《机器学习》理论课教学大纲

## （适用于四年制生物信息学专业）

### 前 言

本课程的理论教学安排全面覆盖了机器学习的核心领域，从经典算法到最新技术，为学生提供了一个深入浅出的学习路径。通过63学时（42学时理论，21学时实验）的紧凑安排，课程旨在培养学生对机器学习领域的深刻理解以及应用这些技术解决问题的能力。

课程从绪论开始，介绍机器学习的基本概念、历史背景以及其在现代科技中的应用，旨在为学生构建一个坚实的知识基础，并激发他们对后续学习的兴趣。接下来，通过模型评估与选择章节，学生将学习如何正确地评估机器学习模型的性能，理解偏差与方差的权衡，以及如何选择最适合特定问题的模型。

在掌握了评估和选择模型的基本方法后，课程进入到具体的机器学习算法，包括线性模型、决策树、支持向量机（SVM）、集成学习中的随机森林等。这些章节不仅讲解了算法的数学原理和计算方法，还强调了在实际应用中的优势和局限性，帮助学生建立起应用机器学习解决问题的直观感觉。

课程还特别强调了数据预处理和特征工程的重要性，通过降维与度量学习以及特征选择与稀疏矩阵的章节，学生能够学习到如何处理高维数据，提取有效特征，以及如何利用稀疏性来提高模型的计算效率和性能。聚类章节将引导学生进入到机器学习的更高级主题，­­不仅扩展了学生的知识范围，也增加了他们解决复杂问题的工具箱。

在课程的后半部分，通过Pytorch深度学习、神经网络的多个子章节，以及Transformer与自注意机制的深入探讨，学生将有机会学习到当前人工智能领域最前沿的技术。这些章节介绍了深度学习的基础架构、关键技术和算法，以及如何应用于图像处理等领域的应用。

总的来说，本课程通过从基础到高级的渐进式教学，不仅使学生能够掌握机器学习的理论和算法，还能够提升他们的实际应用能力。通过随堂测试和课程总复习，学生能够对所学知识进行巩固和综合运用，为未来的学术研究或职业生涯奠定坚实的基础。

本教学大纲所安排的教学内容是为四年制生物信息学专业的学生而制定。

《机器学习》课程为专业必修课，总学时数为63学时，其中理论课教学的学时数为42。

### 第一章 绪论

1. 目的要求：
2. 了解：机器学习的引言、基本术语以及其在人工智能领域的位置和作用。
3. 熟悉：机器学习的假设空间、归纳偏好的概念及其在模型选择中的重要性。
4. 掌握：机器学习的发展历程、应用现状以及未来发展的可能方向；机器学习算法的基本分类和应用场景。
5. 教学内容：
6. 机器学习的引言和基本术语：介绍机器学习的定义、基本术语和工作机制；探讨机器学习与人工智能、数据挖掘、统计学习的关系。
7. 机器学习的假设空间与归纳偏好：解释假设空间在机器学习中的作用；讨论归纳偏好原则和其对算法泛化能力的影响。
8. 机器学习的发展历程：回顾机器学习的历史发展，从早期的符号学习到现代的深度学习；分析关键技术突破和理论进展。
9. 机器学习的应用现状：概述机器学习在各行各业中的应用实例，如图像识别、自然语言处理、推荐系统等；讨论机器学习技术面临的挑战及未来趋势。
10. 教学学时安排：

2学时

1. 教学方法：

主要采用PPT播放来展示理论内容，辅以板书进行关键点强调。在关键时刻，切换到程序执行页面，让学生实时观察程序演示。

### 第二章 模型评估与选择

1. 目的要求：
2. 了解：模型评估与选择的重要性，经验误差与过拟合现象。
3. 熟悉：不同的模型评估方法及其应用场景，性能度量的类型和意义。
4. 掌握：如何通过比较检验对模型进行科学评估，偏差与方差的概念以及它们对模型性能的影响。
5. 教学内容：
6. 经验误差与过拟合：讨论模型在训练集上的表现与实际泛化能力之间的差异；分析过拟合现象及其成因。
7. 评估方法：介绍留出法、交叉验证法、自助法等模型评估方法，讨论它们的优缺点和适用条件。
8. 性能度量：概述性能度量的基本概念，包括准确率、召回率、F1分数、ROC曲线等，解释它们的计算方法和评价标准。
9. 比较检验：讲解如何通过统计检验方法比较不同模型的性能，包括t检验、方差分析等。
10. 偏差与方差：解释偏差与方差对模型泛化能力的影响，讨论如何通过调整模型复杂度来平衡二者，以达到降低泛化误差的目的。
11. 教学学时安排：

4学时

1. 教学方法：

主要采用PPT播放来展示理论内容，辅以板书进行关键点强调。在关键时刻，切换到程序执行页面，让学生实时观察程序演示。

### 第三章 线性模型

1. 目的要求：
2. 了解：线性模型的基本概念及其在回归和分类任务中的应用。
3. 熟悉：线性回归的最小二乘法，对数几率回归与线性判别分析的原理和方法。
4. 掌握：处理多分类任务的策略（一对一、一对其余、多对多）以及类别不平衡问题的基本解决方法。
5. 教学内容：
6. 线性回归与最小二乘法：介绍线性回归模型的定义和最小二乘法的原理，讨论其在实际问题中的应用及其局限性。
7. 二分类任务的线性模型：
   1. 对数几率回归：讲解对数几率回归的基本思想、模型构建和求解过程，以及如何应用于二分类问题。
   2. 线性判别分析（LDA）：介绍LDA的原理，讨论其如何通过最大化类间距离与最小化类内距离来进行数据分类。
8. 多分类任务的处理策略：探讨一对一、一对其余、多对多等策略在处理多分类问题时的原理、优缺点和适用场景。
9. 类别不平衡问题：分析类别不平衡对模型训练和性能的影响，讨论常用的处理不平衡数据的方法，如重采样技术、成本敏感学习等。
10. 教学学时安排：

