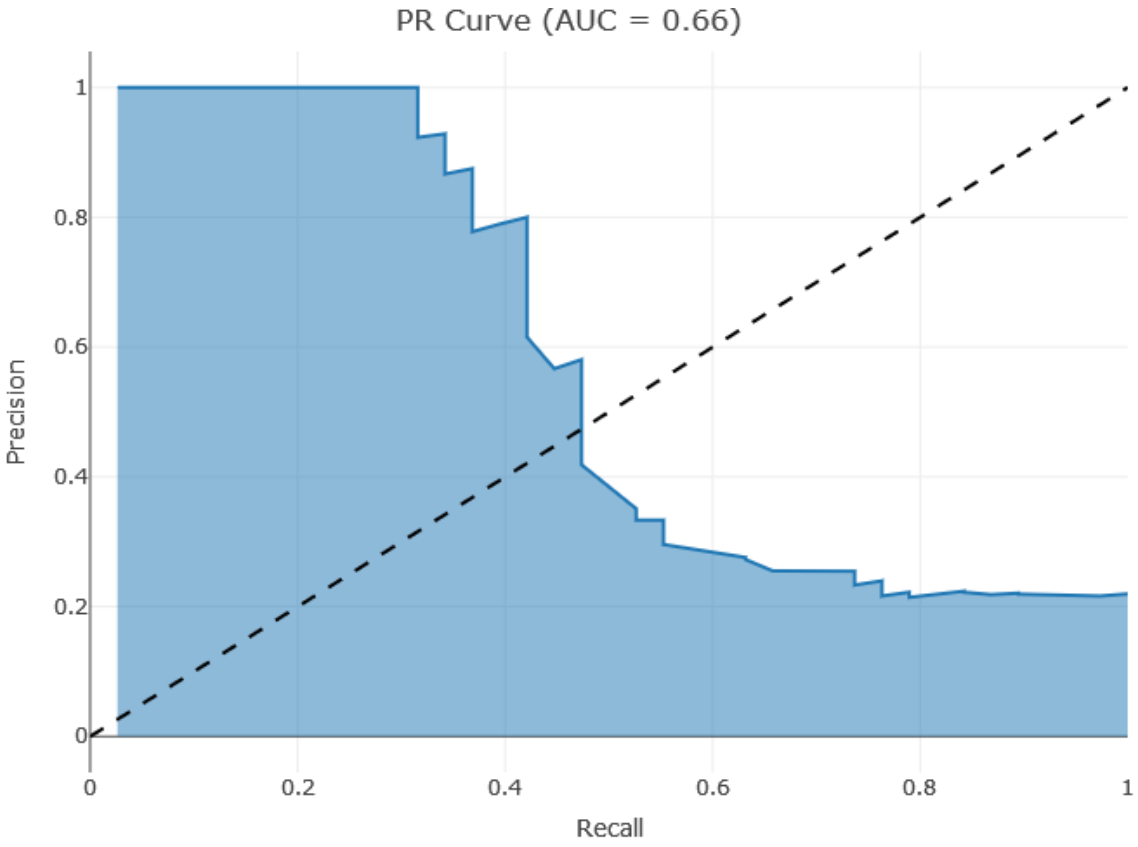
《机器学习》课程第 2 次作业

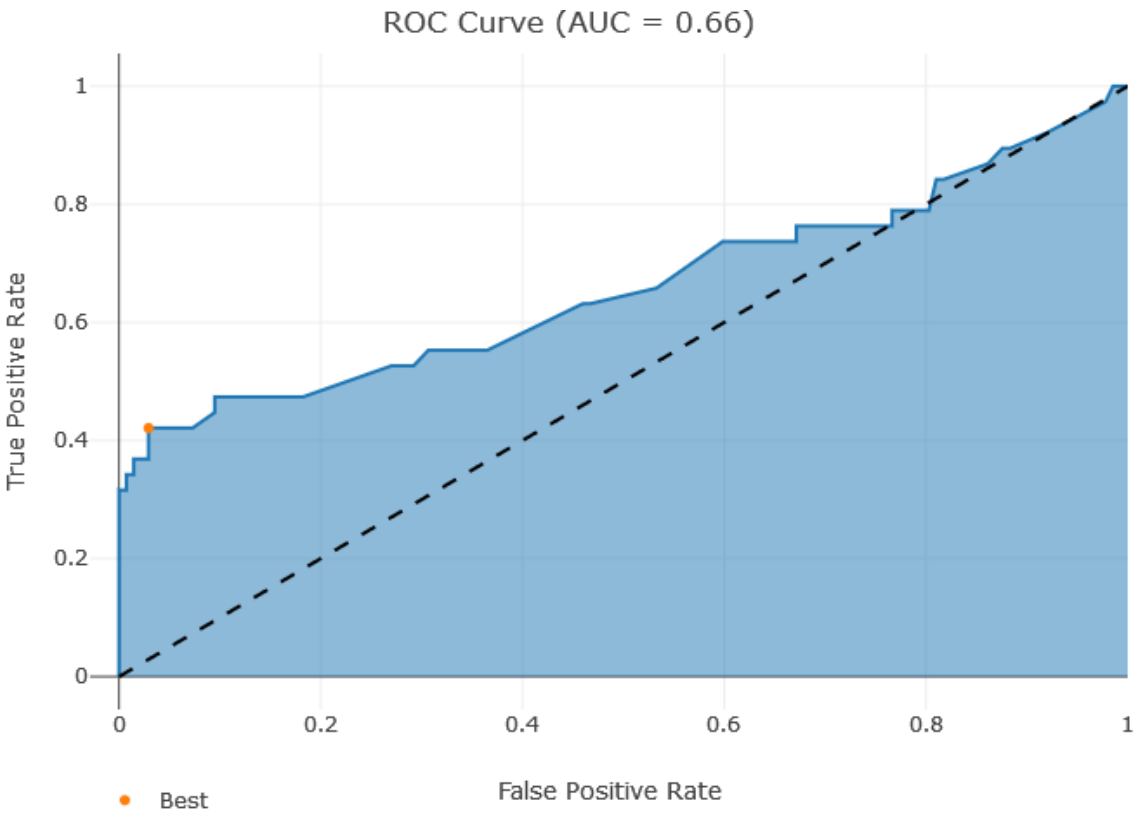
姓名：刘哲　　学号：2022103691

1　钓鱼问题

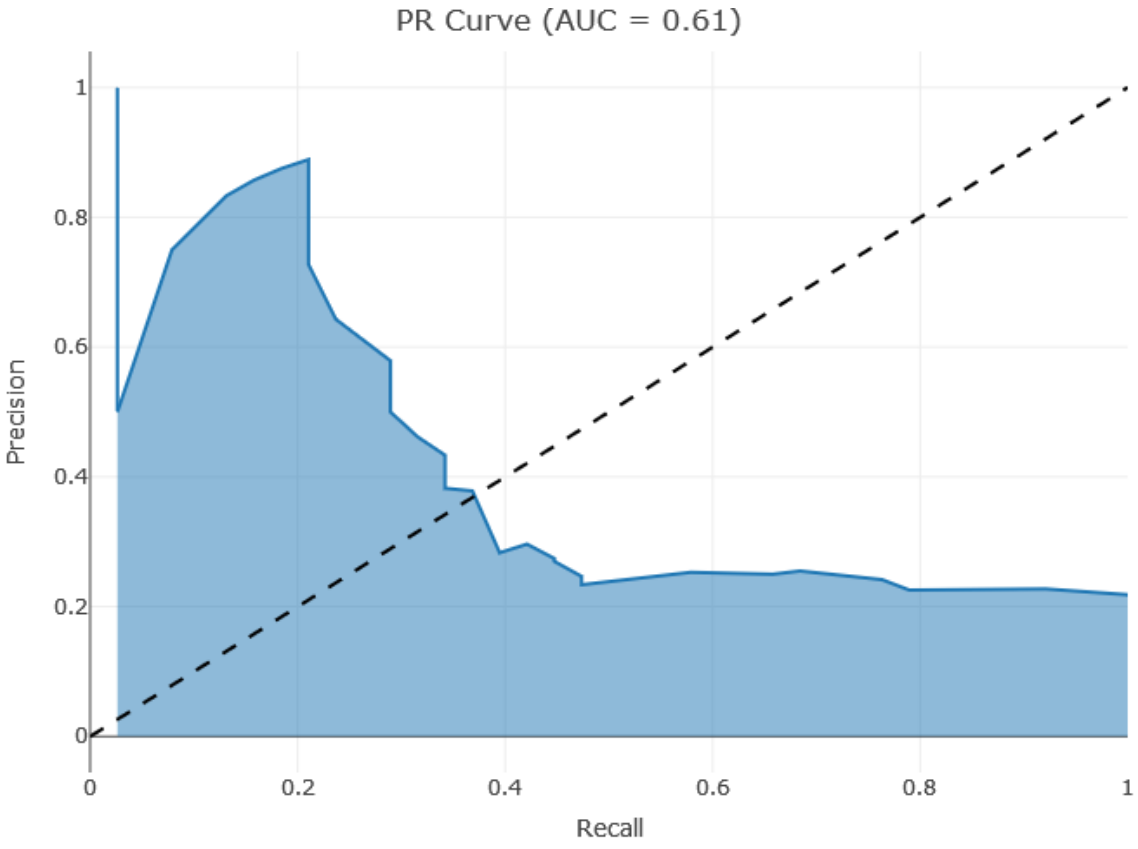
A

