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、填空题(每空2分，共18分)

本题得分	

1. 机器学习是一门研究如何使计算机系统_____的学科，它的目标是通过构建模型和算法来使计算机具备学习能力。

2. 机器学习的一个核心概念是模型，在机器学习中，模型是对_____的一种简化和抽象，用来描述数据的结构和规律。

3. 在监督学习中，训练数据集包含了输入变量和对应的_____标签，用来训练模型预测新的未知数据。

4. 无监督学习与监督学习不同，无监督学习中的训练集只包含_____，没有对应的标签信息。

5. 在机器学习中，泛化能力是指模型在_____上能够很好地预测未知数据的能力。

6. 数据预处理是机器学习中的重要步骤，它包括数据_____、特征选择、特征变换等操作，以提高数据质量和_____表现。

7. 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，过拟合问题常常由于模型过于_____或训练数据过少导致。

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。

本题得分	

二、判断题(每题2分，共24分)

1. 深度学习是一种基于传统机器学习算法的进一步改进和扩展。( )

2. 回归问题是一种机器学习任务，其目标是通过给定的输入预测连续的输出值。( )

3. 在无监督学习中，模型通过学习数据的内在结构和模式，来进行数据分类和预测。( )

4. 在机器学习中，特征选择是一种减少特征维度的方法，以提高模型性能和减少计算负担。( )

5. 交叉验证是一种评估模型性能和选择超参数的方法，通过将数据集划分为多个子集进行模型训练和测试。( )

6. 过拟合是指模型在训练中过度拟合训练数据，导致在新数据上表现不佳的现象。( )

7. 决策树是基于树状图结构进行决策的机器学习模型，仅适用于处理离散型数据。( )

8. 集成学习是一种通过将多个模型组合来提高预测性能和泛化能力的机器学习方法。( )

9. 增大训练数据集的规模可以缓解过拟合问题，提高模型的泛化能力。( )

10. SVM是一种非监督学习算法，通过寻找最优的超平面来实现数据的分类。( )

11. 模型评估通过将数据集分为训练集和测试集来完成，训练集用于模型的训练，而测试集用于评估模型的性能。( )

12. 特征工程是机器学习中的一项关键任务，涉及选择、提取和转换数据的特征，以便更好地训练模型。( )

三、多项选择题(每题3分，共21分)

本题得分	

1. 机器学习包括以下哪些主要任务( )

A. 分类

B. 聚类

C. 回归

D. 降维

E. 数据预处理

2. 在机器学习中，下列哪些算法属于监督学习( )

A. 决策树

B. K均值聚类

C. 支持向量机

D. 朴素贝叶斯分类器

E. 随机森林

3. 下面哪些方法可以用于解决过拟合问题( )

A. 增加训练数据规模

B. 减少模型的复杂度

C. 正则化

D. 特征选择

E. 集成学习

4. 在机器学习中，以下哪些特征选择方法是常用的( )

A. 方差阈值

B. 相关系数

C. 信息增益

D. L1正则化

E. 主成分分析

5. 可以用于评估分类模型的性能包括哪些指标( )

A. 准确率

B. 精确率

C. 召回率

D. F1分数

E. 均方误差

6. 哪些场景适合使用决策树算法( )

A. 处理离散型数据

B. 处理连续型数据

C. 处理高维数据

D. 处理带有缺失值的数据

E. 处理非线性关系的数据

7. 下面哪些方法可以用于特征转换和降维( )

A. 主成分分析

B. 线性判别分析

C. tSNE

D. 随机森林

E. 奇异值分解

四、证明题(每题6分，共18分)

本题得分	

1. 证明朴素贝叶斯分类器在特征独立性假设下的生成模型推导过程，并说明其为何被称为朴素贝叶斯。

2. 证明决策树学习算法的ID3算法的基本原理，推导信息增益公式的计算过程，并讨论其在选择最佳划分属性时的局限性。

3. 证明支持向量机学习算法的原理，推导出支持向量的含义和支持向量机的最优化问题的对偶形式，并解释为何支持向量机可以处理非线性分类问题。

五、计算题(共19分)

本题得分	

你在处理一个房价数据集，其中包含100个房屋的信息，包括房屋面积、卧室数量和房龄，以及对应的房价。你想要使用线性回归模型来进行房价预测。假设模型的形式为：$h_\theta(x)=\theta_0+\theta_1x_1+\theta_2x_2+\theta_3x_3$，其中$x_1$表示房屋面积，$x_2$表示卧室数量，$x_3$表示房龄。

1. 使用梯度下降算法对模型进行训练，并求解出模型的参数θ_0、θ_1、θ_2和θ_3。

2. 在完成模型训练后，使用该模型对下面这个房屋进行预测：$x_1=1650$(平方英尺)、$x_2=3$(卧室数量)、$x_3=20$(房龄)。

3. 根据训练数据集和测试数据集，计算出模型的均方误差和均方根误差。

4. 为了进一步提升模型性能，尝试使用特征缩放方法对数据进行预处理。具体来说，对每个特征进行均值归一化处理，即将每个特征的值减去其均值，再除以其标准差。重新训练模型，并与原模型进行性能比较。