmatrix} = \begin{pmatrix} 11.56 \\ 11.56 \\ 11.56 \end{pmatrix}$

继续运用反向传播法，于是：$\delta_j^{(1)} = \sum_k (\delta_k^{(2)})(w_{jk}^{(2)})(x_j^{(1)})'$，所以：

$$\delta_j^{(1)} = \sum_k (\delta_k^{(2)})(w_{jk}^{(2)}) [\![s_j^{(1)} \geq 0]\!] = (\delta_1^{(2)} w_{j1}^{(2)} + \delta_2^{(2)} w_{j2}^{(2)} + \delta_3^{(2)} w_{j3}^{(2)}) [\![s_j^{(1)} \geq 0]\!]$$

由此可以得到：

$$\vec{\delta}^{(1)} = \begin{pmatrix} \delta_1^{(1)} \\ \delta_2^{(1)} \end{pmatrix} = \begin{pmatrix} [\![s_1^{(1)} \geq 0]\!] & 0 \\ 0 & [\![s_2^{(1)} \geq 0]\!] \end{pmatrix} \begin{pmatrix} w_{11}^{(2)} & w_{12}^{(2)} & w_{13}^{(2)} \\ w_{21}^{(2)} & w_{22}^{(2)} & w_{23}^{(2)} \end{pmatrix} \begin{pmatrix} \delta_1^{(2)} \\ \delta_2^{(2)} \\ \delta_3^{(2)} \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} -0.26 & -0.26 & -0.26 \\ -0.26 & -0.26 & -0.26 \end{pmatrix} \begin{pmatrix} 11.56 \\ 11.56 \\ 11.56 \end{pmatrix} = \begin{pmatrix} -9.02 \\ -9.02 \end{pmatrix}$$

令$\eta = 0.01$，利用梯度下降法进行权系数更新：

$$\mathbf{w}_2^{(1)} = \mathbf{w}_1^{(1)} - \eta \vec{x}_n^{(0)} \overrightarrow{(\delta^{(1)})}^T = \mathbf{w}_1^{(1)} - \eta \begin{pmatrix} 1 \\ x_{n1}^{(0)} \\ x_{n2}^{(0)} \end{pmatrix} \left(\delta_1^{(1)}, \delta_2^{(1)} \right)$$

$$= \begin{pmatrix} w_{01}^{(1)} & w_{02}^{(1)} \\ w_{11}^{(1)} & w_{12}^{(1)} \\ w_{21}^{(1)} & w_{22}^{(1)} \end{pmatrix} - \eta \begin{pmatrix} \delta_1^{(1)} & \delta_2^{(1)} \\ x_{n1}^{(0)} \delta_1^{(1)} & x_{n1}^{(0)} \delta_2^{(1)} \\ x_{n2}^{(0)} \delta_1^{(1)} & x_{n2}^{(0)} \delta_2^{(1)} \end{pmatrix}$$

$$= \begin{pmatrix} -0.26 & -0.26 \\ -0.26 & -0.26 \\ -0.26 & -0.26 \end{pmatrix} - 0.01 \begin{pmatrix} -9.02 & -9.02 \\ (-1) * (-9.02) & (-1) * (-9.02) \\ (-1) * (-9.02) & (-1) * (-9.02) \end{pmatrix}$$

$$= \begin{pmatrix} -0.17 & -0.17 \\ -0.35 & -0.35 \\ -0.35 & -0.35 \end{pmatrix}$$

$$\mathbf{w}_2^{(2)} = \mathbf{w}_1^{(2)} - \eta \vec{x}_n^{(1)} (\overrightarrow{\delta}^{(2)})^T = \mathbf{w}_1^{(2)} - \eta \begin{pmatrix} 1 \\ x_1^{(1)} \\ x_2^{(1)} \end{pmatrix} \left(\delta_1^{(2)}, \delta_2^{(2)}, \delta_3^{(2)}\right)$$

$$= \begin{pmatrix} w_{01}^{(2)} & w_{02}^{(2)} & w_{03}^{(2)} \\ w_{11}^{(2)} & w_{12}^{(2)} & w_{13}^{(2)} \\ w_{21}^{(2)} & w_{22}^{(2)} & w_{23}^{(2)} \end{pmatrix} - \eta \begin{pmatrix} \delta_1^{(2)} & \delta_2^{(2)} & \delta_3^{(2)} \\ x_1^{(1)}\delta_1^{(2)} & x_1^{(1)}\delta_2^{(2)} & x_1^{(1)}\delta_3^{(2)} \\ x_2^{(1)}\delta_1^{(2)} & x_2^{(1)}\delta_2^{(2)} & x_2^{(1)}\delta_3^{(2)} \end{pmatrix}$$

$$= \begin{pmatrix} 0.58 & 0.58 & 0.58 \\ -0.26 & -0.26 & -0.26 \\ -0.26 & -0.26 & -0.26 \end{pmatrix}$$

