第三章 线性模型

一、填空题

1. 生成模型通过估计（类条件概率）进行分类，判别模型通过估计（后验概率）进行分类。
2. 感知器算法通常采用（梯度下降）法求解目标函数的优化问题。
3. 线性回归模型的封闭解又称为（最小二乘）解。
4. 逻辑回归函数的取值区间是（(0, 1)）。

二、判断题

5. 当两个类别均服从正态分布时，根据贝叶斯决策理论计算出的决策面必然是一个线性决策面。(×)
6. 当两个类别均服从正态分布时，根据贝叶斯决策理论计算出的决策面必然是一个二次型决策面。(×)
7. 线性回归模型如存在唯一解，必然可以令模型在训练集上的均方误差为 0.(×)
8. 4 个不具有共线性的三维样本，无法用线性回归模型得到唯一解。(×)
9. 逻辑回归模型无法用于多类分类问题。(×)

三、选择题

10. 假设线性回归问题的训练集为：$\boldsymbol{x}_i \in \Re^3, i = 1, \ldots, 100$，则该线性模型包含多少个参数：（B）

A. 3 个　　B. 4 个　　C. 100 个　　D. 101 个

11. 上题中如果使用线性回归的最小二乘解的封闭解形式，则自相关矩阵R_x的大小为：（B）

A. 3×3　　B. 4×4　　C. 100×100　　D. 101×101

12. 逻辑回归模型中的逻辑回归函数可以看作是对以下哪种概率的描述：（B）

A. $p(x|\omega_1)$　　B. $p(\omega_1|x)$　　C. $p(x)$　　D. $p(\omega_1)$

四、简答题

13. 请写出感知器算法的目标函数的标准数学形式，并解释其中每一个符号的意义与计算方法，说明其合理性。

答：

$$J(\hat{\boldsymbol{w}}) = \sum_{x \in \Omega(\hat{\boldsymbol{w}})} \delta_x \hat{\boldsymbol{w}}^T \hat{\boldsymbol{x}}$$

其中：

$$\delta_x = \begin{cases} -1 & \boldsymbol{x} \in \omega_1 \\ 1 & \boldsymbol{x} \in \omega_2 \end{cases}$$

当样本被错分类时两种情况：

- 实际属于ω_1，被错分ω_2：$\hat{\boldsymbol{w}}^T \hat{\boldsymbol{x}} < 0$，$\delta_x = -1$
- 实际属于ω_2，被错分ω_1：$\hat{\boldsymbol{w}}^T \hat{\boldsymbol{x}} > 0$，$\delta_x = 1$

当存在错误样本时，$\delta_x \hat{\boldsymbol{w}}^T \hat{\boldsymbol{x}}$总是一个整数，因此减少错误样本有利于减少目标函数$J(\hat{\boldsymbol{w}})$。

14. 请写出线性回归模型的封闭解数学形式，及其推导过程，并标明符号意义。

解：线性回归模型封闭解是对代价函数求导等于 0 的方程求解得到的结果。

代价函数为：

$$J(\hat{\boldsymbol{w}}) = \mathbb{E}\left[\left(y - f(\boldsymbol{x})\right)^2\right] = \mathbb{E}[(y - \hat{\boldsymbol{w}}^T \hat{\boldsymbol{x}})^2] \approx \frac{1}{N} \sum_{n=1}^{N} (y_n - \hat{\boldsymbol{w}}^T \hat{\boldsymbol{x}}_n)^2$$

代价函数求导为零，列方程如下：

$$\frac{\partial J(\hat{\boldsymbol{w}})}{\partial \hat{\boldsymbol{w}}} \approx \frac{2}{N} \sum_{n=1}^{N} [-\hat{\boldsymbol{x}}_n (y_n - \hat{\boldsymbol{w}}^T \hat{\boldsymbol{x}}_n)] = \frac{2}{N} \sum_{n=1}^{N} [(\hat{\boldsymbol{x}}_n \hat{\boldsymbol{x}}_n^T) \hat{\boldsymbol{w}} - \hat{\boldsymbol{x}}_n y_n] = 0$$

$$\left[\sum_{n=1}^{N} \hat{\boldsymbol{x}}_n \hat{\boldsymbol{x}}_n^T\right] \hat{\boldsymbol{w}} = \sum_{n=1}^{N} \hat{\boldsymbol{x}}_n y_n$$

