

# 《数据挖掘学习课程》Homework 2.

## 第一题：

(1)、根据题： $Ent(D) = \sum_{k=1}^K P_k \log_2 \frac{1}{P_k} = -\sum_{k=1}^K P_k \log_2 P_k$   
 当  $P_k=1/K$  时，信息熵  $Ent(D)$  取得最大。  
 此时  $Ent(D) = -K \times \left[ \frac{1}{K} \log_2 \frac{1}{K} \right] = \log_2 K$

(2) 假设  $D_1$  对应 outlook 列的 4 个属性值

$$Gain(D, a_1) = Ent(D) - \sum_{i=1}^4 \frac{|D_i|}{|D|} Ent(D_i), \quad Ent(D_1^3) = -\log_2 1 = 0$$

$$Ent(D) = -\frac{3}{10} \log_2 \frac{3}{10} - \frac{7}{10} \log_2 \frac{7}{10} = 0.88129,$$

$$Ent(D_1^1) = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} = -\log_2 \frac{1}{2} = 1$$

$$Ent(D_1^2) = -\frac{1}{4} \log_2 \frac{1}{4} - \frac{3}{4} \log_2 \frac{3}{4} = 0.81127$$

$$Gain(D, a_1) = 0.88129 - \frac{1}{10} \times 1 - \frac{4}{10} \times 0.81127 - 0 = 0.156782$$

假设  $D_2$  对应 temperature 的 3 个属性值

$$hot: \quad Ent(D_2^1) = -\frac{2}{3} \log_2 \frac{2}{3} - \frac{1}{3} \log_2 \frac{1}{3} = 0.918296$$

$$cool: \quad Ent(D_2^2) = -\frac{1}{4} \log_2 \frac{1}{4} - \frac{3}{4} \log_2 \frac{3}{4} = 0.811278$$

$$mild \quad Ent(D_2^3) = 0$$

$$Gain(D, a_2) = 0.88129 - \frac{4}{10} \times 0.811278 - \frac{3}{10} \times 0.918296 \\ = 0.28129$$

假设  $D_3$  对应 humidity 的属性值

$$high \quad Ent(D_3^1) = -\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} = 0.97095$$

$$nor \quad Ent(D_3^2) = -\frac{1}{5} \log_2 \frac{1}{5} - \frac{4}{5} \log_2 \frac{4}{5} = 0.721928$$

$$Gain(D, a_3) = 0.88129 - \frac{1}{2} \times 0.97095 - \frac{1}{2} \times 0.721928 \\ = 0.0348$$

假设  $D_4$  对应 windly 的属性值

$$False: \quad Ent(D_4^1) = -\frac{1}{7} \log_2 \frac{1}{7} - \frac{6}{7} \log_2 \frac{6}{7} = 0.591672$$

$$True: \quad Ent(D_4^2) = -\frac{2}{3} \log_2 \frac{2}{3} - \frac{1}{3} \log_2 \frac{1}{3} = 0.918296$$

$$Gain(D, a_4) = 0.88129 - \frac{7}{10} \times 0.591672 - \frac{3}{10} \times 0.918296 \\ = 0.191637$$

由于  $Gain(D, a_2) > Gain(D, a_4) > Gain(D, a_1) > Gain(D, a_3)$

## 1. 选取 temperature 作为第一个分类属性

13) 处理方法：

- ① 对于决策树来说，采用离散化，将连续属性  $n$  个排序为  $a_1, a_2, \dots, a_n$ ，则形成 **Binning** 成了  $n-1$  个区间，可以用  $\frac{a_i+a_{i+1}}{2}$  代表整个区间，从而将连续型样本转化为  $n-1$  个数据集合；
- 2分法 { 或是二分法，找到  $k$ ，使得  $a_1, \dots, a_n$  可分为  $\{a_i | a_i \leq k\}$  和  $\{a_i | a_i > k\}$ ，选择最大化信息增益的阈值进行分割。
- ② 对于朴素贝叶斯，假设连续属性符合正态分布  $N(\mu, \sigma^2)$ ，用 training data 计算  $\mu$  和  $\sigma^2$ ，基于高斯密度函数测试样本属于某类条件概率。

14)  $Recall = \frac{TP}{TP+FN}$ , 为查全率，代表查到的 正确结果占所有正确结果的比例，适合漏掉正类样本后果严重场景。

$Precision = \frac{TP}{TP+FP}$  为查准率，代表所有被预测为正确结果的中，真正为正确结果的比例，适用于负类样本被误判为正类后果较严重的场景。

