1、总体要求

期末课程设计一共五个题目备选，两到四人组队成一个小组共同完成其中一个题目。每队有一名队长，负责组织讨论，并组织小组同学参与最终课程报告演示与答辩。

2、报告规范

格式规范可以参考学校毕业论文设计格式要求。报告单独存放为 pdf 和 doc 文件，最终提交 PDF 和 doc 电子稿和打印稿，封面写“《机器学习课程设计》课程报告”以及姓名、学号、班级。

每个小组提交一套自己的源码，打包压缩为一个 zip 或 rar 文件，文件名包含组号和全体成员姓名。

3、实践课程设计题目

3.1 温度预报

温度预报对人类生产生活、经济发展和生态环境保护等方面都具有重要意义。随着科技的进步和社会需求的不断提高，温度预报研究也在不断深入和发展。然而，传统方法受限于大气系统的复杂性、观测资料的不足以及模式误差等因素，导致温度预报的精度难以显著提升。近年来，机器学习在气象领域的应用日益广泛，为温度预报带来了新的机遇。机器学习能够从海量气象数据中学习复杂的非线性规律，从而预测未来天气变化。相较于传统方法，机器学习预报具有精度高、时效性强等显著优势，已成为未来温度预报发展的重要方向。

数据说明：采用 Max Planck Institute for Biogeochemistry 的天气时间序列数据集，时间跨度为 2009 年至 2016 年，时间分辨率为 10 分钟。该数据集包含 14 个不同的气象特征，包括气温、大气压力、湿度等，为温度预报模型的训练和验证提供了丰富的数据基础。数据格式：csv。

要求：实现温度预报。

（1）读取对应文件中的数据。

（2）数据预处理、清洗、特征分析、特征表示设计、可视化等。

（3）对数据要素进行统计分析，挖掘不同数据要素间的关联性。

（4）划分训练集和测试集，选取不同模型实现温度预报，并对不同模型得到的预报结果进行比较分析及可视化。

（5）选做：设计实现一个（原型）系统，用于上述功能的展示。

3.2 表情识别

表情识别是计算机理解人类情感的一个重要方向，也是人机交互的一个重要方面。表情识别是指从静态照片或视频序列中选择出表情状态，从而确定人物的情绪与心理变化。20 世纪 70 年代的美国心理学家 Ekman 和 Friesen 通过大量实

验，定义了人类六种基本表情。人脸表情识别（FER）在人机交互和情感计算中有着广泛的研究前景，包括人机交互、情绪分析、智能安全、娱乐、网络教育、智能医疗等。

数据说明：FER-2013 数据集包括七种表情，分别是快乐，气愤，惊讶，害怕，厌恶、悲伤和中性。数据格式：图像。

要求：根据面部图像识别表情。

（1）读取对应文件中的数据。

（2）数据预处理、清洗、特征分析、特征表示设计、可视化等。

（3）设计数据增强功能。

（4）选取不同方法实现表情识别，并对不同方法得到的识别结果进行比较分析及可视化。

（5）选做：设计实现一个（原型）系统，用于上述功能的展示，尝试识别自己的表情。

3.3 学术论文主题聚类与分析

随着全球科研活动的蓬勃发展，学术论文数量呈现爆炸式增长，仅计算机科学领域每年就产生数十万篇研究成果。在这一背景下，学术论文主题聚类与可视化分析技术应运而生，其意义深远。首先，自动化主题聚类能有效组织和结构化海量文献，将相似研究方向的论文归类，帮助研究者快速定位相关文献，避免在浩如烟海的学术资源中迷失方向。其次，主题聚类能揭示学科内隐藏的知识结构和研究热点，展现不同研究方向间的关联与演化，为研究趋势分析和学科发展预测提供数据基础。再者，可视化分析通过将高维复杂的文本数据转化为直观的视觉表达，使研究者能一目了然地把握学术领域全貌，发现常规分析中易被忽视的模式与关联。随着人工智能技术不断进步，学术论文的智能分析与可视化将越发精准和个性化，成为知识发现和科研协作的关键基础设施，为加速科学进步和知识传播提供强大动力。