4学时

1. 教学方法：

主要采用PPT播放来展示理论内容，辅以板书进行关键点强调。在关键时刻，切换到程序执行页面，让学生实时观察程序演示。

### 第四章 决策树

1. 目的要求：
2. 了解：决策树的基本流程及其核心机制。
3. 熟悉：决策树的划分选择方法、剪枝处理技术以及如何处理连续与缺失值。
4. 掌握：构建多变量决策树的原理和方法。
5. 教学内容：
6. 基本流程：

介绍决策树学习的整体流程，包括特征选择、树的构建、划分选择和最终决策的形成。

1. 划分选择：
   1. 讨论不同的划分选择方法，如信息增益（ID3算法）、信息增益率（C4.5算法）和基尼指数（CART算法）。
   2. 分析这些方法在实际应用中的优缺点及其适用场景。
2. 剪枝处理：
3. 解释剪枝的概念，包括预剪枝和后剪枝的原理与实现。
4. 讨论剪枝如何帮助防止过拟合，提高模型的泛化能力。
5. 连续与缺失值：
6. 介绍如何在决策树中处理连续属性，例如通过设定阈值将其转化为离散值。
7. 讲解决策树如何处理缺失值，包括在属性划分和树构建过程中的策略。
8. 多变量决策树：
9. 介绍多变量决策树的概念，它在节点划分时考虑多个属性。
10. 讨论如何构建多变量决策树，以及它与单变量决策树的差异和优势。
11. 教学学时安排：

2学时

1. 教学方法：

主要采用PPT播放来展示理论内容，辅以板书进行关键点强调。在关键时刻，切换到程序执行页面，让学生实时观察程序演示。

### 第五章 神经网络

1. 目的要求：
2. 了解：神经网络的基础概念，包括神经元模型的工作原理及其在模拟人脑处理信息过程中的作用。
3. 熟悉：感知机与多层网络的结构和功能，以及误差逆传播算法在训练神经网络中的关键作用。
4. 掌握：如何利用误差逆传播算法优化多层神经网络，识别全局最小与局部最小的差异，以及其他常见神经网络模型和深度学习的基本概念与应用。
5. 教学内容：
6. 神经元模型：介绍人工神经元的结构与数学模型，包括激活函数的作用和类型。
7. 感知机与多层网络：讲解单层感知机的原理和局限性，以及多层前馈神经网络如何克服这些局限性，实现复杂函数的逼近。
8. 误差逆传播算法：详细介绍误差逆传播算法（Backpropagation）的工作原理，包括梯度下降法在其中的应用，以及如何通过该算法进行网络权重的更新。
9. 全局最小与局部最小：讨论在神经网络训练过程中可能遇到的全局最小和局部最小问题，以及如何采取策略避免陷入局部最小。
10. 其他常见神经网络：概述除多层前馈网络外的其他常见神经网络类型，如卷积神经网络（CNN）、递归神经网络（RNN）等，以及它们的应用场景。
11. 深度学习：简要介绍深度学习的发展历程、基本概念以及在图像处理、语音识别等领域的应用案例，讨论深度学习相较于传统机器学习方法的优势。
12. 教学学时安排：

3学时

1. 教学方法：

主要采用PPT播放来展示理论内容，辅以板书进行关键点强调。在关键时刻，切换到程序执行页面，让学生实时观察程序演示。

### 第六章 支持向量机

1. 目的要求：
2. 了解：支持向量机（SVM）的基本原理，包括间隔、支持向量的概念以及SVM在分类问题中的应用。
3. 掌握：对偶问题的求解方法，核函数的选择及其在非线性分类问题中的应用，软间隔与正则化的概念，以及支持向量回归的基本思想。
4. 教学内容：
5. 间隔与支持向量：介绍SVM的基本概念，包括如何通过最大化间隔来构建决策边界，以及支持向量在模型中的作用。
6. 对偶问题与核函数：讲解SVM的对偶形式的优势，核函数的引入及其如何使SVM能够有效处理非线性数据。
7. 软间隔与正则化：简述在存在噪声数据时软间隔的必要性及正则化在防止过拟合中的作用。
8. 支持向量回归和核方法：概览支持向量回归（SVR）的原理，以及核方法在其他机器学习任务中的应用示例。
9. 教学学时安排：

2学时

1. 教学方法：

主要采用PPT播放来展示理论内容，辅以板书进行关键点强调。在关键时刻，切换到程序执行页面，让学生实时观察程序演示。

### 第七章 贝叶斯分类器

1. 目的要求：
2. 了解：贝叶斯决策论的基本原理及其在决策过程中的应用。
3. 掌握：朴素贝叶斯分类器的原理和在不同类型数据上的应用。
4. 熟悉：贝叶斯网络的构建及EM算法在参数估计中的使用。
5. 教学内容：
6. 贝叶斯决策论：探讨贝叶斯决策论的原理和在分类问题中的应用。
7. 极大似然估计：介绍极大似然估计方法及其在统计参数估计中的重要性。
8. 朴素贝叶斯分类器：详细解析朴素贝叶斯分类器的工作机制及其实际应用。
9. 半朴素贝叶斯分类器：讲述半朴素贝叶斯分类器的概念及其相对于朴素贝叶斯的优势。
10. EM算法：阐明EM算法的过程及其在贝叶斯估计中处理隐变量问题的能力。
11. 教学学时安排：

