太原理工大学

TAIYUAN UNIVERSITY OF TECHNOLOGY

《机器学习课程设计 A》指导书

（2025 版）

太原理工大学

人工智能学院

2025 年 6 月

目录

1.课程基本信息

课程中文名称：机器学习课程设计 A

课程编号：SJ003216

学分：2

总学时（周数）：2 周

适用专业：数据科学与大数据技术专业、人工智能专业

先修课程：机器学习 B、机器学习与模式识别

后续课程：《深度学习》、《深度学习课程设计》

开课单位：人工智能学院

2.课程设计目的

本课程设计旨在提高学生加深对机器学习算法的理解，提高学生分析问题、解决问题的能力。

1、进一步巩固加深学生对机器学习中的基本原理和算法的理解，掌握科学的方法，培养初步应用能力。

2、课程设计明确任务要求，学生自己查阅资料、设计方案和动手实践，并进行结果记录和分析，充分发挥学生的创造性和主动性。

3、熟悉掌握 python 语言，可以进行机器学习的应用处理的开发设计。

3.课程设计选题

课程设计内容分为三个模块，分别为识别系统、预测模型、聚类模型。

学生 5 人为一小组，每组从三个模块中任意选择两个模块，再从已选模块中分别选择一道题作为自己的课程设计题目，每个学生**两道题**（有能力和时间的学生可以选择多道题，不设上限）。

要求：同组学生的选题不能相同。

注意：根据所选模块中的题目，封皮中课程设计名称相应为以下三者之一：

识别系统与预测模型；

识别系统与聚类模型；

预测模型与聚类模型。

模块一 识别系统

1.1 基于二分类算法的谣言检测

1.1.1 数据集介绍

本设计使用的数据集是从新浪微博不实信息举报平台抓取的中文谣言数据，数据集中共包含 1538 条谣言和 1849 条非谣言。每条数据均为 json 格式，其中 text 字段代表微博原文的文字内容。

数据集下载地址：

https://aistudio.baidu.com/datasetdetail/95286

1.1.2 任务目标

根据文本数据，对谣言进行分类。

1.1.3 设计要求

1.在 python 环境下完成对数据的读取。

2.选择 3 种二分类模型，实现谣言的分类。

3.对三种分类模型的分类结果进行可视化的展示、效果的比较。

4.给出模型训练与验证及性能评估。

提示：数据准备阶段，提取 json 文件中 text 字段的文本内容，并设置谣言标签。

1.2 对话情绪识别

1.2.1 数据集介绍

对话情绪识别任务输入是一段用户文本，输出是检测到的情绪类别，包括消极、积极、中性，这是一个经典的短文本三分类任务。训练、预测、评估使用的数据示例如下，数据由两列组成，以制表符（'\t'）分隔，第一列是情绪分类的类别（0 表示消极；1 表示中性；2 表示积极），第二列是以空格分词的中文文本。

数据集下载地址：

https://aistudio.baidu.com/datasetdetail/12605

1.2.2 任务目标

根据文本数据，对话情绪进行多分类。

1.2.3 设计要求

1.在 python 环境下完成对数据的读取。

2.选择 3 种分类模型，得到情绪分类的模型。

3.对三种分类模型的分类结果进行可视化的展示、效果的比较。

4.给出模型训练与验证及性能评估。

提示：通过对文本进行向量化的表示（如词袋模型），得到文本的特征。

1.3 蘑菇分类问题

1.3.1 数据集介绍

数据集源于蘑菇记录摘自《奥杜邦学会北美蘑菇野外指南》，包括对假设样本的描述对应于蘑菇中的 23 种有鳃蘑菇。每个物种都被确定为绝对可食用、绝对有毒或可食用性未知且不推荐。后一类与有毒物质结合在一起一《指南》明确指出确定蘑菇的可食用性。该数据集包含有 8124 个样本，每个样本包含有 22 个属性特征。该数据集模拟可食用蘑菇和不可食用蘑菇的数据，旨在提供一个综合的测试平台，用于测试算法的准确性及有效性。

