期末考试题型：

一、单选题（2 分×15 题）

二、多选题（2 分×5 题）

三、填空题（1 分×15 空）

四、简答题（10 分×2 题）

五、综合应用题（10 分，15 分，共 2 题）

一、单选题

01、熵是为消除不确定性所需要获得的信息量，投掷均匀正六面体骰子的熵是：

____ ___。

A. 1 比特　　B. 2.6 比特　　C. 3.2 比特　　D. 3.8 比特

02、假设属性 income 的最大最小值分别是 12000 元和 98000 元。利用最大最小规范化的方法将属性的值映射到 0 至 1 的范围内。对属性 income 的 73600 元将被转化为：

___ ___。

A. 0.821　　B. 1.224　　C. 1.458　　D. 0.716

03、以下哪些算法是**回归**算法___ ____。

A、DBSCAN　　B. C4.5　　C. K-Mean　　D. Lasso

04、以下哪些分类方法可以较好地避免样本的不平衡问题，____。

A.K 最近邻方法　　B. 支持向量机

C.朴素贝叶斯方法　　D. BP 神经网络

05、以下哪项关于决策树的说法是错误的____ ____。

A. 冗余属性不会对决策树的准确率造成不利的影响

B. 子树可能在决策树中重复多次

C. 决策树算法对于噪声的干扰非常敏感

D. 寻找最佳决策树是 NP 完全问题

06、以下哪些算法是基于规则的分类器 ____ ____。

A. C4.5　　B. KNN　　C. 朴素贝叶斯分类器　　D. ANN

07、以下关于人工神经网络（ANN）的描述**错误**的有____ ____。

A. 神经网络对训练数据中的噪声非常鲁棒　　B. 可以处理冗余特征

C. 训练 ANN 是一个很耗时的过程　　D. 至少含有一个隐藏层的多层神经网络

08、通过聚集多个分类器的预测来提高分类准确率的技术称为 ____ ____。

A. 组合(ensemble)　　B. 聚集(aggregate)

C. 合并(combination)　　D. 投票(voting)

09、简单地将数据对象集划分成不重叠的子集，使得每个数据对象恰在一个子集中，这种聚类类型称作____ ____。

A. 层次聚类　　B. 划分聚类　　C. 非互斥聚类　　D. 模糊聚类

10、不纯性度量的**信息熵**指标的计算公式是____ ____。

A. $-\sum_{i=0}^{c-1} p(i|t)\log_2 p(i|t)$　　B. $1-\sum_{i=0}^{c-1}[p(i|t)]^2$

C. $1-\max_i[p(i|t)]$　　D. $\sum_{i=1}^{K}\sum_{x\in c_i} dist(c_i,x)^2$

11、不纯性度量的 **Gini 指标**的计算公式是____ ____。

A. $-\sum_{i=0}^{c-1} p(i|t)\log_2 p(i|t)$　　B. $1-\sum_{i=0}^{c-1}[p(i|t)]^2$

C. $1-\max_i[p(i|t)]$　　D. $\sum_{i=1}^{K}\sum_{x\in c_i} dist(c_i,x)^2$

12、___ ____是一个观测值，它与其他观测值的差别如此之大，以至于怀疑它是由不同的机制产生的。

A. 边界点　　B. 质心　　C. 离群点　　D. 核心点

13、___ ____将两个簇的邻近度定义为不同簇的所有点对的平均逐对邻近度，它是一种凝聚层次聚类技术。

A. MIN（单链）　　B. MAX（全链）

C. 组平均　　D. Ward 方法

14、____ ____将两个簇的邻近度定义为两个簇合并时导致的平方误差的增量,它是一种凝聚层次聚类技术。

A. MIN（单链）　　B. MAX（全链）

C. 组平均　　D. Ward 方法

15、关于 K 均值和 DBSCAN 的比较，以下说法**不正确**的是 ____。

A. 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。

B. 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。

C. 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。

D. 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。

16、以下两种描述分别对应哪两种对分类算法的评价标准？

(a)警察抓小偷，描述警察抓的人中有多少个是小偷的标准。

(b)描述有多少比例的小偷给警察抓了的标准。

A. Precision, Recall　　B. Recall, Precision

C. Precision, ROC　　D. Recall, ROC

17、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？

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

18、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. 分类（判别分析）　B. 聚类分析　C. 回归分析　D. 隐马尔可夫链