令：

$$X = \begin{bmatrix} \hat{\boldsymbol{x}}_1^T \\ \hat{\boldsymbol{x}}_2^T \\ \vdots \\ \hat{\boldsymbol{x}}_N^T \end{bmatrix}, \qquad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

将其带入方程，可以得到：

$$(X^T X) \hat{\boldsymbol{w}} = X^T \boldsymbol{y}$$

进而推出线性回归模型的封闭解如下：

$$\hat{\boldsymbol{w}} = (X^T X)^{-1} X^T \boldsymbol{y}$$

15. 在线性回归模型中，假设输入样本记为$\boldsymbol{x}_i, i = 1, \dots, N$，相应的类别标签记为$y_i, i = 1, \dots, N$。请给出自相关矩阵和互相关向量的定义（公式与符号表达）。如果样本集$X \in \Re^{N \times (d+1)}$中，$N < d + 1$，应如何处理才能得到合理的模型参

数向量唯一解。

自相关矩阵记为$R_x = X^T X$，其中$X = \begin{bmatrix} \hat{\boldsymbol{x}}_1^T \\ \hat{\boldsymbol{x}}_2^T \\ \vdots \\ \hat{\boldsymbol{x}}_N^T \end{bmatrix}$，

互相关向量记作$X^T\boldsymbol{y}$，其中$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$

当$N < d+1$时，可以通过增加正则化项——单位阵$I \in \Re^{(d+1)\times(d+1)}$，得到新的最小二乘解如下：

$$\hat{\boldsymbol{w}} = (X^T X + \lambda I)^{-1} X^T \boldsymbol{y}$$

16. 请从求解线性方程组的角度说明线性回归模型的无解情况，从矩阵运算角度说明线性回归模型时的唯一解情况，并比较两者之间的联系与区别。

答：

(1) 线性回归模型，相当于求$\boldsymbol{y} = \hat{\boldsymbol{w}}^T \hat{\boldsymbol{x}}$，即求下列方程组的解

$$\begin{cases} w_1 x_1^{(1)} + w_2 x_2^{(1)} + \cdots + w_d x_d^{(1)} + w_0 = y^{(1)} \\ w_1 x_1^{(2)} + w_2 x_2^{(2)} + \cdots + w_d x_d^{(2)} + w_0 = y^{(2)} \\ \cdots \\ w_1 x_1^{(N)} + w_2 x_2^{(N)} + \cdots + w_d x_d^{(N)} + w_0 = y^{(N)} \end{cases}$$

上述方程组共 N 个方程，d+1 个未知数。在所有样本线性不相关的假设下，当$d+1 < N$时，方程组无解。

(2) 从矩阵运算的角度看，线性回归模型的解为$\hat{\boldsymbol{w}} = (X^T X)^{-1} X^T \boldsymbol{y}$，当$d$+1⩽$N$时，矩阵$X^T X$满秩，存在唯一解。

(3) 两者的差别在于当$d+1 = N$时，方程组存在唯一解；当$d+1 < N$时，矩阵运算虽然得到了唯一解，但该唯一解是使得代价函数最小的解而非使得代价函数为 0 的解，因此无法使得方程组成立，因此对于方程组来说此时仍然无解。

17. 标签y是一个随机变量，由函数$\hat{\boldsymbol{w}}^T \hat{\boldsymbol{x}}$加上一个随机噪声生成：$y = \hat{\boldsymbol{w}}^T \hat{\boldsymbol{x}} + \epsilon$；其中噪声$\epsilon$服从正态分布$\mathcal{N}(0, \sigma^2)$，设当前训练样本集为$X = \{\hat{\boldsymbol{x}}_i | i = 1, \dots, N\}, Y = \{y_i | i = 1, \dots, N\}$；