15) 答：由于  $3! = 3 \times 2 = 6$ ，最多可以有 6 种不同决策树结构

## 第二题：

11) 解：设问题为  $\min \frac{1}{2} \|w\|^2$ ,

$$y_1(w^T x_1 + b) \geq 1,$$

将  $(x_1, y_1) = (2, 3, 1)^T$ ,  $(x_2, y_2) = (3, 2, 1)^T$ ,  $(x_3, y_3) = (1, 1, -1)^T$  代入：

$$\begin{cases} 2w_1 + 3w_2 + b \geq 1 \\ 3w_1 + 2w_2 + b \geq 1 \\ -(w_1 + w_2 + b) \geq 1 \end{cases} \quad \text{解得} \quad \begin{cases} w_1 = 2/3, \\ w_2 = 2/3 \\ b = -\frac{7}{3} \end{cases}$$

12) 答：不会，因为该点不是支持向量，SVM 的目标为最大化已存在支持向量到决策边界的距离，而不包括新加入、且非支持向量数据点的距离。

### 第3题

1. ① 以年龄为标准：青年样本5个，放贷2个，错误率  $1 - 0.4 = 0.4$

中年样本5个，放贷3个，错误率  $1 - 0.6 = 0.4$

老年样本5个，放贷4个，错误率  $1 - 0.8 = 0.2$

$$\text{总加权错误率} = \frac{1}{3} \times 0.4 + \frac{1}{3} \times 0.4 + \frac{1}{3} \times 0.2 = \frac{1}{3}$$

② 有工作：

无工作样本10个，放贷4个，错误率  $1 - \frac{4}{10} = 0.4$

有工作样本5个，放贷5个，错误率 0

$$\text{总加权错误率} = \frac{10}{15} \times 0.4 + \frac{5}{15} \times 0 = \frac{4}{5}$$