3学时

1. 教学方法：

主要采用PPT播放来展示理论内容，辅以板书进行关键点强调。在关键时刻，切换到程序执行页面，让学生实时观察程序演示。

### 第八章 集成学习-随机森林

1. 目的要求：
2. 了解：集成学习的基本概念，包括个体学习器与集成学习器的区别及其重要性。
3. 熟悉：主要的集成学习方法，包括Boosting和Bagging的基本原理及其代表算法Adaboost和随机森林。
4. 掌握：集成学习中的结合策略和多样性的概念，包括如何通过多样性提升集成学习的性能。
5. 教学内容：
6. 个体与集成：介绍集成学习的概念，解释为什么多个学习器结合的预测效果通常要优于单个学习器。
7. Boosting与Adaboost：
   1. Boosting：讲解Boosting的核心思想，即如何通过弱学习器的线性组合逐步逼近复杂模型。
   2. Adaboost：详细介绍Adaboost算法的工作原理，包括权重更新机制和如何通过提高被错分数据的权重来增强模型。
8. Bagging与随机森林：
   1. Bagging：讲解Bagging的基本原理，以及如何通过自助采样法减少过拟合，提高模型的泛化能力。
   2. 随机森林： 深入解析随机森林的构建过程，特别是在Bagging基础上引入特征随机选择来增加树之间的多样性。
9. 结合策略与多样性：
   1. 结合策略： 讨论不同的结合策略，如平均法、投票法及其适用场景。
   2. 多样性： 介绍多样性在集成学习中的重要性，包括误差-分歧分解和多样性度量，以及如何通过多样性扰动提高集成学习的性能。
10. 教学学时安排：

2学时

1. 教学方法：

主要采用PPT播放来展示理论内容，辅以板书进行关键点强调。在关键时刻，切换到程序执行页面，让学生实时观察程序演示。

### 第九章 聚类

1. 目的要求：
2. 了解：聚类任务的定义、类型及其在数据挖掘中的应用。
3. 熟悉：聚类性能度量的方法，包括内部评价指标和外部评价指标，以及距离计算在聚类中的作用和不同距离计算方法。
4. 掌握：原型聚类、密度聚类和层次聚类的基本原理，算法实现及其适用场景。
5. 教学内容：
6. 聚类任务：介绍聚类分析的基本任务和目标，讨论聚类与分类的区别，以及聚类在各种实际应用中的作用。
7. 性能度量与距离计算：
   1. 性能度量： 讲解如何评价聚类效果的好坏，包括轮廓系数、Davies-Bouldin指数等内部指标，以及兰德指数等外部指标的计算方法和应用。
   2. 距离计算： 介绍聚类中常用的距离度量方法，如欧氏距离、曼哈顿距离和余弦相似度等，讨论它们的特点和适用条件。
   3. 原型聚类：深入讲解K-means和学习向量量化（LVQ）等原型聚类方法的原理和算法过程，讨论它们的优缺点及适用场景。
8. 密度聚类： 详细介绍基于密度的聚类方法，如DBSCAN和OPTICS算法的工作原理，强调如何根据样本密度进行聚类以及如何处理噪声点。
9. 层次聚类：讲解层次聚类的概念和构建过程，包括AGNES（自底向上）和DIANA（自顶向下）等算法的特点，以及如何通过树状图（树图）展示聚类结构。
10. 教学学时安排：

6学时

1. 教学方法：

主要采用PPT播放来展示理论内容，辅以板书进行关键点强调。在关键时刻，切换到程序执行页面，让学生实时观察程序演示。

### 第十章 降维与度量学习

1. 目的要求：
2. 了解：降维与度量学习的基本概念及其在数据预处理和特征提取中的重要性。
3. 熟悉：主成分分析（PCA）、低维嵌入和流形学习等降维方法的原理及应用场景。
4. 掌握：k近邻学习在度量学习中的作用，以及如何通过度量学习改善数据的相似性度量，提升学习模型的性能。
5. 教学内容：
6. k近邻学习：介绍k近邻算法的基本原理和工作机制，以及在降维和度量学习中的应用。
7. 低维嵌入与主成分分析：
   1. 低维嵌入： 讲解如何将高维数据映射到低维空间，以及这一过程对数据可视化和解释性的重要性。
   2. 主成分分析（PCA）： 详细介绍PCA的数学原理、计算过程及其在数据压缩和去噪中的应用。
8. 流形学习：概述流形学习的概念，探讨如何通过假设数据存在于低维流形上来发现数据的内在结构。
9. 度量学习：讨论度量学习的目的和方法，包括如何学习数据点间的有效距离度量，以改善分类或聚类的性能。。
10. 教学学时安排：

3学时

1. 教学方法：

主要采用PPT播放来展示理论内容，辅以板书进行关键点强调。在关键时刻，切换到程序执行页面，让学生实时观察程序演示。

### 第十一章 特征选择与稀疏矩阵

1. 目的要求：
2. 了解：特征选择的基本概念、目的及其在数据预处理中的重要性。
3. 熟悉：不同特征选择方法的原理和分类，包括过滤式选择、包裹式选择、嵌入式选择及L\_1正则化的应用。
4. 掌握：稀疏表示的基本理论，包括字典学习和压缩感知的概念、原理及其在信号处理和图像处理中的应用。
5. 教学内容：
6. 特征选择的基本概念与方法：
7. 子集选择与评价： 介绍特征子集选择的过程和子集评价的标准。
8. 过滤式选择： 讲解过滤式方法的原理，强调其如何依据统计特性独立于学习算法进行特征选择。
9. 包裹式选择： 详述包裹式方法的工作机制，包括如何围绕学习器性能选择特征子集。
10. 嵌入式选择与L\_1正则化： 解释嵌入式特征选择的概念，特别是L\_1正则化在特征选择中如何促进稀疏性。
11. 稀疏表示与其学习方法：
12. 字典学习： 介绍字典学习的基本原理和目标，讨论其在实现数据的稀疏表示中的应用。
13. 压缩感知： 讲解压缩感知的理论基础，包括稀疏信号的恢复过程和技术要求，以及其在信号处理领域的重要性。
14. 教学学时安排：

