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

试 题 专 用 纸

开课编号:083900M01004H-01
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( )
A. 支持向量机
B. Logistic 回归
C. 层次聚类
D. 决策树

2. 下列方法不受数据归一化影响的是( )
A. SVM
B. 神经网络
C. Logistic 回归
D. 决策树

3. SVM 的原理的简单描述，可概括为( )
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

4. 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

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. 集成学习中基分类器的选择如何，学习效率通常越好( )
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

7. 下面属于 Boosting 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相同
C. 分类器可以并行训练
D. 预测结果时，分类器的比重不同

8. 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

9. 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

10. 关于 CNN，以下结论正确的是（ ）
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

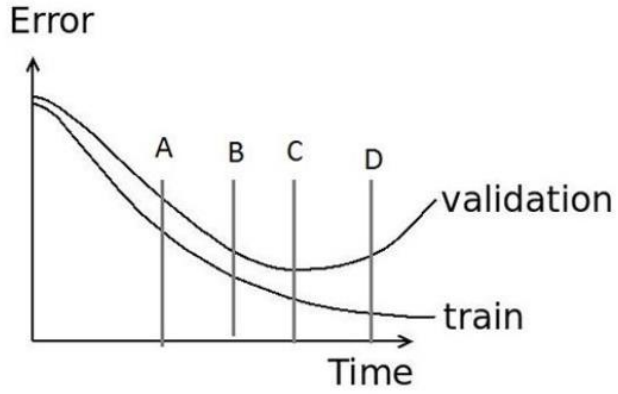
11. 关于 k-means 算法，正确的描述是（ ）
A. 能找到任意形状的聚类
B. 初始值不同，最终结果可能不同
C. 每次迭代的时间复杂度是 O(n^2)，其中 n 是样本数量
D. 不能使用核函数

12. 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

13. 下面关于 Adaboost 算法的描述中，错误的是（ ）
A. 是弱分类器的线性组合
B. 提升树是以分类树或者回归树为基本分类器的提升办法
C. 该算法实际上是前向分步算法的一个实现，在这个方法里，模型是加法模型，损失函数是指数损失，算法是前向分步算法。
D. 同时独立地学习多个弱分类器

14. 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

15. 当训练一个神经网络来作图像识别任务时，通常会绘制一张训练集误差和验证集误差图来进行调试。在下图中，最好在哪个时间停止训练( )



A. A　　B. B　　C. C　　D. D

16. 下列方法中没有考虑先验分布的是( )
A. 最大后验估计
B. 贝叶斯分类器
C. 贝叶斯学习
D. 最大似然估计

17. 下列哪一项主要负责在神经网络中引入非线性？（ ）
A. 随机梯度下降
B. 修正线性单元（ReLU）
C. 输入的加权求和
D .以上都不正确

18. 在一个神经网络中，下面哪种方法可以用来处理过拟合？（ ）
A. Dropout
B. 分批归一化(Batch Normalization)
C. 正则化(regularization)
D. 都可以

19. L1 与 L2 范数在 Logistic Regression 中,如果同时加入 L1 和 L2 范数,会产生什么效果。( )
A. 可以做特征选择,并在一定程度上防止过拟合
B. 能解决维度灾难问题
C. 能加快计算速度
D. 能增加模型的拟合能力

20. 下列模型中属于判别式模型的是（ ）
A. 支持向量机　　B. 隐马尔可夫模型
C. 朴素贝叶斯模型　　D. 高斯混合模型

21. k-NN 方法一般在（ ）的情况下效果较好
A. 样本较多，典型性不好
B. 样本较少，典型性好
C. 样本呈团状分布
D. 样本呈链状分布

22. “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（ ）
A. 对的
B. 错的
C. 偶尔对偶尔错
D. 不一定

23. 关于交叉验证，下列说法中错误的是（ ）
A. 交叉验证能够提升模型的准确率
B. 交叉验证能够让样本数据被模型充分利用
C. 交叉验证搭配网格搜索能够提升我们查找最优超参数组合的效率
D. 使用网格搜索时我们一般会提供超参数的可能取值字典

24. 关于 SVM 的损失函数，下列说法中错误的是：（ ）
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

25. 关于朴素贝叶斯，下列说法错误的是：（ ）
A. 它是一个分类算法
B. 朴素的意义在于它的一个天真的假设：所有特征之间是相互独立的
C. 它实际上是将多条件下的条件概率转换成了单一条件下的条件概率，简化了计算
D. 以贝叶斯估计的角度来看朴素贝叶斯时，其没有估计联合概率

26. 避免直接的复杂非线性变换，采用线性手段实现非线性学习的方法是（ ）
A. 核函数方法
B. 集成学习
C. 线性鉴别分析
D. Logistic 回归

27. 下列关于样本类别不均衡场景的描述正确的是（ ）
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

28. 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

29. 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

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. 以下属于聚类方法的是( )
A. k-means
B. 层次聚类
C. Fisher 鉴别
D. 密度聚类

3. 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

4. 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

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. 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

7. 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

8. 下面属于线性分类方法的是( )
A. 线性回归　　B. 决策树
C. 最近邻　　D. Fisher 鉴别

9. 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

10. 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

11. 在机器学习中，下列关于各算法对应的损失函数正确的是（ ）
A. 最小二乘-Square loss
B. SVM-Hinge Loss
C. Logistic Regression-交叉熵损失函数
D. AdaBoost-指数损失函数

12. 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

13. 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

14. 以下哪些机器学习算法可以不对特征做归一化处理（ ）

A. 随机森林　　B. 逻辑回归　　C. SVM　　D. 决策树

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

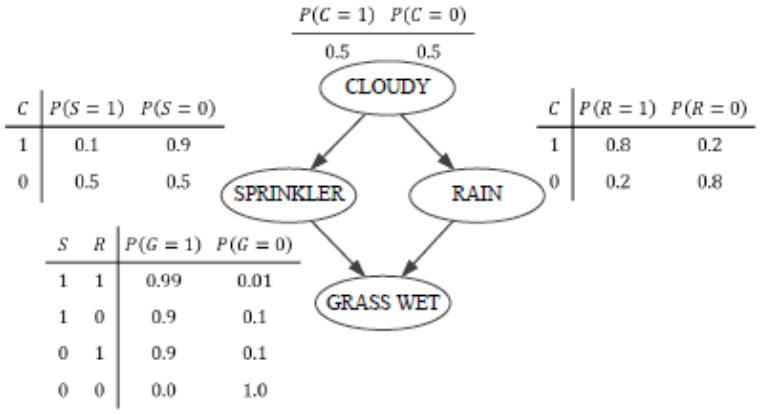
三、简答题(15 分，每题 5 分)

1. 请简要介绍 SVM 的设计思想。
2. 请给出你对泛化误差的理解。
3. 请给出你对生成式模型和判别式模型的理解。

四、计算题(30 分，每题 10 分)

1. 某地气象台对当地晴雨两种天气的统计如下：某月份共计 30 天，其中 12 天晴天，18 天雨天。只考虑两种天气且每日天气情况独立的情况下，试采用极大似然估计，对该地区出现晴天和雨天的概率进行估计。

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 28×28 的特征图进行卷积层操作，卷积核 5 个 4×4，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情感分类任务。一个情感分类数据集中包含 5000 段自然语言形式的电影评价，需要实现一个机器学习模型将这 5000 段电影评价分为正面、中立、负面三类。要求给出设计思想、简要模型结构和参数估计方法。(10 分)