**《机器学习》课程教学大纲**

**一、课程简介**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **课程中文名** | 机器学习 | | | | | | |
| **课程英文名** | **Machine learning** | | | | **双语授课** | | □是 ☑否 |
| **课程代码** | 24122094 | **课程学分** | **4** | **总学时数** | | 64（含实践32） | |
| **课程类别** | □通识教育课程  □公共基础课程  ☑专业教育课程  □综合实践课程  □教师教育课程 | **课程性质** | □必修  ☑选修  □其他 | **课程形态** | | □线上  ☑线下  □线上线下混合式  □社会实践  □虚拟仿真实验教学 | |
| **考核方式** | □闭卷 □开卷 □课程论文 ☑课程作品 ☑汇报展示 □报告  ☑课堂表现 ☑阶段性测试 ☑平时作业 ☑其他 （可多选） | | | | | | |
| **开课学院** | 大数据与智能工程学院 | | **开课**  **系(教研室)** | 计算机科学与技术系 | | | |
| **面向专业** | 数据科学与大数据工程 | | **开课学期** | 第5学期 | | | |
| **课程负责人** | 张素兰 | | **审核人** | 曾俊，邢昌元，黄金龙 | | | |
| **先修课程** | 数据结构与算法、Python程序设计 | | | | | | |
| **后续课程** | 大数据应用开发综合实训 | | | | | | |
| **选用教材** | 1.安德里亚斯·穆勒、莎拉·吉多著，张亮译，Python机器学习基础教程[M]，北京：清华大学出版社，2018. | | | | | | |
| **参考书目** | 1.周志华，机器学习[M]，北京：清华大学出版社，2016  2.塞巴斯蒂安·拉施卡著，陈斌译，Python机器学习[M]，北京：机器工业出版社，2021 | | | | | | |
| **课程资源** | https://www.educoder.net/paths/ijeknthr | | | | | | |
| **课程简介** | 机器学习是数据科学与大数据专业选修课，主要介绍机器学习的基本概念、有监测学习、无监测学习、半监测学习、神经网络等人工智能方法。本课程是《数据结构与算法》课程的延伸，在学生智能信息处理及分析能力培养中处于较为重要地位。通过该门课程的学习，不但可以验证、巩固理论知识，加深对机器学习算法的理解，而且可以提高学生分析问题、解决实际问题的能力，培养实事求是的科学态度和创新精神，为后续课程的学习和将来就业、科研工作打下基础。 | | | | | | |

**二、课程目标**

**表 2-1 课程目标**

|  |  |
| --- | --- |
| **序号** | **具体课程目标** |
| **课程目标 1** | 能够陈述常用机器学习算法的原理，习得机器学习算法构建方法，阐明机器学习过程模型，能根据数据科学与大数据应用领域数据特征，选择特定机器学习模型，对模型进行训练，对未知数据进行预测，对模型进行结果进行评估，并能描述机器学习领域的最新技术和发展趋势。 |
| **课程目标 2** | 具备运用机器学习方法、数据可视化方法及相关分析工具，对数据科学与大数据复杂工程问题进行分析，运用创新思维，构建机器学习模型，逐步形成智能数据分析与处理的能力，具备对机器学习模型结果进行归纳和总结能力，逐步形成科学的学习观和方法论。 |
| **课程目标** 3 | 紧跟机器学习领域的前沿技术，掌握最新工具的使用方法，逐步养成严谨的科学态度、积极向上的价值观和终身学习的精神，为未来的学习、工作和生活奠定基础。 |