(1) 试推导$\hat{\boldsymbol{w}}$的最大似然解的数学形式；

答：根据定义，有$\epsilon = y - \widehat{\boldsymbol{w}}^T\widehat{\boldsymbol{x}} \sim \mathcal{N}(0,\sigma^2)$，其似然性函数为：

$$p(\boldsymbol{x};\delta^2,\widehat{\boldsymbol{w}}) = \frac{1}{\sqrt{2\pi\delta^2}}\exp\left(\frac{(y-\widehat{\boldsymbol{w}}^T\widehat{\boldsymbol{x}})^2}{2\delta^2}\right)$$

对于当前数据集X的整体对数似然性函数可以写为：

$$LL(X;\delta^2,\widehat{\boldsymbol{w}}) = \sum_{i=1}^{N}\ln\left(\frac{1}{\sqrt{2\pi\delta^2}}\exp\left(\frac{(y_i-\widehat{\boldsymbol{w}}^T\widehat{\boldsymbol{x}}_i)^2}{2\delta^2}\right)\right)$$

$$= \sum_{i}^{N}\left(\frac{(y-\widehat{\boldsymbol{w}}^T\widehat{\boldsymbol{x}}_i)^2}{2\delta^2} - \frac{1}{2}\ln 2\pi - \frac{1}{2}\ln\delta^2\right)$$

则线性回归参数$\widehat{\boldsymbol{w}}$的估计问题转化为以下优化问题：

$$\widehat{\boldsymbol{w}}^* = \underset{\widehat{\boldsymbol{w}}}{\operatorname{argmax}}\left(LL(X;\delta^2,\widehat{\boldsymbol{w}})\right)$$

令对数似然函数对线性回归函数的参数$\widehat{\boldsymbol{w}}$求导为 0，则有：

$$\frac{\partial LL(X;\delta^2,\widehat{\boldsymbol{w}})}{\partial\widehat{\boldsymbol{w}}} = \sum_{i=1}^{N}\left(\frac{(y_i-\widehat{\boldsymbol{w}}^T\widehat{\boldsymbol{x}}_i)\widehat{\boldsymbol{x}}_i}{\delta^2}\right) = 0$$

$$\sum_{i=1}^{N}\left((y_i-\widehat{\boldsymbol{w}}^T\widehat{\boldsymbol{x}}_i)\widehat{\boldsymbol{x}}_i\right) = 0$$

$$\left[\sum_{i=1}^{N}\widehat{\boldsymbol{x}}_i\widehat{\boldsymbol{x}}_i^T\right]\widehat{\boldsymbol{w}} = \sum_{i=1}^{N}\widehat{\boldsymbol{x}}_n y_n$$

$$\widehat{\boldsymbol{w}} = (X^TX)^{-1}X^T\boldsymbol{y}$$

(2) 假设$\widehat{\boldsymbol{w}}$的先验分布服从$d+1$元正态分布$\mathcal{N}\left(\boldsymbol{0}, I_{(d+1)\times(d+1)}\right)$,试采用最大后验概率估计法推导$\widehat{\boldsymbol{w}}$的最优解的数学形式.

答：根据最大后验概率估计的定义，则目标函数可以写为：

$$J(X;\delta^2,\widehat{\boldsymbol{w}}) = \sum_{i=1}^{N}\ln\left(\frac{1}{\sqrt{2\pi\delta^2}}\exp\left(\frac{(y_i-\widehat{\boldsymbol{w}}^T\widehat{\boldsymbol{x}}_i)^2}{2\delta^2}\right)\right) + \ln\left(\frac{1}{\sqrt{2\pi}}\exp\left(\frac{\widehat{\boldsymbol{w}}^T\widehat{\boldsymbol{w}}}{2}\right)\right)$$