$$- 0.01 \begin{pmatrix} 11.56 & 11.56 & 11.56 \\ 0.26 * 11.56 & 0.26 * 11.56 & 0.26 * 11.56 \\ 0.26 * 11.56 & 0.26 * 11.56 & 0.26 * 11.56 \end{pmatrix}$$

$$= \begin{pmatrix} 0.46 & 0.46 & 0.46 \\ -0.29 & -0.29 & -0.29 \\ -0.29 & -0.29 & -0.29 \end{pmatrix}$$

$$\mathbf{w}_2^{(3)} = \mathbf{w}_1^{(3)} - \eta \vec{x}_n^{(2)} (\overrightarrow{\delta}^{(3)})^T = \mathbf{w}_1^{(3)} - \eta \begin{pmatrix} 1 \\ x_1^{(2)} \\ x_2^{(2)} \\ x_3^{(2)} \end{pmatrix} \delta_1^{(3)}$$

$$= \begin{pmatrix} 0.58 \\ -1.94 \\ -1.94 \\ -1.94 \end{pmatrix} - 0.01 \begin{pmatrix} -5.96 \\ 0.44 * (-5.96) \\ 0.44 * (-5.96) \\ 0.44 * (-5.96) \end{pmatrix} = \begin{pmatrix} 0.64 \\ -1.91 \\ -1.91 \\ -1.91 \end{pmatrix}$$

t=3，对于第三个样本$\vec{x}_3 = (-1,1)^T$，则第一层神经元的输入为：

$$\begin{pmatrix} s_1^{(1)} \\ s_2^{(1)} \end{pmatrix} = (\mathbf{w}^{(1)})^T \vec{x}_n^{(0)} = \begin{pmatrix} -0.17 & -0.35 & -0.35 \\ -0.17 & -0.35 & -0.35 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} = \begin{pmatrix} -0.17 \\ -0.17 \end{pmatrix}$$

则：$\begin{pmatrix} x_1^{(1)} \\ x_2^{(1)} \end{pmatrix} = \begin{pmatrix} \max(0, s_1^{(1)}) \\ \max(0, s_2^{(1)}) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$

第二层神经元的输入为：

$$\begin{pmatrix} s_1^{(2)} \\ s_2^{(2)} \\ s_3^{(2)} \end{pmatrix} = (\mathbf{w}^{(2)})^T \begin{pmatrix} 1 \\ x_1^{(1)} \\ x_2^{(1)} \end{pmatrix} = \begin{pmatrix} 0.46 & -0.29 & -0.29 \\ 0.46 & -0.29 & -0.29 \\ 0.46 & -0.29 & -0.29 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0.46 \\ 0.46 \\ 0.46 \end{pmatrix}$$

则：$\begin{pmatrix} x_1^{(2)} \\ x_2^{(2)} \\ x_3^{(2)} \end{pmatrix} = \begin{pmatrix} 0.46 \\ 0.46 \\ 0.46 \end{pmatrix}$

则第三层的输入为：

$$s_1^{(3)} = (\mathbf{w}^{(3)})^T \begin{pmatrix} 1 \\ x_1^{(2)} \\ x_2^{(2)} \\ x_3^{(2)} \end{pmatrix} = (0.64 \quad -1.91 \quad -1.91 \quad -1.91) \begin{pmatrix} 1 \\ 0.46 \\ 0.46 \\ 0.46 \end{pmatrix} = -2.00$$

即输出$\hat{y} = s_1^{(3)} = -2.00$

对于样本$\vec{x}_3$，其标签为-1，采用平方误差函数：$e_n = (y_n - \hat{y}_n)^2$，则：

$$\delta_1^{(3)} = -2\left(y_n - s_1^{(3)}\right) = -2\big(-1 - (-2.00)\big) = -2.00$$

运用反向传播法，于是：

$$\delta_j^{(2)} = \sum\nolimits_k (\delta_k^{(3)})(w_{jk}^{(3)}) \left[\!\left[s_j^{(2)} \geq 0 \right]\!\right] = \delta_1^{(3)} w_{j1}^{(3)} \left[\!\left[s_j^{(2)} \geq 0 \right]\!\right]$$

即：$\vec{\delta}^{(2)} = \begin{pmatrix} \delta_1^{(2)} \\ \delta_2^{(2)} \\ \delta_3^{(2)} \end{pmatrix} = \begin{pmatrix} \delta_1^{(3)} w_{11}^{(3)} \left[\!\left[s_1^{(2)} \geq 0 \right]\!\right] \\ \delta_1^{(3)} w_{21}^{(3)} \left[\!\left[s_2^{(2)} \geq 0 \right]\!\right] \\ \delta_1^{(3)} w_{31}^{(3)} \left[\!\left[s_3^{(2)} \geq 0 \right]\!\right] \end{pmatrix} = \begin{pmatrix} (-2.00) * (-1.91) * 1 \\ (-2.00) * (-1.91) * 1 \\ (-2.00) * (-1.91) * 1 \end{pmatrix} = \begin{pmatrix} 3.82 \\ 3.82 \\ 3.82 \end{pmatrix}$