数据集下载地址：

https://archive.ics.uci.edu/dataset/73/mushroom

1.3.2 任务目标

分类检测，识别有毒蘑菇和可食用蘑菇比例以及分类的准确性。

1.3.3 设计要求

1.在 python 环境下完成对数据的读取，并进行数据探索及数据预处理。

2.选择或构造能够有效区分正常与异常行为的特征。进行特征选择，排除低相关性或冗余的特征。

3.至少使用三种分类方法进行分类检测。

4.给出模型训练与验证及性能评估。

1.4 恶性乳腺癌肿瘤预测

1.4.1 数据集介绍

数据来自于威斯康星大学医院，共计 699 例样本，每例样本有 11 列，分别指“样本编号”,“肿块厚度”,“细胞大小均匀性”,“细胞形状均匀性”,“边缘黏性”,“单上皮细胞大小”,“裸核”,“染色体”,“正常核”,“有丝分裂”,“肿瘤性质”，其中肿瘤性质 2 代表良性，4 代表恶性，恶性肿瘤样本 241 例，占 34.5%,良性 458 例，占 65.5%，16 例样本存在丢失的属性值。请采用机器学习算法当中相关回归算法对其进行数据分析。

数据集下载地址：

https://archive.ics.uci.edu/dataset/15/breast+cancer+wisconsin+original

1.4.2 任务目标

学会如何使用分类算法对模型进行训练。

1.4.3 设计要求

1.在 python 环境下完成对数据的读取，并进行数据探索及数据预处理。

2.选择或构造能够有效区分正常与异常行为的特征。进行特征选择，排除低相关性或冗余的特征。

3.使用分类算法（如支持向量机、神经网络等）对其进行分析。

4.给出模型训练与验证及性能评估。

1.5 心脏病诊断

1.5.1 数据集介绍

本数据集是一个公开的心脏病数据集，包含患者的临床指标和心脏病患病情况（有无心脏病）。

数据集下载地址：

https://archive.ics.uci.edu/dataset/45/heart+disease

1.5.2 任务目标

构建一个机器学习分类模型，根据患者的临床指标预测其是否患有心脏病。

1.5.3 设计要求

1.在 python 环境下完成对数据的读取，并进行数据探索及数据预处理

2.至少使用两种算法构建模型，逻辑回归、决策树等

3.使用交叉验证和网格搜索等技术来优化模型的超参数选择。

4.通过 ROC 曲线和精确度-召回率曲线等方式对模型进行评估，并比较不同算法的表现。

1.6 红酒质量分类

1.6.1 数据集介绍

本数据集是一个公开的红酒质量数据集，包含红酒的化学特征以及对应的质量评分。

数据集下载地址：

https://archive.ics.uci.edu/dataset/186/wine+quality

1.6.2 任务目标

构建一个机器学习分类模型，根据红酒的化学特征预测其质量评分等级（如优质、中等、低质等级）。

1.6.3 设计要求

1.在 python 环境下完成对数据的读取，并进行数据探索和可视化分析，以了解红酒化学特征与质量评分等级之间的关系

2.至少使用两种算法构建模型，如逻辑回归、支持向量机（SVM）或随机森林等。

3.尝试进行特征缩放、特征选择和模型调参等步骤，以优化模型性能

4.可以通过混淆矩阵和分类报告等方式对模型进行评估，并分析模型的预测能力。

1.7 森林覆盖类型

1.7.1 数据集介绍

森林覆盖类型数据集来源于美国科罗拉多州罗斯福国家森林的真实数据。这个数据集主要用于预测森林区域的覆盖类型，它由美国国土测绘局和美国森林服务局提供的数据组成。这些数据被广泛用于地理信息系统（GIS）相关的研究和机器学习的分类任务。

数据集包含 581,012 个样本。每个样本有 54 个属性，包括土壤类型、野生动物栖息地以及各种地形和人为特征。共有 7 种不同的森林覆盖类型。

主要特征包括：

高程（Elevation）：海拔高度。

坡度（Aspect）：坡面的方位角。

坡度（Slope）：坡面的倾斜度。

水平距离到最近表面水源点（Horizontal Distance to Hydrology）：从预测点到最近表面水源的水平距离。

垂直距离到最近表面水源点（Vertical Distance to Hydrology）：从预测点到最近表面水源的垂直距离。

水平距离到最近道路（Horizontal Distance to Roadways）：从预测点到最近道路的水平距离。

9AM 阴影指数（Hillshade 9am）：早上 9 点时山坡的阴影指数。

正午阴影指数（Hillshade Noon）：中午时山坡的阴影指数。

3PM 阴影指数（Hillshade 3pm）：下午 3 点时山坡的阴影指数。

水平距离到最近野火点（Horizontal Distance to Fire Points）：从预测点到最近野火发生点的水平距离。

野生动物栖息地类型（Wilderness Area）：4 种不同的野生动物栖息地类型，采用独热编码表示。

土壤类型（Soil Type）：40 种不同的土壤类型，采用独热编码表示。

这个数据集可以用来训练机器学习模型来自动分类和预测给定区域的森林覆盖类型，对于环境科学、生态研究和资源管理具有重要的应用价值。

数据集下载地址：

https://pan.baidu.com/s/1z9Zp0KQ-0cvzzwzv3ToclQ

提取码：0610

1.7.2 任务目标

利用森林覆盖类型数据集（Forest Cover Type dataset），通过机器学习模型预测科罗拉多州荒野区的森林覆盖类型。项目的主要目的是使学生能够熟练掌握数据预处理、特征工程、模型选择和模型评估等关键技术，同时增强学生解决实际问题的能力。