③ 有自己的房子：

无房子样本9个，放贷3个，错误率  $1 - \frac{3}{9} = \frac{2}{3}$

有房子样本6个，不放贷0个，错误率  $1 - \frac{0}{6} = 0$

$$\text{加权错误率} = \frac{9}{15} \times \frac{2}{3} = \frac{1}{5}$$

④ 信贷情况：

一般样本5个，放贷1个，错误率  $1 - \frac{1}{5} = \frac{4}{5}$

好样本6个，放贷5个，错误率  $1 - \frac{5}{6} = \frac{1}{6}$

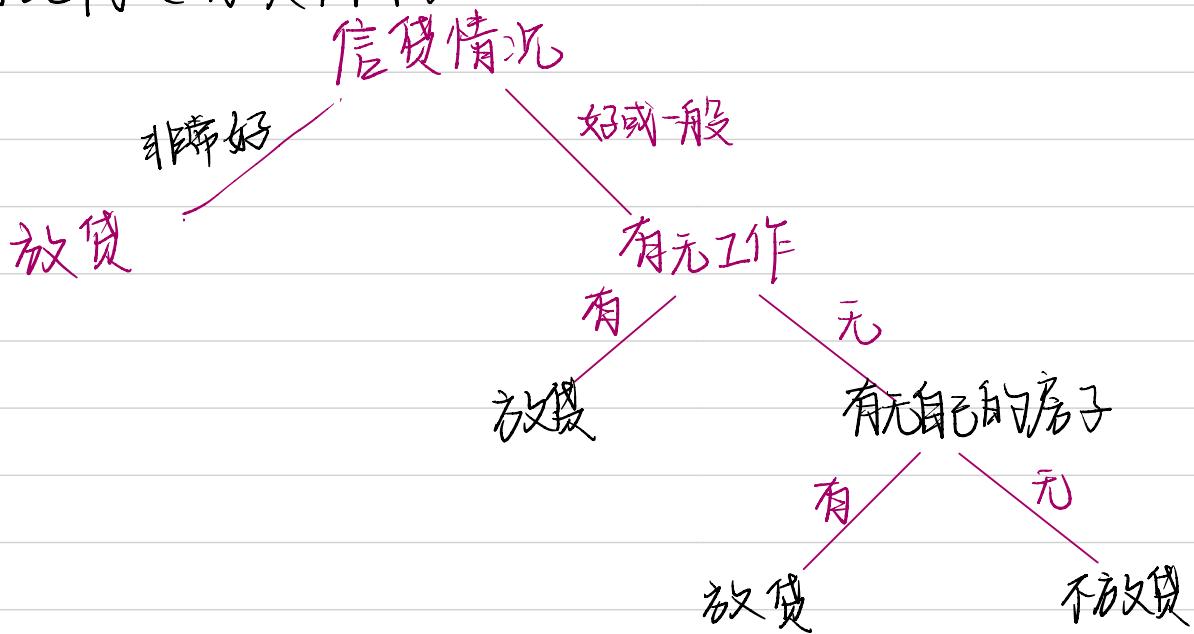
非常好样本4个，放贷4个，错误率为 0

因此选择“信贷情况”作为决策树第一个节点。

按以上思路，第二个节点选择“有无工作”，对有工作的放贷

第三个节点选择“有无自己的房子”，对有房子的放贷

因此构建的决策树为：



2. 按照上面决策树，属性值（中年，无工作，无自己的房子，信贷情况好）的申请者不放贷

3. (过拟合):
- (1) 剪枝: 用预剪枝和后剪枝策略, 提前停止树的生长或删除分支  
可以限制树的复杂度.
  - (2) 集成方法: 将多个决策树集成, 使用如 XGBoost; Random Forest, AdaBoosting 等方法引入多样性与随机性
  - (3) 正则化: 可以通过限制叶子节点或引入 penalty coefficient 避免过多分支, 从而抑制过拟合
  - (4) 增加数据量: 采取数据增强方法等

4. 对于连续属性的样本数据, 通过类似离散属性值的分割方法对离散属性离散化, 依据信息增益等标准选择最优的划分点, 对样本集合进行划分; 一般情况, 使用二元分裂法将连续属性分为“ $\leq$ 划分点”及“ $>$ 划分点”2个部分, 也可选择多个划分点, 将连续属性划分为多个部分.

## 第四题

请评价两个分类器 M1 和 M2 的性能。所选择的测试集包含 26 个二值属性，记作 A 到 Z。表中是模型应用到测试集时得到的后验概率（图中只显示正类的后验概率）。因为这是二类问题，所以  $P(-)=1-P(+)$ ,  $P(-|A, \dots, Z)=1-P(+|A, \dots, Z)$ 。假设需要从正类中检测实例。

1. 画出 M1 和 M2 的 ROC 曲线（画在一幅图中）。哪个模型更好？给出理由。
2. 对模型 M1，假设截止阈值  $t=0.5$ 。换句话说，任何后验概率大于  $t$  的测试实例都被看作正例。计算模型在此阈值下的 precision, recall 和 F-score。
3. 对模型 M2 使用相同的截止阈值重复（2）的分析。比较两个模型的 F-score，哪个模型更好？所得结果与从 ROC 曲线中得到的结论一致吗？
4. 使用阈值  $t=0.1$  对模型 M2 重复（2）的分析。 $t=0.5$  和  $t=0.1$  哪一个阈值更好？该结果和你从 ROC 曲线中得到的一致吗？

```
import numpy as np
import matplotlib.pyplot as plt

def prf(gt,p,th):
    tp = 0
    fp = 0
    tn = 0
    fn = 0
    for i in range(len(gt)):
        if(p[i]>th):
            if(gt[i] == 1):
                tp += 1
            else:
                fp += 1
        else:
            if(gt[i] == 1):
                fn += 1
            else:
                tn += 1
    return [tp/(),]

def fpr(gt,p,th):# 假正率, FP/FP+TN, 横坐标
    fp = 0
    tn = 0
    for i in range(len(gt)):
        if(p[i]>th):#判为阳性
            if(gt[i] == 1):
                pass
                #tp += 1
        else:
            fp += 1
    else:#判为阴性
        if(gt[i] == 1):
            pass
            #fn += 1
        else:#真阴性
            tn += 1
```

