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

试 题 专 用 纸

开课编号:083900M01004H-02
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. 属于无监督学习的机器学习算法是( ) C
A. 支持向量机
B. Logistic 回归
C. 层次聚类
D. 决策树

2. SVM 的原理的简单描述，可概括为( ) C
A. 最小均方误差分类
B. 最小距离分类
C. 最大间隔分类
D. 最近邻分类

3. SVM 的算法性能取决于( ) D
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. 以下描述中，属于决策树策略的是( ) D
A. 最优投影方向
B. 梯度下降方法
C. 最大特征值
D. 最大信息增益

5. 集成学习中基分类器的选择如何，学习效率通常越好( ) D
A. 分类器相似
B. 都为线性分类器
C. 都为非线性分类器
D. 分类器多样，差异大

6. 集成学习中，每个基分类器的正确率的最低要求( ) A
A. 50%以上 B. 60%以上
C. 70%以上 D. 80%以上

7. 回归问题和分类问题的区别( ) A
A. 前者预测函数值为连续值，后者为离散值
B. 前者预测函数值为离散值，后者为连续值
C. 前者是无监督学习
D. 后者是无监督学习

8. 主成分分析方法是一种什么方法( ) C
A. 分类方法　　B. 回归方法
C. 降维方法　　D. 参数估计方法

9. 多层感知机方法中，可用作神经元的非线性激活函数( ) A
A. logistic 函数
B. 范数
C. 线性内积
D. 加权求和

10. 梯度下降算法的正确步骤是什么( ) B
(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

11. 假如使用一个较复杂的回归模型来拟合样本数据，使用岭回归，调试正则化参数λ，来降低模型复杂度。若λ较大时，关于偏差（bias）和方差（variance），下列说法正确的是（ ） C
A. 若 λ 较大时，偏差减小，方差减小
B. 若 λ 较大时，偏差减小，方差增大
C. 若 λ 较大时，偏差增大，方差减小
D. 若 λ 较大时，偏差增大，方差增大

12. 以下关于深度网络训练的说法正确的是：( ) D
A. 训练过程需要用到梯度，梯度衡量了损失函数相对于模型参数的变化率
B. 损失函数衡量了模型预测结果与真实值之间的差异
C. 训练过程基于一种叫做反向传播的技术
D. 其他选项都正确

13. 关于 CNN，以下结论正确的是（ ） C
A. 在同样层数、每层神经元数量一样的情况下，CNN 比全连接网络拥有更多的参数
B. CNN 可以用于非监督学习，但是普通神经网络不行
C. Pooling 层用于减少图片的空间分辨率
D. 接近输出层的 filter 主要用于提取图像的边缘信息

14. 关于 k-means 算法，正确的描述是（ ）B
A. 能找到任意形状的聚类
B. 初始值不同，最终结果可能不同
C. 每次迭代的时间复杂度是 O(n^2)，其中 n 是样本数量
D. 不能使用核函数

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（ ） D
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（ ） D
A. 欧式距离
B. 海明距离
C. Jaccard 距离
D. 余弦距离

17. 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
A. 梯度裁剪　　B. 所有方法都不行　　C. Dropout　　D. 加入正则项

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

19. “过拟合”只在监督学习中出现，在非监督学习中没有“过拟合”，这种说法是（ ）B
A. 对的
B. 错的
C. 偶尔对偶尔错
D. 不一定

20. 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？（ ） D
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

21. 下列选项中属于机器学习可解决的问题的有（ ） D
A. 分类　　B. 聚类
C. 回归　　D. 以上均可

22. 下列选项中，关于 KNN 算法说法不正确的是（D）
A. 能找出与待测样本相近的 K 个样本
B. 可以使用欧氏距离度量相似度
C. 实现过程相对简单，但是可解释性不强
D. 效率很高

23. 关于 SVM 的损失函数，下列说法中错误的是：（D）
A. SVM 适用于多种损失函数
B. 0/1 损失函数的最终结果只有两个，0 代表分类正确，1 代表分类错误
C. 合页损失(Hinge loss)衡量了被误分类的样本离分割超平面的距离的大小程度
D. 分类 SVM 常用平方误差损失来衡量模型的好坏

24. 关于 SVM 核函数，下列说法中错误的是：（ ）C
A. 核函数的引入提升了 SVM 在线性不可分场景下的模型的稳健性
B. 核函数就是一类具有将某一类输入映射为某一类输出的函数
C. 核函数把特征映射到的空间维度越高越好
D. 常见的核函数有线性核、高斯核、多项式核、sigmoid 核

