一、（30 分，总共 30 题，每题答对得 1 分，答错得 0 分）单选题

1、当不知道数据所带标签时，可以使用哪种技术促使带同类标签的数据与带其他标签的数据相分离（ A ）

A、聚类； B、关联分析； C、分类； D、隐马尔科夫

2、朴素贝叶斯是一种特殊的贝叶斯分类器，特征变量是 X，类别标签是 C，它的一个假定是：（ C ）

A、各类别的先验概率 P(C)是相等的

B、以 0 为均值，sqr(2)/2 为标准差的正态分布

C、特征变量 X 的各个维度是类别条件独立随机变量

D、P(X|C)是高斯分布

3、下列说法错误的是 （ A ）

A. K-means 算法能够解决有离群点的聚类问题

B. K-modes 能够解决离散数据的聚类问题

C. K-means++能够解决初始点影响聚类效果的问题

D. K 中心点能够解决有离群点的聚类问题

4、只有非零值才重要的二元属性被称作：（ C ）

A、计数属性 B、离散属性

C、非对称的二元属性 D、对称属性

5、以下哪些方法不可以直接来对文本分类（ A ）

A、Kmeans B、决策树 C、支持向量机 D、KNN

6、在 logistic 分类中，L1 正则化和 L2 正则化的引入为了解决什么问题？（ C ）

A、数据量不充分 B、训练数据不匹配

C、训练过拟合 D、训练速度太慢

7、标称类型数据的可以实现数学计算（ A ）

A、众数 B、中位数 C、均值 D、方差

8、对于数据组：200,300,500,700,1000，使用最小-最大规范化，将数据规约到区间[5,10]，其中数据 500 将变换为（ C ）

A、7.375 B、5.52 C、6.875 D、7

9、主成分分析（PCA）中各因子的关系是（ A ）

A、互相独立 B、线性相关 C、非线性相关 D、都有可能

10、数据科学家可能会同时使用多个算法（模型）进行预测，并且最后把这些算法的结果集成起来进行最后的预测（集成学习），以下对集成学习说法正确的是（ B ）

A、单个模型之间有高相关性

B、单个模型之间有低相关性

C、在集成学习中使用“平均权重”而不是“投票”会比较好

D、单个模型都是用的一个算法

11、训练神经网络时，以下哪种激活函数最容易造成梯度消失：（ B ）

A、tanh　　B、sigmoid　　C、ReLU　　D、leaky ReLU

12、在训练Lasso回归模型时，训练数据集有N个特征（X1，X2，…，XN）。在模型调优阶段的数据预处理时，无意将某个特征XX扩大了20倍，然后用相同的正则化参数对 Lasso 回归进行修正。那么，下列说法正确的是？（　B　）
A、特征XX很可能被排除在模型之外
B、特征XX很可能还包含在模型之中
C、无法确定特征XX是否被舍弃
D、其他选项说法都不对

13、以下模型中，在数据预处理时，不需要考虑归一化处理的是：（　C　）
A、logistic回归　　B、SVM　　C、树形模型　　D、神经网络

14、关于数据预处理，以下说法错误的是（　B　）
A、可以通过聚类分析方法找出离群点。
B、数据质量的三个基本属性（指标）是：正确性、精确性和完整性。
C、聚类和回归算法可在数据预处理中做数据规约操作。
D、数据集成包括内容集成和结构集成。

15、如果对相同的数据进行逻辑回归，将花费更少的时间，并给出比较相似的精度（也可能不一样），怎么办？（假设在庞大的数据集上使用Logistic回归模型。可能遇到一个问题，Logistic回归需要很长时间才能训练。）（　D　）
A、降低学习率，减少迭代次数　　B、降低学习率，增加迭代次数
C、提高学习率，增加迭代次数　　D、增加学习率，减少迭代次数

16、小明想使用逻辑回归预测用户点击某广告可能性，使用了变量x1，x2作为输入特征，两个变量量纲差异巨大，且x1本身呈双峰分布，两个分布中心数值差异巨大，请问小明应该怎么做特征工程（　D　）
A、对x1，x2做min-max归一化
B、对x1做z-score归一化，对x2做min-max归一化
C、对x1，x2做z-score归一化
D、以上皆不对

