中国科学院大学
University of Chinese Academy of Sciences

试 题 专 用 纸

开课编号:083900M01004H-03
课程名称:机器学习
任课教师:周晓飞

学生姓名: _____ 学号:_____ 培养单位: _____ 分数: ____

满分 100 分，试题双面打印，请不要遗漏答题！ **答案写在答题纸上。**

一、单项选择题(30 分，每题 1 分)

1. 下列属于线性分类方法的是( )
A. 决策树
B. 感知机
C. 最近邻
D. 集成学习

2. 关于线性鉴别分析的描述最准确的是，找到一个投影方向，使得( )
A. 类内距离最大，类间距离最小
B. 类内距离最小，类间距离最大
C. 类内距离最大，类间距离最大
D. 类内距离最小，类间距离最小

3. SVM 的算法性能取决于( )
A. 核函数的选择
B. 核函数的参数
C. 软间隔参数 C
D. 以上所有

4. 支持向量机的对偶问题是( )
A. 线性优化问题
B. 二次优化
C. 凸二次优化
D. 有约束的线性优化

5. 以下对支持向量机中的支撑向量描述正确的是( )
A. 最大特征向量
B. 最优投影向量
C. 最大间隔支撑面上的向量
D. 最速下降方向

6. 下面属于 Bagging 方法的特点是( )
A. 构造训练集时采用 Bootstraping 的方式
B. 每一轮训练时样本权重不同
C. 分类器必须按顺序训练
D. 预测结果时，分类器的比重不同

7. 软间隔 SVM 的阈值趋于无穷，下面哪种说法正确( )
A. 只要最佳分类超平面存在，它就能将所有数据全部正确分类
B. 软间隔 SVM 分类器将正确分类数据
C. 会发生误分类现象
D. 以上都不对

8. 正则化的回归分析，可以避免( )
A. 线性化
B. 过拟合
C. 欠拟合
D. 连续值逼近

9. 混合高斯聚类中，运用了以下哪种过程( )
A. EM 算法　　B. 集合运算
C. 密度可达　　D. 样本与集合运算

10. PCA 在做降维处理时，优先选取哪些特征( )
A. 中心化样本的协方差矩阵的最大特征值对应特征向量
B. 最大间隔投影方向
C. 最小类内聚类
D. 最速梯度方向

11. 梯度下降算法的正确步骤是什么( )
(1)计算预测值和真实值之间的误差
(2)迭代更新，直到找到最佳权重
(3)把输入传入网络，得到输出值
(4)初始化随机权重和偏差
(5)对每一个产生误差的神经元，改变相应的（权重）值以减小误差
A. 1, 2, 3, 4, 5
B. 4, 3, 1, 5, 2
C. 3, 2, 1, 5, 4
D. 5, 4, 3, 2, 1

12. 以下哪种方法会增加模型的欠拟合风险（ ）
A. 添加新特征
B. 增加模型复杂度
C. 减小正则化系数
D. 数据增强

13. 以下说法正确的是（ ）
A. Boosting 和 Bagging 都是组合多个分类器投票的方法，二者都是根据单个分类器的正确率决定其权重
B. 梯度下降有时会陷于局部极小值，但 EM 算法不会
C. 除了 EM 算法，梯度下降也可求混合高斯模型的参数
D. 基于最小二乘的线性回归问题中，增加 L2 正则项，总能降低在测试集上的 MSE 误差

14. 在其他条件不变的前提下，以下哪种做法容易引起机器学习中的过拟合问题（ ）
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 算法中使用高斯核代替线性核

15. 在 HMM 中,如果已知观察序列和产生观察序列的状态序列,那么可用以下哪种方法直接进行参数估计（ ）
A. EM 算法
B. 维特比算法
C. 前向后向算法
D. 极大似然估计

16. 以下哪种距离会侧重考虑向量的方向（ ）
A. 欧式距离
B. 海明距离
C. Jaccard 距离
D. 余弦距离

17. 解决隐马模型中预测问题的算法是（ ）
A. 前向算法
B. 后向算法
C. Baum-Welch 算法
D. 维特比算法

18. 梯度爆炸问题是指在训练深度神经网络的时候，梯度变得过大而损失函数变为无穷。在 RNN 中，下面哪种方法可以较好地处理梯度爆炸问题( )
A. 梯度裁剪　　B. 所有方法都不行　　C. Dropout　　D. 加入正则项