继续运用反向传播法，于是：$\delta_j^{(1)} = \sum_k (\delta_k^{(2)})(w_{jk}^{(2)})(x_j^{(1)})'$，所以：

$$\delta_j^{(1)} = \sum\nolimits_k (\delta_k^{(2)})(w_{jk}^{(2)}) \left[\!\left[s_j^{(1)} \geq 0 \right]\!\right] = (\delta_1^{(2)} w_{j1}^{(2)} + \delta_2^{(2)} w_{j2}^{(2)} + \delta_3^{(2)} w_{j3}^{(2)}) \left[\!\left[s_j^{(1)} \geq 0 \right]\!\right]$$

由此可以得到：

$$\vec{\delta}^{(1)} = \begin{pmatrix} \delta_1^{(1)} \\ \delta_2^{(1)} \end{pmatrix} = \begin{pmatrix} \left[\!\left[s_1^{(1)} \geq 0 \right]\!\right] & 0 \\ 0 & \left[\!\left[s_2^{(1)} \geq 0 \right]\!\right] \end{pmatrix} \begin{pmatrix} w_{11}^{(2)} & w_{12}^{(2)} & w_{13}^{(2)} \\ w_{21}^{(2)} & w_{22}^{(2)} & w_{23}^{(2)} \end{pmatrix} \begin{pmatrix} \delta_1^{(2)} \\ \delta_2^{(2)} \\ \delta_3^{(2)} \end{pmatrix}$$

$$= \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} -0.29 & -0.29 & -0.29 \\ -0.29 & -0.29 & -0.29 \end{pmatrix} \begin{pmatrix} 3.82 \\ 3.82 \\ 3.82 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

令$\eta = 0.01$，利用梯度下降法进行权系数更新：

$$\mathbf{w}_3^{(1)} = \mathbf{w}_2^{(1)} - \eta \vec{x}_n^{(0)} (\overrightarrow{\delta}^{(1)})^T = \mathbf{w}_2^{(1)} - \eta \begin{pmatrix} 1 \\ x_{n1}^{(0)} \\ x_{n2}^{(0)} \end{pmatrix} \left(\delta_1^{(1)}, \delta_2^{(1)}\right)$$

$$= \begin{pmatrix} w_{01}^{(1)} & w_{02}^{(1)} \\ w_{11}^{(1)} & w_{12}^{(1)} \\ w_{21}^{(1)} & w_{22}^{(1)} \end{pmatrix} - \eta \begin{pmatrix} \delta_1^{(1)} & \delta_2^{(1)} \\ x_{n1}^{(0)}\delta_1^{(1)} & x_{n1}^{(0)}\delta_2^{(1)} \\ x_{n2}^{(0)}\delta_1^{(1)} & x_{n2}^{(0)}\delta_2^{(1)} \end{pmatrix}$$

$$= \begin{pmatrix} -0.17 & -0.17 \\ -0.35 & -0.35 \\ -0.35 & -0.35 \end{pmatrix} - 0.01 \begin{pmatrix} 0 & 0 \\ (-1)*0 & (-1)*0 \\ (+1)*0 & (+1)*0 \end{pmatrix}$$

$$= \begin{pmatrix} -0.17 & -0.17 \\ -0.35 & -0.35 \\ -0.35 & -0.35 \end{pmatrix}$$

$$\mathbf{w}_3^{(2)} = \mathbf{w}_2^{(2)} - \eta \vec{x}_n^{(1)} (\overrightarrow{\delta}^{(2)})^T = \mathbf{w}_2^{(2)} - \eta \begin{pmatrix} 1 \\ x_1^{(1)} \\ x_2^{(1)} \end{pmatrix} \left(\delta_1^{(2)}, \delta_2^{(2)}, \delta_3^{(2)}\right)$$

$$= \begin{pmatrix} w_{01}^{(2)} & w_{02}^{(2)} & w_{03}^{(2)} \\ w_{11}^{(2)} & w_{12}^{(2)} & w_{13}^{(2)} \\ w_{21}^{(2)} & w_{22}^{(2)} & w_{23}^{(2)} \end{pmatrix} - \eta \begin{pmatrix} \delta_1^{(2)} & \delta_2^{(2)} & \delta_3^{(2)} \\ x_1^{(1)}\delta_1^{(2)} & x_1^{(1)}\delta_2^{(2)} & x_1^{(1)}\delta_3^{(2)} \\ x_2^{(1)}\delta_1^{(2)} & x_2^{(1)}\delta_2^{(2)} & x_2^{(1)}\delta_3^{(2)} \end{pmatrix}$$

