试卷

一、（30 分，总共 30 题，每题答对得 1 分，答错得 0 分）单选题

1、在 ID3 算法中信息增益是指（ D ）

A、信息的溢出程度　　B、信息的增加效益

C、熵增加的程度最大　　D、熵减少的程度最大

2、下面哪种情况不会影响 K-means 聚类的效果？（ B ）

A、数据点密度分布不均

B、数据点呈圆形状分布

C、数据中有异常点存在

D、数据点呈非凸形状分布

3、下列哪个不是数据对象的别名（ C ）

A、样品　B、实例　C、维度　D、元组

4、人从出生到长大的过程中，是如何认识事物的？（ D ）

A、聚类过程

B、分类过程

C、先分类，后聚类

D、先聚类，后分类

5、决策树模型中应如何妥善处理连续型属性：（ C ）

A、直接忽略　　B、利用固定阈值进行离散化

C、根据信息增益选择阈值进行离散化　D、随机选择数据标签发生变化的位置进行离散化

6、假定用于分析的数据包含属性 age。数据元组中 age 的值如下（按递增序）：13，15，16，16，19，20，20，21，22，22，25，25，25，30，33，33，35，35，36，40，45，46，52，70。问题：使用按箱平均值平滑方法对上述数据进行平滑，箱的深度为 3。第二个箱子值为：（ A ）

A、18.3　B、22.6　C、26.8　D、27.9

7、建立一个模型，通过这个模型根据已知的变量值来预测其他某个变量值属于数据挖掘的哪一类任务？（ C ）

A、根据内容检索　　B、建模描述

C、预测建模　　D、寻找模式和规则

8、如果现在需要对一组数据进行样本个体或指标变量按其具有的特性进行分类，寻找合理的度量事物相似性的统计量，应该采取（ A ）

A、聚类分析　　B、回归分析　　C、相关分析　　D、判别分析

9、时间序列数据更适合用（ A ）做数据规约。

A、小波变换　　B、主成分分析　　C、决策树　　D、直方图

10、下面哪些场景合适使用 PCA？（ A ）

A、降低数据的维度，节约内存和存储空间

B、降低数据维度，并作为其它有监督学习的输入

C、获得更多的特征

D、替代线性回归

11、数字图像处理中常使用主成分分析（PCA）来对数据进行降维，下列关于 PCA 算法错误的是：（ C ）
A、PCA 算法是用较少数量的特征对样本进行描述以达到降低特征空间维数的方法；
B、PCA 本质是 KL-变换；
C、PCA 是最小绝对值误差意义下的最优正交变换；
D、PCA 算法通过对协方差矩阵做特征分解获得最优投影子空间，来消除模式特征之间的相关性、突出差异性；

12、将原始数据进行集成、变换、维度规约、数值规约是在以下哪个步骤的任务？（ C ）
A、频繁模式挖掘　　B、分类和预测
C、数据预处理　　D、数据流挖掘

13、假设使用维数降低作为预处理技术，使用 PCA 将数据减少到 k 维度。然后使用这些 PCA 预测作为特征，以下哪个声明是正确的？（ B ）
A、更高的“k”意味着更正则化　　B、更高的“k”意味着较少的正则化
C、都不对　　D、都正确

14、为节省神经网络模型的训练时间，神经网络模型的权重和偏移参数一般初始化为（ D ）
A、0　　B、0.5　　C、1　　D、随机值

15、在逻辑回归输出与目标对比的情况下，以下评估指标中哪一项不适用？（ D ）
A、AUC-ROC　　B、准确度　　C、Logloss　　D、均方误差

16、假设对数据提供一个逻辑回归模型，得到训练精度 X 和测试精度 Y。在数据中加入新的特征值，则下列哪一项是正确的？提示：其余参数是一样的。（ B ）
A、训练精度总是下降　　B、训练精度总是上升或不变
C、测试精度总是下降　　D、测试精度总是上升或不变

17、SVM（支持向量机）与 LR（逻辑回归）的数学本质上的区别是什么？（ A ）
A、损失函数　　B、是否有核技巧　　C、是否支持多分类　　D、其余选项皆错

18、逻辑回归为什么是一个分类算法而不是回归算法？（ A ）
A、是由于激活函数 sigmod 把回归问题转化成了二分类问题
B、是由于激活函数 maxsoft 把回归问题转化成了二分类问题
C、是由于激活函数 tanh 把回归问题转化成了二分类问题
D、是由于激活函数 Relu 把回归问题转化成了二分类问题

19、以下关于逻辑回归说法错误的是：（ C ）
A、特征归一化有助于模型效果
B、逻辑回归是一种广义线性模型
C、逻辑回归相比最小二乘法分类器对异常值更敏感
D、逻辑回归可以看成是只有输入层和输出层且输出层为单一神经元的神经网络

