

2024秋 人工智能导论 第二次作业

- 作业完成后，请将电子版通过邮件发送至课程邮箱：hepengustc@163.com (文件命名为：**学号-姓名**，不备注学号姓名将影响作业上交情况的统计)
- 截至时间：**2024年12月10日 23:59**，**无需上交纸质版作业**

1、准备采用两套系统做患病辅助诊断（判断是否患病），假设共有 100 名受测者，其中 80 名是正常人，20 名是某病患者。

系统 1 检测结果：检出 82 正常人，其中 78 名为正常人，4 名为患者

系统 2 检测结果：检出 88 正常人，其中 80 名正常人，8 名患者

以正常人为正类：

- (1) 计算系统 1 的查准率，查全率，F1；
- (2) 计算系统 2 的查准率，查全率，F1；
- (3) 请讨论：这两个系统哪个更优秀？

2、在一组测试中，根据每个测试样本属于正样本的概率值从大到小排序，表中共有20个样本，“CLASS”表示样本真正的标签（P为正样本，N为负样本），SCORE表示每个样本属于正样本的概率。当测试样本的概率大于或等于阈值时我们认为他为正样本，否则为负样本。

Inst#	Class	Score
1	p	.9
2	p	.8
3	n	.7
4	p	.6
5	p	.55
6	p	.54
7	n	.53
8	n	.52
9	p	.51
10	n	.505

- (1) 当选取阈值为 0.7 时，列出混淆矩阵，计算此时对应ROC曲线的点的坐标；
- (2) 当选取阈值为 0.55 时，列出混淆矩阵，计算此时对应ROC曲线的点的坐标。

3、（1）Logistic 回归是在回归模型中引入 Sigmoid 函数的一种非线性回归模型，对于一个二分类问题，定义线性函数 $z = \omega^T \mathbf{x} + b$ （ \mathbf{x} 为输入数据， ω 和 b 为模型的参数），设输出为 y ，写出 Logistic 回归模型的表达式；

（2）在该二分类问题中，如何理解 y 对应的实际含义；

（3）简要绘制 Sigmoid 函数的示意图，并简述它的数学性质具有哪些主要优势。

4、下表是一个由 15 个样本组成的贷款申请训练数据。数据包括贷款申请人的 4 个特征（属性）：第 1 个特征是年龄，有 3 个可能值：青年，中年，老年；第 2 个特征是有工作，有 2 个可能值：是，否；第 3 个特征是有自己的房子，有 2 个可能值：是，否；第 4 个特征是信贷情况，有 3 个可能值：非常好，好，一般。表的最后一列是类别，是否同意贷款，取 2 个值：是，否。

ID	年龄	有工作	有自己的房子	信贷情况	类别
1	青年	否	否	一般	否
2	青年	否	否	好	否
3	青年	是	否	好	是
4	青年	是	是	一般	是
5	青年	否	否	一般	否
6	中年	否	否	一般	否
7	中年	否	否	好	否
8	中年	是	是	好	是
9	中年	否	是	非常好	是
10	中年	否	是	非常好	是
11	老年	否	是	非常好	是
12	老年	否	是	好	是
13	老年	是	否	好	是
14	老年	是	否	非常好	是
15	老年	否	否	一般	否

- (1) 根据表中所给的训练数据集，应用ID3算法搭建决策树；
- (2) 根据表中所给的训练数据集，应用CART算法搭建决策树。

5、已知一个如下图所示的训练数据集，其正例点为 $x_1 = (4,2)^T$ ，负例点为 $x_2 = (1,1)^T$ ， $x_3 = (2,3)^T$ ，试求：

(1) 请通过转化为对偶问题的思路，求解约束最优问题；

(2) 求解分离超平面和分类决策函数。