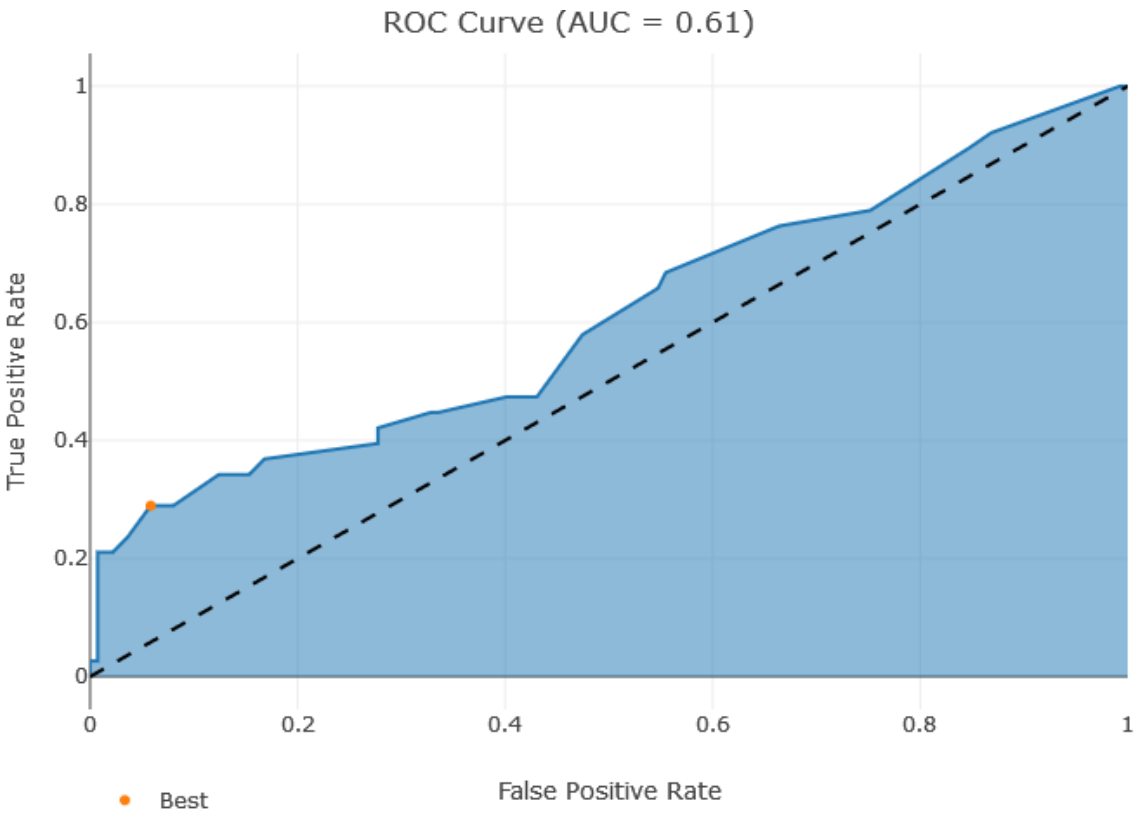
训练数据模型1的P-R图和RUC曲线图：



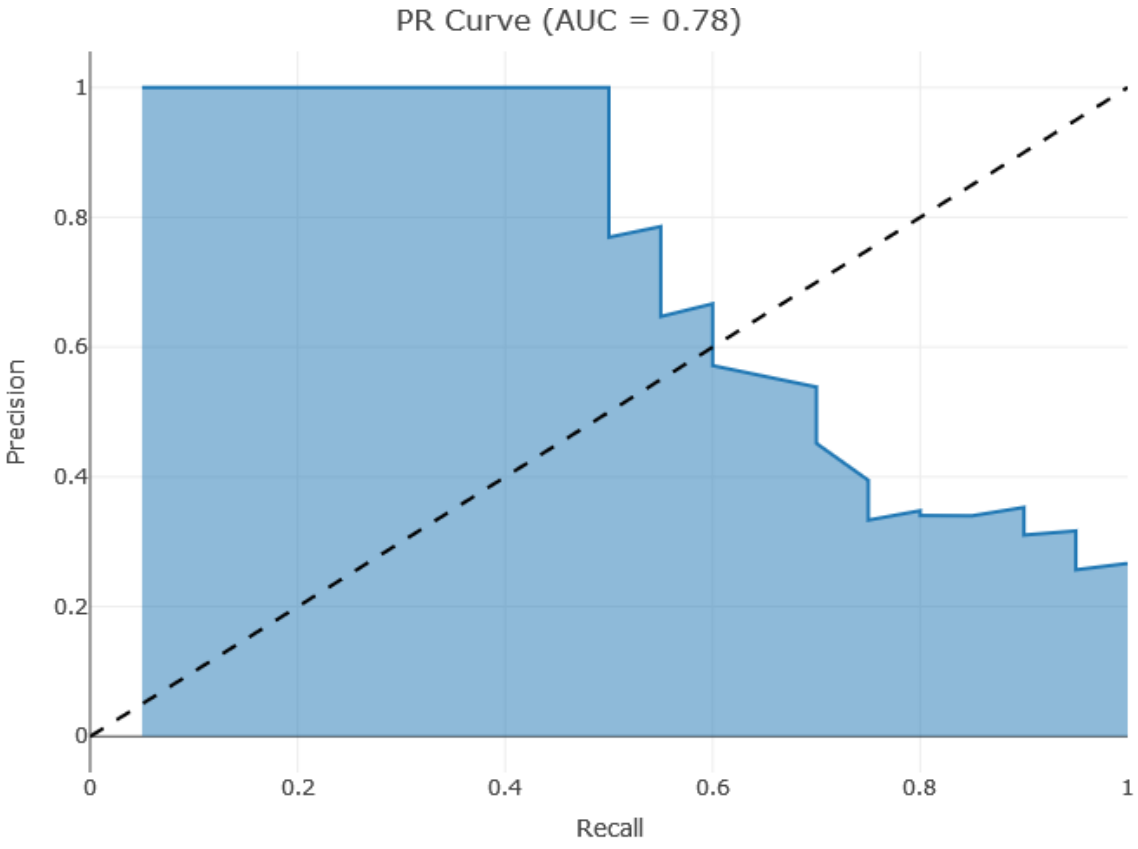


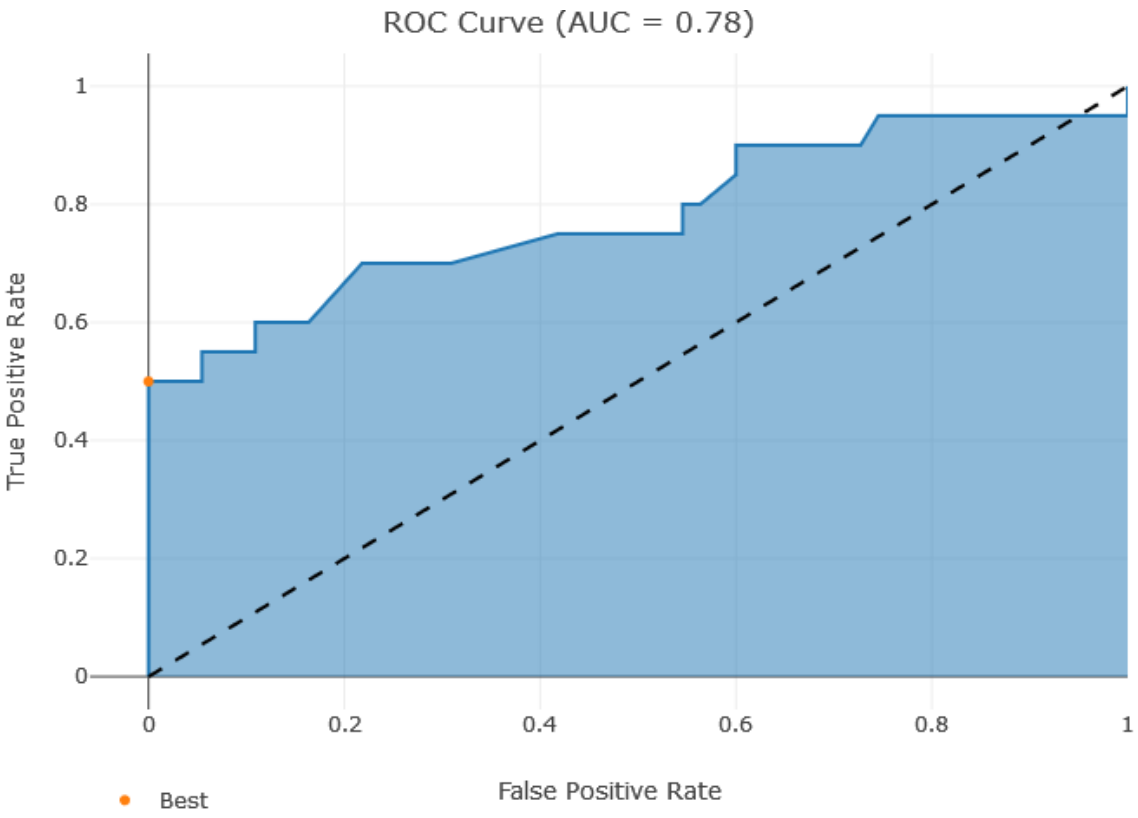
训练数据模型2的P-R图和RUC曲线图：



