机器学习

作业一

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

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)

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)

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)