$$= \begin{pmatrix} 0.46 & 0.46 & 0.46 \\ -0.29 & -0.29 & -0.29 \\ -0.29 & -0.29 & -0.29 \end{pmatrix} - 0.01 \begin{pmatrix} 3.82 & 3.82 & 3.82 \\ 0*3.82 & 0*3.82 & 0*3.82 \\ 0*3.82 & 0*3.82 & 0*3.82 \end{pmatrix}$$

$$= \begin{pmatrix} 0.42 & 0.42 & 0.42 \\ -0.29 & -0.29 & -0.29 \\ -0.29 & -0.29 & -0.29 \end{pmatrix}$$

$$\mathbf{w}_3^{(3)} = \mathbf{w}_2^{(3)} - \eta \vec{x}_n^{(2)} (\overrightarrow{\delta}^{(3)})^T = \mathbf{w}_2^{(3)} - \eta \begin{pmatrix} 1 \\ x_1^{(2)} \\ x_2^{(2)} \\ x_3^{(2)} \end{pmatrix} \delta_1^{(3)}$$

$$= \begin{pmatrix} 0.64 \\ -1.91 \\ -1.91 \\ -1.91 \end{pmatrix} - 0.01 \begin{pmatrix} -2.00 \\ 0.46*(-2.00) \\ 0.46*(-2.00) \\ 0.46*(-2.00) \end{pmatrix} = \begin{pmatrix} 0.66 \\ -1.90 \\ -1.90 \\ -1.90 \end{pmatrix}$$

t=4，对于第四个样本$\vec{x}_2 = (1,-1)^T$，则第一层神经元的输入为：

$$\begin{pmatrix} s_1^{(1)} \\ s_2^{(1)} \end{pmatrix} = (\mathbf{w}^{(1)})^T \vec{x}_n^{(0)} = \begin{pmatrix} -0.17 & -0.35 & -0.35 \\ -0.17 & -0.35 & -0.35 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix} = \begin{pmatrix} -0.17 \\ -0.17 \end{pmatrix}$$

则：$\begin{pmatrix} x_1^{(1)} \\ x_2^{(1)} \end{pmatrix} = \begin{pmatrix} \max(0, s_1^{(1)}) \\ \max(0, s_2^{(1)}) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$

第二层神经元的输入为：

$$\begin{pmatrix} s_1^{(2)} \\ s_2^{(2)} \\ s_3^{(2)} \end{pmatrix} = (\mathbf{w}^{(2)})^T \begin{pmatrix} 1 \\ x_1^{(1)} \\ x_2^{(1)} \end{pmatrix} = \begin{pmatrix} 0.42 & -0.29 & -0.29 \\ 0.42 & -0.29 & -0.29 \\ 0.42 & -0.29 & -0.29 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0.42 \\ 0.42 \\ 0.42 \end{pmatrix}$$

则：$\begin{pmatrix} x_1^{(2)} \\ x_2^{(2)} \\ x_3^{(2)} \end{pmatrix} = \begin{pmatrix} 0.42 \\ 0.42 \\ 0.42 \end{pmatrix}$

则第三层的输入为：

$$\mathrm{s}_1^{(3)} = (\mathbf{w}^{(3)})^T \begin{pmatrix} 1 \\ x_1^{(2)} \\ x_2^{(2)} \\ x_3^{(2)} \end{pmatrix} = \begin{pmatrix} 0.66 & -1.90 & -1.90 & -1.90 \end{pmatrix} \begin{pmatrix} 1 \\ 0.42 \\ 0.42 \\ 0.42 \end{pmatrix} = -1.73$$

即输出$\hat{y} = \mathrm{s}_1^{(3)} = -1.73$

对于样本$\vec{x}_4$，其标签为-1，采用平方误差函数：$e_n = (y_n - \hat{y}_n)^2$，则：

$$\delta_1^{(3)} = -2\left(y_n - \mathrm{s}_1^{(3)}\right) = -2\left(-1 - (-1.73)\right) = -1.46$$

运用反向传播法，于是：

$$\delta_j^{(2)} = \sum_k (\delta_k^{(3)})(w_{jk}^{(3)}) \left[\!\left[s_j^{(2)} \geq 0 \right]\!\right] = \delta_1^{(3)} w_{j1}^{(3)} \left[\!\left[s_j^{(2)} \geq 0 \right]\!\right]$$

即：$\vec{\delta}^{(2)} = \begin{pmatrix} \delta_1^{(2)} \\ \delta_2^{(2)} \\ \delta_3^{(2)} \end{pmatrix} = \begin{pmatrix} \delta_1^{(3)} w_{11}^{(3)} \left[\!\left[s_1^{(2)} \geq 0 \right]\!\right] \\ \delta_1^{(3)} w_{21}^{(3)} \left[\!\left[s_2^{(2)} \geq 0 \right]\!\right] \\ \delta_1^{(3)} w_{31}^{(3)} \left[\!\left[s_3^{(2)} \geq 0 \right]\!\right] \end{pmatrix} = \begin{pmatrix} (-1.46) * (-1.90) * 1 \\ (-1.46) * (-1.90) * 1 \\ (-1.46) * (-1.90) * 1 \end{pmatrix} = \begin{pmatrix} 2.77 \\ 2.77 \\ 2.77 \end{pmatrix}$

