5. 某二类数据集的标签分别为p（阳性）和n（阴性）。现有一模型对该数据集样本的阳性预测概率如下：// 如果题中未给排序则先进行排序（依据输出概率）

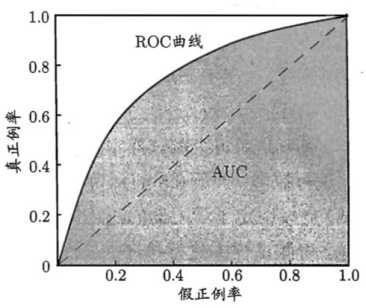
| 样本序号 | 真实标签 | 模型输出概率 | 样本序号 | 真实标签 | 模型输出概率 |
|---|---|---|---|---|---|
| 1 | p | 0.9 | 11 | p | 0.4 |
| 2 | p | 0.8 | 12 | n | 0.39 |
| 3 | n | 0.7 | 13 | p | 0.38 |
| 4 | p | 0.6 | 14 | n | 0.37 |
| 5 | p | 0.55 | 15 | n | 0.36 |
| 6 | p | 0.54 | 16 | n | 0.35 |
| 7 | n | 0.53 | 17 | p | 0.34 |
| 8 | n | 0.52 | 18 | n | 0.33 |
| 9 | p | 0.51 | 19 | p | 0.30 |
| 10 | n | 0.505 | 20 | n | 0.1 |

（1）请绘制此模型的ROC曲线。（15’）

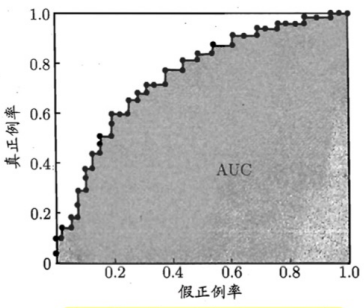
（2）请求出AUC值。（5’）



补充




(a) ROC 曲线与 AUC




(b) 基于有限样例绘制的 ROC 曲线与 AUC

**图 2.4** ROC 曲线与 AUC 示意图

现实任务中通常是利用有限个测试样例来绘制 ROC 图, 此时仅能获得有限个(真正例率, 假正例率)坐标对, 无法产生图 2.4(a)中的光滑 ROC 曲线, 只能绘制出如图 2.4(b)所示的近似 ROC 曲线. 绘图过程很简单: 给定 $m^+$ 个正例和 $m^-$ 个反例, 根据学习器预测结果对样例进行排序, 然后把分类阈值设为最大, 即把所有样例均预测为反例, 此时真正例率和假正例率均为 0, 在坐标 $(0,0)$ 处标记一个点. 然后, 将分类阈值依次设为每个样例的预测值, 即依次将每个样例划分为正例. 设前一个标记点坐标为 $(x,y)$, 当前若为真正例, 则对应标记点的坐标为 $(x, y+\frac{1}{m^+})$; 当前若为假正例, 则对应标记点的坐标为 $(x+\frac{1}{m^-}, y)$, 然后用线段连接相邻点即得.

从定义可知, AUC 可通过对 ROC 曲线下各部分的面积求和而得. 假定 ROC 曲线是由坐标为 $\{(x_1,y_1),(x_2,y_2),\ldots,(x_m,y_m)\}$ 的点按序连接而形成$(x_1=0, x_m=1)$, 参见图 2.4(b), 则 AUC 可估算为

$$\mathrm{AUC}=\frac{1}{2}\sum_{i=1}^{m-1}(x_{i+1}-x_i)\cdot(y_i+y_{i+1})$$

| 真实情况 | 预测结果 | | 含义 | 统计量 |
|---|---|---|---|---|
| | 正例 | 反例 | | |
| 正例 | TP(将正例正确预测为正例) | FN(将正例错误预测为负例) | TP + FN 表示实际数据集中正样本的数量 | 召回率Recall / 灵敏度Sensitivity / TPR = TP/(TP+FN)， 漏诊率 = 1 - 灵敏度 |
| 反例 | FP(将负例错误的预测为正例) | TN(将负例正确的预测为负例) | FP + TN 表示实际数据集中负样本的数量 | FPR = FP/(FP+TN), 特异度(Specificity) = 1 - FPR = TN/(FP+TN) |
| 加和含义 | TP + FP 表示预测的正类样本数 | FN + TN 表示预测的负类样本数 | TP + FN + FP + TN 表示样本总数 | |
| 统计量 | 精确率Precision = TP/(TP+FP) | | | 正确率Accuracy = (TP+TN)/(TP+TN+FP+FN), 错误率 = (FP+FN)/(TP+TN+FP+FN), F-measure = 2*(Precision*Recall)/(Precision+Recall) |



ROC曲线的绘制步骤如下：

1,假设已经得出一系列样本被划分为正类的概率Score值，按照大小排序。

2,从高到低，依次将“Score”值作为阈值threshold，当测试样本属于正样本的概率大于或等于这个threshold时，我们认为它为正样本，否则为负样本。举例来说，对于某个样本，其“Score”值为0.6，那么“Score”值大于等于0.6的样本都被认为是正样本，而其他样本则都认为是负样本。