**表2-2 课程目标与毕业要求对应关系**

| **毕业要求** | **指标点** | **课程目标** |
| --- | --- | --- |
| **1.工程知识：**能够将数学、自然科学、工程基础、数据科学与大数据技术专业知识用于解决大数据应用领域的复杂工程问题。【L】 | 1.3 能够运用数学、自然科学和专业知识对大数据应用领域相关工程问题进行分析推理，综合给出适当的解决方案。 | 课程目标3 |
| **3.设计/开发解决方案：**能够针对大数据应用领域的复杂工程问题，综合应用数据科学与大数据技术基本原理和方法，设计、开发满足特定应用需求的系统方案，并能够在设计环节中体现创新意识，考虑社会、健康、安全、法律、文化以及环境等多维度协同发展因素。【H】 | 3.2 能够根据实际目标和解决方案，设计或开发出大数据应用系统及说明文档，并能够在设计和开发过程中体现创新性。 | 课程目标1 |
| **4.研究：**能够基于数据科学与大数据技术的相应原理，采用科学方法对大数据应用领域中的复杂工程问题进行研究，制定技术路线、设计实验方案并开展实验，通过实验分析得到合理有效的结论。【M】 | 4.1 能够综合应用数据科学与大数据技术基本原理，针对大数据应用领域复杂工程问题，通过文献研究、调研和分析，设计合适的解决案。 | 课程目标2 |