继续运用反向传播法，于是：$\delta_j^{(1)} = \sum_k (\delta_k^{(2)})(w_{jk}^{(2)})(x_j^{(1)})'$，所以：

$$\delta_j^{(1)} = \sum_k (\delta_k^{(2)})(w_{jk}^{(2)}) \left[\!\left[s_j^{(1)} \geq 0 \right]\!\right] = (\delta_1^{(2)} w_{j1}^{(2)} + \delta_2^{(2)} w_{j2}^{(2)} + \delta_3^{(2)} w_{j3}^{(2)}) \left[\!\left[s_j^{(1)} \geq 0 \right]\!\right]$$

由此可以得到：

$$\vec{\delta}^{(1)} = \begin{pmatrix} \delta_1^{(1)} \\ \delta_2^{(1)} \end{pmatrix} = \begin{pmatrix} [\![s_1^{(1)} \geq 0]\!] & 0 \\ 0 & [\![s_2^{(1)} \geq 0]\!] \end{pmatrix} \begin{pmatrix} w_{11}^{(2)} & w_{12}^{(2)} & w_{13}^{(2)} \\ w_{21}^{(2)} & w_{22}^{(2)} & w_{23}^{(2)} \end{pmatrix} \begin{pmatrix} \delta_1^{(2)} \\ \delta_2^{(2)} \\ \delta_3^{(2)} \end{pmatrix}$$

$$= \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} -0.29 & -0.29 & -0.29 \\ -0.29 & -0.29 & -0.29 \end{pmatrix} \begin{pmatrix} 2.77 \\ 2.77 \\ 2.77 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

令$\eta = 0.01$，利用梯度下降法进行权系数更新：

$$\mathbf{w}_4^{(1)} = \mathbf{w}_3^{(1)} - \eta \vec{x}_n^{(0)} (\overrightarrow{\delta}^{(1)})^T = \mathbf{w}_3^{(1)} - \eta \begin{pmatrix} 1 \\ x_{n1}^{(0)} \\ x_{n2}^{(0)} \end{pmatrix} \left(\delta_1^{(1)}, \delta_2^{(1)} \right)$$

$$= \begin{pmatrix} w_{01}^{(1)} & w_{02}^{(1)} \\ w_{11}^{(1)} & w_{12}^{(1)} \\ w_{21}^{(1)} & w_{22}^{(1)} \end{pmatrix} - \eta \begin{pmatrix} \delta_1^{(1)} & \delta_2^{(1)} \\ x_{n1}^{(0)} \delta_1^{(1)} & x_{n1}^{(0)} \delta_2^{(1)} \\ x_{n2}^{(0)} \delta_1^{(1)} & x_{n2}^{(0)} \delta_2^{(1)} \end{pmatrix}$$

$$= \begin{pmatrix} -0.17 & -0.17 \\ -0.35 & -0.35 \\ -0.35 & -0.35 \end{pmatrix} - 0.01 \begin{pmatrix} 0 & 0 \\ (+1) * 0 & (+1) * 0 \\ (-1) * 0 & (-1) * 0 \end{pmatrix}$$

$$= \begin{pmatrix} -0.17 & -0.17 \\ -0.35 & -0.35 \\ -0.35 & -0.35 \end{pmatrix}$$

$$\mathbf{w}_4^{(2)} = \mathbf{w}_3^{(2)} - \eta \vec{x}_n^{(1)} (\overrightarrow{\delta}^{(2)})^T = \mathbf{w}_3^{(2)} - \eta \begin{pmatrix} 1 \\ x_1^{(1)} \\ x_2^{(1)} \end{pmatrix} \left(\delta_1^{(2)}, \delta_2^{(2)}, \delta_3^{(2)} \right)$$

$$= \begin{pmatrix} w_{01}^{(2)} & w_{02}^{(2)} & w_{03}^{(2)} \\ w_{11}^{(2)} & w_{12}^{(2)} & w_{13}^{(2)} \\ w_{21}^{(2)} & w_{22}^{(2)} & w_{23}^{(2)} \end{pmatrix} - \eta \begin{pmatrix} \delta_1^{(2)} & \delta_2^{(2)} & \delta_3^{(2)} \\ x_1^{(1)} \delta_1^{(2)} & x_1^{(1)} \delta_2^{(2)} & x_1^{(1)} \delta_3^{(2)} \\ x_2^{(1)} \delta_1^{(2)} & x_2^{(1)} \delta_2^{(2)} & x_2^{(1)} \delta_3^{(2)} \end{pmatrix}$$