1.7.3 设计要求

1.在 python 环境下完成对数据的读取，并进行数据探索及数据预处理。

2.可视化展示数据的分布和特征间的关系，统计分析，理解数据集的主要特征和结构。

3.选择至少三种适合分类任务的模型，如决策树、随机森林、支持向量机（SVM）、K最近邻（KNN）或神经网络等。使用交叉验证技术选择模型参数，确保模型的泛化能力。

4.使用准确率、精确率、召回率、F1 分数和混淆矩阵等指标评估模型性能。分析模型在不同森林覆盖类型上的预测表现，识别模型的强项和弱项。

1.8 信用卡欺诈检测

1.8.1 数据集介绍

这个数据集包含了欧洲持卡人在两天内通过信用卡进行的交易。数据集包含了284,807 笔交易的信息。共有 31 个特征，其中 28 个是通过 PCA 变换得到的匿名特征。

'Time'特征包含了每笔交易与第一笔交易之间的时间差（单位为秒）。'Amount'特征是交易金额，这是唯一一个未经 PCA 处理的特征。'Class'是响应变量，取值为 1 代表欺诈交易，0 代表非欺诈交易。此数据集的一个重要特点是高度不平衡，只有 492 笔交易是欺诈的，占总样本的 0.172%，这反映了实际环境中信用卡欺诈交易的稀少性。

数据集下载地址：

https://pan.baidu.com/s/1bzV8-rNkLk1BcsFdqKw4Cw

提取码：0610

1.8.2 任务目标

让学生能够通过实际操作信用卡欺诈检测数据集，深入理解和掌握机器学习在实际问题中的应用。学生将学习如何处理高度不平衡的数据、选择和训练适合的机器学习模型，以及如何评估模型的性能，从而有效预测信用卡交易是否为欺诈行为。

1.8.3 设计要求

1.在 python 环境下完成对数据的读取，并进行数据探索及数据预处理。由于数据集中欺诈交易和非欺诈交易的比例极度不平衡，学生需要采取适当的技术如过采样、欠采样或合成少数类过采样技术（SMOTE）来处理这一问题。

2.通过可视化手段探索数据，了解不同特征与欺诈行为之间的关系。并进行基本的统计分析，以识别数据中的潜在模式和异常。

3.基于数据特性选择至少三种机器学习模型进行训练。常见的选择包括逻辑回归、支持向量机、决策树、随机森林和梯度提升等。使用处理过的数据集进行模型训练，调整参数以获得最佳性能。

4.使用混淆矩阵、精确率、召回率、F1 分数和 ROC 曲线等评估指标来衡量模型的性能。通过交叉验证方法确保模型的稳健性和泛化能力。

模块二 预测模型

2.1 隧道拱顶沉降变形预测

2.1.1 数据集介绍

数据集包含 2016 年 4 月 19 日-2016 年 5 月 22 日期间，拱顶沉降变形/mm 和上台阶周边收敛变形/mm 的数据。

数据集下载地址：

https://html.rhhz.net/GLJTKJ/20171213.htm

2.1.2 任务目标

构建拱顶沉降预测模型或上台阶收敛预测模型。

2.1.3 设计要求

1.在 python 环境下完成对数据的读取。

2.可查询相关文献，采用适合的回归模型。

3.以数据的前 28 个数据作为建模样本，剩余 6 个数据作为预测检验样本对三种模型的结果进行可视化的展示、效果的比较。

4.给出模型训练与验证及性能评估。

2.2 睡眠质量预测

2.2.1 数据集介绍

以下（从左到右）哺乳动物数据集的变量：

动物种类、体重（公斤）、脑重（克）、慢波（“无梦”）睡眠（小时/天）、矛盾的（“做梦”）睡眠（小时/天）、总睡眠（小时/天）（慢波和反常睡眠之和）、最大寿命（年）、妊娠时间（天）、捕食指数（1-5）1 =最小（最有可能被捕食）5 =最大（最有可能被捕食）、睡眠暴露指数（1-5）1 =接触最少（例如，在良好保护的巢穴中进行动物睡眠）5 =接触最多、总体危险指数（1-5）（基于上述两个指数和其他信息）1 =最低危险（来自其他动物）5 =最高危险（来自其他动物）

