姓名：曾闻天　学号：231880147　2025 年 9 月 26 日

一. (30 points) 性能度量

学习器 $\mathcal{L}$ 在某个多分类任务数据集上的预测混淆矩阵如表 1 所示，请回答下列问题。本题的答案请以分式或者小数点后两位的形式给出，比如 P=0.67.

真实情况	预测结果		
	第 1 类	第 2 类	第 3 类
第 1 类	8	0	4
第 2 类	2	6	4
第 3 类	2	2	8

表 1: 学习器 $\mathcal{L}$ 在某个多分类任务数据集上的预测混淆矩阵

1. 该学习器的预测准确率是多少? (5 points)

2. 计算该学习器的微查准率 (micro-P)、宏查准率 (macro-P)、微查全率 (micro-R) 和宏查全率 (macro-R)。(10 points)

3. 计算该学习器的微 F1 (micro-F1) 和宏 F1 (macro-F1)。(5 points)

4. 在多分类中，每个样例只有一个标签。而在多标签分类中，每个样例可以有多个标签。如表 2 所示，一共有 3 个标签，样例 x_1 的标签是 1 和 3，学习器的预测是 1 和 2。请根据教材 2.3.2 查准率、查全率与 F1 章节的描述和表 2 的样例，计算学习器 $\mathcal{L}_1$ 的微查准率、微查全率、微 F1、宏查准率、宏查全率和宏 F1。[提示: 依然是利用各类的混淆矩阵计算微 F1 和宏 F1.] (10 points)

样例	x_1	x_2	x_3	x_4
标签	1,3	1,2	2,3	1,2,3
学习器预测	1,2	2,3	2,3	1,3

表 2: 学习器 $\mathcal{L}_1$ 的样例表

解:

1.

$$\text{正确预测数} = 8+6+8 = 22,$$

$$\text{总样本数} = 8+0+4+2+6+4+2+2+8 = 36,$$

$$\text{Accuracy} = \frac{\text{正确预测数}}{\text{总样本数}} = \frac{8+6+8}{36} = \frac{22}{36} \approx 0.61.$$

2.

先列出各类的 TP, FP, FN：

类别	TP	FP	FN
1	8	4	4
2	6	2	6
3	8	8	4

微平均

$$\text{micro-P} = \frac{\sum \text{TP}}{\sum(\text{TP}+\text{FP})} = \frac{8+6+8}{(8+4)+(6+2)+(8+8)} = \frac{22}{36} \approx 0.61,$$

$$\text{micro-R} = \frac{\sum \text{TP}}{\sum(\text{TP}+\text{FN})} = \frac{8+6+8}{(8+4)+(6+6)+(8+4)} = \frac{22}{36} \approx 0.61.$$

宏平均

$$\text{macro-P} = \frac{1}{3}\left(\frac{8}{12}+\frac{6}{8}+\frac{8}{16}\right) = \frac{1}{3}\cdot\frac{23}{12} \approx 0.64,$$

$$\text{macro-R} = \frac{1}{3}\left(\frac{8}{12}+\frac{6}{12}+\frac{8}{12}\right) = \frac{1}{3}\cdot\frac{22}{12} \approx 0.61.$$

3.

$$F_1 = 2PR/(P+R)$$

$$\text{micro-F}_1 = \frac{2\times\frac{11}{18}\times\frac{11}{18}}{\frac{11}{18}+\frac{11}{18}} = \frac{11}{18} \approx 0.61,$$

$$\text{macro-F}_1 = \frac{2\times\frac{23}{36}\times\frac{11}{18}}{\frac{23}{36}+\frac{11}{18}} = \frac{253}{405} \approx 0.62.$$

4.

构建逐标签混淆矩阵对每个标签 $k \in \{1, 2, 3\}$ 统计 TP、FP、FN。

标签 1

$$\mathrm{TP}_1 = 2, \mathrm{FP}_1 = 0, \mathrm{FN}_1 = 1.$$

标签 2

$$\mathrm{TP}_2 = 2, \mathrm{FP}_2 = 1, \mathrm{FN}_2 = 1.$$

标签 3

$$\mathrm{TP}_3 = 2, \mathrm{FP}_3 = 1, \mathrm{FN}_3 = 1.$$

标签	TP	FP	FN
1	2	0	1
2	2	1	1
3	2	1	1

微平均

$$\sum \mathrm{TP} = 6, \sum \mathrm{FP} = 2, \sum \mathrm{FN} = 3.$$

$$\text{micro-P} = \frac{6}{6+2} = \frac{6}{8} = 0.75,$$

$$\text{micro-R} = \frac{6}{6+3} = \frac{6}{9} \approx 0.67.$$

$$\text{micro-F}_1 = \frac{2 \times 0.75 \times \frac{2}{3}}{0.75 + \frac{2}{3}} \approx 0.71.$$