17、关于逻辑回归和SVM算法，说法不正确的是（　A　）
A、逻辑回归的目标是最小化后验概率
B、逻辑回归可以用于预测事件发生概率的大小
C、SVM的目标是最小化结构风险
D、SVM可以有效避免模型过拟合

18、以下关于逻辑回归的说法不正确的是？（　C　）
A、逻辑回归必须对缺失值做预处理；
B、逻辑回归要求自变量和目标变量是线性关系；
C、逻辑回归比决策树，更容易过度拟合；
D、逻辑回归只能做2值分类，不能直接做多值分类；

19、有如下 6 条记录的数据集：t1=[O,P,B]，t2=[P,B,M]，t3=[M,A]，t4=[O,P,M]，t5=[O,P,B,A]，t6=[O,P,B,M,]。则支持度大于50%的频繁3项集为（　A　）

A、OPB　　B、OPM　　C、PBM　　D、OBM

20、通常可以通过关联规则挖掘来发现啤酒和尿布的关系，那么如果对于一条规则 A →B, 如果同时购买 A 和 B 的顾客比例是 4/7，而购买 A 的顾客当中也购买了 B 的顾客比例是 1/2，而购买 B 的顾客当中也购买了 A 的顾客比例是 1/3，则以下对于规则 A →B 的支持度(support)和置信度(confidence)分别是多少？（　C　）

A、4/7，1/3　　B、3/7，1/2
C、4/7，1/2　　D、4/7，2/3

21、下面关于关联规则的描述错误的是（　D　）

A、关联规则经典的算法主要有 Apriori 算法和 FP-growth 算法
B、FP-growth 算法主要采取分而治之的策略
C、FP-growth 对不同长度的规则都有很好的适应性
D、Apriori 算法不需要重复的扫描数据库

22、DBSCAN 算法适用于哪种样本集（　C　）

A、凸样本集　　B、非凸样本集　　C、凸样本集与非凸样本集　　D、无法判断

23、在 k-均值算法中，以下哪个选项可用于获得全局最小？（　D　）

A、尝试为不同的质心（centroid）初始化运行算法
B、调整迭代的次数
C、找到集群的最佳数量
D、以上所有

24、两个种子点 A(-1,0)，B(-1,6)，其余点为(0,0)，(2,0)，(0,6)，(2,6)，利用 Kmeans 算法，点群中心按坐标平均计算。最终同类点到种子点 A 和同类点到种子点 B 的距离和分别为（　B　）

A、1，1　　B、2，2　　C、4，4　　D、6，6

25、一般情况下，KNN 最近邻方法在（　D　）情况下效果最好

A、样本呈现团状分布　　B、样本呈现链状分布
C、样本较多但典型性不好　　D、样本较少但典型性好

26、在使用朴素贝叶斯进行文本分类时，待分类语料中，有部分语句中的某些词汇在训练语料中的 A 类中从未出现过，下面哪些解决方式是正确的（　C　）

A、按照贝叶斯公式计算，这些词汇并未在 A 类出现过，那么语句属于 A 类的概率为零。
B、这种稀疏特征属于噪音，它们的加入会严重影响到分类效果，把这类特征从所有类别中删掉。
C、这种特征可能会起到作用，不易简单删掉，使用一些参数平滑方式，使它起到作用。
D、这种稀疏特征出现在的类别，该句更有可能属于该类，应该把特征从它未出现的类别中删掉。

27、下面关于贝叶斯分类器描述错误的是（　B　）

A、以贝叶斯定理为基础
B、是基于后验概率，推导出先验概率
C、可以解决有监督学习的问题
D、可以用极大似然估计法解贝叶斯分类器

28、我们想在大数据集上训练决策树，为了使用较少时间，我们可以（　C　）

A、增加树的深度　　B、增加学习率（learning rate）
C、减少树的深度　　D、减少树的数量

29、在使用数据挖掘解决现实问题时，有时出现分类问题的正负样本集不均衡的现象，在这种情况下，以下哪种指标不合理？（　B　）
A、F-measure　　B、Accuracy　　C、AUC　　D、G-mean