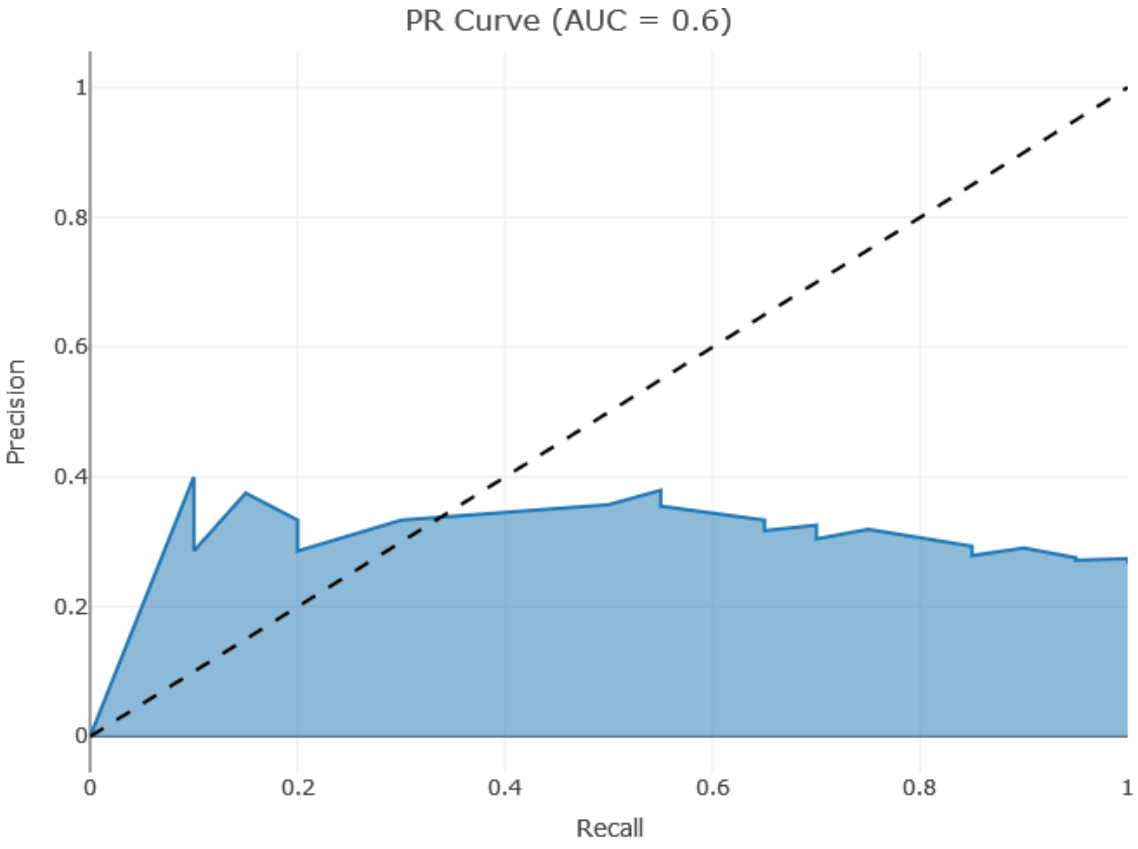


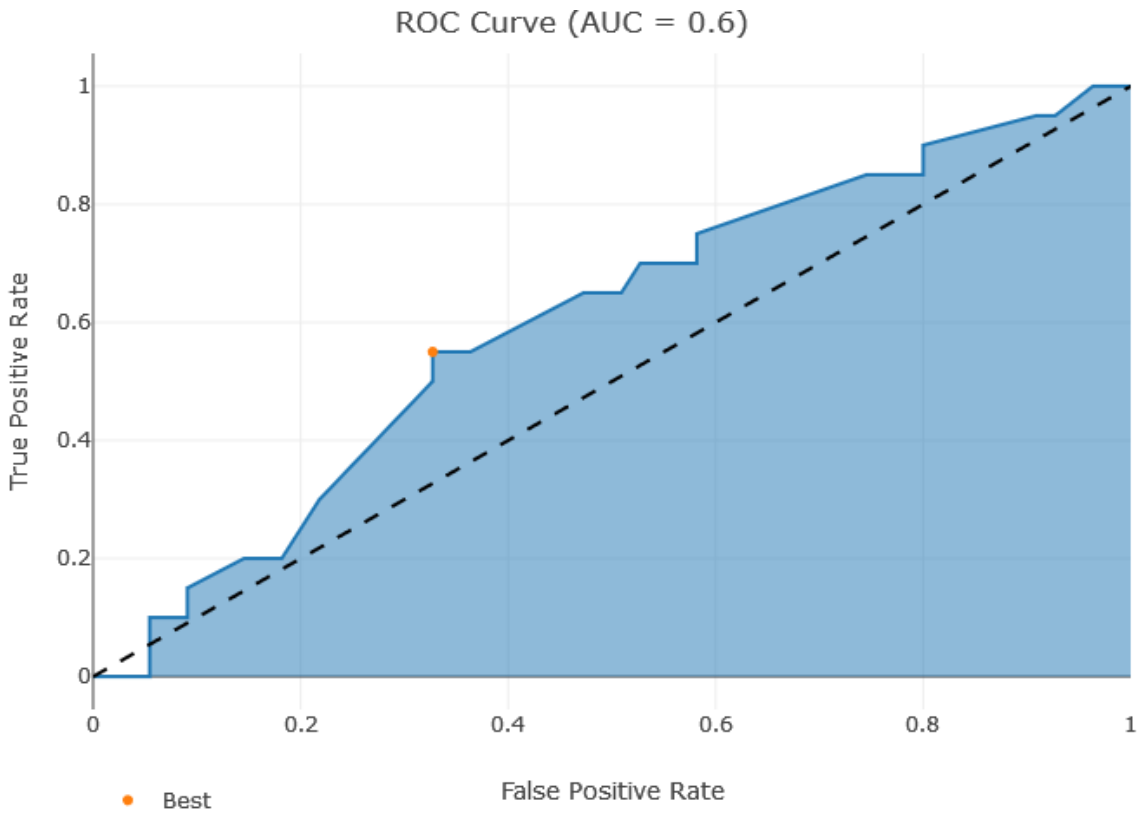
测试数据模型1的P-R图和RUC曲线图：





测试数据模型2的P-R图和RUC曲线图：





各个模型的AUC如图所示，因为模型1在测试数据上具有最高的AUC=0.78，所以认为模型1的整体预测效果最佳。

选取预测阈值依据的是Youden指数：

$$Youden's\ Index = sensitivity + specificity - 1 \tag{1}$$

Youden指数最大的点，即为最佳阈值点。由此可知，模型1的最佳阈值为3.3354，模型2的最佳阈值为1.13828。

B

根据A中给出的各模型的阈值，计算模型1的混淆矩阵为

	Truth 0	Truth 1
Prediction 0	133	22
Prediction 1	4	16

模型2的混淆矩阵为

	Truth 0	Truth 1
Prediction 0	129	27
Prediction 1	8	11

2 似然比方法

论文主要介绍了特征变量型似然比和评分法似然比两种似然比的计算方法。假设有两组数据$\boldsymbol{y}_1$和$\boldsymbol{y}_2$，要检验它们是否属于同一个的数据源，提出假设检验如下：

$$H_p:\boldsymbol{y}_1和\boldsymbol{y}_2同源\ vs\ H_d:\boldsymbol{y}_1和\boldsymbol{y}_2不同源 \tag{2}$$