宏平均

逐标签计算 P_k, R_k 后平均：

标签	P_k	R_k
1	1	2/3
2	2/3	2/3
3	2/3	2/3

$$\text{macro-P} = \frac{1}{3}(1 + \frac{2}{3} + \frac{2}{3}) \approx 0.78,$$

$$\text{macro-R} = \frac{1}{3}(\frac{2}{3} + \frac{2}{3} + \frac{2}{3}) \approx 0.67,$$

$$\text{macro-F}_1 = \frac{2 \times \frac{7}{9} \times \frac{2}{3}}{\frac{7}{9} + \frac{2}{3}} \approx 0.72.$$

二. (30 points) 性能度量

假设数据集包含 10 个样例, 其对应的真实标签和学习器的输出值 (从大到小排列) 如表 3 所示。该任务是一个二分类任务, 标签 1 或 0 表示真实标签为正例或负例。学习器的输出值代表学习器对该样例是正例的置信度 (认为该样例是正例的概率).

样例	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
标签	1	1	0	0	1	0	1	0	0	0
学习器输出值	0.9	0.75	0.62	0.55	0.49	0.4	0.31	0.28	0.2	0.1

表 3: 样例表

1. 计算 P-R 曲线每一个端点的坐标并绘图。(10 points)

2. 计算 ROC 曲线每一个端点的坐标并绘图。(10 points)

3. 基于上一问，计算 AUC 的值。注：AUC 值请以小数点后两位的形式给出。(4 points)

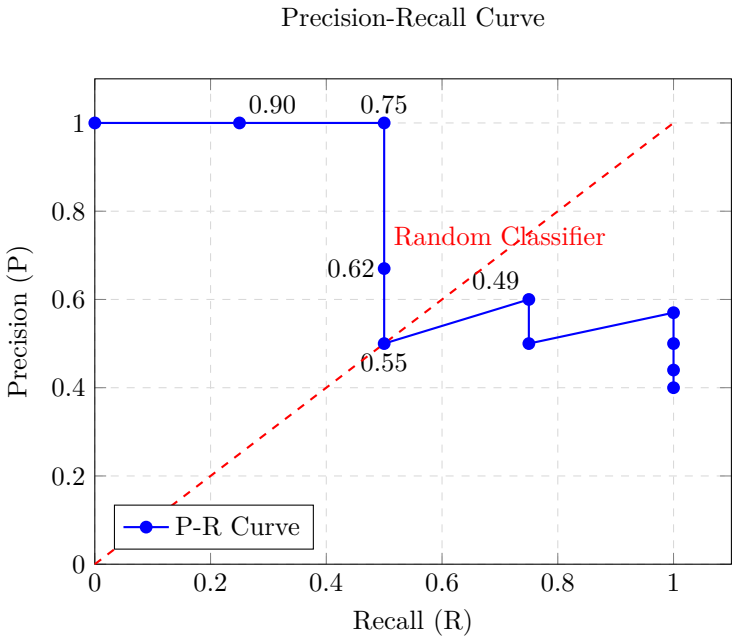
4. FPR95 是一个常见的性能度量指标, 它指的是当真正例率 (true positive rate) 为 95% 时, 假正例率 (false positive rate) 的数值。请问该指标越高学习性能越好还是越低性能越好，并且求解 FPR75 为多少。[提示: FPR75 和 FPR95 类似，FPR75 是真正例率为 75%。](6 points)

解:

1.