19、在基本 K 均值算法里，当邻近度函数采用____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

20、用于对数据分布对称性的测度指标是：____。

A. 均值　　B. 方差　　C.偏态　　D.峰态

21、分类数据的离散程度的度量指标：_____。

A. 极差　　B. 标准差　　C.异众比率　　D.四分位差

22、列联分析（独立性检验）是：_____。

A. 一个分类型变量对一个数值型变量的影响

B.两个数值型变量的分析

C.两个分类型变量的分析

D.一个分类型变量的检验

二、多选题

1、方差分析中的基本假定是：_____。

A、每个总体都就服从正态分布　　B、各个总体的方差必须相同

C、各个总体的均值必须相同　　D、观测值是独立的

2、时间序列的构成要素：_____。

A、长期趋势　　B、季节变动

C、循环变动（周期性）　　D、不规则变动

3、在多元回归线性回归模型中，对误差项ε的基本假定是：_____。

A、误差项ε是一个期望值为 0 的随机变量，即$E(\varepsilon)=0$

B、对于处变量x_1，x_2，...，x_k的所有值，ε的方差σ^2都相同

C、误差项ε是不能由x_1，x_2，...，x_k与 y 线性关系所解释的变异性

D、误差项ε是一个服从正态分布的随机变量，且相互独立，即$\varepsilon\sim N(0,\sigma^2)$

4、以下关于 Pearson 相关系数 r 的描述正确的是：_____。

A、r 的数值大小与 x 和 y 的原点及尺度相关

B、r 的取值范围是[-1,1]

C、r 具有对称性

D、r 仅仅是 x 与 y 之间的线性关系的一个度量，它不能用于描述非线性关系

5、下表是分析道路通行时间与时段（行因素）和路段（列因素）的方差分析结果，正确选项是：_____.

方差分析

差异源	SS	df	MS	F	P-value	F crit
样本	174.05	1	174.05	44.0633	5.7E-06	4.494
列	92.45	1	92.45	23.4051	0.00018	4.494
交互	0.05	1	0.05	0.01266	0.91182	4.494
内部	63.2	16	3.95			
总计	329.75	19				

A.时段（行因素）对行车时间有显著影响；

B.路段（列因素）对行车时间有显著影响；

C.时段（行因素）和路段（列因素）的交互作用对行车时间有显著影响；

D.时段（行因素）和路段（列因素）的交互作用对行车时间无显著影响；

6、在假设检验中，当我们作出检验统计量的观测值为落入原假设的拒绝域时，表示____.

A.没有充足的理由否定原假设

B.原假设是成立的

C.检验的 P 值较大

D.若拒绝原假设,犯第--类错误的概率超过允许限度

7、数值型变量可采用的集中趋势的度量有：_____。

A.标准分数

B.众数

C.中位数

D.平均数

8、无重复双因素方差分析的总误差平方和由_____构成。

A.行变量平方和（SSR）

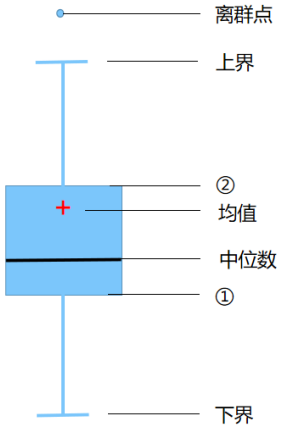
B.列变量平方和（SSC）

C.随机误差平方和（SSE）

D.交互作用平方和（SSRC）

二、填空题

1、如右图所示的箱线图中



①表示：_____.

②表示：_____.

②-①称为：_____.

2、方差分析的基本思想：通过对数据_____的分析来判断_____是否相等，进而分析自变量对因变量是否有显著影响。

3、方差分析中_____的作用是通过对总体均值之间的配对比较来进一步检验到底哪些均值之间存在差异。

4、方差分析中的_____是指每个水平或组的各样本数据与其平均值误差的平方和，反映了每个样本各观测值的离散状况；_____是指各组平均值与总平均值的误差平方各，反映各样本均值之间的差异程度。

5、神经元及其突触是神经网络的基本器件。因此，模拟生物神经网络应首先模拟生物神经元——人工神经元（节点），人工神经网络从三个方面进行模拟：_____、_____、_____。

6、进行数据预处理时所使用的主要方法包括：_____、_____、_____、_____。

7、处理噪声数据的方法主要包括：_____、_____、_____、_____。

8、_____是基本上不存在趋势的序列，各观察值基本上在某个固定的水平上波动或虽有波动，但并不存在某种规律，而其波动可以看成是随机的。

9、时间序列中_____（T）是指现象在较长时期内受某种根本性因素作用而形成的总的变动趋势；_____（C）是指现象以若干年为周期所呈现出的波浪起伏形态的有规律的变动；（S）是指现象在一年内随着季节的变化而发生的有规律的周期性变动。