3学时

1. 教学方法：

主要采用PPT播放来展示理论内容，辅以板书进行关键点强调。在关键时刻，切换到程序执行页面，让学生实时观察程序演示。

### 提升一 Pytorch 预训练网络

1. 目的要求：
2. 了解：预训练模型的基本概念、优势及其在深度学习应用中的重要性。
3. 熟悉：PyTorch Hub的功能、用途和操作流程，以及如何通过PyTorch Hub访问和使用预训练模型。
4. 掌握：使用预训练模型进行图像分类、对象检测和语义分割的基本方法，包括模型的加载、应用和简单调整。
5. 教学内容：
6. 预训练模型的概念和优势：
7. 概念介绍： 解释什么是预训练模型，以及为何在深度学习项目中使用预训练模型。
8. 优势讨论： 讨论使用预训练模型的优势，如节省训练时间、需要的数据量少、在特定任务上可能达到更好的性能等。
9. PyTorch Hub简介：
10. 功能和用途： 介绍PyTorch Hub的主要功能和用途，包括它如何帮助研究者和开发者访问和共享模型。
11. 操作指南： 讲解如何使用PyTorch Hub搜索、下载和加载预训练模型的步骤和技巧。
12. 使用预训练模型进行深度学习应用：
13. 图像分类： 介绍如何使用ResNet、AlexNet等预训练模型进行图像分类任务，包括如何处理输入和解释输出。
14. 对象检测和语义分割： 解释对象检测和语义分割的基本概念，讨论如何应用预训练模型完成这些任务，以及对应模型的选择和应用示例。
15. 常用预训练模型介绍：
16. 模型概览： 提供ResNet、AlexNet、VGGNet等常见预训练模型的简介，包括它们的特点、适用场景和性能优势。
17. 教学学时安排：

4学时

1. 教学方法：

主要采用PPT播放来展示理论内容，辅以板书进行关键点强调。在关键时刻，切换到程序执行页面，让学生实时观察程序演示。

### 提升二 transformer 与自注意力机制

1. 目的要求：
2. 了解：Transformer模型的基本架构及其在自然语言处理（NLP）领域的革命性影响。
3. 熟悉：自注意力机制的工作原理及其如何使Transformer能够高效处理序列数据。
4. 掌握：Transformer模型的关键技术细节，包括位置编码、多头注意力机制、前馈网络、层归一化和残差连接。
5. 应用：理解Transformer模型的多种应用场景，包括机器翻译、文本摘要、问答系统等，并学习如何通过现有的深度学习框架实现Transformer模型。
6. 教学内容：
7. Transformer模型介绍：
8. 背景和动机： 介绍Transformer的发展背景，包括处理序列数据的传统方法（如RNN、LSTM）的局限性，以及为何需要Transformer。
9. 基本架构： 解析Transformer的整体架构，包括编码器和解码器的设计及其功能。
10. 自注意力机制：
11. 原理和计算过程： 详细讲解自注意力机制的原理，包括如何计算注意力权重和如何根据这些权重来更新序列信息。
12. 多头注意力： 解释多头注意力的概念及其在提高模型捕捉序列中不同子空间信息能力中的作用。
13. Transformer模型关键技术：
14. 位置编码： 讲解位置编码的必要性及其如何使模型能够理解序列中的顺序信息。
15. 前馈网络、层归一化和残差连接： 分析这些技术在模型中的作用及其对提高模型性能和稳定性的贡献。
16. Transformer模型的应用与实现：
17. 介绍Transformer在NLP各领域的应用。
18. 提供基于流行深度学习框架PyTorch，实现Transformer模型的基本步骤和建议。
19. 教学学时安排：

4学时

1. 教学方法：

主要采用PPT播放来展示理论内容，辅以板书进行关键点强调。在关键时刻，切换到程序执行页面，让学生实时观察程序演示。

# 《机器学习》实验课教学大纲

## （适用于四年制生物信息专业）

### 前 言

本实验课程旨在加深学生对机器学习理论的理解，并提升他们应用这些理论解决实际问题的能力。课程内容涵盖了从基础工具的使用、数据处理与可视化技术，到经典机器学习算法的应用，最后深入到深度学习的核心技术和框架，基本覆盖了数据科学与机器学习领域的关键技能。

课程开始于数据科学项目必备的环境配置和工具，为学生进行后续实验操作提供了坚实的基础。随后，学生将通过实践学习如何利用主流的数据处理库进行数据的预处理、分析和可视化，这为理解数据特性和后续模型训练奠定了基础。在掌握了数据处理的基本技能后，课程逐渐引导学生进入机器学习模型的世界。通过一系列实验，学生将依次实践分类、回归、决策树、随机森林、支持向量机等多种机器学习算法，每个实验都旨在帮助学生理解算法的工作原理及其在实际中的应用场景。

进入到课程的高级部分，学生将探索降维、特征选择、聚类等高级技术，学习如何在面对高维数据和复杂问题时，有效地提取信息、降低计算复杂度和发现数据内在的结构。最终，课程深入深度学习领域，通过深入浅出的方式学习深度神经网络的构建、训练方法，包括但不限于卷积神经网络和反向传播算法。学生将通过使用先进的深度学习框架，如Pytorch，来实现复杂的网络模型，并解决实际问题。

整个实验课程的设计旨在通过实践操作强化学生的技术技能，促进他们对机器学习理论与应用的深入理解。实验课教学的学时数为21学时。

### 实验一 MobaXterm, Conda, Jupyter notebook

一、实验目的

1. 了解：Mobaterm工具的基本功能及其在远程登录中的应用。
2. 熟悉：如何使用Mobaterm远程登录到教学服务器，并理解远程操作的基本概念。
3. 掌握：在远程服务器上下载、安装Conda的步骤，以及配置Jupyter Notebook环境的方法，为《机器学习》课程的学习和实验准备必要的软件环境。