首先，假设背景数据$\boldsymbol{x}$不变，特征变量型似然比的计算公式为

$$LR=\frac{f\left(\bar{\boldsymbol{y}}_1,\bar{\boldsymbol{y}}_2|\boldsymbol{U},\boldsymbol{C},\bar{\boldsymbol{x}},H_p\right)}{f\left(\bar{\boldsymbol{y}}_1,\bar{\boldsymbol{y}}_2|\boldsymbol{U},\boldsymbol{C},\bar{\boldsymbol{x}},H_d\right)} \tag{3}$$

如果两组数据$\boldsymbol{y}_1$和$\boldsymbol{y}_2$同源，则它们的均值之差$(\bar{\boldsymbol{y}}_1-\bar{\boldsymbol{y}}_2)$应该服从正态分布，且它们对背景数据$\boldsymbol{x}$具有相同的影响，即

$$\begin{aligned}
&f\left(\bar{\boldsymbol{y}}_1,\bar{\boldsymbol{y}}_2|\boldsymbol{U},\boldsymbol{C},\bar{\boldsymbol{x}},H_p\right)=f\left(\bar{\boldsymbol{y}}_1-\bar{\boldsymbol{y}}_2,\bar{\boldsymbol{y}}|\boldsymbol{U},\boldsymbol{C},\bar{\boldsymbol{x}},H_p\right)\\
&=(2\pi)^{-p}\left|\frac{\boldsymbol{U}}{n_1}+\frac{\boldsymbol{U}}{n_2}\right|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left(\bar{\boldsymbol{y}}_1-\bar{\boldsymbol{y}}_2\right)^T\left(\frac{\boldsymbol{U}}{n_1}+\frac{\boldsymbol{U}}{n_2}\right)^{-1}\left(\bar{\boldsymbol{y}}_1-\bar{\boldsymbol{y}}_2\right)\right\}\\
&\times\left|\frac{\boldsymbol{U}}{n_1+n_2}+\boldsymbol{C}\right|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left(\bar{\boldsymbol{y}}-\bar{\boldsymbol{x}}\right)^T\left(\frac{\boldsymbol{U}}{n_1+n_2}+\boldsymbol{C}\right)^{-1}\left(\bar{\boldsymbol{y}}-\bar{\boldsymbol{x}}\right)\right\}
\end{aligned} \tag{4}$$

如果两组数据$\boldsymbol{y}_1$和$\boldsymbol{y}_2$不同源，则它们的均值之间相互独立，且它们各自独立对背景数据$\boldsymbol{x}$产生影响，即

$$\begin{aligned}
&f\left\{\bar{\boldsymbol{y}}_1,\bar{\boldsymbol{y}}_2|\boldsymbol{U},\boldsymbol{C},\bar{\boldsymbol{x}},H_d\right\}=f\left(\bar{\boldsymbol{y}}_1|\boldsymbol{U},\boldsymbol{C},\bar{\boldsymbol{x}},H_d\right)f\left(\bar{\boldsymbol{y}}_2|\boldsymbol{U},\boldsymbol{C},\bar{\boldsymbol{x}},H_d\right)\\
&=(2\pi)^{-p}\left|\frac{\boldsymbol{U}}{n_1}+\boldsymbol{C}\right|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left(\bar{\boldsymbol{y}}_1-\bar{\boldsymbol{x}}\right)^T\left(\frac{\boldsymbol{U}}{n_1}+\boldsymbol{C}\right)^{-1}\left(\bar{\boldsymbol{y}}_1-\bar{\boldsymbol{x}}\right)\right\}\\
&\times\left|\frac{\boldsymbol{U}}{n_2}+\boldsymbol{C}\right|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left(\bar{\boldsymbol{y}}_2-\bar{\boldsymbol{x}}\right)^T\left(\frac{\boldsymbol{U}}{n_2}+\boldsymbol{C}\right)^{-1}\left(\bar{\boldsymbol{y}}_2-\bar{\boldsymbol{x}}\right)\right\}
\end{aligned} \tag{5}$$