则目标函数导数为 0 的方程为：

$$\frac{\partial J(X;\delta^2,\widehat{\boldsymbol{w}})}{\partial\widehat{\boldsymbol{w}}} = \sum_{i=1}^{N}\left(\frac{(y_i-\widehat{\boldsymbol{w}}^T\widehat{\boldsymbol{x}}_i)\widehat{\boldsymbol{x}}_i}{\delta^2} + \widehat{\boldsymbol{w}}\right) = 0$$

$$\sum_{i=1}^{N}\left((y_i-\widehat{\boldsymbol{w}}^T\widehat{\boldsymbol{x}}_i)\widehat{\boldsymbol{x}}_i + \delta^2\widehat{\boldsymbol{w}}\right) = 0$$

$$\left[\sum_{i=1}^{N}(\widehat{\boldsymbol{x}}_i\widehat{\boldsymbol{x}}_i^T + \delta^2 I)\right]\widehat{\boldsymbol{w}} = \sum_{i=1}^{N}\widehat{\boldsymbol{x}}_n y_n$$

$$\widehat{\boldsymbol{w}} = (X^TX + N\delta^2 I)^{-1}X^T\boldsymbol{y}$$

令$N\delta^2 = \lambda$，则有

$$\widehat{\boldsymbol{w}} = (X^TX + \lambda I)^{-1}X^T\boldsymbol{y}$$

18. 如何防止线性回归模型出现过拟合现象，请给出具体的方案

答:在均方误差目标函数中加入正则化项$\lambda\|\widehat{\boldsymbol{w}}\|^2$,从而使得封闭解形式转为:

$$\widehat{\boldsymbol{w}} = (X^T X + \lambda I)^{-1} X^T \boldsymbol{y}$$

上述方法使得当样本数量N小于参数维度$d+1$导致的过拟合现象，同时保证矩阵求逆有唯一解。

19. 请给出逻辑回归模型中 sigmoid 函数形式的推导过程，并说明其概率解释。

假设有一个二分类问题，后验概率：$P(\omega_1|\boldsymbol{x}) = y$，$P(\omega_2|\boldsymbol{x}) = 1 - y$

根据贝叶斯公式，有:

$$y = P(\omega_1|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|\omega_1)P(\omega_1)}{P(\boldsymbol{x})}, \qquad 1 - y = P(\omega_2|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|\omega_2)P(\omega_2)}{P(\boldsymbol{x})}$$

则根据最大后验概率分类准则，判别函数为:

$$\ln\frac{y}{1-y} = \ln\frac{P(\boldsymbol{x}|\omega_1)P(\omega_1)}{P(\boldsymbol{x}|\omega_2)P(\omega_2)}$$

假设两类样本服从正态分布：$\omega_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)$，$\omega_2 \sim \mathcal{N}(\boldsymbol{\mu}_2, \Sigma_2)$

$$p(\boldsymbol{x}|\omega_i) = \frac{1}{2\pi^{\frac{l}{2}}|\Sigma_i|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)^T \Sigma_i^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i)\right]$$

假设两类具有相同的先验概率和协方差矩阵：$P(\omega_1) = P(\omega_2)$，$\Sigma_1 = \Sigma_2 = \Sigma$

可以推出:

$$\ln\frac{y}{1-y} = \ln P(\boldsymbol{x}|\omega_1) - \ln P(\boldsymbol{x}|\omega_2) = \boldsymbol{w}^T \boldsymbol{x} + w_0 = \widehat{\boldsymbol{w}}^T \widehat{\boldsymbol{x}}$$

其中：$\boldsymbol{w} = \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$，$w_0 = -\frac{1}{2}(\boldsymbol{\mu}_1^T \Sigma^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^T \Sigma^{-1} \boldsymbol{\mu}_2)$

进一步化简，可以得到:

$$\ln\frac{y}{1-y} = \widehat{\boldsymbol{w}}^T \widehat{\boldsymbol{x}} \Longrightarrow y = \frac{1}{1 + e^{-\widehat{\boldsymbol{w}}^T \widehat{\boldsymbol{x}}}} = P(\omega_1|\boldsymbol{x})$$

基于上述推导，可以看出逻辑回归函数是样本 x 在具有相同先验概率和协方差矩阵的正态分布假设下从属于ω_1类的后验概率。

20. 请给出逻辑回归模型梯度下降法更新公式的推导过程。

答：逻辑回归函数是一种判别模型，可以视为寻找一组参数，使得后验概率最大化。后验概率原始形式为：

$$P(\boldsymbol{\theta}|\boldsymbol{x}) = \begin{cases} P(\omega_1|\boldsymbol{x}) & y = 1 \\ 1 - P(\omega_1|\boldsymbol{x}) & y = 0 \end{cases}$$