二、实验内容

1. 下载和安装Mobaterm：
2. 工具介绍： 简要介绍Mobaterm的主要功能和使用场景。
3. 安装指南： 指导如何根据电脑型号下载和安装Mobaterm软件。
4. 使用Mobaterm远程登录教学服务器：
5. 远程登录流程： 详细讲解使用Mobaterm远程登录到教学服务器的具体步骤。
6. 基本操作： 简介远程环境下的基本命令操作，为后续实验操作做准备。
7. 在服务器上安装Conda：
8. Conda简介： 解释Conda的作用及其在数据科学和机器学习中的重要性。
9. 安装步骤： 逐步指导如何在服务器上下载并安装Conda，包括选择合适的版本和配置环境变量。
10. 配置Jupyter Notebook环境：
11. 环境配置： 讲解如何使用Conda创建一个包含必要库的虚拟环境，并在该环境中安装Jupyter Notebook。
12. 启动和访问： 指导如何启动Jupyter Notebook服务，并通过浏览器访问Jupyter Notebook，准备开始《机器学习》课程的学习和实验。

通过本次实验课程，学生将能够独立完成远程登录、软件安装和环境配置，为《机器学习》课程的后续学习打下坚实的基础。

三、学时安排

1 学时

#### 四、实验教学方法

在机房中，教师讲解实验内容后，学生独立操作。教师巡视提供辅导，如遇普遍问题，再为全班集体讲解，确保实践与理论相结合。

### 实验二 Numpy, Matplotlib, Scipy, Pandas

#### 一、实验目的

1. 了解：为何Python及其库scikit-learn成为机器学习领域的首选工具。
2. 熟悉：Python机器学习生态系统中的核心库和工具，包括Jupyter Notebook、Numpy、SciPy、matplotlib、pandas和mglearn。
3. 掌握：如何配置和使用这些工具进行基本的数据处理、分析和机器学习模型的构建。

#### 二、实验内容

1. 选择Python的原因：
2. Python的优势： 讨论Python语言在简洁性、易读性和社区支持方面的优势。
3. scikit-learn的重要性： 介绍scikit-learn库在提供高效、可访问的机器学习算法方面的作用。
4. 核心库和工具简介：
5. Jupyter Notebook： 演示如何使用Jupyter Notebook进行交互式编程和文档化。
6. Numpy和SciPy： 介绍这两个库在数值计算和科学计算中的应用。
7. matplotlib和pandas： 展示如何利用matplotlib进行数据可视化，以及如何使用pandas进行数据处理和分析。
8. mglearn： 说明mglearn库如何辅助演示机器学习概念和算法。
9. 实验操作：
10. 环境配置： 指导如何安装Python和上述库，包括推荐的安装方法和环境设置。
11. 基本操作演练： 通过简单示例演示如何使用Jupyter Notebook创建和运行代码单元，如何使用Numpy进行数组操作，以及如何用matplotlib绘制图表。
12. 初步应用： 引导学生使用这些工具加载数据集、进行简单的数据分析和可视化，以及运行一个简单的scikit-learn机器学习模型。

通过本次实验，学生将能够理解Python及其关键库在机器学习领域内的应用价值，掌握基础的数据处理、分析和模型构建技能。

#### 三、实验学时安排

1 学时

#### 四、实验教学方法

在机房中，教师讲解实验内容后，学生独立操作。教师巡视提供辅导，如遇普遍问题，再为全班集体讲解，确保实践与理论相结合。

### 实验三 鸢尾花分类

#### 一、实验目的

1. 了解：机器学习项目的基本流程，以鸢尾花分类问题为例。
2. 熟悉：如何对数据进行初步的探索和处理，以及如何划分训练数据和测试数据。
3. 掌握：构建基于K近邻算法的机器学习模型，进行预测，并通过实际结果评估模型的性能。

#### 二、实验内容

1. 初识数据：
2. 数据集简介： 介绍鸢尾花数据集的基本情况，包括数据特征和目标分类。
3. 数据观察： 指导如何使用pandas和matplotlib查看数据分布和特征关系，以获得对数据的直观理解。
4. 衡量模型是否成功：
5. 训练数据和测试数据： 讲解如何使用scikit-learn划分训练集和测试集，并说明为什么这样做是必要的。
6. 观察数据的重要性： 强调在构建模型前对数据进行观察和分析的重要性，以及如何从观察中获取模型构建的线索。
7. 构建第一个模型：
8. K近邻算法介绍： 简要介绍K近邻算法的原理及其在分类问题中的应用。
9. 模型构建： 指导如何使用scikit-learn库构建K近邻模型，并选择合适的K值。
10. 做出预测，评估模型：
11. 进行预测： 演示如何使用训练好的模型对测试数据进行分类预测。
12. 评估模型： 介绍如何使用准确率等指标评估模型性能，并通过实验结果讨论模型表现及可能的改进方向。

通过本次实验，学生将通过鸢尾花分类问题的实际操作，了解和掌握使用机器学习算法解决分类问题的基本步骤和方法，为后续深入学习机器学习算法打下坚实的基础。

#### 三、实验学时安排

1 学时

#### 四、实验教学方法

在机房中，教师讲解实验内容后，学生独立操作。教师巡视提供辅导，如遇普遍问题，再为全班集体讲解，确保实践与理论相结合。

### 实验四 k近邻算法

#### 一、实验目的

1. 了解：不同类型的数据集（模拟和实际）及其在机器学习中的应用。
2. 熟悉：k近邻（k-NN）学习算法的原理及如何在不同数据集上应用k-NN进行分类和回归任务。
3. 掌握：如何使用Python和scikit-learn库加载数据集、应用k-NN模型，并对模型性能进行评估。

#### 二、实验内容

1. 数据集介绍：
2. 模拟数据集： 介绍forge（分类问题）和wave（回归问题）数据集的特点和用途。
3. 实际数据集： 解释威斯康辛州乳腺癌数据集（cancer）和波士顿房价数据集的来源、特征及其在机器学习中的重要性。
4. k近邻学习算法原理：
5. k-NN概述： 讲解k-NN算法的工作原理，包括距离度量、如何选择k值以及分类与回归的决策规则。
6. k-NN的优缺点： 讨论k-NN在不同类型数据集上的表现及其优缺点。
7. 实验操作：
8. 加载数据集： 指导如何使用scikit-learn库加载模拟和实际数据集。
9. 应用k-NN模型： 演示如何使用scikit-learn实现k-NN算法，包括模型的初始化、训练和预测。
10. 模型评估： 介绍如何使用准确率（对于分类问题）和均方误差（对于回归问题）等指标评估k-NN模型的性能。
11. 参数调优： 实验不同的k值对模型性能的影响，引导学生理解如何选择最优的k值。