注意：缺少值-999.0。

数据集下载地址：

https://aistudio.baidu.com/datasetdetail/107977

2.2.2 任务目标

预测哺乳动物睡眠质量和测试指标之间的关系。

2.2.3 设计要求

1.在 python 环境下完成对数据的读取和数据的预处理。

2.选取 3 种预测模型，选取对睡眠时间（数据的最后一列）进行预测

3.对三种模型的结果进行可视化的展示、效果的比较。

4.给出模型训练与验证及性能评估。

2.3 预测客户流失

2.3.1 数据集介绍

网络视频商家之间存在着巨大的竞争。如果商家想增加收入，他们需要更多的订阅者，但保持现有客户比拥有新客户更重要。 所以商家想知道哪个客户可能会取消他的服务，也就是客户流失。 如果知道谁会流失，也许商家可以通过促销来抓住他们。请采用适合的回归模型对客户的流失情况进行预测分析。数据集描述如下：

features

name	detail	含义
id	unique subscriber id	客户唯一ID
is_tv_subscriber	customer has a tv subscription ?	是否订阅电视
is_movie_package_subscriber	is he/she has a sinema movie package subs	他/她有电影套餐吗
subscription_age	how many year has the customer use our service	服务年限
bill_avg	last 3 months bill avg	过去 3 个月账单平均值
reamining_contract	how many year remaining for customer contract. if null; customer hasnt have a contract. the customer who has a contract time have to use their service until contract end. if they canceled their service before contract time end they pay a penalty fare.	客户合同还剩多少年。 如果为空； 客户还没有合同。有合同时间的客户必须使用他们的服务，直到合同结束。 如果他们在合同期限结束前取消服务，他们将支付罚款。
service_failure_count	customer call count to call center for service failure for last 3 months	过去 3 个月因服务失败而致电呼叫中心的客户呼叫次数
download_avg	last 3 months internet usage (GB)	
upload_avg	last 3 months upload avg (GB)	过去 3 个月平均上传量 (GB)
download_over_limit	most of customer has a download limit. if they reach this limit they have to pay for this. this column contain "limit over count" for last 9 months	大多数客户都有下载限制。 如果他们达到这个限制，他们必须为此付费。 此列包含过去 9 个月的“限制计数”
churn	Whether the customer is churn	客户是否流失

数据集下载地址：

https://pan.baidu.com/s/1PBOS8qnkJjzihMSdwgpOLw

提取码: 3674

2.3.2 任务目标

逻辑回归对模型进行训练预测。

2.3.3 设计要求

1.在 python 环境下完成对数据的读取，并进行数据探索及数据预处理。

2.选择或构造能够有效区分正常与异常行为的特征。进行特征选择，排除低相关性或冗余的特征。

3.使用一种回归方法对其进行分析。

4.给出模型训练与验证及性能评估。

2.4 心力衰竭患者的生存率

2.4.1 数据集介绍

心血管疾病每年导致全球约 1700 万人死亡，主要表现为心肌梗塞和心力衰竭。当心脏无法泵出足够的血液来满足身体的需要时，就会发生心力衰竭（HF）。现有的患者电子病历可量化症状、身体特征和临床实验室测试值，可用于进行生物统计分析，旨在突出显示医生无法检测到的模式和相关性。机器学习，可以根据患者的数据预测患者的生存情况，并可以从患者的病历中找出最重要的特征。通过收集的 299 名心力衰竭患者的数据集，这些患者是在随访期间收集的，每个患者档案都有 13 个临床特征。