25. 下列选项中，关于逻辑回归的说法不正确是：（B）
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. 下列关于样本类别不均衡场景的描述正确的是（ ）A
A. 样本类别不均衡会影响分类模型的最终结果
B. 样本类别不均衡场景下我们没有可行的解决办法
C. 欠采样是复制类别数较少的样本来进行样本集的扩充
D. 过采样会造成数据集部分信息的流失

27. 下列关于无监督学习描述错误的是（ ）C
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

28. 下列关于有监督学习描述错误的是（ ）C
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

29. 支持向量机的对偶问题是( )C
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

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. 以下选项中可用于实现层次聚类的方法有（ ）
A. 自左向右
B. 从右到左
C. 自底向上
D. 自顶向下

3. 以下选项中属于 K 均值聚类方法流程中步骤的有（ ）
A. 初始化类心
B. 利用标签将样本分类
C. 按当前类心对样本归类
D. 迭代类心

4. 以下可行的最近邻分类的加速方案( )
A. 分层搜索
B. 训练样本缩减
C. 样本增加
D. 非线性投影

5. Adaboost 方法中，需要迭代调整的两个重要参数是( )
A. 样本权重　　B. 分类器权重
C. 梯度变化率　　D. 梯度

6. 以下模型中属于贝叶斯网络的有( )
A. 马尔可夫随机场
B. 隐马尔可夫模型
C. 条件随机场
D. 朴素贝叶斯分类器

7. 下面关于集成学习的描述，正确的是( )
A. Bagging 方法可以并行训练
B. Bagging 方法基学习器的比重不同
C. Boosting 方法可以并行训练
D. Boosting 方法基学习器的比重不同

8. 如果 SVM 模型欠拟合，以下方法哪些可以改进模型( )
A. 增大惩罚参数 C 的值
B. 减小惩罚参数 C 的值
C. 减小核系数(gamma 参数)
D. 增大核系数(gamma 参数)

9. 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

10. 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

11. 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

12. 类别不平衡就是指分类问题中不同类别的训练样本相差悬殊的情况，例如正例有 900 个，而反例只有 100 个，这个时候我们就需要进行相应的处理来平衡这个问题,下列方法正确的是（ ）
A. 在训练样本较多的类别中进行欠采样
B. 在训练样本较多的类别中进行过采样
C. 直接基于原数据集进行学习，对预测值进行再缩放处理
D. 通过对反例中的数据进行插值，来产生额外的反例

13. 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

14. 以下选项中可以用来降低欠拟合的方法有（ ）

A. 获取更多训练数据　　B. 添加有效的数据特征

C. 增加模型复杂度　　D. 添加正则化方法

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

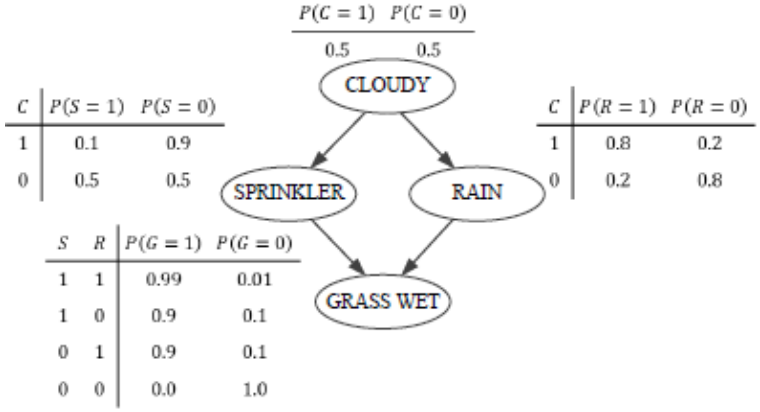
三、简答题(15 分，每题 5 分)

1. 请给出 L1 范数和 L2 范数的计算方法及他们的使用场景。
2. 请给出你对过拟合和欠拟合的理解，并给出两种降低欠拟合的手段。
3. 请给出你对有监督学习和无监督学习的理解。

四、计算题(30 分，每题 10 分)

1. 某同学在商场中开盲盒，购买的 50 个盲盒中 15 个是自己喜欢的款式，35 个是不喜欢的款式。考虑只有喜欢和不喜欢两类且该同学每次开盲盒的行为是独立的情况下，用极大似然估计，估计该同学在该商场中开出喜欢和不喜欢款式盲盒的概率。

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 28×28 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、利用机器学习相关技术实现图像分类任务。一个图像分类数据集中包含 10000 张彩色图像以 RGB 形式记录，需要实现一个机器学习模型将这 10000 张彩色图像分为猫、狗、鸟三类。要求给出设计思想、简要模型结构和参数估计方法。(10 分)