通过本次实验，学生将能够深入理解k近邻算法的原理和应用，掌握如何在Python环境下使用scikit-learn库对不同类型的数据集进行机器学习任务的处理和分析。

#### 三、学时安排

1 学时

#### 四、实验教学方法

在机房中，教师讲解实验内容后，学生独立操作。教师巡视提供辅导，如遇普遍问题，再为全班集体讲解，确保实践与理论相结合。

### 实验五 线性模型

#### 一、实验目的

1. 了解：线性模型在回归和分类问题中的应用，包括线性回归、岭回归、Lasso回归以及线性模型在分类和多分类问题中的使用。
2. 熟悉：线性回归（最小二乘法）的原理，以及岭回归和Lasso回归如何通过正则化来解决过拟合问题。
3. 掌握：如何使用scikit-learn库实现这些线性模型，并应用它们解决实际问题，包括参数选择、模型训练、预测以及性能评估。

#### 二、实验内容

1. 用于回归的线性模型：
2. 线性回归（最小二乘法）：介绍线性回归模型的数学原理和应用，演示如何使用scikit-learn实现线性回归，并在一个回归问题上应用它。
3. 岭回归：讲解岭回归在最小二乘法基础上如何通过L2正则化解决高维数据的过拟合问题，演示如何选择合适的正则化参数。
4. Lasso回归：介绍Lasso回归如何使用L1正则化进行特征选择，从而产生一个稀疏模型，以及如何在实践中应用Lasso回归。
5. 用于分类的线性模型：
6. 逻辑回归：解释逻辑回归在解决二分类问题中的原理和应用，包括如何使用scikit-learn实现逻辑回归。
7. 线性支持向量机（Linear SVM）：讲解线性SVM用于二分类问题的原理，演示如何使用scikit-learn实现线性SVM，并讨论参数调整的策略。
8. 用于多分类的线性模型：
9. 多分类策略：介绍逻辑回归和线性SVM如何扩展到多分类问题，包括一对多（One-vs.-Rest）和一对一（One-vs.-One）策略。
10. 实验操作：演示如何在一个多分类问题上应用逻辑回归和线性SVM，包括数据准备、模型训练、预测和性能评估。

通过本次实验，学生将获得对线性模型在回归和分类问题中应用的深入理解，掌握使用scikit-learn实现这些模型的技能，并能够在实际数据集上训练、预测和评估模型的性能。

#### 三、学时安排

1 学时

#### 四、实验教学方法

在机房中，教师讲解实验内容后，学生独立操作。教师巡视提供辅导，如遇普遍问题，再为全班集体讲解，确保实践与理论相结合。

### 实验六 决策树

#### 一、实验目的

1. 了解：决策树模型的基本原理及其在分类和回归问题中的应用。
2. 熟悉：如何使用scikit-learn构造决策树模型，包括选择合适的参数来控制模型的复杂度。
3. 掌握：分析决策树模型的能力，包括如何解释模型的决策过程和评估特征的重要性。

#### 二、实验内容

1. 构造决策树：
2. 基本概念：介绍决策树的工作原理，包括如何基于数据特征进行递归分裂来构建树。
3. 实验操作：演示如何使用scikit-learn中的DecisionTreeClassifier或DecisionTreeRegressor在给定数据集上构建决策树模型。
4. 控制决策树的复杂度：
5. 参数调整：讲解如何通过调整参数（如max\_depth、min\_samples\_split等）控制决策树的复杂度，以防止过拟合。
6. 实验操作：指导学生在实际数据集上调整这些参数，并观察模型性能的变化。
7. 分析决策树：
8. 树的可视化：介绍如何使用工具（如plot\_tree函数）可视化决策树，以理解模型的决策过程。
9. 实验操作：学生将练习生成决策树的可视化图，并学习如何解读这些图。
10. 树的特征重要性：
11. 重要性评估：解释什么是特征重要性以及如何使用决策树模型评估特征的重要性。
12. 实验操作：指导学生如何获取和解释模型的特征重要性，以及如何利用这些信息进行特征选择或模型解释。

通过本次实验，学生将能够独立构建和调优决策树模型，理解模型的工作原理，以及如何利用决策树模型进行数据分析和特征重要性评估。此外，通过实际操作，加深对决策树模型在机器学习中应用的理解。

#### 三、学时安排

1 学时

#### 四、实验教学方法

在机房中，教师讲解实验内容后，学生独立操作。教师巡视提供辅导，如遇普遍问题，再为全班集体讲解，确保实践与理论相结合。

### 实验七 随机森林

#### 一、实验目的

1. 了解：随机森林算法的基本原理及其如何通过集成学习提高模型性能。
2. 熟悉：如何使用scikit-learn构建随机森林模型，并对数据进行自助采样（Bootstrap sampling）。
3. 掌握：在不同数据集上应用随机森林模型的方法，分析模型的优缺点以及如何调整随机森林的参数以获得最佳性能。

#### 二、实验内容

1. 构造随机森林：
2. 自助采样：介绍自助采样的概念及其在构造随机森林中的作用。
3. 实验操作：演示如何使用scikit-learn的RandomForestClassifier或RandomForestRegressor在给定数据集上构建随机森林模型。
4. 分析随机森林：
5. 应用于two\_moons数据集：使用一个较小的随机森林（例如5棵树）在two\_moons数据集上进行训练和预测，观察模型的决策边界。
6. 应用于乳腺癌数据集：使用一个较大的随机森林（例如100棵树）在乳腺癌数据集上进行训练和预测，分析模型的性能。
7. 实验操作：学生将分别在这两个数据集上应用随机森林模型，并通过实验比较不同数量的树对模型性能的影响。
8. 优缺点和参数分析：
9. 优缺点讨论：讨论随机森林算法的优点（如鲁棒性、减少过拟合风险等）与可能的缺点（如模型复杂度、计算成本等）。
10. 参数调整：介绍影响随机森林性能的关键参数（如n\_estimators、max\_depth、max\_features等），并指导学生如何调整这些参数以优化模型。