3,每次选取一个不同的threshold，得到一组FPR和TPR，以FPR值为横坐标和TPR值为纵坐标，即ROC曲线上的一点。

4,根据3中的每个坐标点，画图。

首先了解一下ROC曲线图上很重要的四个点：

第一个点( 0 , 1 ) (0,1)(0,1)，即FPR=0, TPR=1，这意味着FN（False Negative）=0，并且FP（False Positive）=0。意味着这是一个完美的分类器，它将所有的样本都正确分类。
第二个点( 1 , 0 ) (1,0)(1,0)，即FPR=1，TPR=0，意味着这是一个糟糕的分类器，因为它成功避开了所有的正确答案。
第三个点( 0 , 0 ) (0,0)(0,0)，即FPR=TPR=0，即FP（False Positive）=TP（True Positive）=0，可以发现该分类器预测所有的样本都为负样本（Negative）。
第四个点( 1 , 1 ) (1,1)(1,1)，即FPR=TPR=1，分类器实际上预测所有的样本都为正样本。

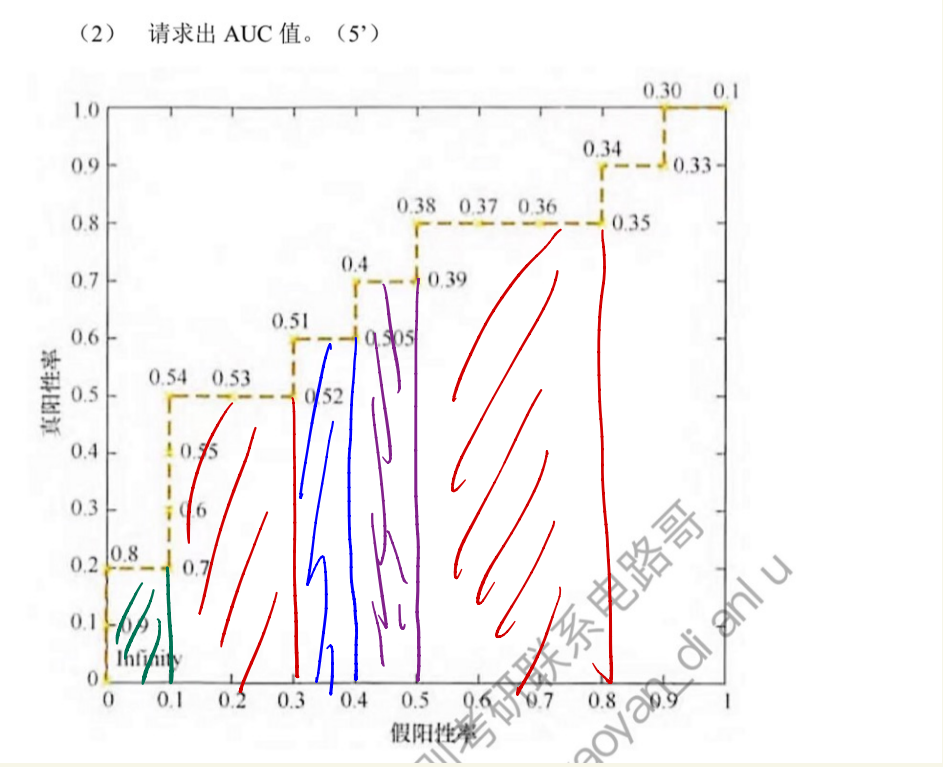
从上面给出的四个点可以发现，ROC曲线图中，越靠近(0,1)的点对应的模型分类性能越好，所以，可以确定的是ROC曲线图中的点对应的模型，它们的不同之处仅仅是在分类时选用的阈值(Threshold)不同，每个点所选用的阈值都对应某个样本被预测为正类的概率值

解：依次选用各样本被预测为正类的概率值作为阈值，设为 $\lambda$

依据 $\lambda$ 依次计算 ROC 曲线的横纵坐标 $(1-Sp, Sn)$

| 则有 $\lambda$ | $(x, y)$ $(1-Sp, Sn)$ |
|---|---|
| 0.9 | (0, 0.1) |
| 0.8 | (0, 0.2) |
| 0.7 | (0.1, 0.2) |
| 0.6 | (0.1, 0.3) |
| 0.55 | (0.1, 0.4) |
| 0.54 | (0.1, 0.5) |
| 0.53 | (0.2, 0.5) |
| 0.52 | (0.3, 0.5) |
| 0.51 | (0.3, 0.6) |
| 0.505 | (0.4, 0.6) |
| 0.4 | (0.4, 0.7) |
| 0.39 | (0.5, 0.7) |
| 0.38 | (0.5, 0.8) |
| 0.37 | (0.6, 0.8) |
| 0.36 | (0.7, 0.8) |
| 0.35 | (0.8, 0.8) |
| 0.34 | (0.8, 0.9) |
| 0.33 | (0.9, 0.9) |
| 0.30 | (0.9, 1) |
| 0.1 | (1, 1) |

把 p 排序后，每更新一个阈值，则点会连续移动




AUC 由面积求得

$AUC = 0.1\times0.2 + 0.2\cdot0.5 + 0.1\cdot0.6 + 0.1\cdot0.7 + 0.3\cdot0.8 + 0.09 + 0.1$

$= 0.02 + 0.1 + 0.06 + 0.07 + 0.24 + 0.19$

$= 0.12 + 0.13 + 0.24 + 0.19$

$= 0.25 + 0.24 + 0.19 = 0.68$