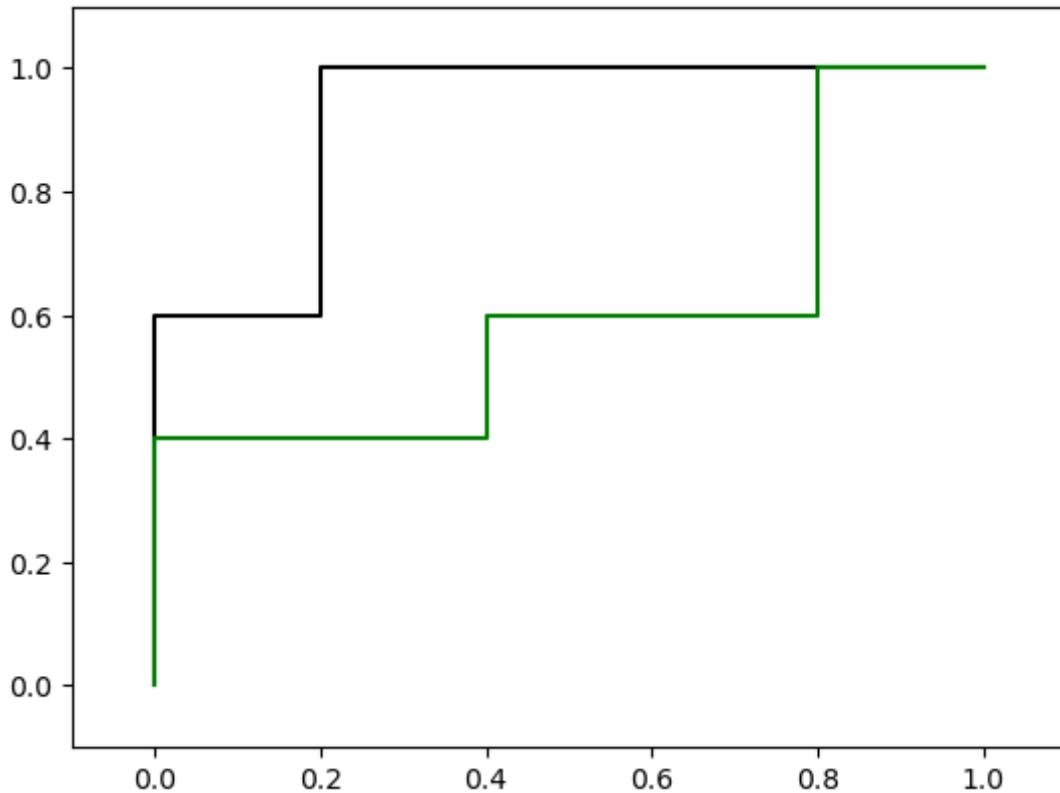
```

    return fp/(fp+tn)

def tpr(gt,p,th):
    tp = 0
    fn = 0
    for i in range(len(gt)):
        if(p[i]>th):
            if(gt[i] == 1):
                tp += 1
            else:
                pass
                #fp += 1
        else:
            if(gt[i] == 1):
                fn += 1
            else:#真阴性
                pass
                #tn += 1
    return tp/(tp+fn)

def corr(gt,p):
    N = 101
    cors = np.zeros((2,N))
    th = np.linspace(0,1,N)
    for i in range(N):
        cors[0][i]=fpr(gt,p,th[i])
        cors[1][i]=tpr(gt,p,th[i])
    return cors
gt = [1,1,0,0,1,1,0,0,1,0]
p_m1 = np.array([0.73,0.69,0.44,0.55,0.67,0.47,0.08,0.15,0.45,0.35])
p_m2_ = np.array([0.61,0.03,0.68,0.31,0.45,0.09,0.38,0.05,0.01,0.04])
p_m2 = 1-p_m2_
print(p_m2)
c1 = corr(gt,p_m1)
c2 = corr(gt,p_m2)
plt.plot(c1[0], c1[1],color='black')
plt.plot(c2[0], c2[1],color='green')
plt.xlim([-0.1, 1.1])
plt.ylim([-0.1, 1.1])
plt.show()

```



1. M1更多。上图中，黑色线条为M1，绿色线条为M2，相同阈值时M1表现更好，其ROC曲线下面积AUC也超过M2，一般认为AUC数值越大，则模型的性能表现更佳。

2.