数据集下载地址：

https://archive.ics.uci.edu/dataset/519/heart+failure+clinical+records

2.4.2 任务目标

采用回归算法对患者的生存情况进行预测，并对与最重要的风险因素相对应的特征进行排序。

2.4.3 设计要求

1.在 python 环境下完成对数据的读取，并进行数据探索及数据预处理。

2.选择或构造能够有效区分正常与异常行为的特征。进行特征选择，排除低相关性或冗余的特征。

3.使用合适的回归算法对其进行检测。

4.给出模型训练与验证及性能评估。

2.5 汽车燃油效率预测

2.5.1 数据集介绍

本数据集是一个公开的汽车燃油效率数据集，包含了不同汽车的多个特征（如气缸数、排量、马力等）以及对应的燃油效率（公里/升）。

数据集下载地址：

https://archive.ics.uci.edu/dataset/9/auto+mpg

2.5.2 任务目标

构建机器学习回归模型，根据汽车的特征预测其燃油效率。

2.5.3 设计要求

1.在 python 环境下完成对数据的读取，进行数据清洗和特征工程，处理缺失值和类别特征，并对特征进行适当的缩放和转换。

2.至少使用两种算法构建模型，如线性回归、支持向量回归（SVR）或随机森林回归等。

3.尝试使用交叉验证和网格搜索等技术来优化模型的超参数配置。

4.通过均方误差（MSE）等指标对模型进行评估，并进行结果分析和可视化。

2.6 空气质量预测

2.6.1 数据集介绍

该数据是一个公开的美国大气质量数据集，包含了不同城市的空气质量指数（AQI）以及各种大气污染物的浓度数据（如 PM2.5、臭氧等）。

数据集下载地址：

https://www.kaggle.com/datasets/epa/epa-historical-air-quality

2.6.2 任务目标

构建一个机器学习回归模型，根据各种污染物的浓度预测该城市的空气质量指数（AQI）。

2.6.3 设计要求

1.在 python 环境下完成对数据的读取，进行数据清洗和特征选择。

2.对数据进行可视化分析，探索污染物浓度与 AQI 之间的关系，选取与 AQI 相关性较高的污染物特征进行建模。

3.至少使用两种算法构建模型，如线性回归、岭回归或梯度提升等算法进行空气质量的预测。

4.评估模型的性能，并比较不同算法的表现，以进一步优化模型。

2.7 加利福尼亚房价预测

2.7.1 数据集介绍

加利福尼亚房价数据集（California Housing Prices dataset）来源于 1990 年的美国加利福尼亚州的人口普查数据。数据集包含加利福尼亚州的住房数据，包括特征如经度、纬度、

房龄、总房屋面积、总卧室面积、建筑物内人口、建筑物内家庭数、中位收入、中位房价、靠海情况和卧室数量。数据集共有 20640 个样本。。数据集包括了房屋的多种属性以及相关的地理位置信息，被广泛用于研究房价与各种因素之间的关系，特别是用于房价预测的回归分析。

数据集下载地址：

https://pan.baidu.com/s/1FlSWjqMS5B90RIuxbx1l2Q

提取码：0610

2.7.2 任务目标

使用加利福尼亚房价数据集，学生需要构建一个回归模型，来预测加利福尼亚州各地区的房屋中值价格。模型将根据数据集提供的多个特征进行预测，这些特征包括地理位置（经纬度）、房龄、房屋大小、房间数量、卧室数量、区域人口、家庭收入中位数、靠海情况等。

2.7.3 设计要求

1.在 python 环境下完成对数据的读取，并进行数据探索及数据预处理。

2.选择和构建对预测房价有帮助的特征，也可考虑创建新的特征。

3.使用三种回归方法进行房价预测。

4.给出模型训练与验证及性能评估。

2.8 洛杉矶犯罪率影响因素的回归分析

2.8.1 数据集介绍

洛杉矶犯罪数据（2000 年至现在）集，收集了从 2000 年开始至今的洛杉矶地区的详细犯罪记录，这些数据由洛杉矶警察局提供。数据集包含从 2000 年开始的洛杉矶犯罪事件记录，包含 1048575 个样本，涉及犯罪发生的日期、时间、地点、犯罪类型、受害者信

息、嫌疑人描述以及事件结果等详细信息，覆盖了多种类型的犯罪。此数据集为研究人员、数据科学家、政策制定者提供了一个深入分析洛杉矶犯罪状况的基础，也可以帮助公共安全机构更好地了解犯罪活动的动态，以便制定有效的对策。

数据集下载地址：

https://pan.baidu.com/s/1grcyn94O0xGkzrD2oj6wig

提取码：0610

2.8.2 任务目标

通过洛杉矶犯罪数据集进行回归分析，预测不同区域的犯罪率，并探索影响犯罪率的主要因素。通过这一分析，可以帮助当地政府和执法机构更有效地分配资源，针对性地预防和减少犯罪。