实验操作：学生将通过调整随机森林的参数，实践如何提高模型在特定数据集上的表现。

通过本次实验，学生将获得对随机森林算法深入的理解，包括其工作原理、如何实现、以及如何通过参数调整优化模型性能。此外，还将通过对比不同数据集和模型配置的实验，更全面地理解随机森林在实际应用中的优势和局限性。

#### 三、学时安排

1 学时

#### 四、实验教学方法

在机房中，教师讲解实验内容后，学生独立操作。教师巡视提供辅导，如遇普遍问题，再为全班集体讲解，确保实践与理论相结合。

### 实验八 支持向量机

#### 一、实验目的

1. 了解：支持向量机（SVM）的基本原理及其如何通过添加非线性特征和使用核技巧处理线性不可分问题。
2. 熟悉：SVM的核技巧在实际应用中的作用，以及如何选择合适的核函数。
3. 掌握：如何对SVM模型进行调参（包括正则化参数C和核参数），以及如何为SVM模型进行适当的数据预处理。

#### 二、实验内容

1. 线性模型与非线性特征：在简单的线性可分数据集上应用线性SVM模型，然后通过添加交互项和多项式特征，观察和分析模型性能的变化。
2. 核技巧：使用SVM的不同核函数（如线性核、多项式核和RBF核）在非线性可分数据集上训练模型，并比较它们的性能差异。
3. 理解SVM：构建并分析SVM模型的决策边界，讨论支持向量在模型中的作用和重要性。
4. SVM调参：调整SVM模型的正则化参数C和核参数（如gamma），使用网格搜索等技术找到最佳参数组合，分析参数变化对模型性能的影响。
5. 为SVM预处理数据：在应用SVM模型之前，对数据进行标准化处理，比较预处理前后模型的性能，理解数据预处理对SVM性能的影响。
6. 优缺点总结：总结SVM的主要优点（如在高维空间中的有效性、适用于复杂问题）和缺点（如对参数选择敏感、计算成本高），通过实际例子讨论SVM在特定场景下的适用性。

通过完成这些实验内容，学生将能够深入理解支持向量机的工作原理，掌握如何在实际问题中有效地应用SVM，包括如何处理非线性问题、进行模型调参和数据预处理。

#### 三、学时安排

1 学时

#### 四、实验教学方法

在机房中，教师讲解实验内容后，学生独立操作。教师巡视提供辅导，如遇普遍问题，再为全班集体讲解，确保实践与理论相结合。

### 实验九 神经网络模型与调参

#### 一、实验目的

1. 了解：多层感知机（MLP）的基本结构和工作原理，以及如何在实际问题中应用MLP模型。
2. 熟悉：神经网络的调参技巧，包括隐层的数量、每个隐层的单元数以及正则化参数（alpha）对模型性能的影响。
3. 掌握：通过实验观察和分析不同参数设置下神经网络学到的决策边界的变化，以及如何选择合适的参数以优化模型性能。

#### 二、实验内容

1. 神经网络模型：构建一个包含两个隐层的多层感知机模型，使用scikit-learn的MLPClassifier在简单数据集上训练并分析模型性能。
2. 将MLPClassifier应用到two\_moons数据集上，分别构建包含不同数量隐单元和隐层的神经网络模型，通过实验图解展示不同模型学到的决策边界。
3. 在two\_moons数据集上继续实验，调整隐单元个数与alpha参数的不同设定，通过实验图解展示不同参数配置下模型的决策函数和性能变化。
4. 基于实验结果，分析隐层数量、隐单元个数和正则化参数（alpha）对模型性能的具体影响，总结在调整神经网络参数时的最佳实践和策略。

通过本次实验，学生将通过实际操作深入了解神经网络（特别是多层感知机）的基本概念和调参技巧。学习如何构建和调整MLP模型，以及通过实验观察和分析模型学到的决策边界和性能。

#### 三、学时安排

2 学时

#### 四、实验教学方法

在机房中，教师讲解实验内容后，学生独立操作。教师巡视提供辅导，如遇普遍问题，再为全班集体讲解，确保实践与理论相结合。

### 实验十 Pytorch 预训练网络

#### 一、实验目的

1. 了解：如何利用PyTorch框架获取并使用预训练网络进行图像识别任务。
2. 熟悉：深度学习中的预训练模型，如AlexNet和ResNet，以及它们在图像识别中的应用。
3. 掌握：生成对抗网络（GAN）和CycleGAN的基本原理及其在图像转换任务中的应用。
4. 应用：通过实际操作，探索预训练模型在不同深度学习任务中的效果，包括图像识别和图像风格转换。

#### 二、实验内容

1. 一个识别图像主体的预训练网络：获取一个预训练网络，如AlexNet或ResNet，介绍这些网络的基本架构和特点；运行模型进行图像识别，展示模型在标准图像数据集（如ImageNet的一部分）上的识别效果。
2. 一个足以以假乱真的预训练模型：探索GAN游戏的基本原理，使用预训练的GAN模型，如CycleGAN，进行图像风格转换的实验。介绍GAN的工作机制，包括生成器和判别器的角色。
3. 一个把马变为斑马的网络：使用CycleGAN进行特定图像转换任务——将马变为斑马。介绍CycleGAN的架构及其如何从ImageNet数据集中训练得到。演示图像转换过程，并分析转换效果。

通过本次实验，学生将通过实际操作深入理解预训练网络和生成对抗网络（GAN）的工作原理及其在实际深度学习任务中的应用。学会在PyTorch环境中加载和使用预训练模型，以及如何调整和应用这些模型解决图像识别和图像转换等问题。

