

南京大学 人工智能学院 全日制统招本科生

《机器学习导论》期末考试试卷 闭卷

任课教师姓名: _____

考试日期: _____ 考试时长: _____ 小时 _____ 分钟

考生年级 _____ 考生专业 _____ 考生学号 _____ 考生姓名 _____

题号	一	二	三	四	五	总分
得分						

一、单项选择题(每题2分, 共16分)

本题得分	
------	--

- 什么是“机器学习”()
 - 一种与人工智能无关的技术
 - 让机器在没有人类明确指示的情况下自主地改进其性能的算法
 - 一种将数据的特征转化为输出的算法
 - 为机器提供不断优化和变化自身体系结构的框架
- 监督学习和非监督学习的区别是什么()
 - 监督学习依赖于标记数据, 非监督学习不依赖
 - 监督学习中的结果可以被解释, 而非监督学习中不可以
 - 监督学习需要事先定义好学习目标, 而非监督学习则不需要
 - 监督学习只适用于分类和回归问题, 而非监督学习适用于聚类和降维问题
- 线性回归和岭回归的主要区别是什么()
 - 线性回归只能处理两个变量之间的线性关系, 岭回归对多个变量之间的线性关系进行建模
 - 线性回归可以捕捉所有可能的线性关系, 而岭回归只能捕捉一部分
 - 线性回归可以推广到非线性情况, 岭回归无法推广
 - 线性回归对多重共线性非常敏感, 岭回归可以一定程度上解决共线性问题
- 决策树算法是一种常用的分类算法。以下对决策树算法描述最准确的是()

- A. 将一个数据集分成两部分，使得每一部分都尽可能简单
- B. 基于一个特征对数据集进行切割，将数据集分成尽可能多的子集
- C. 一种基于一阶Predicate Logic的分类器
- D. 一种类似于集成学习的算法，由多棵决策树构成
5. 在机器学习中，交叉验证用于()
- A. 在训练集中寻找最佳参数
- B. 在模型上评估未使用的数据
- C. 选择最佳的预测模型
- D. 评估大型数据集的有效性
6. 支持向量机是一种非常常见的分类算法。SVM区别于其他分类器最主要的特点是()
- A. SVM使用统计学习理论来选择复杂度
- B. SVM优化了模型和数据性能
- C. SVM通过使用最优化算法解决了问题
- D. SVM使用几何理论选取最优决策边界
7. 超参数的作用是什么()
- A. 优化算法的选择和调整
- B. 定义模型中的层数和神经元数
- C. 确定模型预测准确性的参数值
- D. 设置模型训练过程的相关参数
8. 在机器学习中，人们通常会将原始数据转换为一组特征，以便机器可以使用这些特征来做出预测。这个过程称为()
- A. 特征选择
- B. 特征提取
- C. 特征缩放
- D. 特征映射

二、填空题(每空2分，共22分)

本题得分	
------	--

1. 深度学习是一种机器学习方法，它模仿人类大脑神经元的工作原理，通过构建多层神经网络来学习和表示_____。

2. 在机器学习中, _____指识别模式、规律和趋势的能力, 用于对未知数据进行预测和分类。
3. 机器学习的一个关键要素是_____, 也就是将收集的数据转化为可用于学习的形式。
4. 在监督学习中, 训练数据集包含了_____和对应的标签信息, 用于训练模型和评估模型的准确性。
5. 无监督学习与监督学习不同, 它的训练数据集中只有_____, 没有对应的标签信息。
6. 在机器学习中, _____是指模型对未知数据的泛化能力和预测准确性。
7. 数据预处理是机器学习中的重要步骤, 它包括_____, 特征选择、特征变换等操作, 以提高数据质量和模型表现。
8. 交叉验证是一种常用的模型评估方法, 它将数据划分为_____个互不重叠的子集, 依次使用其中一个子集作为测试集, 其他子集作为训练集。
9. 过拟合是指模型在训练数据上表现良好, 但在新数据上表现较差, 解决过拟合的常见方法包括_____, 正则化等。
10. 集成学习是一种将多个基学习器组合起来共同进行决策的方法, 常见的集成学习方法包括_____。
11. 特征工程是机器学习中的重要环节, 它包括特征选择、特征变换等操作, 以提取和构造对模型有用的_____。

三、判断题(每题2分, 共22分)

本题得分	
------	--

1. 机器学习是一门通过算法和模型使计算机自动学习的学科。()
2. 机器学习的目标是使计算机具备像人类一样的自主决策能力。()
3. 监督学习是一种机器学习方法, 需要为算法提供带有标签的训练样本。()
4. 无监督学习是不使用标签信息, 而是寻找数据的内在结构和模式。()
5. 模型可以通过训练数据学习到特定的函数关系, 用于预测新数据的输出。()
6. 机器学习算法通常可以直接处理原始数据, 无需进行预处理和特征选择。()
7. 欠拟合指模型对训练数据过度拟合, 导致在新数据上表现较差。()
8. 特征工程是机器学习中的一项重要任务, 它涉及选择、提取和转换数据的特征。
()

9. 模型评估是通过将数据集分为训练集和测试集来完成的，测试集用于评估模型的性能。()

10. 深度学习是基于神经网络的机器学习方法，可以处理复杂的图像、语音和自然语言等任务。()

11. 在机器学习中，过拟合可以通过增加训练数据数量来缓解，从而提高模型的泛化能力。()

四、简答题(每题6分，共24分)

本题得分	
------	--

1. 简要概述机器学习的定义和基本原理，并解释监督学习、无监督学习和强化学习之间的区别。

2. 机器学习算法的性能评估指标有哪些？分别介绍准确率、召回率、F1 值、ROC 曲线和AUC 的含义和应用场景。

3. 说明过拟合和欠拟合在机器学习中的含义和影响，并阐述如何解决这两个问题。

4. 过程梯度下降法和随机梯度下降法是常用的优化算法，比较这两种算法的原理、优缺点和适用场景。

五、应用题(共16分)

本题得分	
------	--

你是一个数据科学家，受雇于一家电商公司，需要设计一个机器学习模型来预测用户的购买行为。公司给你提供了一个包含100,000个用户的数据集，每个用户有以下特征：

年龄：连续变量，表示用户的年龄。

性别：离散变量，表示用户的性别，取值为0(女性)或1(男性)。

收入：连续变量，表示用户的年收入。

购买次数：离散变量，表示用户的购买次数。

1. 选择合适的性能度量指标，用于评估分类模型的性能，并解释你的选择。
2. 根据给定的数据集，选择合适的特征工程方法，用于处理特征数据，并解释你的选择。
3. 选择适合的机器学习算法来构建分类模型，并解释你的选择。
4. 假设你训练了一个分类模型，它在训练集上的性能非常好，但在测试集上的性能较差。分析可能导致这种情况发生的原因，并提出改进模型性能的方法。