上式等价于：

$$\ln P(\boldsymbol{\theta}|\boldsymbol{x}) = y\ln P(\omega_1|\boldsymbol{x}) + (1-y)\ln\big(1 - P(\omega_1|\boldsymbol{x})\big)$$

根据逻辑回归函数定义：

$$P(\omega_1|\boldsymbol{x}) = h(x) = \frac{1}{1 + e^{-\widehat{\boldsymbol{w}}^T\widehat{\boldsymbol{x}}}}$$

带入上式，将后验概率最大化问题转化为目标函数$J(\widehat{\boldsymbol{w}})$的最小化问题：

$$\begin{aligned}
J(\widehat{\boldsymbol{w}}) &= -\textstyle\sum_{x\in X}\big(y\ln P(\omega_1|\boldsymbol{x}) + (1-y)\ln\big(1 - P(\omega_1|\boldsymbol{x})\big)\big) \\
&= -\sum_{x\in X}\left(y\ln\frac{1}{1+\exp(-\widehat{\boldsymbol{w}}^T\widehat{\boldsymbol{x}})} + (1-y)\ln\left(\frac{\exp(-\widehat{\boldsymbol{w}}^T\widehat{\boldsymbol{x}})}{1+\exp(-\widehat{\boldsymbol{w}}^T\widehat{\boldsymbol{x}})}\right)\right) \\
&= -\sum_{x\in X}\big(y\ln[h(x)] + (1-y)\ln[h(x)] + (1-y)(-\widehat{\boldsymbol{w}}^T\widehat{\boldsymbol{x}})\big) \\
&= -\sum_{x\in X}\left(\ln\left[\frac{\exp(\widehat{\boldsymbol{w}}^T\widehat{\boldsymbol{x}})}{1+\exp(\widehat{\boldsymbol{w}}^T\widehat{\boldsymbol{x}})}\right] + (1-y)(-\widehat{\boldsymbol{w}}^T\widehat{\boldsymbol{x}})\right) \\
&= -\sum_{x\in X}\big(-\ln[1+\exp(\widehat{\boldsymbol{w}}^T\widehat{\boldsymbol{x}})] + \widehat{\boldsymbol{w}}^T\widehat{\boldsymbol{x}} + (1-y)(-\widehat{\boldsymbol{w}}^T\widehat{\boldsymbol{x}})\big) \\
&= \sum_{x\in X}\big(-y\widehat{\boldsymbol{w}}^T\widehat{\boldsymbol{x}} + \ln(1+\exp(\widehat{\boldsymbol{w}}^T\widehat{\boldsymbol{x}}))\big)
\end{aligned}$$

利用梯度下降法，先求$J(\widehat{\boldsymbol{w}})$的导数：

$$\begin{aligned}
\frac{\partial J(\widehat{\boldsymbol{w}})}{\partial\widehat{\boldsymbol{w}}} &= \sum_{x\in X}\left(-y\widehat{\boldsymbol{x}} + \frac{\exp(\widehat{\boldsymbol{w}}^T\widehat{\boldsymbol{x}})\,\widehat{\boldsymbol{x}}}{1+\exp(\widehat{\boldsymbol{w}}^T\widehat{\boldsymbol{x}})}\right) \\
&= \sum_{x\in X}\left(\frac{1}{1+\exp(-\widehat{\boldsymbol{w}}^T\widehat{\boldsymbol{x}})} - y\right)\widehat{\boldsymbol{x}} \\
&= \sum_{x\in X}(h(\boldsymbol{x}) - y)\,\widehat{\boldsymbol{x}}
\end{aligned}$$

最后的梯度下降法更新公式：

$$\widehat{\boldsymbol{w}}(t+1) = \widehat{\boldsymbol{w}}(t) - \rho\sum_{x\in X}\big(h\big(\boldsymbol{x};\widehat{\boldsymbol{w}}(t)\big) - y\big)\,\widehat{\boldsymbol{x}}$$