蓝色为M1-ROC，红色为M2-ROC

```
#2,3,4
import numpy as np
# 以1代表+, 0代表-, 输入p均为P(+)
def prf(gt,p,th):
    tp = 0
    fp = 0
    tn = 0
    fn = 0
    for i in range(len(gt)):
        if(p[i]>th):#判为阳性
            if(gt[i] == 1):#真阳性
                tp += 1
            else:#假阳性
                fp += 1
        else:#判为阴性
            if(gt[i] == 1):#假阴性
                fn += 1
            else:#真阴性
                tn += 1
    precision = tp/(tp+fp)
    recall = tp/(tp+fn)
    fscore = 2*precision*recall/(precision+recall)
    return precision,recall,fscore

gt = [1,1,0,0,1,1,0,0,1,0]
p_m1 = np.array([0.73,0.69,0.44,0.55,0.67,0.47,0.08,0.15,0.45,0.35])
```

```

p_m2_ = np.array([0.61,0.03,0.68,0.31,0.45,0.09,0.38,0.05,0.01,0.04])
p_m2 = 1-p_m2_
print("#2 P={0[0]},R={0[1]},F={0[2]}".format(prf(gt,p_m1,0.5)))
print("#3 P={0[0]},R={0[1]},F={0[2]}".format(prf(gt,p_m2,0.5)))
print("#4 P={0[0]},R={0[1]},F={0[2]}".format(prf(gt,p_m2,0.1)))

```

由打印出来的结果得，precision为0.75，recall为0.6，F-score为0.667.

3. 由第二题的代码打印结果，得：precision为0.5，recall为0.8，F-score为0.6154；此时模型M1的性能表现得更好，与ROC曲线一致。
4. 模型M2在阈值t=0.1下的precision为0.5，recall为1.0，F-score为0.667，模型M1更好，与ROC曲线一致。

## 第五题

下图中数据元组已经按分类器返回概率值的递减顺序排序。对于每个元组，计算真正例 (TP)、假正例 (FP)、真负例 (TN) 和假负例 (FN) 的个数。计算真正例率 (TPR) 和假正例率 (FPR)。为该数据绘制 ROC 曲线。

元组号	类	概率
1	P	0.95
2	N	0.85
3	P	0.78
4	P	0.66
5	N	0.60
6	P	0.55
7	N	0.53
8	N	0.52
9	N	0.51
10	P	0.40

```

import numpy as np
import matplotlib.pyplot as plt
from sklearn.metrics import roc_curve

# 二分类结果：1 表示正类，0 表示负类
a = [1, 0, 1, 1, 0, 1, 0, 0, 0, 1]

b = np.array([0.95, 0.85, 0.78, 0.66, 0.60, 0.55, 0.53, 0.52, 0.51, 0.40])

# 计算 FPR, TPR 和 阈值
fpr, tpr, thresholds = roc_curve(a, b)