其中$\boldsymbol{y}$、$\boldsymbol{y}_1$、$\boldsymbol{y}_2$和$\boldsymbol{x}$均服从正态假设，且$\boldsymbol{y}$、$\boldsymbol{y}_1$和$\boldsymbol{y}_2$同方差。 $\boldsymbol{U}$表示类内协方差矩阵，$\boldsymbol{C}$表示类间协方差矩阵，由背景数据$\boldsymbol{x}$获得其估计，计算公式为

$$\hat{\boldsymbol{U}}=\frac{\boldsymbol{S}_w}{m\left(n-1\right)} \tag{6}$$

$$\hat{\boldsymbol{C}}=\frac{\boldsymbol{S}^*}{m-1}-\frac{\boldsymbol{S}_w}{nm\left(n-1\right)} \tag{7}$$

其中，$\boldsymbol{S}_w=\sum_{i=1}^{m}\sum_{j=1}^{n}\left(x_{ij}-\bar{x}_i\right)\left(x_{ij}-\bar{x}_i\right)^T$，$\boldsymbol{S}^*=\sum_{i=1}^{m}\left(x_i-\bar{x}\right)\left(x_i-\bar{x}\right)^T$，$m$为类的个数，$n$为样本中每个类的样品数。特征变量型似然比方法以两种不同的前提假设分别计算了联合密度函数，将它们的比值作为似然比统计量。

其次，评分法似然比的计算公式为

$$S\hat{L}R=\frac{f\left(s_p|H_p,I\right)}{f\left(s_d|H_d,I\right)} \tag{8}$$

其中s_p表示同源比对分数（组内），s_d表示非同源比对分数（组间），I为背景信息。在分数为连续型变量时，f表示概率密度函数；在分数为离散型变量时，f表示概率分布函数。对分数的分布进行非参数检验，如果分布为参数分布，可以使用其概率函数的比值作为似然比；如果分布不符合常见的参数分布，则使用合适的非参数分布计算似然比。评分法似然比方法建立了一个评分体系，在这个评分体系下给出数据同源或不同源的比对分数，进而形成两个分数的分布，代表两种假设成立的可能性，使用分数分布的比值作为似然比统计量。

对于似然比检验来说，如果似然比大于一个不小于1的阈值，认为同源假设下的似然函数显著大于不同源假设下的似然函数，则两组数据同源或相关的可能性较大；如果似然比小于一个不大于1的阈值，认为不同源假设下的似然函数显著大于同源假设下的似然函数，则两组数据不同源且相互独立的可能性较大。

研究人员常使用直方图、Tippett图、DET图和ECE图来评价似然比的计算结果，评价指标为识别力和区分力。其中，识别力反映似然比模型区分两种假设对应的似然比数值的能力，区分力反映似然比模型在支持某一假设时的正确程度。直方图是将支持两种假设的对数似然比数值分布以直方图的形式在数轴上表示出来，两组直方在数轴上的重叠程度越小，似然比模型的识别能力越强。 Tippett图是支持两组假设的对数似然比数值的累积分布图，两条累积分布曲线在垂直方向上的分离程度越大，似然比模型的识别能力越强。 DET图展示的是似然比模型的假阳性率与假阴性率之间的相互关系，曲线越靠近坐标轴零点位置，似然比模型的识别能力越强。 ECE图是通过设置罚函数，结合PAV（Pool Adjacent Violators）算法绘制的图形，它是唯一一个能同时反映似然比模型识别力（discrimination）和区分力（calibration）的图形，经过PAV算法的似然比数值曲线越靠近横轴识别力越强，且真实似然比数值曲线与经过PAV算法的似然比数值曲线的距离越小区分力越强。

3 计算似然比

计算似然比的函数见代码。

在计算时发现，公式(7)可能会导致$\hat{\boldsymbol{C}}$非正定，这使得似然比计算公式中的部分行列式为负值，无法求其$-1/2$次方。因此，暂且将行列式取绝对值，得到墨迹鉴定数据的似然比如下：

Object	1	2	3	4	5
LR	5.359×10^{-11}	2.897×10^{-8}	1.154×10^{2}	4.744×10^{-20}	3.316×10^{-9}