10、根据复合的形式，时间序列主要有两种结构：_____，即假设各构成部分对时间序列的影响是相互独立的；_____，即假设各构成部分时间序列的影响均按比例而变化。

11、对多元线性回归方程，需要用_____来评价其拟合程度，它是多元回归中_____和占总平方和的比例，反映了在因变量 y 的变差中被估计的回归方程所解释的比例。

12、当回归模型中两个或两个以上的自变量彼此相关时，则称回归模型中存在_____，最简单的一种检测方法是计算模型中各对自变量之间的_____，并进行显著性检验。

13、一个好的聚类分析方法会产生高质量的聚类，具有两个特征：
和_____。

14、常用的聚类分析方法包括：_____、_____、_____、基于网格的方法和基于模型的方法。

15、Python 主要数据预处理函数
8 个

16、卷积神经网络是一类包含卷积计算且具有深度结构的前馈神经网络，卷积神经网络的隐含层包含_____、_____和_____3 类常见神经网络层构筑。

三、简答题

1、简述聚类分析与分类的不同，现代聚类分析方法的划分？

2、简述数据预处理中异常值分析的概念及方法？

3、简述集成学习的概念和构建组合分类器的几种方法。

四、综合应用题

1、从 3 个总体中各抽取容量不同的样本数据，结果如下。检验 3 个总体的均值之间是否有差异（$\alpha = 0.01$，$F_{0.01} = 8.02$）

样本 1	样本 2	样本 3
158	153	169
148	142	158
161	156	180
154	149	
169		

解：

2、使用下表中相似度矩阵进行单链和全链层次聚类。

（1）绘制树状图显示结果。树状图应当清楚地显示合并的次序。（6 分）

	P1	P2	P3	P4	P5
P1	1.00	0.10	0.41	0.55	0.35
P2	0.10	1.00	0.64	0.47	0.98
P3	0.41	0.64	1.00	0.44	0.85
P4	0.55	0.47	0.44	1.00	0.76
P5	0.35	0.98	0.85	0.76	1.00

（2）Scikit 库导入层次聚类的 Ward 方法和画树状图方法，完成以下程序：

```
import numpy as np
import matplotlib.pyplot as plt
from _____import _____.
linkage = _____#对数据集 X_blobs 采用 Ward 方法层次聚类
_____#生成树状图
```

3、PCA 算法作为一个非监督学习的降维方法，它只需要特征值分解，就可以对数据进行压缩，去噪。

（1）采用 UCI 葡萄酒数据集（wine）为例，进行归一化后再采用 PCA 生成两个主成分，完成以下程序：

_____#导入 wine 数据集方法

from sklearn.preprocessing import StandardScaler #导入 Z-score 标准化工具

_____ #导入 PCA 方法

wine =_____#导入数据预处理工具

scaler = _____#生成 Z-score 标准化对象

X = wine.data

y = wine.target

X_scaled = scaler.fit_transform(X)

print (X_scaled.shape)

_____#设置主成分数量为 2 以便我们进行可视化

_____#训练 PCA 模型

X_pca =_____#由 X_scaled 数据生成主成分

print(X_pca.shape)

（2）简述 PCA 算法的步骤。

（3）简述 PCA 算法的**优点**。

4、以 UCI 葡萄酒数据集（wine）为例，采用多层感知器（MLP）分类器模型进行学习。

（1）完成以下程序：

_____#导入 MLP 神经网络

_____#导入红酒数据集

from sklearn.model_selection import train_test_split#导入数据拆分工具

wine = _____#导入 wine 数据集

X = wine.data[:,:2]

y = wine.target

X_train, X_test, y_train, y_test =

_____#拆分数据集

mlp = _____#定义分类器，采用 quasi-Newton 方法的优化器

_____#训练 MLP 模型

代码运行结果：

```
MLPClassifier(activation='relu', alpha=0.0001, batch_size='auto', beta_1=0.9,
       beta_2=0.999, early_stopping=False, epsilon=1e-08,
       hidden_layer_sizes=(100,), learning_rate='constant',
       learning_rate_init=0.001, max_iter=200, momentum=0.9,
       nesterovs_momentum=True, power_t=0.5, random_state=None,
       shuffle=True, solver='lbfgs', tol=0.0001, validation_fraction=0.1,
       verbose=False, warm_start=False)
```

（2）从运行结果可以看出，Scikit 库中的多层感知器 MLPClassifier 模型，默认的**隐含层数量**及**隐含层神经元数量**、**神经元的激活函数类型**、**最大迭代步数**和**终止误差**分别是什么？