五、计算（画图）题

21. 已知两个类别ω_1和ω_2，其先验概率相等，两类的类条件概率密度服从正态分布，有$\mathrm{p}(\boldsymbol{x}|w_1) = \mathcal{N}\left(\mu_1, \Sigma\right)$和$\mathrm{p}(\boldsymbol{x}|w_2) = \mathcal{N}\left(\mu_2, \Sigma\right)$，且$\mu_1 = \begin{bmatrix}0\\0\end{bmatrix}$，$\mu_2 = \begin{bmatrix}3\\3\end{bmatrix}$，

$\sum = \begin{bmatrix} 1.1 & 0.3 \\ 0.3 & 1.9 \end{bmatrix}$，试写出上述问题每一类的判别函数$g_i(\boldsymbol{x}), i = 1,2$的线性形式以及整个分类问题的决策面方程的线性形式。并判断样本$\boldsymbol{x} = \begin{bmatrix} 1.2 \\ 1.9 \end{bmatrix}$属于哪一类，给出计算过程。

答：判别函数是后验概率的某种单调递增函数，考虑到两类服从高斯分布。可以首先使用后验概率的对数函数作为初始的判别函数：

$$g_i(x) = \ln\big(P(\omega_i|x)\big) = \ln\left(\frac{P(x|\omega_i)P(\omega_i)}{P(x)}\right)$$

题中条件先验概率相等，边缘概率$P(x)$不改变两类之间的相对大小，因此可以简化为：

$$\begin{aligned} g_i(x) &= \ln\big(P(x|\omega_i)\big) \\ &= \ln\left(\frac{1}{2\pi|\Sigma|^{\frac{1}{2}}}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right]\right) \end{aligned}$$

考虑到两类的协方差矩阵相等，可以去除包含协方差$\sum$的常数项和二次项，得到：

$$g_i(x) = \boldsymbol{\mu}_i^T\Sigma^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_i^T\Sigma^{-1}\boldsymbol{\mu}_i$$

已知:$\sum = \begin{bmatrix} 1.1 & 0.3 \\ 0.3 & 1.9 \end{bmatrix}$,可以求得$\Sigma^{-1} = \begin{bmatrix} 0.95 & -0.15 \\ -0.15 & 0.55 \end{bmatrix}$,带入均值向量$\boldsymbol{\mu}_i$得:

$$g_1(x) = 0$$
$$g_2(x) = 2.4x_1 + 1.2x_2 - 5.4$$

决策面方程为:

$$g(x) = g_1(x) - g_2(x) = 0$$

推出决策面方程为:

$$-4x_1 - 2x_2 + 9 = 0$$

将样本$\mathbf{x} = \begin{bmatrix} 1.2 \\ 1.9 \end{bmatrix}$，带入决策面方程等号右侧得到$0.4 > 0$,因此为$\omega_1$类

22. 已知决策面方程：$5x_1 - x_2 - 1 = 0$,当前错分样本为:(0.4,0.6)，(0.1 − 0.25)。

(1) 写出参数向量w ，并设定合适的学习步长η;

