2024 秋机器学习平台课
大作业

一、信用卡反诈预测

1.目标

本作业旨在构建一个信用卡反欺诈预测模型，基于信用卡的历史交易数据，通过机器学习方法提前识别可能的欺诈行为。在现实场景中，信用卡欺诈检测对于金融机构和持卡人都至关重要，尤其是在交易频繁且快速的环境中。通过检测模型能够及时发现异常交易，减少潜在的经济损失，提升客户账户的安全性。

2.数据集

该数据集包含了 284,807 条信用卡交易记录，其中仅有 492 条为欺诈交易，反映了信用卡欺诈行为的稀少性，数据分布高度不平衡。数据特征包括交易时间、交易金额，以及经过主成分分析（PCA）处理的 28 个数值特征（V1-V28），以保护用户隐私。数据集可从 Kaggle 获取，用于模型训练和评估。

下载链接：https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

3.数据预处理与分析

首先，通过数据打印了解数据的基本特征。打印数据表格的部分记录，查看 28 个数据维度的分布情况，以便初步了解数据结构。在预处理过程中，针对规格与其他特征差异较大的数据进行标准化处理，例如通过特征缩放减小量纲差异对模型的影响。

其次，使用数据可视化方法来深入分析正负样本的分布情况。通过直方图或扇形图展示正负样本的分布特点，有助于更直观地理解数据集特性，并为后续模型训练提供参考。

接着，进行特征相关性分析，探索在信用卡被盗刷事件中各变量间的关系。例如，分析欺诈行为与交易金额的关系、消费行为与时间之间的关系等，以识别与欺诈事件更相关的特征，从而提升模型的预测能力。

最后，为处理类别不平衡问题，使用欠采样或过采样方法调整样本比例，以应对欺诈交易数量极少的情况。此外，应用 SMOTE 等方法生成合成的欺诈交易样本，进一步增强模型对少数类的识别能力，提高检测效果。

4.模型选择

选择下面的至少三种模型进行实现

- Logistic 回归
- 随机森林
- 朴素贝叶斯
- 线性判别分析
- 决策树

5.模型评估

使用准确率、精确率、召回率和 F1 分数评估不同模型，并将结果存储。

6.结果可视化并分析

将不同模型的预测指标结果进行可视化，并分析造成该结果的原因。

二、网络入侵检测

1.目标

本作业的目标是利用网络连接的数值型数据，构建一个入侵检测分类模型，用于识别网络入侵行为。通过机器学习算法及时检测潜在的异常连接，可以帮助企业主动预防和应对网络安全威胁，从而提高网络环境的安全性和稳定性。

2.数据集

数据集包含了多种网络连接记录的数值型特征，每条记录标注为正常或异常（入侵），用于模型训练和评估。特征包括连接的持续时间（duration）、协议类型（protocol_type，如 TCP、UDP）、源到目标的数据字节数（src_bytes）、目标到源的数据字节数（dst_bytes）等，全面描述了网络连接的特性。数据集可从 KDD Cup 1999 获取，提供丰富的网络入侵检测样本。

下载链接：http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

3.数据预处理与分析

首先，通过数据打印了解数据基本信息：随机打印部分数据记录，查看各特征的数据类型、值范围以及是否存在缺失值。然后，将类别特征（如 protocol_type）转换为数值编码，以便模型能有效处理这些特征。对于范围差异较大的特征，如 duration、src_bytes 和 dst_bytes，可以进行标准化或归一化处理，以减少量纲差异对模型的影响。

其次，使用数据可视化来更直观地理解数据特征。通过柱状图或饼图展示协议分布（如 TCP、UDP），观察是否某些协议类型在入侵事件中占比更高。此外，绘制入侵和正常连接的 duration 特征直方图，比较其持续时间分布，判断入侵连接是否存在明显的时间特征差异。

最后，进行数据分析以深入挖掘特征与目标变量之间的关系。首先，通过计算各特征与目标变量（如入侵与否）的相关性，识别对模型性能影响较大的特征（如 src_bytes 和 dst_bytes）。接着，对异常值进行分析，例如高流量连接是否更可能为入侵连接。此外，若类别不平衡（如正常流量远多于入侵流量），可采用欠采样或 SMOTE 方法进行平衡，以确保模型在各类别上的表现稳定。

4.模型选择

选择下面的至少三种模型进行实现

- K-means 聚类
- 随机森林
- 朴素贝叶斯
- 决策树

5.模型评估

使用准确率、精确率、召回率和 F1 分数对模型进行评估，并将结果记录。

6.结果可视化并分析

将不同模型的评价指标可视化，使用柱状图展示准确率、精确率、召回率和 F1 分数，分析各模型在入侵检测任务中的表现差异，并讨论模型选择的原因和数据特征对模型结果的影响。