$$= \begin{pmatrix} 0.42 & 0.42 & 0.42 \\ -0.29 & -0.29 & -0.29 \\ -0.29 & -0.29 & -0.29 \end{pmatrix} - 0.01 \begin{pmatrix} 2.77 & 2.77 & 2.77 \\ 0 * 2.77 & 0 * 2.77 & 0 * 2.77 \\ 0 * 2.77 & 0 * 2.77 & 0 * 2.77 \end{pmatrix}$$

$$= \begin{pmatrix} 0.39 & 0.39 & 0.39 \\ -0.29 & -0.29 & -0.29 \\ -0.29 & -0.29 & -0.29 \end{pmatrix}$$

$$\mathbf{w}_4^{(3)} = \mathbf{w}_3^{(3)} - \eta \vec{x}_n^{(2)} (\overrightarrow{\delta}^{(3)})^T = \mathbf{w}_3^{(3)} - \eta \begin{pmatrix} 1 \\ x_1^{(2)} \\ x_2^{(2)} \\ x_3^{(2)} \end{pmatrix} \delta_1^{(3)}$$

$$= \begin{pmatrix} 0.66 \\ -1.90 \\ -1.90 \\ -1.90 \end{pmatrix} - 0.01 \begin{pmatrix} -1.46 \\ 0.42 * (-1.46) \\ 0.42 * (-1.46) \\ 0.42 * (-1.46) \end{pmatrix} = \begin{pmatrix} 0.67 \\ -1.89 \\ -1.89 \\ -1.89 \end{pmatrix}$$

1. 对于两分类问题，证明最小风险贝叶斯规则可表示为

若 $l(x)=\dfrac{p(\mathrm{X}|\omega_1)}{p(\mathrm{X}|\omega_2)}>\dfrac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}}\dfrac{P(\omega_2)}{P(\omega_1)}$，则决策$\mathrm{X}\in\omega_1$；否则$\mathrm{X}\in\omega_2$。

解：计算条件风险

$$
\begin{aligned}
R(\alpha_1|x) &= \sum_{j=1}^{2}\lambda_{1j}P(w_j|x) \\
&= \lambda_{11}P(w_1|x)+\lambda_{12}P(w_2|x)
\end{aligned}
$$

$$
\begin{aligned}
R(\alpha_2|x) &= \sum_{j=1}^{2}\lambda_{2j}P(w_j|x) \\
&= \lambda_{21}P(w_1|x)+\lambda_{22}P(w_2|x)
\end{aligned}
$$

如果$R(\alpha_1|x)<R(\alpha_2|x)$，则$x\in w_1$。

$$
\begin{aligned}
\lambda_{11}P(w_1|x)+\lambda_{12}P(w_2|x) &< \lambda_{21}P(w_1|x)+\lambda_{22}P(w_2|x) \\
(\lambda_{21}-\lambda_{11})P(w_1|x) &> (\lambda_{12}-\lambda_{22})P(w_2|x) \\
(\lambda_{21}-\lambda_{11})P(w_1)p(x|w_1) &> (\lambda_{12}-\lambda_{22})P(w_2)p(x|w_2) \\
\frac{p(x|w_1)}{p(x|w_2)} &> \frac{(\lambda_{12}-\lambda_{22})P(w_2)}{(\lambda_{21}-\lambda_{11})P(w_1)}
\end{aligned}
$$

所以，如果$\dfrac{p(x|w_1)}{p(x|w_2)}>\dfrac{(\lambda_{12}-\lambda_{22})P(w_2)}{(\lambda_{21}-\lambda_{11})P(w_1)}$，则$x\in w_1$。反之则$x\in w_2$。

2. 有两类样本 ω1 和 ω2，已知先验概率 P(ω1)=0.2 和 P(ω2)=0.8，类概率密度函数如下：

$$
p(x\,|\,\omega_1)=\begin{cases} x & 0\le x<1 \\ 2-x & 1\le x\le 2 \\ 0 & \text{其他} \end{cases}
\qquad
p(x\,|\,\omega_2)=\begin{cases} x-1 & 1\le x<2 \\ 3-x & 2\le x\le 3 \\ 0 & \text{其他} \end{cases}
$$