30、神经网络模型是受人脑的结构启发发明的。神经网络模型由很多的神经元组成，每个神经元都接受输入，进行计算并输出结果，那么以下选项描述正确的是（　D　）
A、每个神经元只有一个单一的输入和单一的输出
B、每个神经元有多个输入而只有一个单一的输出
C、每个神经元只有一个单一的输入而有多个输出
D、每个神经元有多个输入和多个输出

二、（20 分，总共 10 题，每题全对得 2 分，漏选得 1 分，错选得 0 分）不定项选择题
1、采用决策树分类算法，连续数据如何处理？（　AB　）
A、连续数据离散化　　B、选择最佳划分点分裂
C、连续数据每 2 个值之间形成分裂　　D、以上均不正确

2、主成分分析（PCA）是一种重要的降维技术，以下对于 PCA 的描述正确的是：（　ABC　）
A、主成分分析是一种无监督方法
B、主成分数量一定小于等于特征的数量
C、各个主成分之间相互正交
D、原始数据在第一主成分上的投影方差最小

3、影响基本 K-均值算法的主要因素有（　ABD　）。
A、样本输入顺序　　B、模式相似性测度
C、聚类准则　　D、初始类中心的选取

4、关于 K 均值和 DBSCAN 的比较，以下说法正确的是（　ABC　）
A、K 均值使用簇的基于原型的概念，而 DBSCAN 使用基于密度的概念
B、K 均值很难处理非球形的簇和不同大小的簇，DBSCAN 可以处理不同大小和不同形状的簇。
C、K 均值可以发现不是明显分离的簇，即便簇有重叠也可以发现，但是 DBSCAN 会合并有重叠的簇
D、K 均值丢弃被它识别为噪声的对象，而 DBSCAN 一般聚类所有对象

5、贝叶斯分类器的训练中，最大似然法估计参数的过程包括以下哪些步骤（　ABCD　）
A、写出似然函数
B、求导数，令偏导数为 0，得到似然方程组
C、对似然函数取对数，并整理
D、解似然方程组，得到所有参数即为所求

6、决策树中属性选择的方法有？（　BCD　）
A、信息值　　B、信息增益
C、信息增益率　　D、GINI 系数

7、在数据挖掘中需要划分数据集，常用的划分测试集和训练集的划分方法有哪些（　ABC　）

A、留出法　　B、交叉验证法　　C、自助法　　D、评分法

8、下列有关机器学习中 L1 正则化和 L2 正则化说法正确的是？（　A D　）
A、使用 L1 可以得到稀疏的权值
B、使用 L2 可以得到稀疏的权值
C、使用 L1 可以得到平滑的权值
D、使用 L2 可以得到平滑的权值

9、下列哪些因素会对 BP 神经网络的训练效果产生影响（　ABCD　）
A、权值初始值　　B、阈值初始值
C、学习率　　D、隐层神经元个数

10、下列关于随机森林和 Adaboost 说法正确的是（　ACD　）
A、和 Adaboost 相比，随机森林对错误和离群点更鲁棒
B、随机森林准确率不依赖于个体分类器的实例和他们之间的依赖性
C、随机森林对每次划分所考虑的属性数很偏感
D、Adaboost 初始时每个训练元组被赋予相等的权重

三、（10 分，总共 10 题，每题答对得 1 分，答错得 0 分）判断题，正确的用“T”，错误的用“F”
1、具有较高的支持度的项集具有较高的置信度。（　错　）

2、利用先验原理可以帮助减少频繁项集产生时需要探查的候选项个数。（　对　）

3、可以利用概率统计方法估计数据的分布参数，再进一步估计待测试数据的概率，以此来实现贝叶斯分类。（　对　）

4、数据库中某属性缺失值比较多时，数据清理可以采用忽略元组的方法。（　错　）

5、K-means++能够解决初始点影响聚类效果的问题。（　对　）

6、逻辑回归等同于一个使用交叉熵 loss，且没有隐藏层的神经网络。（　对　）

7、朴素贝叶斯分类器不存在数据平滑问题。（　错　）

8、逻辑回归分析需要对离散值做预处理，决策树则不需要。（　对　）

9、在 AdaBoost 算法中，所有被分错的样本的权重更新比例相同。（　对　）

10、分类和回归都可用于预测，分类的输出是连续数值，而回归的输出是离散的类别值。（　错　）

四、（10 分）假设正常对象被分类为离群点的概率是 0.01，而离群点被分类为离群点概率为 0.99，如果 99%的对象都是正常的，那么检测率和假警告率各为多少？**（使用下面的定义）**