20、Apriori 算法的计算复杂度受（ D ）影响
A、项数（维度）　　B、事务平均宽度　　C、事务数　　D、支持度阀值

21、考虑下面的频繁3-项集的集合：{1. 2. 3}，{1. 2. 4}，{1. 2. 5}，{1. 3. 4}，{1. 3. 5}，{2. 3. 4}，{2. 3. 5}，{3. 4. 5}。假定数据集中只有5个项，采用合并策略，由候选产生过程得到4-项集不包含（ C ）
A、1. 2. 3. 4 B、1. 2. 3. 5 C、1. 2. 4. 5 D、1. 3. 4. 5

22、在关联规则中，有三个重要的指标，支持度（support），置信度（confident），作用度（lift），则对于规则X->Y的三个指标说法错误的是(N表示所有的样本item数目)：（ C ）
A、support = freq(X,Y)/N
B、confident = freq(X,Y)/freq(x)
C、lift = freq(X,Y)/freq(Y)
D、lift = freq(X,Y)*N/(freq(X)*freq(Y))

23、在基本K均值算法里，当邻近度函数采用（ A ）的时候，合适的质心是簇中各点的中位数。
A、曼哈顿距离 B、平方欧几里德距离 C、余弦距离 D、Bregman散度

24、一共5个点A(0,0)， B(1,0.3)，C(3,0.5)，D(2,1)，E(1.8,1.5)，采用Kmeans方法如果选取A，D为种子点，B，C，E分别属于（ A ）种子点
A、A,D,D B、A,A,D C、D,D,A D、D,A,D

25、图像中应用的kmeans算法，以下说法错误的是：（ D ）
A、kmeans算法有效的前提假设是数据满足高斯分布
B、kmeans需要手工指定类别的数目K
C、对于多维实数数据，kmeans算法最终一定是收敛的
D、kmeans算法可以直接得到类别分布的层级关系

26、以下关于KNN的描述，不正确的是（ A ）
A、KNN算法只适用于数值型的数据分类
B、KNN算法对异常值不敏感
C、KNN算法无数据输入假定
D、其他说法都正确

27、假定某同学使用贝叶斯分类模型时，由于失误操作，致使训练数据中两个维度重复表示。下列描述中正确的是：（ B ）
A、被重复的在模型中作用被加强
B、模型效果精度降低
C、如果所有特征都被重复一遍，则预测结果不发生变化
D、以上均正确

28、在其他条件不变的前提下，以下哪种做法容易引起模型中的过拟合问题？（ D ）
A、增加训练集量
B、减少神经网络隐藏层节点数
C、删除稀疏的特征
D、SVM算法中使用高斯核/RBF核代替线性核

29、下列哪一项在神经网络中引入了非线性（ B ）
A、SGD B、激活函数 C、卷积函数 D、都不正确

30、下列哪个神经网络结构会发生权重共享（ D ）
A、卷积神经网络 B、循环神经网络
C、全连接神经网络 D、选项 A 和 B

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、下列哪些是非监督数据离散化方法（ ABC ）
A、等宽法 B、等频法 C、聚类法 D、决策树法

2、在现实世界的数据中，元组在某些属性上缺少值是常有的。描述处理该问题的各种方法有：（ ABCD）
A、忽略元组 B、使用属性的平均值填充空缺值
C、使用一个全局常量填充空缺值 D、使用最可能的值填充空缺值

3、序数类型数据的可以实现数学计算 （ AB ）
A、众数 B、中位数 C、均值 D、方差

4、应用 PCA 后，以下哪项可以是前两个主成分？（ CD ）
A、(0.5,0.5,0.5,0.5)和(0.71,0.71,0,0)
B、(0.5,0.5,0.5,0.5)和(0,0,-0.71,0.71)
C、(0.5,0.5,0.5,0.5)和(0.5,0.5,-0.5,-0.5)
D、(0.5,0.5,0.5,0.5)和(-0.5,-0.5,0.5,0.5)

5、贝叶斯分类器是一种（ AC ）
A、基于贝叶斯公式的分类器
B、是一种无监督的学习
C、是一种概率预测模型
D、可处理小样本数据的方法

6、下面关于贝叶斯分类器说法正确的是（ AC ）
A、贝叶斯的思想是“由因推果”
B、贝叶斯的思想是“执果溯因”
C、可以用极大似然估计法解贝叶斯分类器
D、可以解决无监督学习的问题

7、对于信息增益，决策树分裂节点，下面说法正确的是（ BC ）
A、纯度高的节点需要更多的信息去区分
B、信息增益可以用“1 比特-熵”获得
C、如果选择一个属性具有许多归类值，那么这个信息增益是有偏差的
D、上述均错误

8、下列哪些机器学习算法不需要做数据归一化处理（ CD ）
A、K 均值 B、线性回归 C、决策树 D、朴素贝叶斯

9、对于 PCA 说法正确的是：（　ABD　）

A、我们必须在使用 PCA 前规范化数据
B、我们应该选择使得模型有最大 variance 的主成分
C、我们应该选择使得模型有最小 variance 的主成分
D、我们可以使用 PCA 在低维度上做数据可视化