```

```

plt.figure(figsize=(8, 6))
plt.plot(fpr, tpr, color='darkorange', linewidth=2, label='ROC curve')

plt.plot([0, 1], [0, 1], color='navy', linestyle='--', linewidth=1.5,
label='Random classifier')

# 设置坐标轴范围
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])

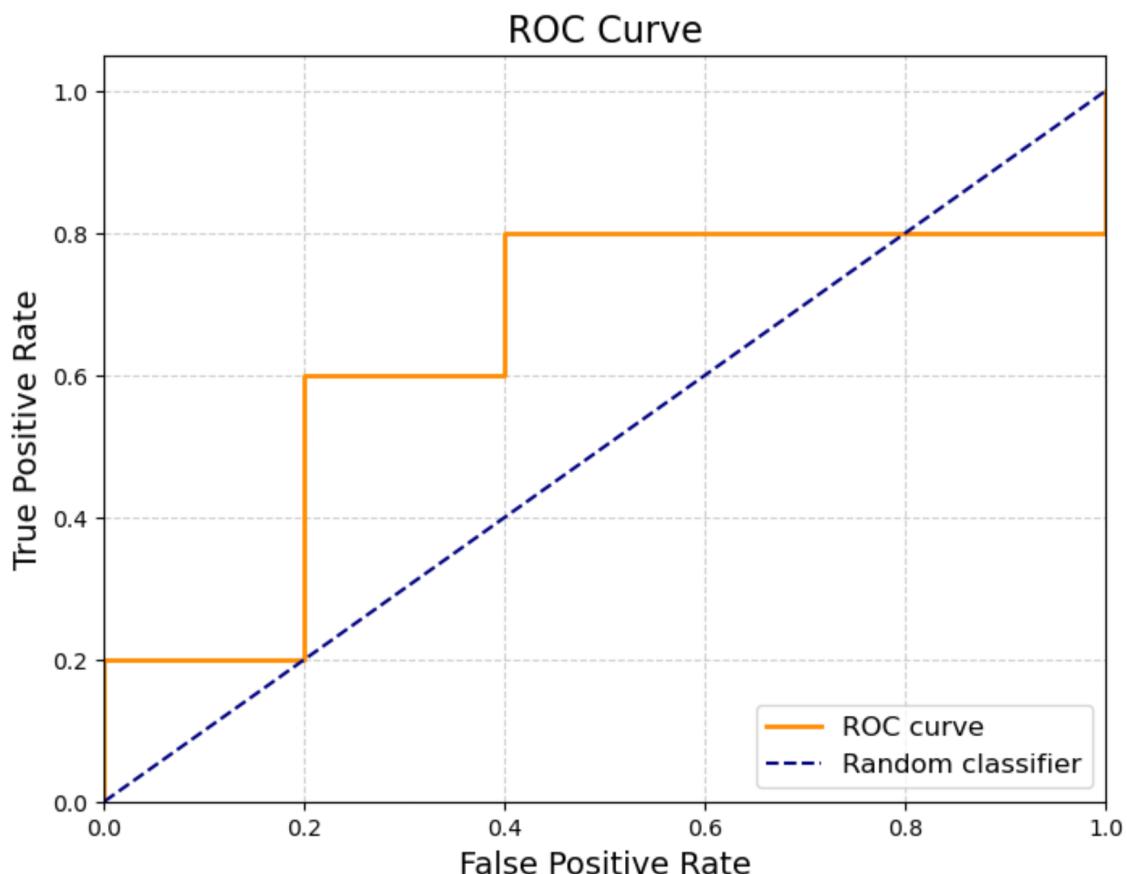
# 设置坐标轴标签和标题
plt.xlabel('False Positive Rate', fontsize=14)
plt.ylabel('True Positive Rate', fontsize=14)
plt.title('ROC Curve', fontsize=16)

# 显示网格
plt.grid(True, linestyle='--', alpha=0.6)

# 添加图例
plt.legend(loc='lower right', fontsize=12)

# 显示图形
plt.show()

```



元组的真正例 (TP) 、假正例 (FP) 、真负例 (TN) 和假负例 (FN) 的个数的计算：

阈值编号	TP	FP	TN	FN	TPR	FPR
1	1	0	5	4	0.20	0.00
2	1	1	4	4	0.20	0.20
3	2	1	4	3	0.40	0.20
4	3	1	4	2	0.60	0.20
5	3	2	3	2	0.60	0.40
6	4	2	3	1	0.80	0.40
7	4	3	2	1	0.80	0.60
8	4	4	1	1	0.80	0.80
9	4	5	0	1	0.80	1.00
10	5	5	0	0	1.00	1.00

## 第6题

解：设  $\Sigma M = [230.5, 32.2, 20.7, 20.6, \dots, 28.0]$ ,

$\Sigma N = [22.4, 14.5, 22.4, 19.6, \dots, 35]$

$$\text{则 } \Sigma M - \Sigma N = \Sigma = [8.1, 17.7, -1.1, 1, 10.3, 20.6, 5.6, 8.6, 3.3, -7]$$

$$M = \frac{\sum \Sigma_i}{|\Sigma|} = 6.65, \quad S^2 = \frac{\sum (\Sigma_i - M)^2}{|\Sigma|} = 63.6225,$$

$$\text{则 } \left| \frac{\sqrt{10} \times M}{S} \right| = 2.6364, \text{ 使用双边检验: } t_{0.01/2, 9} = 3.25$$

因此  $2.6364$  落入了接受域  $(-t_{0.01/2, 9}, t_{0.01/2, 9})$  内，

认为  $M$  显著优于  $N$

## 第7题

$$\frac{1}{a\frac{P}{R} + (1-a)\frac{R}{P}} = \frac{PR}{aR + (1-a)P} = \frac{\frac{PR}{a}}{R + \frac{1-a}{a}P} = \frac{(B^2+1)PR}{BP+R},$$

$$\text{则 } \begin{cases} B^2 = \frac{1-a}{a} \\ B^2+1 = \frac{1}{a} \end{cases} \rightarrow a = \frac{1}{B^2+1}$$

由于  $P = \frac{TP}{TP+FP}$ ,  $R = \frac{TP}{TP+FN}$ , 代入:

$$F = \frac{(B^2+1) \times \frac{TP}{TP+FP} \times \frac{TP}{TP+FN}}{B^2 \left( \frac{TP}{TP+FP} \right) + \frac{TP}{TP+FN}} = \frac{(B^2+1) TP^2}{B^2 \times TP \times (TP+FN) + TP(TP+FP)}$$

$$= \frac{(B^2+1) TP}{B^2(TP+FN) + TP+FP}$$

$$= \frac{(B^2+1) TP}{(B^2+1) TP + B^2 FN + FP}$$

## 第八题

设  $x_0$  在平面  $S$  反影为  $x_1$ ,  $\vec{w}x_1 + b = 0$ .

