DSCI 6001P 数据科学基础

作业 2 分类方法

提交截止日期：2024.10.20 号（周日）晚上 24 点之前
提交方式：电子版发送至 方羿(peterfang@mail.ustc.edu.cn)
邮件主题和 PDF 命名格式：HW2-姓名-学号，如 HW2-张三-SA24123123

1. 如下表数据，前四列是天气情况（阴晴 outlook，气温 temperature，湿度 humidity，风 windy）；最后一列是类标签，表示根据天气情况是否出去玩。
   (1) “信息熵”是度量样本集合纯度最常用的一种指标，假定当前样本集合 D 中第 k 类样本所占的比例为 p_k（k=1, 2, ⋯, K），请问当什么条件下，D 的信息熵 Ent(D)取得最大，最大值为多少？
   (2) 根据表中训练数据，基于信息增益决策树应该选哪个属性作为第一个分类属性？
   (3) 对于含有连续型属性的样本数据，决策树和朴素贝叶斯分类能有哪些处理方法？
   (4) 在分类算法的评价指标中，recall 和 precision 分别是什么含义？
   (5) 若一批数据中有 3 个属性特征，2 个类标记，则最多可能有多少种不同的决策树？（不同决策树指同一个样本在两个两个决策下可能得到不同的类标记）

outlook	temperature	humidity	windy	play
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
rainy	cool	normal	TRUE	no
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
rainy	mild	normal	FALSE	yes
overcast	cool	normal	TRUE	yes
rainy	cool	normal	FALSE	yes
rainy	mild	high	FALSE	yes
overcast	hot	high	FALSE	yes

2. 已知正例点 x1 = (2，3)T，x2 = (3, 2)T，负例点 x3 = (1, 1)T
   (1) 试用 SVM 对其进行分类，求最大间隔分离超平面，并指出所有的支持向量。
   (2) 现额外有一个点能被 SVM 正确分类且远离决策边界，如果将该点加入到训练集，SVM 的决策边界会受影响吗？为什么？

3. 下表是一个由 15 个贷款申请训练数据，数据包括贷款申请人的四个特征属性：分别是年龄，是否有工作，是否有自己的房子以及信贷情况，表的最后一列为类别，是否同意贷款。

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）请根据上表的训练数据，以错误率作为划分标准来构建预测是否进行放贷的决策树。
2）按照所构建的决策树，对属性值为（中年，无工作，无自己的房子，信贷情况好）的申请者是否进行放贷
3） 在构建决策树的时候，可能会出现过拟合的问题，有哪些方法可以避免或者解决？
4） 对于含有连续型属性的样本数据，决策树有哪些处理方法？

4. 请评价两个分类器 M1 和 M2 的性能。所选择的测试集包含 26 个二值属性，记作 A 到 Z。 表中是模型应用到测试集时得到的后验概率（图中只显示正类的后验概率）。因为这是二类问题，所以 P(-)=1-P(+),P(-|A,⋯,Z)=1-P(+|A,⋯,Z)。假设需要从正类中检测实例
    1) 画出 M1 和 M2 的 ROC 曲线（画在一幅图中）。哪个模型更多？给出理由。
    2) 对模型 M1，假设截止阈值 t=0.5。换句话说，任何后验概率大于 t 的测试实例都被看作正例。计算模型在此阈值下的 precision，recall 和 F-score。
    3) 对模型 M2 使用相同的截止阈值重复（2）的分析。比较两个模型的 F-score，哪个模型更好？所得结果与从 ROC 曲线中得到的结论一致吗？
    4) 使用阈值 t=0.1 对模型 M2 重复（2）的分析。t=0.5 和 t=0.1 哪一个阈值更好？该结果和你从 ROC 曲线中得到的一致吗？

实例	真实类	P(+|A,⋯,Z,M1)	P(-|A,⋯,Z,M2)
1	+	0.73	0.61
2	+	0.69	0.03
3	-	0.44	0.68
4	-	0.55	0.31
5	+	0.67	0.45
6	+	0.47	0.09

7	-	0.08	0.38
8	-	0.15	0.05
9	+	0.45	0.01
10	-	0.35	0.04

5. 下图中数据元组已经按分类器返回概率值的递减顺序排序。对于每个元组，计算真正例（TP）、假正例（FP）、真负例（TN）和假负例（FN）的个数。计算真正例率（TPR）和假正例率（FPR）。为该数据绘制 ROC 曲线。

元组号	类	概率
1	P	0.95
2	N	0.85
3	P	0.78
4	P	0.66
5	N	0.60
6	P	0.55
7	N	0.53
8	N	0.52
9	N	0.51
10	P	0.40

6. 假设两个预测模型 M 和 N 之间进行选择。已经在每个模型上做了 10 轮 10-折交叉验证，其中在第 i 轮，M 和 N 都是用相同的数据划分。M 得到的错误率为 30.5、32.2、20.7、20.6、31.0、41.0、27.7、28.0、21.5、28.0。N 得到的错误率为 22.4、14.5、22.4、19.6、20.7、20.4、22.1、19.4、18.2、35.0。评述在 1%的显著水平上，一个模型是否显著地比另一个好。

7. F 值（F-measure）是准确率和召回率的加权调和平均数，定义如下：

$$F=\frac{1}{\alpha\frac{1}{P}+(1-\alpha)\frac{1}{R}}=\frac{(\beta^2+1)PR}{\beta^2P+R}$$

请推导出β与α的关系，并把 F_β 写成真正例、假负例和假正例的函数。

8. [支持向量机] 在样本空间中，超平面通过如下线性方程来描述：

$$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b=0\ ,$$

请计算样本空间中任意点 $\boldsymbol{x}_0$ 到超平面的距离。

9. [最近邻算法] 根据下表中给定的数据集用 3-近邻对数据点 x1=4, x2=1 进行分类，其中 x1 和 x2 为属性特征，y 为标签。

X1	1	2	3	3	5	4	1	2	4	5
X2	5	3	1	4	1	2	2	4	3	5
Y	-	+	-	+	-	+	-	-	+	+

10. [支持向量机] 已知一个训练数据集如下，两个正例点：x1= (3 , 3)T, x2 = (4 , 3)T , 一个负例点: x3 = (1 , 1)T 试求最大间隔分离超平面（在二维空间即为一条直线）。