以每个不同输出值为阈值，计算对应的查准率 P 与查全率 R。(正例总数 $\text{Pos} = 4$)

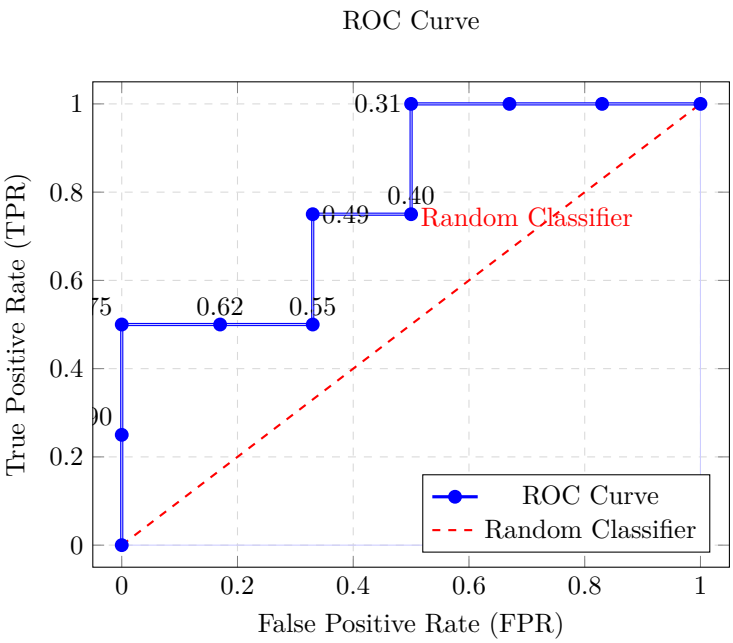
阈值	P	R
∞	$0/0 = 1.00$	$0/4 = 0.00$
0.90	$1/1 = 1.00$	$1/4 = 0.25$
0.75	$2/2 = 1.00$	$2/4 = 0.50$
0.62	$2/3 = 0.67$	$2/4 = 0.50$
0.55	$2/4 = 0.50$	$2/4 = 0.50$
0.49	$3/5 = 0.60$	$3/4 = 0.75$
0.40	$3/6 = 0.50$	$3/4 = 0.75$
0.31	$4/7 = 0.57$	$4/4 = 1.00$
0.28	$4/8 = 0.50$	$4/4 = 1.00$
0.20	$4/9 = 0.44$	$4/4 = 1.00$
0.10	$4/10 = 0.40$	$4/4 = 1.00$



2.

负例总数 Neg = 6。定义 TPR = $\frac{\text{TP}}{4}$, FPR = $\frac{\text{FP}}{6}$。

阈值	TPR	FPR
∞	0.00	0.00
0.90	0.25	0.00
0.75	0.50	0.00
0.62	0.50	$1/6 = 0.17$
0.55	0.50	$2/6 = 0.33$
0.49	0.75	$2/6 = 0.33$
0.40	0.75	$3/6 = 0.50$
0.31	1.00	$3/6 = 0.50$
0.28	1.00	$4/6 = 0.67$
0.20	1.00	$5/6 = 0.83$
0.10	1.00	$6/6 = 1.00$



3.

采用梯形法求 ROC 曲线下面积：

$$
\begin{aligned}
\text{AUC} &= \frac{1}{2}\sum_{i=1}^{10}(\text{TPR}_i + \text{TPR}_{i-1})(\text{FPR}_i - \text{FPR}_{i-1}) \\
&= \frac{1}{2}\Big(0 + 0 + 1.00 \times 0.17 + 1.00 \times 0.16 + 0 + 1.50 \times 0.17 + 0 + 2.00 \times 0.17 + 2.00 \times 0.16 + 2.00 \times 0.17\Big) \\
&= \frac{1}{2}\Big(0.17 + 0.16 + 0.255 + 0.34 + 0.32 + 0.34\Big) \\
&= \frac{1}{2} \times 1.575 \\
&= 0.7875 \\
&\approx 0.79
\end{aligned}
$$

4.

FPR95 越低，学习性能越好。

理由：

FPR95 = FPR | TPR = 0.95

较低的 FPR95 表示模型在检测出 95% 正样本的同时，误报率较低

这反映了模型更好的分类性能和更强的区分能力

根据给定的 ROC 数据：

阈值	TPR	FPR
0.49	0.75	0.33
0.40	0.75	0.50

$$\begin{aligned}\text{FPR75} &= \min\{\text{FPR} \mid \text{TPR} = 0.75\} \\ &= \min\{0.33, 0.50\} \\ &= 0.33\end{aligned}$$

- **FPR95** 和 **FPR75** 都是越低越好，表示模型性能越优
- 在当前数据中，**FPR75** = 0.33

三. (15 points) 评估方法

留出法（hold-out）和交叉验证法（cross validation）是两种常用的评估方法。请回答跟这两种评估方法相关的题目：

(1) 如果不考虑时间开销，哪种评估方法是更稳定和有效的评估方式？(3 points)

(2) 如果采用留出法作为评估方法，产生测试集的过程要特别注意什么？交叉验证法需要注意吗？(6 points)

(3) 请描述留出法和交叉验证法的之间的联系。(6 points)

解:

1.