10、逻辑回归有哪些处理非线性关系特征的方法？（　ABCD　）
A、特征离散化　　B、特征交叉
C、引入高阶项　　D、引入核函数

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、“飞机的飞行高度 3000 米”表示信息。　　（　对）

2、皮尔逊相关系数可用来判断 X 和 Y 之间的因果关系。（　错　）

3、熵衡量的是系统的不确定性，熵值越大（接近于 1）说明系统的不确定性越低。（　错　）

4、样品是数据对象的别名。（　对　）

5、在决策树中，随着树中结点数变得太大，即使模型的训练误差还在继续减低，但是检验误差开始增大，这是出现了模型拟合不足的问题。（　错　）

6、杰卡德系数用来度量非对称的二进制属性的相似性。（　对　）

7、K 均值聚类的核心目标是将给定的数据集划分为 K 个簇，并给出每个数据对应的簇中心点。（　对　）

8、决策树算法只能做 2 值分类，不能做多值分类。（　错　）

9、决策树通过预剪枝和后剪枝提升模型的泛化能力。（　对　）

10、杰卡德系数用来度量非对称的二进制属性的相似性。（　对　）

四、（10 分）已知两个一维模式类别的类概率密度函数为：

$$p(x/\omega_1)=\begin{cases} x & 0\le x<1 \\ 2-x & 1\le x\le 2 \\ 0 & \text{其它} \end{cases}$$

$$p(x/\omega_2)=\begin{cases} x-1 & 1\le x<2 \\ 3-x & 2\le x\le 3 \\ 0 & \text{其它} \end{cases}$$

先验概率 P(ω1)=0.6；P(ω2)=0.4，**则样本{x=1.65}属于哪一类别？**

五、（10 分）对于数据：{12，9，7，6，20，100，35，21，11，18，25，37}
（1）计算它的平均值，20%的截断均值和中位数。
（2）使用 MIN-MAX 规范方法将值其中的 6，100，35 转换到[0,1]。
（3）对数据按照深度为 4 进行划分，再写出按边界值进行平滑后的结果。

六、假设我们手上有 60 个正样本，40 个负样本，我们要找出所有的正样本，系统查找出 50 个，其中只有 40 个是真正的正样本，计算上述各指标。
请计算：
（1）TP：将正类预测为正类数?
（2）FN：将正类预测为负类数?
（3）FP：将负类预测为正类数?
（4）TN：将负类预测为负类数?
（5）准确率(accuracy)?
（6）精确率(precision)?
（7）召回率(recall)?

七、逻辑回归中，常用优势比 OR(odds ratio)衡量因素作用大小的比数比例指标：

$$OR_j = \frac{P_1/(1-P_1)}{P_0/(1-P_0)}$$ 式中 P_1 和 P_0 分别表示在 X_j 取值为 c_1 及 c_0 时的发病概率，OR_j 称作多变量调整后的优势比，表示扣除了其他自变量影响后危险因素的作用。

在一个具有 17 个家庭的样本里，共有 3 家的收入为￥10000，5 家的收入为￥11000，9 家的收入为￥12000。在收入为￥10000 的家庭里，1 个主妇不工作，2 个主妇工作；在收入为￥11000 的家庭里，1 个主妇不工作，4 个主妇工作；在收入为￥12000 的家庭里，1 个主妇不工作，8 个主妇工作。

收入（单位：千）	**主妇工作状况**		**总计**
	0（不工作）	**1（工作）**	
10 11 12	1 1 1	2 4 8	3 5 9
总计	3	14	17

令收入为变量 X，类别标签为工作状态。
（1）计算 X 为 10 和 11 时，优势比 OR 等于多少?
（2）计算 X 为 11 和 12 时，优势比 OR 等于多少?

四、解：
属于 w1 的概率：（2-1.65）*0.6=0.21[4 分]
属于 w2 的概率：0.65*0.4=0.26[4 分]
因此，属于 w2 类[2 分]

五、解：

（1）平均值是25.08，20%的截断均值（两端各去掉两个数以后的平均值）是18.8，中位数是19。[3分]
（2）规范化后，转换后的值为（0，1，0.31）[3分]
（3）深度为4进行划分，得到三个箱（6，7，9，11），（12，18，20，21），（25，35，37，100）；边界值平滑后的结果为（6，6，11，11），（12，12，21，21），（25，25，25，100）[4分]

六、解：
（1）[1分]TP：将正类预测为正类数：40
（2）[1分]FN：将正类预测为负类数：20（60-40，剩余没正确分类的正样本）
（3）[1分]FP：将负类预测为正类数：10
（4）[1分]TN：将负类预测为负类数：30
（5）[2分]准确率(accuracy) = 预测对的/所有 = (TP+TN)/(TP+FN+FP+TN) = 70%
（6）[2分]精确率(precision) = TP/(TP+FP) = 80%
（7）[2分]召回率(recall) = TP/(TP+FN) = 2/3

七、解：
X分别取10和11时，or=4/2=2 [5分]
X分别取11和12时，or=4/2=2 [5分]