2.8.3 设计要求

1.在 python 环境下完成对数据的读取，并进行数据探索及数据预处理。

2.选择和构建对预测犯罪率有帮助的特征，也可考虑创建新的特征。

3.至少使用三种回归方法进行犯罪率预测。

4.给出模型训练与验证及性能评估。

5.提出基于模型结果的实际应用建议，如何更有效地部署警力或进行城市规划。

模块三 聚类模型

3.1 基于聚类的轴承故障检测

3.1.1 数据集介绍

凯西西储大学轴承数据集，分为训练样本 2000 个，测试样本 400 个。每个样本采集点数为 64*64=4096 个。

数据集下载地址：

https://aistudio.baidu.com/datasetdetail/113834

3.1.2 任务目标

轴承故障检测。

3.1.3 设计要求

1.在 python 环境下完成对数据的读取和处理。

2.选择 3 种聚类模型，对数据集分为 10 类，完成轴承故障的检测。

3.对三种模型的结果进行可视化的展示、效果的比较。

4.给出模型训练与验证及性能评估。

3.2 客户人群分析

3.2.1 数据集介绍

俗话说，“物以类聚，人以群分”，聚类算法其实就是将一些具有相同内在规律或属性的样本划分到一个类别中，这里，使用聚类算法去分析超市购物中心客户的一些基本数据，把客户分成不同的群体，供营销团队参考并相应地制定营销策略数据集包含有关客户的信息，包括

CustomerID：客户 ID。

Gender：性别。

Age：年龄。

Annual Income (k$)：年收入。

Spending Score (1-100)：消费指数，2000 个数据实例。

数据集下载地址：

https://aistudio.baidu.com/datasetdetail/107018

3.2.2 任务目标

对客户人群进行聚类。

3.2.3 设计要求

1.在 python 环境下完成对数据的读取和清洗。

2.选择 3 种聚类模型，实现对数据的聚类。

3.对三种聚类模型的结果进行可视化的展示、效果的比较。

3.3 精准定位营销策略

3.3.1 数据集介绍

在当今社会，大数据已经成为了企业决策的重要依据。通过对客户进行细分分析，企业可以更好地了解客户的需求和行为，从而制定更加精准的营销策略，提高市场竞争力。要达到的数据分析目标是通过对客户数据的分析，找出不同客户群体的特征和需求，为企业提供有针对性的营销建议。从社会、经济、技术、数据来源等方面来看，随着互联网和移动互联网的快速发展，企业和个人产生的数据量呈现爆炸式增长，这为大数据分析提供了丰富的数据来源；同时，大数据技术的发展也为数据分析提供了强大的技术支持。本数据集共有 1599 个样本，其中每个样本数据包括客户 ID，性别，客户年龄、客户年收入、客户的消费习惯，5 个属性特征

数据集下载地址：

https://www.kaggle.com/datasets/govindkrishnadas/segment/data

3.3.2 任务目标

通过聚类算法将客户划分为不同的群体，找出每个群体的特征和需求。

3.3.3 设计要求

1.在 python 环境下完成对数据的读取，并进行数据探索及数据预处理。

2.选择或构造能够有效区分正常与异常行为的特征。进行特征选择，排除低相关性或冗余的特征。

3.采用合适的聚类方法进行分析检测。

4.给出模型训练与验证及性能评估。

3.4 产品的聚类

3.4.1 数据集介绍

该数据集是从流行的产品比较平台 PriceRunner 收集的。它包括 10 个类别的 35311 个产品优惠，由 306 个不同的商家提供。该数据集为评估聚类和实体匹配算法提供了理想的基础。

数据集下载地址：

https://archive.ics.uci.edu/dataset/837/product+classification+and+clustering

3.4.2 任务目标

实现产品聚类和实体匹配。

3.4.3 设计要求

1.在 python 环境下完成对数据的读取，并进行数据探索及数据预处理。

2.选择或构造能够有效区分正常与异常行为的特征。进行特征选择，排除低相关性或冗余的特征。

3.使用合适的分类聚类方法对商品实现聚类及实体匹配。

4.给出模型训练与验证及性能评估。

3.5 客户消费行为分析

3.5.1 数据集介绍

本数据集是一个公开的客户消费数据集，记录了客户的消费行为。

数据集下载地址：

https://archive.ics.uci.edu/dataset/352/online+retail

3.5.2 任务目标

构建一个机器学习模型，对客户基于其消费行为进行聚类分析，识别不同的客户群体。

3.5.3 设计要求

1.在 python 环境下完成对数据的读取，并进行数据探索及数据预处理。

2.实现算法并对结果进行可视化展示，以便于理解和解释不同客户群体的特征和行为。

3.6 异常检测

3.6.1 数据集介绍

本数据集是公开的异常检测数据集（如 Numenta Anomaly Benchmark 数据集）。

数据集下载地址：

https://github.com/numenta/NAB

3.6.2 任务目标

通过密度聚类算法，如 DBSCAN（Density-Based Spatial Clustering of Applications with Noise），对数据集中的样本进行聚类分析和异常检测。