19. 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 B. 聚类 C. 关联分析 D. 隐马尔可夫链

20. 下列哪一种架构有反馈连接并常被用来处理序列数据？（ ）
A. 循环神经网络
B. 卷积神经网络
C. 全连接网络
D. 都不是

21. 以下方法中不是解决样本类别不平衡的手段的是（ ）
A. 欠采样
B. 加深神经网络的层数
C. 过采样
D. 使用 focal loss

22. 神经网络模型（Neural Network）因受人类大脑的启发而得名。神经网络由许多神经元（Neuron）组成，每个神经元接受一个输入，对输入进行处理后给出一个输出。请问下列关于神经元的描述中，哪些是正确的？（ ）
A. 每个神经元有多个输入和一个输出
B. 每个神经元有一个输入和多个输出
C. 每个神经元有多个输入和多个输出
D. 以上所有

23. 以下不属于贝叶斯分类器参数估计的准则的是（ ）
A. 最大高斯后验
B. 最大 beta 后验
C. 最大间隔
D. 极大似然

24. EM 算法（Expectation Maximization Algorithm）是机器学习领域的一个经典算法，下面关于 EM 算法的说法中不正确的有：（ ）
A. EM 算法属于一种分类算法
B. EM 算法可用于隐马尔科夫模型的参数估计
C. EM 算法可以分为 E-step 和 M-step 两步
D. EM 算法可用于从不完整的数据中计算最大似然估计

25. 下列选项中，关于逻辑回归的说法不正确是：（ ）
A. 逻辑回归是监督学习
B. 逻辑回归利用了回归的思想
C. 逻辑回归是一个分类模型
D. 逻辑回归使用 sigmoid 函数作为激活函数对回归的结果做了映射

26. 下列关于无监督学习描述错误的是（ ）
A. 无标签信息
B. 聚类是其中一个应用
C. 不能使用降维
D. 在现实生活中有广泛的应用

27. 下列关于聚类说法错误的是（ ）
A. 无需样本有标签
B. 可用于抽取一些特征
C. 可提取关于数据的结构信息
D. 同一个类内的样本之间差异较大

28. 下列关于 k-means 说法不正确的是（ ）
A. 算法有可能终止于局部最优解
B. 簇的数目需要事先给定
C. 对噪声和离群点敏感
D. 适合处理非凸型数据

29. 下列关于有监督学习描述错误的是（ ）
A. 有标签信息
B. 分类是其中一个分支
C. 所有数据都相互独立
D. 分类原因不透明

30. 在机器学习中，当模型的参数量大于样本量时参数估计使用（ ）
A. 解析法
B. 穷举法
C. 集成法
D. 梯度下降法

二、多项选择题(15 分，每题 1 分)

1. 可用于贝叶斯决策的函数( )
A. $\omega^* = \arg\max_{\omega_i} p(x|\omega_i)p(\omega_i)$

B. $g(x) = p(\omega_1|x) - p(\omega_2|x)$

C. $g(x) = \ln\frac{p(x|\omega_1)}{p(x|\omega_2)} + \ln\frac{p(\omega_1)}{p(\omega_2)}$

D. $p(\omega_1|x)$

2. 对聚类问题描述不正确的( )
A. 监督学习　　B. 无监督学习
C. 线性决策　　D. 增量学习

3. 以下选项中属于聚类问题可用的相似性度量准则有（ ）
A. 样本-样本距离
B. 样本-集合距离
C. 集合-集合距离
D. 集合内样本间距

4. 支持向量机可能解决的问题( )
A. 线性分类
B. 非线性分类
C. 回归分析
D. BP 算法

5. 下面属于非线性模型的机器学习的方法( )
A. 决策树
B. PCA
C. 多层感知机
D. 单层感知机

6. 下列选项中属于实现决策树分类方法时的常见组件有（ ）
A. 基分类器　　B. 激活函数
C. 剪枝方法　　D. 划分目标

7. 给定两个特征向量，以下哪些方法可以计算这两个向量相似度（ ）
A. 欧式距离　　B. 夹角余弦
C. 信息熵　　D. 曼哈顿距离

8. 影响 K-Means 聚类算法结果的主要因素有( )
A. 样本顺序
B. 相似性度量
C. 初始聚类中心
D. 样本类别

9. 影响深度神经网络训练效果的因素有( )
A. 学习率
B. 训练集规模
C. 网络深度
D. 激活函数

10. 下面关于特征选择和特征提取的描述正确的是( )
A. Relief 算法属于特征提取方法
B. 特征选择的目标是从原始的 d 个特征中选择 k 个特征
C. 特征提取的目标是根据原始的 d 个特征的组合形成 k 个新的特征
D. PCA 属于特征选择方法