数据说明：UCI 公开的 AAAI 接收论文数据集（包含约 400 篇论文的标题、作者、关键词和摘要）。数据格式：csv。

要求：根据论文主题进行聚类分析。

（1）读取对应文件中的数据。

（2）数据预处理、清洗、特征分析、可视化，选择合适的向量化方法构建论文特征表示。

（3）选取不同算法进行主题聚类分析，并对不同模型得到的预报结果进行比较分析及可视化。

（4）通过提取代表性关键词分析每个聚类的主题特点。

（5）选做：从零开始实现一种聚类方法，即不使用封装好的聚类函数库实现上述功能。

3.4 字符匹配验证码识别

验证码（CAPTCHA）最初作为区分人类与机器的安全机制而诞生，用于防止恶意程序的自动化攻击。然而，随着人工智能技术的飞速发展，验证码识别技术已成为计算机视觉和深度学习领域的重要研究方向。验证码识别在多个方面具有重要意义。首先，它推动了 OCR（光学字符识别）技术的进步，促使识别算法在处理扭曲、噪声干扰等复杂文本场景时更加精准。其次，在自动化测试领域，验证码识别能够提高网站和应用程序的测试效率，减少人工干预，加速软件开发周期。在学术研究方面，验证码识别是衡量人工智能系统能力的重要标准，研究人员通过不断突破验证码识别的难题，推动了计算机视觉算法的创新。同时，这种“攻防博弈”也促使安全研究人员开发更先进的验证机制，形成技术上的良性竞争。总之，验证码识别技术在安全研究、人工智能发展和提升用户体验等方面具有深远意义，成为数字化时代不可或缺的技术领域。

数据说明：训练集共有 9000 个样本，序号 0000~9999，每个样本对应一个文件夹，文件夹中包含：一个与文件夹同名的图片，为验证码图片，此验证码包含 4 个中文字符；9 个单字图片，序号 0~8，其中包含验证码图片中的 4 个字符；train_label.txt 对应样例序号和标签。数据格式：图像。

要求：验证码中包含 4 个中文汉字和 9 个中文单字，要求从 9 个单字中按顺序选出验证码中的汉字，实现对 test 数据集的字符匹配。

（1）读取对应文件中的数据。

（2）数据预处理、特征分析、特征表示设计、可视化等。

（3）针对字符匹配验证码，选取不同的机器学习算法，进行验证码识别，并对不同模型得到的结果进行比较分析及可视化。

（4）结合不同的算法的特点使用准确率作为评价指标分析结果之间的差异。

（5）选做：设计实现一个（原型）系统，用于上述功能的展示。

3.5 基于回归分析的大学综合得分预测

大学排名是一个非常重要同时也是具有挑战性与争议性的问题，一所大学的综合实力涉及科研、师资、学生等方方面面。目前全球有上百家评估机构会评估大学的综合得分进行排序，而这些机构的打分也往往并不一致。根据世界各地知名大学各方面的排名（师资、科研等），一方面通过数据可视化的方式观察不同大学的特点，另一方面希望构建机器学习模型（线性回归）预测一所大学的综合得分。

数据说明：基本输入特征有 8 个：world_rank，institution，region，national_rank，quality_of_education，alumni_employment，quality_of_faculty，publications，influence，citations，broad_impact，patents，score，year；预测目标为 score。数据格式：csv。

要求：对数据中的大学综合得分预测。

（1）读取对应文件中的数据。

（2）数据预处理、清洗、特征分析、特征表示设计、可视化等。

（3）随机划分训练集和测试集，使用不同的回归模型进行综合得分预测，用 RMSE 作为评价指标，并对回归模型的系数进行分析。
（4）尝试将离散的地区特征融入线性回归模型，并对结果进行对比分析。
（5）选做：从零开始实现一种回归分析方法，即不使用封装好的函数库实现上述功能。