截止时间： 2024.12.29 23:59

1. 下表的数据集包含两个属性 X 和 Y，两个类标签“+”和“−”。每个属性取 3 个不同的值：0，1 或 2。

X	Y	实例数	
		+	−
0	0	0	100
1	0	0	0
2	0	0	100
0	1	10	100
1	1	10	0
2	1	10	100
0	2	0	100
1	2	0	0
2	2	0	100

(a) 使用分类误差建立该数据集的决策树。

(b) 决策树的准确率、精度、召回率和 F1 度量各是多少?(注意，精度、召回率和 F1 度量均是对“+”类定义的。)要给出混淆矩阵和具体的计算过程。

2. 考虑下表中的数据集

(a)估计条件概率P(A|+)、P(B|+)、P(C|+)、P(A| −)、P(B|−)和 P(C|−)。

(b)根据(a)中的条件概率，使用朴素贝叶斯方法预测测试样本(A=0，B=1，C=0)的类标签。

样本	A	B	C	类标号
1	0	0	0	+
2	0	0	1	−
3	0	1	1	−
4	0	1	1	−
5	0	0	1	+
6	1	0	1	+
7	1	0	1	−
8	1	0	1	−
9	1	1	1	+
10	1	0	1	+

3. 某产品的广告费支出x(单位:百万元) 与销售额y(单位: 百万元)之间有如下数据：

(1)使用最小二乘法求y关于x的线性回归方程。需要给出具体的计算过程。

(2)预测广告费为 9 百万元时的销售额是多少。

x	2	4	5	6	8
y	30	40	60	50	70

4. 设有如下所示交易数据库：

TID	购买商品
T1	{a,b}
T2	{b,c,d}
T3	{c}
T4	{b,d}

（1） 令 min_sup = 0.5，试用 Apriori 算法求出其所有的频繁项集。需要给出具体过程。

（2） 令 min_conf = 0.7，在所得频繁项集的基础上，求出所有的强关联规则。

（3） 为每一条强关联规则，绘制出对应的列联表，并计算对应的提升度。

5. 一维点的集合是:{6，12，18，24，30，42，48}，执行 K 均值算法

(a)对于下列每组初始质心，将每个点指派到最近的质心，创建两个簇，然后对两个簇的每组质心分别计算总平方误差。对每组质心，给出这两个簇和总平方误差

i.(18，45)

ii.(15，40)

(b)两组质心代表稳定解吗，即如果在该数据集上，使用给定的质心作为初始质心运行 K 均值，所产生的簇会有改变吗?