3.6.3 设计要求

1.通过确定样本的密度和邻域关系，将正常样本聚集在一起，并将异常样本标记为噪声。

2.通过对数据进行可视化展示，绘制样本的散点图或密度图，以便于理解聚类结果和异常检测效果。

3.尝试调整聚类算法的参数，比如邻域半径和最小样本数，以观察对结果的影响，并与其它常见的聚类算法进行比较。

3.7 社交媒体用户行为分析

3.7.1 数据集介绍

数据集是一个包含了美国航空公司在 Twitter 上的用户评论数据的数据集。该数据集被广泛用于情感分析和舆情监测的研究与实践。这个数据集包含了包括了约 14,000 条关于美国航空公司的推文，每条推文包括了以下字段：

tweet_id： 推文的唯一标识符。

airline_sentiment： 推文的情感极性，可以是积极、中立或消极。

airline_sentiment_confidence： 推文情感的置信度，表示系统对情感的预测可信度。

negativereason： 如果推文是负面的，负面情感的原因。

negativereason_confidence： 负面情感原因的置信度。

airline： 推文涉及的航空公司。

airline_sentiment_gold： 人工标注的推文情感。

text： 推文的文本内容。

tweet_coord： 推文的地理坐标。

tweet_created： 推文的创建时间。

tweet_location： 推文的地理位置。

user_timezone： 推文的用户时区。

该数据集通常被用于训练情感分析模型，评估航空公司的声誉以及研究用户对航空服务的态度和反馈。

数据集下载地址：

https://pan.baidu.com/s/1NxXYUiUa3dWL-4g8NkZbOA

提取码：0610

3.7.2 任务目标

通过分析 Twitter 上航空公司的用户情感数据，找出不同情感类型的用户群体，并了解其行为特征。

3.7.3 设计要求

1.在 python 环境下完成对数据的读取，并进行数据探索及数据预处理。

2.使用机器学习技术对用户情感数据进行分析。

3.通过对用户的情感数据进行聚类来识别不同的用户群体。

4.分析不同情感类型的用户在社交媒体上的行为特征，例如发帖频率、内容主题等。

3.8 健康群体聚类分析与个性化健康管理

3.8.1 数据集介绍

NHANES（National Health and Nutrition Examination Survey）是美国疾病控制和预防中心（CDC）进行的一个重要调查项目，旨在评估美国民众的健康和营养状况。该调查

项目通过对代表性样本的美国人口进行问卷调查、身体检查和实验室检测，收集各种健康相关数据。NHANES 数据集是由这些数据组成的大型综合数据库，包含了丰富的健康指标、人口统计学特征以及慢性疾病诊断等信息。

NHANES 数据集的主要特点包括：

代表性：NHANES 采用多阶段随机抽样设计，确保样本代表性和可靠性。

多维度数据：NHANES 收集了包括人口统计学特征、健康指标（如身高、体重、血压、血糖、血脂等）、营养摄入、生活方式等多个维度的数据。

长期跟踪：NHANES 定期进行数据采集，允许研究人员跟踪和分析美国人口健康状况的变化趋势。

公开透明：NHANES 数据集是公开可用的，任何人都可以在 CDC 的网站上获取和下载数据，用于科学研究和政策制定。

NHANES 数据集在医学研究、公共卫生政策制定、临床实践等领域具有重要的价值，为了解美国人口健康状况、探索健康与疾病的关系提供了重要的数据支持。

数据集下载地址：

https://pan.baidu.com/s/1XTGyCEzCcPKUnMXxZ3I1cg

提取码：0610

3.8.2 任务目标

通过对 NHANES（National Health and Nutrition Examination Survey）数据集进行聚类分析，识别不同的健康群体，包括青年健康群体、老年健康群体、肥胖健康群体、代谢综合征群体、慢性疾病群体等，以便深入了解各群体的健康状况及特征。并为不同健康群体提供个性化的健康管理建议，以促进公众健康意识和健康管理水平的提升。

3.8.3 设计要求

1.在 python 环境下完成对数据的读取，并了解 NHANES 数据集中的字段含义和数据结构，包括人口统计学特征（如年龄、性别、种族）、健康指标（如体重、身高、血压、血糖、血脂等）、慢性疾病诊断等。