交叉验证法更稳定和有效。

因为交叉验证（尤其是 k 折交叉验证）通过多次划分训练集和测试集并取平均性能，可以减少因单次划分数据不同而导致的评估偏差，更充分地利用数据，评估结果通常更稳定可靠。留出法只进行一次划分，评估结果受划分方式影响较大。

2.

保持训练/测试集数据分布一致性（例如: 分层采样）

多次随机划分、重复实验取平均值 (例如: 100 次随机划分)

测试集不能太大、不能太小 (例如：1/5 1/3)

3.

根本目的相同：都是用于评估模型泛化性能的基于采样的方法。

交叉验证可看作留出法的推广：留出法是只划分一次（可视为 2 折交叉验证的一次特例），交叉验证是多次划分、多次训练测试并取平均。

两者均需保持数据分布一致性，且测试集（或验证折）不参与训练，遵循相同的数据使用原则。

四. (25 points) 假设检验

在一个二分类任务中，我们使用成对 t 检验（paired t-tests）比较两种学习器 A 和 B 的性能。为此，我们在同一个数据集上进行了 $k = 10$ 折交叉验证，记录了每一折的分类准确率。结果如下表所示：

折数	1	2	3	4	5	6	7	8	9	10
学习器 A	0.82	0.88	0.79	0.91	0.86	0.84	0.90	0.87	0.85	0.89
学习器 B	0.80	0.85	0.78	0.88	0.83	0.82	0.86	0.85	0.83	0.87

1. 在显著性水平 $\alpha = 0.05$ 下，计算出 τ_t。(6 points)

2. 在上一问的基础上，判断学习器 A 是否优于学习器 B。(6 points)

3. 显著水平为 0.01 和 0.1 呢? (5 points)

4. 如果我们只使用前 5 次的结果，而不是完整 10 次结果来进行假设检验。在显著性水平 $\alpha = 0.05$ 下，判断学习器 A 是否优于学习器 B。(8 points)

解:

1.

计算差值 $d_i = A_i - B_i$：

$$d = [0.02, 0.03, 0.01, 0.03, 0.03, 0.02, 0.04, 0.02, 0.02, 0.02]$$

计算均值：

$$\bar{d} = \frac{0.02 + 0.03 + 0.01 + 0.03 + 0.03 + 0.02 + 0.04 + 0.02 + 0.02 + 0.02}{10} = \frac{0.24}{10} = 0.024$$

计算标准差：

$$\begin{aligned} s_d^2 &= \frac{\sum_{i=1}^{10}(d_i - \bar{d})^2}{9} \\ &= \frac{6.4 \times 10^{-4}}{9} \approx 7.11 \times 10^{-5} \end{aligned}$$

计算 t 统计量：

$$\tau_t = \frac{\bar{d}}{s_d/\sqrt{n}} \approx 9.0$$

2.

自由度 $df = 9$，双尾检验临界值 $t_{0.025,9} \approx 2.262$。

由于 $|8.891| > 2.262$，拒绝原假设 $H_0 : \mu_d = 0$。

均值差 $\bar{d} = 0.024 > 0$，因此学习器 A 优于 B。

3. 显著性水平 $\alpha = 0.01$ 和 $\alpha = 0.1$ 的情况

- $\alpha = 0.01$：$t_{0.005,9} \approx 3.250$，$8.891 > 3.250$，显著
- $\alpha = 0.1$：$t_{0.05,9} \approx 1.833$，$8.891 > 1.833$，显著

两种显著性水平下均拒绝原假设，学习器 A 优于 B。

4. 仅使用前 5 次结果 ($\alpha = 0.05$)

前 5 次差值：

$$d = [0.02, 0.03, 0.01, 0.03, 0.03]$$

计算：

$$\bar{d} = \frac{0.12}{5} = 0.024$$

$$s_d^2 = \frac{0.00032}{4} = 0.00008, \quad s_d \approx 0.008944$$

$$\tau_t = \frac{0.024}{0.008944/\sqrt{5}} \approx \frac{0.024}{0.004} = 6.0$$

自由度 $df = 4$，$t_{0.025,4} \approx 2.776$，$6.0 > 2.776$，显著。