（1）求贝叶斯最小误判概率准则下的判决域，并判断样本 x=1.5 属于哪一类；

（2）求总错误概率 P(e)；

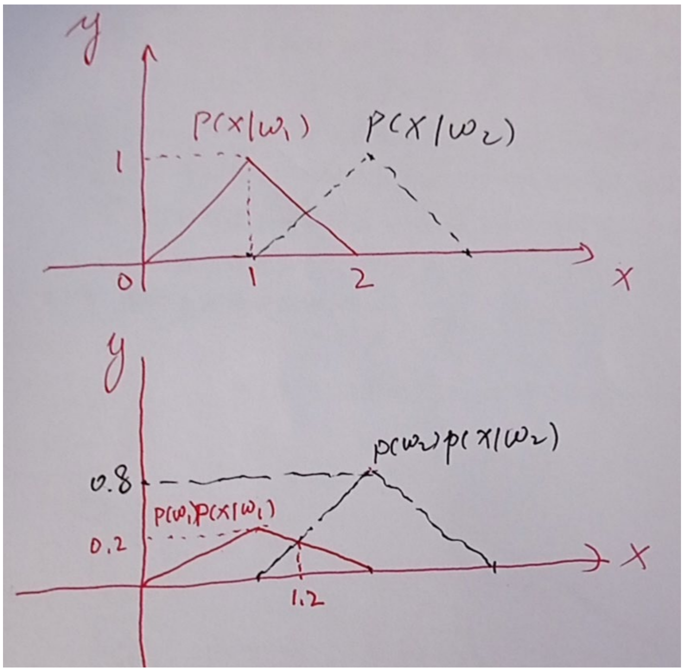
（3）假设正确判断的损失 $\lambda_{11}=\lambda_{22}=0$，误判损失分别为 λ_{12} 和 λ_{21}，若采用最小损失判决准则，λ_{12} 和 λ_{21} 满足怎样的关系时，会使上述对样本 x=1.5 的判断相反？

解：

（1）在$1\le x\le 2$范围内两个类别才有交集，所以只需要考虑这一段的决策域。

g（x）=0.2×(2-x)-0.8×(x-1)=1.2-x=0

所以，决策面为 x=1.2，如下图所示。x<1.2,判决为第一类，否则就判决为第二类。所以样本 x=1.5 属于第 2 类.



（2）
$$P(e)=P(\omega_1)\int_{\Gamma_2} p(\mathrm{X}\mid\omega_1)d\mathrm{X}+P(\omega_2)\int_{\Gamma_1} p(\mathrm{X}\mid\omega_2)d\mathrm{X}$$
=(2-1.2)*0.5*0.16+(1.2-1)*0.5*0.16=0.08

（3）$\lambda_{11}P(\omega_1\mid x)+\lambda_{12}P(\omega_2\mid x)\underset{>}{<}\lambda_{21}P(\omega_1\mid x)+\lambda_{22}P(\omega_2\mid x),\ \text{then}\ x\in\begin{cases}\omega_1\\ \omega_2\end{cases}$

有 $\lambda_{11}=\lambda_{22}=0$，

$\lambda_{12}P(\omega_2\mid x)=\lambda_{12}\times 0.8\times(x-1)<\lambda_{21}P(\omega_1\mid x)=\lambda_{21}\times 0.2\times(2-x),\ \text{then}\ x\in\omega_1$，带入 x=1.5，有 $\lambda_{12}<\frac{1}{4}\lambda_{21},\ \text{then}\ x\in\omega_1$

3. 已知两类问题，有 $\mu_1=\begin{bmatrix}4\\4\end{bmatrix}$，$\mu_2=\begin{bmatrix}8\\8\end{bmatrix}$，$\sum_1=\begin{bmatrix}2&0\\0&2\end{bmatrix}$，$\sum_2=\begin{bmatrix}2&0\\0&2\end{bmatrix}$，先验概率为 P(ω1) =0.8, P(ω2) = 0.2。

(1) 求两类的贝叶斯决策分界面。

(2) 要求根据 Bayes 决策，对样本 x (3, 3)进行分类。

解： 情况 1

$g_i(x)=w_i^T x+w_{i0}$

$w_1=\frac{1}{\sigma^2}\mu_1=\begin{bmatrix}2\\2\end{bmatrix}$ ， $w_{10}=-\frac{1}{2\sigma^2}\mu_i^T\mu_i+InP(w_i)=-8-0.22=-8.22$ ，

$g_1(x)=2x_1+2x_2-8.22$

$w_2=\frac{1}{\sigma^2}\mu_1=\begin{bmatrix}4\\4\end{bmatrix}$ ， $w_{20}=-\frac{1}{2\sigma^2}\mu_i^T\mu_i+InP(w_i)=-32-1.61=-33.61$ ，

$g_2(x)=4x_1+4x_2-33.61$

$g(x)=g_1(x)-g_2(x)=-2x_1-2x_2+25.39$

$g(x)=g_1(x)-g_2(x)=-2x_1-2x_2+25.39=13.39>0$ 属于 ω1 类

4. 设 x 服从概率密度函数 $p(x|\theta)=\begin{cases}(\theta+1)x^{(\theta+1)} & (0<x<1)\\ 0 & (otherwise)\end{cases}$，样本 $x_1,\dots,x_n$ 是从分布 $p(x|\theta)$ 中独立抽取，试用最大似然估计参数 θ。

解：

$$L=\prod^n(\theta+1)x_i^{(\theta+1)}=(\theta+1)^n\prod^n x_i^{(\theta+1)}$$

$$\ln L=n\ln(\theta+1)+(\theta+1)\sum_n \ln x_i$$

$$\frac{d\ln L}{d\theta}=\frac{n}{\theta+1}+\sum_n \ln x_i=0$$

$$\theta=-\frac{n}{\sum_n \ln x_i}-1$$

5.Consider Hidden Markov Model. The hidden states are $\{\omega_1,\omega_2,\omega_3\}$, and the visible states are $\{v_1,v_2,v_3\}$. The transition probabilities are $a_{ij}=\begin{bmatrix}1&0&0\\0.3&0.3&0.4\\0.2&0.4&0.4\end{bmatrix}$,

$b_{jk}=\begin{bmatrix}1&0&0\\0&0.6&0.4\\0&0.2&0.8\end{bmatrix}$.