**三、课程学习内容与方法**

**（一）理论学习内容及要求**

**表3-1 课程目标、学习内容和教学方法对应关系**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **序号** | **课程模块** | **学习内容** | **学习任务** | **课程目标** | **学习重点难点** | **教学方法** | **学时** |
| 1 | 机器学习基础 | 1.机器学习概述 | 1.掌握机器学习相关概念  2.掌握sklearn构建机器学习模型过程  3.能够对不同模型使用相应的方法进行评估 | 1 | **重点**：  1.学生对机器学习相关概念有初步认识  2.学生对机学习框架有初步了解  **难点：**  1.学生掌握机器学习在实际工程中的运用能力 | 1.讲授法：引导学生掌握机器学习的概念。  2.专题研讨：促进学生理解机器学习模型构建过程。 | 2 |
| 2.机器学习框架 | 1/2 |
| 3.模型的评判 | 1/2 |
| 4.过拟合问题 | 1/2 |
| 2 | 相关工具 | 1.Numpy基本操作 | 1.掌握Pandas库中serial和data frame数据类型及常用函数  2.掌握Pandas相关数据分析及可视化方法  3.能够使用sklearn进模型的构建 | 2/3 | **重点：**  1.学生掌握Numpy工具的使用  2.学生掌握Pandas工具的使用  **难点：**  1.学生掌握使用sklearn构建机器学习模型的过程 | 1.讲授法：引导学生掌握Numpy、Pandas工具的使用。  2.专题研讨：促进学生理解sklearn构建机器学习模型的方法。 | 4 |
| 2.Pandas基本操作 | 2/3 |
| 3. Pandas数据分析 | 1/2/3 |
| 4.sklearn工具的使用 | 1/2/3 |
| 5.回归模型的实现 | 1/2/3 |
| 6.线性回归诊断 | 2/3 |
| 7.混淆矩阵、ROC、AUC | 2/3 |
| 3 | 分类 | 1.分类相关概念 | 1.掌握k最近邻分类器原理  2.掌握支持向量机分类器原理  3.掌握贝叶斯分类器原理  4.能够使用Python语言实现分类模型  5.能够对分类模型进行性能评估 | 1 | **重点：**  1.K-最近邻分类器原理  2.支持向量机分类器原理  3.贝叶斯分类器原理  **难点：**  1.回归模型实现  2.回归模型性能评估 | 1.讲授法：引导学生掌握各类分类器原理。  2.专题研讨：促进学生掌握分类模型的构建及性能评估。 | 6 |
| 2.k最近邻 | 1/2 |
| 3.支持向量机 | 1/2 |
| 4.贝叶斯分类器 | 1/2 |
| 5.分类器实现 | 1/2/3 |
| 6.分类器性能评估 | 2/3 |
| 4 | 回归 | 1.回归相关概念 | 1.掌握线性回归原理  2.掌握岭回归、Lasso回归原理  3.掌握逻辑回归的原理  4.能够使用Python语言实现回归模型  5.能够对回归模型进行诊断与评估 | 1 | **重点：**  1.线性回归模型的原理  2.岭回归与Lasso回归的原理  3.逻辑回归的原理  **难点：**  1.回归模型的实现  2.回归模型的性能评估 | 1.讲授法：引导学生掌握机回归分析的原理。  2.专题研讨：促进学生掌握回归分析模型的构建及性能评估。 |  |
| 2.线性回归原理 | 1/2 |
| 3.岭回归与Lasso回归原理 | 1/2 |
| 4.逻辑回归的原理 | 1/2 |
| 5.回归模型的实现 | 1/2/3 |
| 6.线性回归诊断 | 2/3 |
| 7.混淆矩阵、ROC、AUC | 2/3 |
| 5 | 聚类 | 1.聚类相关概念 | 1.掌握k-means聚类算法原理  2.掌握DBSCAN聚类算法原理  3.掌握层次聚类算法原理  4.能够使用Python语言实现聚类模型  5.能够对聚类模型进行性能评估 | 1 | **重点：**  1.K-means聚类算法原理  2.DBSCAN聚类算法原理  3.层次聚类算法原理  **难点：**  1.聚类算法实现  2.聚类算法性能评估 | 1.讲授法：引导学生掌握各类聚类原理。  2.专题研讨：促进学生掌握聚类模型的构建及性能评估。 | 4 |
| 2.原型聚类 | 1/2 |
| 3.密度聚类 | 1/2 |
| 4.层次聚类 | 1/2 |
| 5.聚类算法实现 | 1/2/3 |
| 6.聚类算法性能评估 | 2/3 |
| 6 | 降维 | 1.降维相关概念 | 1.掌握降维原理  2.掌握主成分分析算法原理  3.掌握流形学习算法原理  4.能够使用Python语言实现聚类模型  5.能够对聚类模型进行性能评估 | 1 | **重点：**  1.主成分分析算法原理  2.流形学习算法原理  **难点：**  1.降维算法实现  2.降维算法性能评估 | 1.讲授法：引导学生掌握各类降维算法原理。  2.专题研讨：促进学生掌握降维算法的构建及性能评估。 | 4 |
| 2.主成分分析原理 | 1/2 |
| 3.流形学习原理 | 1/2 |
| 4.降维算法实现 | 1/2/3 |
| 5.降维算法性能评估 | 2/3 |
| 7 | 推荐 | 1.关联规则定义 | 1.掌握关联规则概念  2.掌握Apriori算法原理  3.掌握协同过滤算法原理  4.能够使用Python语言实现推荐算法  5.能够对推荐算法进行性能评估 | 1 | **重点：**  1.Apriori算法原理  2.协同过滤算法算法原理  **难点：**  1.推荐算法实现  2.推荐算法性能评估 | 1.讲授法：引导学生掌握各类推荐算法原理。  2.专题研讨：促进学生掌握推荐算法的构建及性能评估。 | 4 |
| 2.Apriori算法 | 1/2 |
| 3.协同过滤算法原理 | 1/2 |
| 4.推荐算法实现 | 1/2/3 |
| 5.推荐算法性能评估 | 2/3 |
| 8 | 神经网络 | 1.神经网络概述 | 1.理解神经元模型  2.理解并掌握感知机学习模型  3.理解神经网络  4.理解BP算法  5.理解不同网络结构、激活函数、训练轮等条件下，多层神经网络的不同表现 | 1 | **重点：**  1.神经网络模型  2.神经网络运算  3.三层感知机的误差反向传播学习过程  4.误差反向学习算法  **难点：**  1.感知机学习算法  2.误差反向学习过程 | 1.讲授法：引导学生掌握神经网络原理。  2.专题研讨：促进学生掌握BP神经网络算法的构建及应用。 | 4 |
| 2.神经元模型 | 1/2 |
| 3.感知机 | 1/2 |
| 4.反向传播 | 1/2 |
| 5.BP神经网络 | 1/2 |
| 6.BP算法实现 | 1/2/3 |
| 7.BP算法在分类和回归问题上的应用 | 1/2/3 |
| 合计 |  |  |  |  |  |  | 32 |