2.根据需要进行特征选择、转换或创建新特征，以便于聚类分析。

3.对数据进行标准化或归一化等预处理操作，以确保不同特征具有相似的尺度。

4.使用聚类算法（如 K-means、层次聚类等）对 NHANES 数据进行分析，识别不同的健康群体。

5.解释每个聚类群体的特征，包括其健康指标的平均值、分布情况等，以及与各种健康状况相关的慢性疾病的发病率。

6.根据不同健康群体的特征，提出相关的健康管理建议，如生活方式改变、预防措施、医疗监测等，以改善人们的健康状况和生活质量。

4.课程设计方案制定

1、程序运行环境是 Windows 平台。

2、开发工具推荐选用 python 作为编程开发工具。

3、熟悉 sklearn 程序包。

5.课程设计的一般步骤

1、选题与搜集资料：根据课题，进行系统调查，搜集资料。

2、分析与设计：根据搜集的资料，进行功能分析，并对系统功能与模块划分等设计。

3、程序设计：运用掌握的语言，编写程序，实现所设计的功能。

4、调试与测试：自行调试程序，同学之间交叉测试程序，并记录测试情况。

5、验收与评分：指导教师对每个成员开发的程序进行综合验收，结合设计报告，根据课程设计成绩的评定方法，评出成绩。

6、课程设计报告的撰写。撰写详细的设计报告，包括数据分析、模型建立和评估结果，及后续改进的建议。

7、提供清晰的设计代码，包括所有数据处理和模型训练步骤。编写详细的文档和注释，说明代码的功能和操作方法。

6.要求

6.1 总体要求

1、要充分认识课程设计对培养自己的重要性，认真做好设计前的各项准备工作。尤其是对编程软件的使用有基本的认识。

2、既要虚心接受老师的指导，又要充分发挥主观能动性。结合课题，独立思考，努力钻研，勤于实践，勇于创新。

3、独立按时完成规定的工作任务，不得弄虚作假，不准抄袭他人内容，否则成绩以不及格计。

4、在设计过程中，要严格要求自己，树立严肃、严密、严谨的科学态度，必须按时、按质、按量完成课程设计。

6.2 实施要求

1、理解各种机器学习算法确切意义。

2、独立进行方案的制定，算法结构设计要合理。

3、在程序开发时，则必须清楚主要实现函数的目的和作用，需要在程序书写时说明做适当的注释。实验使用 python 语言进行开发，要理解每个函数的具体意义和适用范围，在写课程设计报告时，必须要将主要函数的功能和参数做详细的说明。

4、通过多次实验来检测该算法的稳定性和正确性。

6.3 课程设计报告的内容及要求

在完成课题验收后，学生应在规定的时间内完成课程设计报告一份，报告的内容和要求如下。

6.3.1 报告的格式内容如下：

一、报告的字体段落格式如下：标题 1：黑体三号

标题 2：宋体加粗四号标题 3：宋体加粗小四正文字体：宋体小四

段落：固定值 20 磅、两端对齐图表标题：宋体 11 磅，居中

文中英文、数字：Times New Roman

二、报告内容安排如下：

1、设计的背景与内容

2、总体方案设计

根据课程设计的具体情况，描述算法的具体构架，包括：功能模块的划分、系统运行的环境、选用的工具及主要实现功能的原理。

3、功能模块的实现及分析

主要的功能实现和函数要进行详细的说明，包括其用法，使用范围，及参数等。

4、出现的问题

按课程设计要求，选用多个数据对程序进行测试，并提供机器学习算法的主要功能实现的效果图。并对调试中发现的问题做说明。

5、总结与体会

主要说明设计中学到的东西和取得的经验总结，心得体会。

6、参考文献

写出具体的主要参考文献，标明其作者、出处、年代、若是期刊文章，还需要给出期刊名。网络的文章要给出网址。要求：至少列出 6 个参考文献。

6.3.2 报告要求

1、必须按照以上格式书写报告。

2、必须对课程设计总体方案进行详细地说明。

3、详细说明各个功能模块的具体实现，对用到的主要函数及参数要做具体的说明，同时要有必要的实现流程图。

4、程序代码后必须贴上主要步骤实现的效果图。

7.课程设计的质量标准与成绩评定

本课程的考核方式为：对学生单独进行验收和答辩，学生必须演示程序，并回答教师提出的问题。根据验收答辩的情况和课程设计报告的质量综合给出成绩。

成绩构成为：平时成绩 10%，课程报告 70%，答辩 20%。