The initial hidden state is ω_2, and initial visible state is v_2. Try to get the probability to generate the particular visible sequence $V^3 = \{v_2, v_3, v_1\}$.

解：估值问题，采用前向算法计算生成特定可见状态序列的概率！

	v2	v3	v1
ω_1	0	0	0.1
ω_2	1	0.12	0
ω_3	0	0.32	0
t	1	2	3

由上表可得观测到 $V^3 = \{v_2, v_3, v_1\}$ 的概率为 0.1

6. 当你在数据中发现噪声时，你将在 k-NN 中考虑以下哪个选项？

A）增加 k 的值

B）减少 k 的值

C）噪声不能取决于 k

D）这些都不是

A

7. 有数据样本 x1 =2, x2 = 2.5, x3 = 3, x4 = 1 和 x5 = 6, 用 Parzen 估计在 x = 3 处的概率密度值, 采用方差 σ ^{2}=1 的 Gaussian 函数作为窗函数。

答：核函数为 $k(x, x_1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-x_1)^2}{2}}$

$$P(x=3) = \frac{1}{5} \cdot \frac{1}{\sqrt{2\pi}} \left[e^{-\frac{(2-3)^2}{2}} + e^{-\frac{(2.5-3)^2}{2}} + e^{-\frac{(3-3)^2}{2}} + e^{-\frac{(1-3)^2}{2}} + e^{-\frac{(6-3)^2}{2}} \right] = 0.228$$

8.

例: 已知两类样本

$\omega_1:\{(-5,-5)',(-5,-4)',(-4,-5)',(-5,-6)',(-6,-5)'\}$

$\omega_2:\{(5,5)',(5,6)',(6,5)',(5,4)',(4,5)'\}$

试用PCA变换做一维特征提取。

解： $\because \hat{P}(\omega_1)=\hat{P}(\omega_2)=5/10=1/2$

（1） $\therefore R=E[xx']=\sum_{i=1}^{2}\hat{P}(\omega_i)E[x^{(i)}x^{(i)}{}']=\frac{1}{2}[\frac{1}{5}\sum_{i=1}^{5}x_i^{(1)}x_i^{(1)}{}']+\frac{1}{2}[\frac{1}{5}\sum_{i=1}^{5}x_i^{(2)}x_i^{(2)}{}']$

$$=\begin{pmatrix}25.4 & 25\\ 25 & 25.4\end{pmatrix}$$

（2）求R的特征值、特征矢量

$$|R-\lambda I|=(25.4-\lambda)^2-25^2=0 \ \Rightarrow\ \lambda_1=50.4,\ \lambda_2=0.4$$

$$Rt_j=\lambda_j t_j, j=1,2\ \Rightarrow\ t_1=\frac{1}{\sqrt{2}}\begin{pmatrix}1\\1\end{pmatrix},\quad t_2=\frac{1}{\sqrt{2}}\begin{pmatrix}1\\-1\end{pmatrix}$$

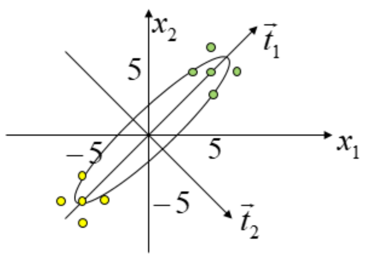
(4) 选λ_1对应的 $\vec{t}_1$作为变换矩阵 $U=[\vec{t}_1]=\frac{1}{\sqrt{2}}\begin{pmatrix}1\\1\end{pmatrix}$

由 $y=T'\vec{x}$ 得变换后的一维模式特征为

$$\vec{y}_1^{(1)}=U^T\vec{x}_1^{(1)}=\frac{1}{\sqrt{2}}(1,1)\begin{pmatrix}-5\\-5\end{pmatrix}=-\frac{10}{\sqrt{2}}$$

$\vdots$

$$y_5^{(1)}=U^T\vec{x}_5^{(1)}=-\frac{11}{\sqrt{2}}$$



得 $\omega_1:\{-\frac{10}{\sqrt{2}},-\frac{9}{\sqrt{2}},-\frac{9}{\sqrt{2}},-\frac{11}{\sqrt{2}},-\frac{11}{\sqrt{2}}\}$

$\omega_2:\{\frac{10}{\sqrt{2}},\frac{11}{\sqrt{2}},\frac{11}{\sqrt{2}},\frac{9}{\sqrt{2}},\frac{9}{\sqrt{2}}\}$