三、人脸图像补全

1.目标

本作业旨在利用不同的机器学习方法完成图像补全任务，借助模型对缺失的图像部分进行填补，恢复人脸图像的完整性。通过本作业，深入分析不同补全方法在实现方式和结果表现上的差异，有助于理解图像补全技术的优缺点。

2.数据集

数据使用 CelebA 人脸数据集，包含大量人脸图像，适用于图像补全任务。数据集可从 Kaggle 获取，为模型训练和测试提供了充足的样本。

下载链接：https://www.kaggle.com/datasets/jessicali9530/celeba-dataset

3.数据预处理与分析

首先，自行设计图像掩盖方法，为每张人脸图像增加遮挡区域，以形成待补全的图像输入。通过不同类型和大小的掩盖方式，形成多样化的输入，便于模型学习图像补全的泛化能力。

4.模型选择

选择以下至少两种模型实现图像补全任务：

- 变分自编码器
- 扩散模型
- 掩码自编码器

5.模型评估

使用均方误差（MSE）等定量指标评估图像补全的效果，衡量模型在还原图像细节上的表现。

6.结果可视化并分析

将补全后的图像与原始图像进行可视化对比，分析不同模型的补全效果和优劣，以更好地理解各模型在还原人脸特征方面的性能。

四、服装图像分类

1.目标

本作业旨在构建图像分类模型，对不同类型的服装进行分类。通过不同的机器学习方法完成分类任务，并分析不同模型在实现方式和结果表现上的差异，以

提高模型在服装图像分类任务中的效果。

2.数据集

数据使用 Fashion MNIST 服装图像数据集，涵盖多个类别的服装图像样本，适合图像分类任务。数据集可从 Kaggle 获取，为模型训练和测试提供丰富的数据资源。

下载链接：https://www.kaggle.com/datasets/zalando-research/fashionmnist

3.数据预处理与分析

根据数据集特点和模型要求，自行设计数据预处理方法，以优化数据在不同模型中的表现，包括图像尺寸调整、归一化等操作。

4.模型选择

选择以下至少三种模型进行分类任务：

- 结合降维方法的线性模型
- 朴素贝叶斯
- 层次聚类
- Adaboost
- 对比学习中的 SimCLR

5.模型评估

使用准确率、精确率、召回率和 F1-score 评估模型性能，比较不同模型的分类效果。

6.结果可视化并分析

通过混淆矩阵、ROC 曲线、PR 曲线等可视化分类结果，分析各模型的表现差异，总结模型选择和数据特征对分类结果的影响。

五、交通标志图像分类

1.目标

本作业旨在构建一个图像分类模型，对不同类型的交通标志进行分类。通过不同的机器学习方法完成分类任务，并对不同模型的实现方式和结果表现进行深入分析，以帮助提升交通标志识别的准确性。

2.数据集

数据使用 GTSRB - German Traffic Sign Recognition Benchmark 数据集，涵盖多种交通标志类别，适用于图像分类任务。数据集可从 Kaggle 获取，为模型训练和测试提供了丰富的数据支持。

下载链接：

https://www.kaggle.com/datasets/meowmeowmeowmeowmeow/gtsrb-german-traffic-sign

3.数据预处理与分析

基于数据集特征和后续使用模型的需求，自行设计数据预处理方法，包括图像归一化、尺寸调整等操作，确保数据与模型要求一致。

4.模型选择

选择以下至少三种模型进行实现：

- 结合降维方法的线性模型
- 朴素贝叶斯
- 层次聚类
- Adaboost
- 对比学习中的 SimCLR

5.模型评估

使用准确率、精确率、召回率和 F1-score 评估模型性能，全面分析模型在分类任务中的表现。

6.结果可视化并分析

通过混淆矩阵、ROC 曲线、PR 曲线等图表可视化分类结果，分析不同模型在分类任务中的表现差异，探讨模型选择和数据特征对结果的影响。

六、BBC 新闻文本分类

1.目标

本作业旨在构建一个新闻文本分类模型，基于不同的机器学习方法对新闻文本进行分类，并深入分析各模型在实现方式和分类结果上的差异。通过对新闻文本进行自动分类，有助于快速组织和检索信息，在新闻推荐、舆情监控等实际应用场景中具有重要价值。

2.数据集

使用 BBC 新闻分类数据集，该数据集包含 2225 篇文章，每篇文章被标记为以下五个类别之一：商业（business）、娱乐（entertainment）、政治（politics）、