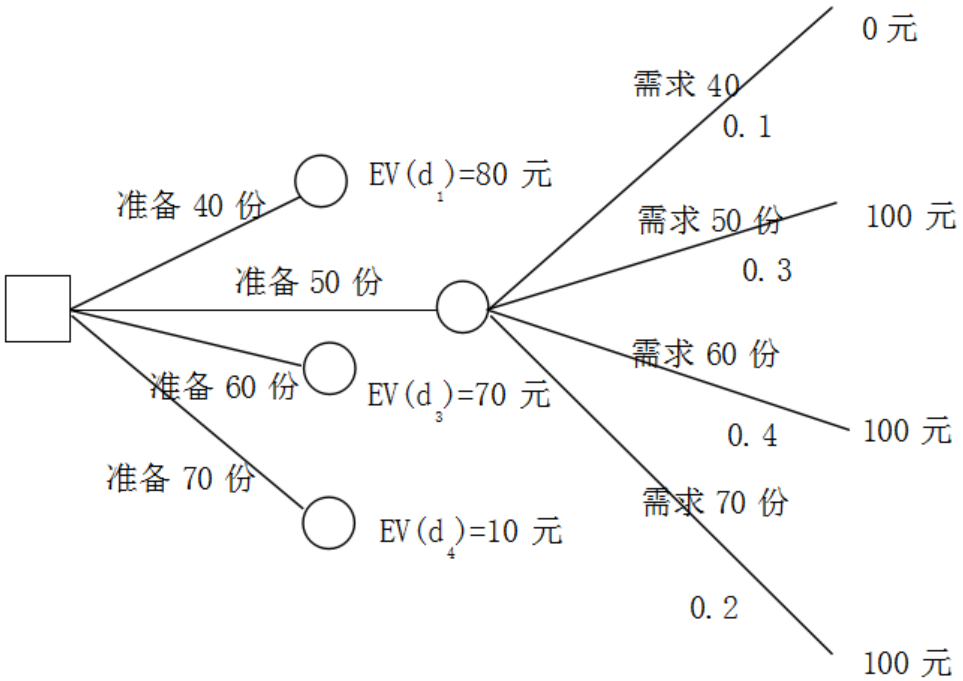
$$检测率=\frac{检测出的离群点个数}{离群点的总数}$$

$$假警告率=\frac{假离群点的个数}{被分类为离群点的个数}$$

五、（10 分）小明开家餐厅卖炒饭，每份炒饭售价 10 元，成本 8 元，每天需要以 10 份为单位提前准备炒饭，按每天可能需求 40，50，60，70 份炒饭做出下方支付矩阵：

| 炒饭的需求量 | | 炒饭的供应量 | | | |
|---|---|---|---|---|---|
| | | 40 份 | 50 份 | 60 份 | 70 份 |
| | 40 份 | 80 元 | 0 元 | -80 元 | -160 元 |
| | 50 份 | 80 元 | 100 元 | 20 元 | -60 元 |
| | 60 份 | 80 元 | 100 元 | 120 元 | 40 元 |
| | 70 份 | 80 元 | 100 元 | 120 元 | 140 元 |

观察发现，每天有 10%概率需求 40 份，30%概率需求 50 份，40%概率需求 40 份，20%概率需求 70 份，做出下方部分决策树：




**请计算准备 50 份炒饭的利润的期望值？**

六、（10 分）从某超市顾客中随机抽取 5 名，他们的购物篮数据的二元 0/1 表示如下：

| | 面包 | 牛奶 | 尿布 | 啤酒 | 鸡蛋 | 可乐 |
|---|---|---|---|---|---|---|

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| 2 | 1 | 0 | 1 | 1 | 1 | 0 |
| 3 | 0 | 1 | 1 | 1 | 0 | 1 |
| 4 | 1 | 1 | 1 | 1 | 0 | 0 |
| 5 | 1 | 1 | 1 | 0 | 0 | 1 |