1. **实验学习内容及要求**
2. **表3-2 课程目标、学习内容和教学方法对应关系**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **序号** | **项目名称** | **项目来源** | **教学目标（观测点、重难点）** | **学时数** | **项目类型** | **要求** | **每组人数** | **教学方法** | **课程目标** |
| 1 | Numpy、Pandas数据分析及可视化 | 实验教材 | 1.掌握Numpy、Pandas工具的使用（重点） | 4 | 验证性实验 | 必做 | 1 | 实验指导 | 2/3 |
| 2.掌握数据可视化工具及方法（重点） |
| 3.掌握数据预处理及简单分析方法（难点） |
| 2 | 基于sklearn的机器学习模型构建 | 实验教材 | 1.掌握KNN算法原理（重点） | 4 | 验证性实验 | 必做 | 1 | 实验指导 | 1/2/3 |
| 2.掌握sklearn下KNN模型构建（重点） |
| 3.使用分类精度、分类误差等对模型进行评估（难点）（难点） |
| 3 | 基于SVM分类器的鸢尾花分类预测 | 实验教材 | 1.掌握SVM分类算法原理（重点） | 4 | 设计性实验 | 必做 | 1 | 实验指导 | 1/2/3 |
| 2.掌握SVM分类模型的构建（重点） |
| 3.使用分类精度、交叉验证精度等对分类模型性能进行评估（难点） |
| 4 | 基于朴素贝叶斯分类器的幸存者预测 | 实验教材 | 1.掌握朴素贝叶斯分类器原理（重点） | 4 | 设计性实验 | 必做 | 1 | 实验指导 | 1/2/3 |
| 2.能够建立朴素贝叶斯分类器，对幸存者进行预测（重点） |
| 3.学习绘制ROC曲线，比较predict和predict\_prob的区别（难点） |
| 5 | 基于回归模型的汽车油耗预测 | 实验教材 | 1.掌握回归算法的原理（重点） | 4 | 设计性实验 | 必做 | 1 | 实验指导 | 1/2/3 |
| 2.掌握回归模型的构建（重点） |
| 3.使用R2、MSE、AIC、BIC、Prob(F-statistic)等指标对对回归模型性能进行评估（难点） |
| 6 | 基于逻辑回归的非线性分类预测 | 实验教材 | 1.掌握逻辑归算法的原理（重点） | 4 | 设计性实验 | 必做 | 1 | 实验指导 | 1/2/3 |
| 2.掌握逻辑回归模型的构建（重点） |
| 3.使用混淆矩阵、精确率、召回率、准确率、f1-score可以量化衡量模型 |
| 4.能够建立决策树分类器，对幸存者进行预测（重点） |
| 5.使用LabelEncoder对离散变量进行编码和简单的缺失值处理方法。（难点） |
| 7 | k-means、层次聚类和DBSCAN聚类对比实验 | 实验教材 | 1.掌握聚类算法原理（重点） | 4 | 设计性实验 | 必做 | 1 | 实验指导 | 1/2/3 |
| 2.能够构建k-means、层次聚类、DBSCAN聚类模型（重点） |
| 3.能够分析kmeans、层次聚类和DBSCAN聚类模型效果差异（难点） |
| 8 | 基于PCA模型的鸢尾花数据可视化 | 实验教材 | 1.掌握PCA算法的原理（重点） | 2 | 设计性实验 | 选做 | 1 | 实验指导 | 1/2/3 |
| 2.能够通过线性组合特征值为“主成分”实现降维（重点） |
| 3.能够使用降维后的数据进行可视化（难点） |
| 9 | 基于FP树的商品推荐模型 | 实验教材 | 1.掌握关联规则挖掘原理（重点） | 2 | 设计性实验 | 选做 | 1 | 实验指导 | 1/2/3 |
| 2.能够将结构化数据转化为“购物篮”数据（重点） |
| 3.能够建立FP关联规则模型，生成规则（难点） |
| 10 | 基于BP神经网络的分类/回归 | 实验教材 | 1.掌握神经网络模型的结构（重点） | 2 | 设计性实验 | 选做 | 1 | 实验指导 | 1/2/3 |
| 2.能够构建BP神经网络模型（重点） |
| 3.能够利用BP神经网络模型解决分类或回归（难点） |