由于  $\vec{w}x_1$  与  $S$  的法向量平行, 且  $|\vec{w} \cdot \vec{x}_0 x_1| = ||w|| \times d = ||w|| \times |\vec{x}_0 x_1|$ .

$$\text{而 } \vec{w} \cdot \vec{x}_0 x_1 = w^1(x_0^1 - x_1^1) + w^2(x_0^2 - x_1^2) + \dots + w^n(x_0^n - x_1^n)$$

$$= w^1 x_0^1 + w^2 x_0^2 + \dots + w^n x_0^n - (w^1 x_1^1 + w^2 x_1^2 + \dots + w^n x_1^n)$$

$$= w^1 x_0^1 + w^2 x_0^2 + \dots + w^n x_0^n + b$$

$$\text{即 } ||w|| d = |\vec{w}x_0 + b|,$$

$$d = \frac{1}{||w||} |\vec{w}x_0 + b|$$

## 第9題

設表中點為  $P_1, P_2, \dots, P_{10}$ , 記  $(x_1=4, x_2=1)$  為點  $q_n$ .

$$\begin{aligned} d(P_1, q_n) &= \sqrt{9+16} = 5, & d(P_2, q_n) &= \sqrt{4+4} = 2\sqrt{2}, & d(P_3, q_n) &= \sqrt{1+0} = 1 \\ d(P_4, q_n) &= \sqrt{1+9} = \sqrt{10}, & d(P_5, q_n) &= 1, & d(P_6, q_n) &= \sqrt{0+1} = 1 \\ d(P_7, q_n) &= \sqrt{1+9} = \sqrt{10}, & d(P_8, q_n) &= \sqrt{4+9} = \sqrt{13}, & d(P_9, q_n) &= \sqrt{4} = 2 \\ d(P_{10}, q_n) &= \sqrt{16+1} = \sqrt{17} \end{aligned}$$

最近的为  $P_5, P_3, P_6$ , 最遠的为 " $-$ " - "类

二 將  $(x_1=4, x_2=1)$  分為 " $-$ " - "类

## 第10題

解：模型為  $m \geq \|W\|^2$

$$\left\{ y_i (W^T x_i + b) \geq 1, \quad \rightarrow \quad \begin{array}{l} \text{將 } x_1 = [1, 3]^T, x_2 = [4, 3]^T \\ x_3 = [1, 1]^T \text{ 代入:} \end{array} \right.$$

$$\begin{cases} 3w_1 + 3w_2 + b \geq 1 \\ 4w_1 + 3w_2 + b \geq 1 \\ -w_1 - w_2 - b \geq 1 \end{cases}, \text{構造輔助函數:}$$

$$L(W, b, a) = \frac{1}{2}(w_1^2 + w_2^2) + a_1(1 - 3w_1 - 3w_2 - b) + a_2(1 + w_1 + w_2 + b)$$

→ 对  $a_1, a_2$  求偏导:

$$\begin{cases} w_1 = 3a_1 - a_2 \\ w_2 = 3a_1 - a_2 \\ 0 = -a_1 + a_2 \end{cases}$$

将  $\begin{cases} w_1 = 3a_1 - a_2 \\ w_2 = 3a_1 - a_2 \\ 0 = -a_1 + a_2 \end{cases}$  代入辅助函数:

$$f(x_1) = -4a_1^2 + 2a_1$$

$$\text{且 } f'(a_1) = -8a_1 + 2 = 0 \rightarrow a_1 = 1/4,$$

此时  $a_2 = 1/4$

$$w_1 = \frac{3}{4} - \frac{1}{4} = \frac{1}{2} = w_2, b \leq -2,$$

则最大间隔超平面为  $\frac{1}{2}x^{(1)} + \frac{1}{2}x^{(2)} - 2 = 0$ .