某学生依据此数据做关联分析，考虑规则{牛奶，尿布}->{啤酒}，**请计算该规则的支持度（support）、置信度（confidence）**。

七、（10 分）下表的数据集包含两个属性 X 与 Y，两个类标号“+”和“-”。每个属性取三个不同值策略：0，1 或 2。“+”类的概念是 Y=1，“-”类的概念是 X=0 and X=2。

| X | Y | 实例数 | |
|---|---|---|---|
| | | + | - |
| 0 | 0 | 0 | 100 |
| 1 | 0 | 0 | 0 |
| 2 | 0 | 0 | 100 |
| 1 | 1 | 10 | 0 |
| 2 | 1 | 10 | 100 |
| 0 | 2 | 0 | 100 |
| 1 | 2 | 0 | 0 |
| 2 | 2 | 0 | 100 |

（1）建立该数据集的决策树。该决策树能捕捉到“+”和“-”的概念吗？（**注意：纯性度量采用 Classification Error，** $Error(t) = 1 - \max_i P(i \mid t)$）

（2）决策树的准确率、精度、召回率和 F1 各是多少？（注意：精度，召回率和 F1 量均是对“+”类的定义）

四、解：

假警告率 =（99%*1%）/(99%*1%+1%*99%)=50%　　5

检测率 =（1%99%）/(1%)=99%　　5

五、解：

需求 40 的利润：0.1*0=0　　2

需求 50 的利润：0.3*100=30　　2
需求 60 的利润：0.4*100=40　　2
需求 70 的利润：0.2*100=20　　2
利润期望值：30+40+20=90　　2
六、解：
支持度：{牛奶，尿布，啤酒}都出现的个数/事务数=2/5　　5
置信度：{牛奶，尿布，啤酒}都出现的个数/{牛奶，尿布}出现的个数=2/3　　5
七、解：

（1）4 在数据集中有 20 个正样本和 500 个负样本，因此在根结点处错误率为

$$E=1-\max\left(\frac{20}{520},\frac{500}{520}\right)=\frac{20}{520}$$ （1 分）

如果按照属性 X 划分，则：

| | X=0 | X=1 | X=2 |
|---|---|---|---|
| + | 0 | 10 | 10 |
| − | 200 | 0 | 300 |

Ex=0=0/310=0
Ex=1=0/10=0
Ex=2=10/310

$$\triangle X = E - \frac{200}{520} * 0 - \frac{10}{520} * 0 - \frac{310}{520} * \frac{10}{310} = \frac{10}{520}$$ （1 分）

如果按照属性 Y 划分，则：

| | Y=0 | Y=1 | Y=2 |
|---|---|---|---|
| + | 0 | 20 | 0 |
| − | 200 | 100 | 200 |

EY=0=0/200=0
EY=1=20/120
EY=2=0/200=0

$$\triangle X = E - \frac{120}{520} * \frac{20}{120} = 0$$ （1 分）

因此 X 被选为第一个分裂属性，因为 X=0 和 X=1 都是纯节点，所以使用 Y 属性去分割不纯节点 X=2。

Y=0 节点包含 100 个负样本，Y=1 节点包含 10 个正样本和 100 个负样本，Y=2 节点 100 个负样本，所以子节点被标记为“_”。整个结果为：**（2 分）**

$$\text{类标记}=\begin{cases}+, X=1 \\ -, \text{其他}\end{cases}$$

（2）**（每个 1 分，总计 5 分）**

| | | 预测类 | |
|---|---|---|---|
| | | + | − |
| 实际类 | + | 10 | 10 |
| | − | 0 | 500 |

accuracy: $\frac{510}{520}$=0.9808，　　precision: $\frac{10}{10}$=1.0

recall: $\frac{10}{20}$=0.5，　　F-measure: $\frac{2*0.5*1.0}{1.0+0.5}$=0.6666