## 四、课程考核

**（一）考核内容与考核方式**

**表4-1 课程目标、考核内容与考核方式对应关系**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **课程目标** | **考核内容** | **所属**  **学习模块/项目** | **考核占比** | **考核方式** |
| 课程  目标 1 | 1.机器学习概念 | 学习模块1 | 50% | 课堂表现  实验项目  阶段测验  课程设计 |
| 2.分类相关概念 | 学习模块3 |
| 2.k最近邻 | 学习模块3  项目2 |
| 3.支持向量机 | 学习模块3  项目3 |
| 4.贝叶斯决策论 | 学习模块3  项目4 |
| 5.回归分析概念 | 学习模块4 |
| 6.线性回归 | 学习模块4  项目5 |
| 7.岭回归及Lasso回归 | 学习模块4  项目5 |
| 8.逻辑回归 | 学习模块4  项目6 |
| 9.聚类相关概念 | 学习模块5 |
| 10.K-means | 学习模块5  项目7 |
| 11.层次聚类 | 学习模块5  项目7 |
| 12.密度聚类 | 学习模块5  项目7 |
| 13.降维相关念 | 学习模块6 |
| 14.主成分分析 | 学习模块6  项目8 |
| 15.流形学习 | 学习模块6  项目8 |
| 16.推荐分析概念 | 学习模块8 |
| 17.关联规则挖掘 | 学习模块8  项目9 |
| 18.协同过滤 | 学习模块7 |
| 19.神经网络 | 学习模块8  项目10 |
| 课程  目标 2 | 1.数据预测处理 | 学习模块2  项目1,3-10 | 31% | 课堂表现  实验项目  阶段测验  课程设计 |
| 2.数据可视化 | 学习模块2  项目3-8,10 |
| 3.分类算法实现 | 学习模块3  项目3-4 |
| 4.回归算法实现 | 学习模块4  项目5-6 |
| 5.聚类算法实现 | 学习模块5  项目7 |
| 6.降维算法实现 | 学习模块6  项目8 |
| 7.推荐算法实现 | 学习模块7  项目9 |
| 8.BP神经网络算法实现 | 学习模块8  项目10 |
| 课程  目标 3 | 1.Numpy、Pandas数据分析 | 学习模块2  项目1 | 19% | 课堂表现  实验项目  阶段测验  课程设计 |
| 2.sklearn机器学习模型 | 学习模块2  项目2 |
| 3.算法性能评估 | 学习模块3-8  项目3-10 |

**表4-2 课程目标与考核方式矩阵关系**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 课程  目标 | 考核方式 | | | | | | | 考核占比 |
| 平时考核（50%） | | | | | | 期末考核（50%） |
| 课堂表现（10%） | | 阶段测验  （15%） | | 实验项目  （25%） | |
| 课程目标1 | 60% | 60% | | 40% | | 50% | | 50%=10%\*60%+25%\*40%+15%\*60%+50%\*50% |
| 课程目标2 | 30% | 20% | | 40% | | 30% | | 31%=10%\*30%+25%\*30%+15%\*20%+50%\*30% |
| 课程目标3 | 10% | 20% | | 20% | | 20% | | 19%=10%\*10%+25%\*30%+15%\*20%+50%\*20% |