11. 以下关于正则化的描述正确的是（ ）
A. 正则化可以防止过拟合
B. L1 正则化能得到稀疏解
C. L2 正则化约束了解空间
D. Dropout 也是一种正则化方法

12. 以下选项中可以用来降低过拟合的方法有（ ）
A. 获取更多训练数据
B. 减少使用训练样本的量
C. 增加模型复杂度
D. 添加正则化方法

13. 以下哪些模型是分类模型（ ）
A.最近邻　　B.K 均值　　C.朴素贝叶斯　　D.逻辑回归

14. 在某神经网络的隐层输出中，包含-1.5，那么该神经网络采用的激活函数不可能是（ ）
A. Sigmoid　　B. Tanh　　C. Relu　　D. Leaky Relu

15. 最近邻分类中测度度量，经常采用范数距离，以下属于范数距离的是（ ）

A. $D(x,y)=\sum_i |x_i-y_i|$

B. $D(x,y)=\max_i |x_i-y_i|$

C. $D(x,y)=[(x-y)^T(x-y)]^{1/2}$

D. $D(x,y)=(x-y)^T\Sigma^{-1}(x-y)$

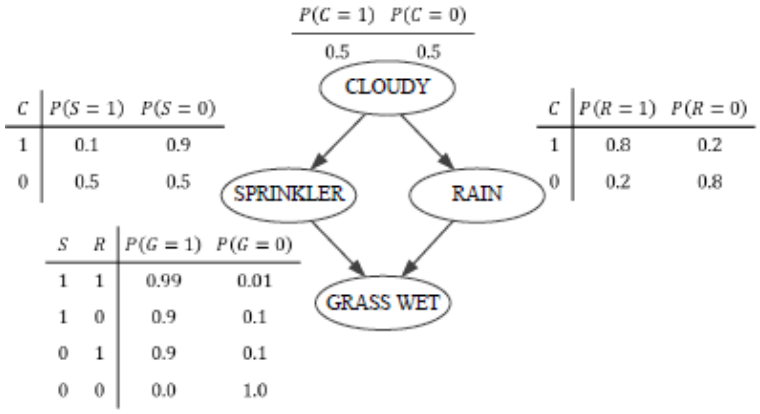
三、简答题(15 分，每题 5 分)

1. 简要介绍在深度神经网络中引入非线性激活函数的作用。
2. 描述主成分分析的主要步骤。
3. 请给出分类和聚类要实现的目标，并各举两个代表性算法。

四、计算题(30 分，每题 10 分)

1. 某同学经常在网络购物平台购买超低价图书，截止目前总共买了 100 本书，其中的 99 本是盗版书，1 本是正版书。考虑购买到的图书只有正版和盗版之分且该同学每次购买超低价图书的行为是独立的情况下，利用极大似然估计，对该同学用超低价在网络购物平台买到盗版书和正版书的概率进行估计。

2. 已知四个随机变量 C、S、R、G，分别代表 CLOUDY、SPRINKLER、RAIN 和 GRASS WET，它们之间构成的贝叶斯网络如图所示。



计算：（1）在 G=1 的条件下，S=0 的概率；（2）在 G=1 的条件下，R=0 的概率。

3. 对 3 个 256×256 的特征图进行卷积层操作，卷积核 10 个 8×8，Stride 是 1，pad 为 2，输出特征图的尺度是多少？卷积层的参数是多少？

五、请利用机器学习相关技术实现潜在用户分类任务。一个潜在用户分类数据集中包含 20000 名用户的特征以 300 维向量的形式给出，需要实现一个机器学习模型根据这 20000 名用户的特征向量分为潜在用户、非潜在用户和暂不明确三类。要求给出设计思想、简要模型结构和参数估计方法。(10 分)