#### 三、学时安排

2 学时

#### 四、实验教学方法

在机房中，教师讲解实验内容后，学生独立操作。教师巡视提供辅导，如遇普遍问题，再为全班集体讲解，确保实践与理论相结合。

### 实验十一 降维与度量学习

#### 一、实验目的

1. 深入理解和实践字典及其应用
2. 掌握递归函数的原理和应用
3. 通过实验加深对于divide-and-conquer过程的理解

#### 二、实验内容

1. 字典操作：按根据学生的分数返回相应的字母成绩，要求完成letter\_grades函数，并测试。
2. 递归函数I：使用递归计算给定值在列表中出现的次数，完成sum\_list函数。
3. 递归函数II：使用递归在列表中替换所有给定值，完成replace函数；使用递归从列表中删除给定值的第一个出现，完成remove\_first函数。

#### 三、学时安排

3 学时

#### 四、实验教学方法

在机房中，教师讲解实验内容后，学生独立操作。教师巡视提供辅导，如遇普遍问题，再为全班集体讲解，确保实践与理论相结合。

### 实验十二 聚类（1）

#### 一、实验目的

1. 了解：K均值聚类算法的基本原理及其在无监督学习中的应用。
2. 熟悉：K均值聚类的三个基本步骤，以及如何通过scikit-learn实现K均值聚类。
3. 掌握：K均值聚类在不同数据分布情况下的性能，包括其优点和局限性。
4. 应用：利用K均值进行矢量量化，探索使用更多簇来对数据进行编码的效果。

#### 二、实验内容

1. 输入数据与K均值算法的三个步骤：选择一个适合的数据集，展示如何应用K均值聚类算法进行数据聚类。详细介绍算法初始化、迭代更新簇中心以及分配数据点到最近簇中心的步骤。
2. K均值算法找到的簇中心和簇边界：使用scikit-learn的KMeans类对数据集进行聚类，可视化聚类结果，包括簇中心的位置和簇边界。
3. 簇的密度不同，K均值找到的簇分配：在密度不同的数据集上应用K均值聚类，观察并分析K均值如何将数据点分配到各个簇，讨论其对密度不同簇的处理能力。
4. K均值无法识别非球形簇：选择一个包含非球形簇的数据集，应用K均值聚类并可视化结果，分析K均值聚类在处理非球形簇时的局限性。
5. K均值做矢量量化：通过设置比输入维度更多的簇数，展示如何使用K均值聚类对数据进行矢量量化。讨论矢量量化在数据压缩和特征提取中的应用。

通过本次实验，学生将深入理解K均值聚类算法的工作原理和应用场景，包括其在处理不同类型数据时的优势和局限。通过一系列的实验操作，学生将学会如何在实际问题中应用K均值聚类，如何解释聚类结果，以及如何根据具体的应用需求调整聚类策略。此外，通过矢量量化的实验，学生将能够探索K均值聚类在数据编码和特征提取方面的潜在用途。

#### 三、学时安排

2学时

#### 四、实验教学方法

在机房中，教师讲解实验内容后，学生独立操作。教师巡视提供辅导，如遇普遍问题，再为全班集体讲解，确保实践与理论相结合。

### 实验十三 聚类（2）

#### 一、实验目的

1. 掌握面向对象编程的基本概念：理解类、对象、属性和方法的核心思想；学习如何在Python中定义和使用类。
2. 掌握生物信息学中DNA序列分析的基本功能：让学生对生物信息学中DNA序列分析有个初步了解，知道其需要实现的核心功能。

#### 二、实验内容

1. 类与对象的定义：定义一个名为 DNASequence 的类，代表DNA序列。
2. 基本属性与方法：属性：为 DNASequence 类定义基本属性，如序列字符串 sequence 和序列名称 name。方法：为 DNASequence 类定义方法，如计算GC含量的 gc\_content 方法和获取反向互补序列的 reverse\_complement 方法。
3. 测试与调试：实例化 DNASequence 类的对象，并调用其方法以确保代码无误。

#### 三、学时安排

2学时

#### 四、实验教学方法

在机房中，教师讲解实验内容后，学生独立操作。教师巡视提供辅导，如遇普遍问题，再为全班集体讲解，确保实践与理论相结合。

### 实验十四 特征选择

#### 一、实验目的

1. 了解：如何对分类数据进行预处理，包括One-Hot编码和虚拟变量的创建，以及数字编码分类变量的方法。
2. 熟悉：分箱（离散化）技术对于线性模型和树模型性能的影响，以及如何在实际数据集上应用这些技术。
3. 掌握：如何通过添加交互特征和多项式特征增强模型的表示能力，以及这些技术对模型性能的具体影响。

#### 二、实验内容

1. 分类数据处理：选择一个包含分类特征的数据集，演示如何使用One-Hot编码和虚拟变量转换这些特征。讨论数字编码分类变量时的注意事项。
2. 分箱、离散化、线性模型与树：在wave数据集上比较线性回归和决策树的性能；对wave数据集的特征进行分箱处理，再次比较线性回归和决策树在分箱特征上的性能差异。
3. 交互特征与多项式特征：在包含分箱特征的wave数据集上应用线性回归，首先不考虑交互特征，观察模型性能；引入交互特征，使每个箱子具有不同的斜率，再次应用线性回归，分析性能变化；对wave数据集应用具有10次多项式特征的线性回归，比较模型性能，并讨论多项式特征的优缺点。

通过本次实验，学生将能够深入理解特征预处理和特征工程的重要性及其在提升模型性能中的作用。学生将学会如何针对不同类型的数据和模型选择合适的数据转换技术，并通过实际操作来掌握这些技术的应用。

#### 三、学时安排

2 学时

#### 四、实验教学方法

在机房中，教师讲解实验内容后，学生独立操作。教师巡视提供辅导，如遇普遍问题，再为全班集体讲解，确保实践与理论相结合。