## （二）成绩评定

**1. 平时成绩评定**

（1）课堂表现（20分）：通过学生课堂发言、提问及小组讨论情况，评价学生的课程学习态度和参与能力、专业认同感和终身学习意识。

（2）阶段性测验（30分）：以阶段性测验方式进行，评价学生当前的学习情况。

（3）实验项目（50分）：实验课程的实验项目及实验报告完成情况，主要评价学生的实际操作能力。

**2.期末成绩评定**

主要考察学生掌握机器学习的基本概念和基础知识情况，对分类、回归、聚类、降维、推荐、神经网络等智能模型的理解和应用，考核学生进行数据预处理、数据分析、数据可视化的能力。

**3.总成绩评定**

总成绩（100%）=平时成绩（50%）+期末成绩（50%）

**（三）评分标准**

1.平时成绩

（1）课堂表现：通过学生在课堂上的表现来评价学生对知识的掌握情况及相关能力水平。包括课堂发言、提问、回答问题、测验、练习等，每人每期参加上述课堂活动不得少于两次，少于两次者，课堂表现成绩记为0分（目标1：60%；目标2：30%；目标3：10%)。

（2）阶段测验：根据教学任务，在教学实验系统中安排阶段测验，考察学生对相应章节理论知识的掌握情况，对实际问题分析能力，使用相应工具能力和对结果归纳总结能力。取学生阶段测验平均分，未做测验按“0”分计算，成绩由教学实验系统记录确定（目标1：60%；目标2：20%；目标3：20%)。

（3）实验项目：通过实验教学中的实际案例，考察学生使用相应工具和可视化方法，进行数据回归、分类、聚类、降维、推荐、离群分析的能力，以及对结果进行归纳总结、撰写文档能力。根据学生每个实验项目实验过程的合理性、实验结果正确性、归纳总结的合理性给定成绩（目标1：40%；目标2：40%；目标3：20%)。

2.期末成绩

期末课程设计：课程设计结果为可运行的项目和文本，考察学生机器学习算法的分析、设计、实现、评估能力，要求学生能综合运用机器学习知识，针对数据科学与大数据专业实际工程问题，设计相应的机器学习算法，并利用相关的工具和平台，实现相应的机器学习算法，从而求解相关问题。

**表5-1 课程设计评分标准**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 课程设计作品 | 内容健康、充实且积极向上，设计思路清晰，算法设计合理，运行结果正确，可视化效果好。实现代码的可读性好、程序注释的合理性、命名的规范。能正确回答老师就该设计提出的问题并且项目有很好地创新。 | 内容健康、充实且积极向上，设计思路清晰，算法设计较为合理，运行结果正确，可视化效果较好。程序的可读性较好、程序注释的合理性、命名的规范。能正确回答老师就该设计提出大部分问题并且项目有一定的创新。 | 内容健康、充实且积极向上，设计思路清晰，算法设计基本合理，运行结果基本正确，能进行部分数据的可视化。程序有一部分注释、命名基本规范。基本能正确回答老师就该设计提出的问题。 | 内容健康、充实且积极向上，设计思路清晰，算法设计有部分错误，运行结果大部分合理，能进行少量数据的可视化。程序注释和命名欠规范。基本能正确回答老师就该设计提出的问题。 | 未能按时完成项目，答辩时不能正确演示项目和回答老师提问。 |
| 课程设计报告 | 课程设计报告数据特征分析合理、算法选型正确、图形图表规范、处理步骤清晰明了、文档符合数据处理流程规范。 | 课程设计报告数据特征分析较合理、算法选型正确、图形图表较规范、处理步骤清晰明了、文档符合数据处理流程规范。 | 课程设计报告数据特征基本合理、算法选型基本正确、图形图表有些不规范、处理步骤有一定问题、文档基本符数据处理流程规范，有少部分格式欠规范。 | 课程设计数据特征分析不到位、不能选择合适算法、图形图表不太规范、处理步骤不太清晰，格式欠规范。 | 课程设计报告格式混乱，不符合数据处理流程规范。 |

## 五、其它说明

本课程大纲依据2023版人才培养方案，由大数据与智能工程学院计算机科学与技术系讨论制定，大数据与智能工程学院教学工作委员会审定，教务处审核批准，自2023级开始执行。