体育（sport）、科技（tech）。数据集划分为 1490 条记录的训练集和 735 条记录的测试集，适用于模型的训练和评估。数据集可从 Kaggle 获取。

下载链接：https://www.kaggle.com/competitions/learn-ai-bbc/data

3.数据预处理与分析

首先，通过解压数据和读取文件对数据进行准备，打印数据集的前几条记录以查看数据内容。在数据预处理过程中，进行以下步骤：

- 文本清洗：去除标点符号、数字及多余空格以规范化文本内容。
- 分词：将文本拆分为单词，提高模型对文本信息的理解。
- 特征向量化：使用 TF-IDF 方法将文本转换为数值格式，以便模型处理文本特征。

4.模型选择

选择以下三种模型实现文本分类任务：

- Softmax 多分类
- 朴素贝叶斯
- 随机森林

5.模型评估

计算准确率、精确率、召回率、F1-score 等指标评估模型性能，并绘制 ROC 曲线和 PR 曲线分析分类效果。除此之外，对一篇数据集外的文章进行分类，观察模型的分类效果，以评估模型的泛化能力。

6.结果可视化并分析

将不同模型的分类结果进行可视化，使用混淆矩阵、ROC 曲线和 PR 曲线

展示各指标，分析各模型的优劣及其对不同类别的表现，探讨模型选择对结果的影响。

七、微博谣言检测

1.目标

本作业旨在利用新浪微博不实信息举报平台抓取的中文谣言数据，通过机器学习方法进行谣言检测。随着社交媒体的广泛应用，虚假信息的传播速度越来越快，因此，建立一个高效的谣言检测系统对防止信息误导和维护网络安全具有重要意义。

2.数据集

该数据集包含 1538 条谣言和 1849 条非谣言数据，数据来源于新浪微博不实信息举报平台。每条数据均以 JSON 格式提供，其中 text 字段代表微博原文的文字内容。数据集可从 百度 AI Studio 获取，用于模型的训练和评估。

下载链接：https://aistudio.baidu.com/datasetdetail/20519

3.数据预处理与分析

首先，通过解压数据文件并读取解析数据，生成 all_data.txt 文件和数据字典（dict.txt），并划分训练集和测试集。在数据准备阶段，打印训练集的前几条数据，以查看数据格式和内容，为后续处理提供参考。

4.模型选择

选择以下至少三种模型进行谣言检测任务：

- 逻辑回归

- 决策树
- 随机森林
- 朴素贝叶斯

5.模型评估

使用至少两种评价指标（如准确率、精确率、召回率等）对模型进行评估，并通过可视化方法展示模型的评估结果。可以通过绘制混淆矩阵、ROC 曲线等方式分析模型在谣言检测任务中的表现。

6.结果可视化并分析

对不同模型的评估结果进行可视化，分析各模型在谣言检测中的优劣，讨论模型选择和数据特征对检测效果的影响。

八、北京空气污染序列预测

1.目标

本作业旨在利用北京的历史空气污染数据，通过机器学习方法预测未来一天中特定时刻的 PM2.5 浓度值。准确的空气质量预测能够帮助公众提前做好健康防护，并为相关部门提供环境管理的参考数据，有助于改善城市空气质量。

2.数据集

数据集包含了 2010 年 1 月 1 日至 2014 年 12 月 31 日北京的空气污染数据，数据格式为结构化表格，包含时间信息（如 year、month、day、hour）以及空气质量相关指标（如 pm2.5、DEWP、TEMP、PRES、cbwd、Iws、Is、Ir 等）。数据集可从 百度 AI Studio 获取，为模型训练和评估提供了丰富的样本。

下载链接：https://aistudio.baidu.com/datasetdetail/55547

3.数据预处理与分析

首先，读取数据集并检查数据内容，查看缺失值并对其进行填充处理。接下来，移除索引值，并将时间字段作为索引，以便更好地处理时间序列数据。随后，将数据划分为训练集（80%）和测试集（20%），并对数据进行标准化处理，确保模型输入的稳定性。

4.模型选择

选择以下三种模型实现空气污染预测任务：

- 线性模型
- 决策树
- 随机森林

5.模型评估

使用至少两种评价指标（如均方误差 MSE、均方根误差 RMSE 等）评估模型的预测效果，并通过可视化展示模型在预测任务中的表现。

6.结果可视化并分析

通过可视化方法展示不同模型的预测结果，分析各模型在 PM2.5 浓度预测中的优劣。结合实际数据讨论模型选择、时间序列特征对预测结果的影响，并总结模型的适用性。