(2) 列出迭代公式和计算结果

(3) 列出迭代后的决策面方程

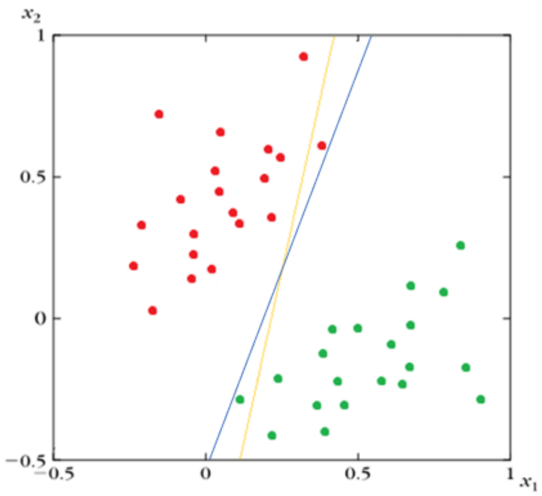
(4) 画出相应的决策面（虚线）

(5) 给出结论与分析

(1) $\boldsymbol{w}=\begin{bmatrix}5\\-1\\-1\end{bmatrix}$，$\eta=0.9$

(2) $\widehat{\boldsymbol{w}}(t+1)=\begin{bmatrix}5\\-1\\-1\end{bmatrix}-0.9(+1)\begin{bmatrix}0.4\\0.6\\1\end{bmatrix}-0.9(-1)\begin{bmatrix}0.1\\-0.3\\1\end{bmatrix}=\begin{bmatrix}4.73\\-1.81\\-1\end{bmatrix}$

(3) 迭代后的决策面方程：$4.73x_1-1.81x_2-1=0$



(4) 通过梯度下降法，经过一次迭代，决策面的位置发生了变化。将两个错分样本(0.4,0.6)，(0.1 − 0.25)分别代入新的决策面方程，可以发现分类结果正确，说明感知器算法实现了对决策面的优化，提升了在训练集上的正确率。

23. 以 Iris 数据库每类的 70%的样本为训练集，基于线性回归模型，使用花萼长度、花萼宽度和花瓣宽度特征来估计其花瓣长度特征，请写出线性回归模型参数的封闭解的数学形式，标明每个变量的矩阵大小或矢量维度。

24. 已知 3 个样本坐标为$\boldsymbol{x}^{(1)}=[1,0],\boldsymbol{x}^{(2)}=[0,1],\boldsymbol{x}^{(3)}=[1,1]$，其中$\boldsymbol{x}^{(1)},\boldsymbol{x}^{(2)}\in\omega_1,\boldsymbol{x}^{(3)}\in\omega_2$,初始化逻辑回归模型对应的线性决策面为：$x_1=0.5$。设学习步长为 0.2。请计算基于逻辑回归梯度下降法进行一次迭代后的决策面方程，并在下图中画出相应的决策面直线。

根据题意，定义以下变量并且初始化：

步长：$\rho=0.2$

模型参数：$\widehat{\boldsymbol{w}}=[w_1,w_2,w_0]=[1,0,-0.5]$

标签：$\boldsymbol{y} = [1,1,0]^T$

训练样本：$\hat{\boldsymbol{x}}^{(1)} = [1,0,1], \hat{\boldsymbol{x}}^{(2)} = [0,1,1], \hat{\boldsymbol{x}}^{(3)} = [1,1,1]$

计算梯度：

$$\frac{\partial J(\hat{\boldsymbol{w}})}{\partial \hat{\boldsymbol{w}}} = \frac{1}{M}\sum_{i=1}^{M}\left(h(\hat{\boldsymbol{x}}^{(i)}) - y^{(i)}\right)\hat{\boldsymbol{x}}^{(i)}$$

$$= \frac{1}{3}\left\{\left(\frac{1}{1+e^{-[1,0,-0.5][1,0,1]^T}} - 1\right)\begin{bmatrix}1\\0\\1\end{bmatrix} + \left(\frac{1}{1+e^{-[1,0,-0.5][0,1,1]^T}} - 1\right)\begin{bmatrix}0\\1\\1\end{bmatrix} + \left(\frac{1}{1+e^{-[1,0,-0.5][1,1,1]^T}} - 0\right)\begin{bmatrix}1\\1\\1\end{bmatrix}\right\}$$

$$= [0.082,\ 0, -0.13]^T$$

更新模型参数：

$$\hat{\boldsymbol{w}}^{(1)} = \hat{\boldsymbol{w}}^{(0)} - \rho\frac{\partial J(\hat{\boldsymbol{w}})}{\partial \hat{\boldsymbol{w}}} = \begin{bmatrix}1\\0\\-0.5\end{bmatrix} - 0.2 \times \begin{bmatrix}0.082\\0\\-0.13\end{bmatrix} = \begin{bmatrix}0.98\\0\\-0.47\end{bmatrix}$$

迭代一次后的决策面方程为：

$$0.98x_1 - 0.47 = 0$$
$$x_1 = 0.48$$