SLDS 2023: Project 2

2023.5.15

SUSTech Southern University of Science and Technology

项目1：信用评分

SUSTech Southern University of Science and Technology

- 项目简介：运用所学的统计学习方法，设计并实现一个信用评分模型。信用评分模型广泛应用于金融行业，用于评估客户的信用风险。项目使用真实的金融数据集，对客户的信用评分进行预测。可以尝试使用不同的机器学习算法，如逻辑回归、决策树、随机森林等，并比较各算法的预测性能。

- 数据集：German Credit Risk, Default of Credit Card Clients Dataset

- 项目提交内容：
  - 一份包含完整Python代码的Jupyter Notebook或.py文件，其中应包括注释以解释代码的作用。
  - 项目报告，包括以下内容：项目背景与目的；数据集描述；数据预处理方法与过程；特征选择依据；选用的算法及原因；模型评估方法与结果；模型优化策略与过程；可视化结果展示；结论与建议

项目1：信用评分

SUSTech Southern University of Science and Technology

- 项目要求：
  - 数据预处理：清洗数据，处理缺失值、异常值、离群值等。对数据进行归一化或标准化处理，使其满足统计学习算法的要求。
  - 特征选择：根据业务背景和数据特点，选择合适的特征变量。
  - 模型选择与训练：尝试使用至少三种不同的机器学习算法，对数据进行训练。
  - 模型评估：使用交叉验证等方法，对模型的预测性能进行评估，包括准确率、召回率、F1分数等指标。
  - 模型优化：根据评估结果，对模型进行调参优化，提高预测性能。
  - 可视化：通过可视化展示模型的预测结果，以及各特征对模型的影响程度。
  - 文档撰写：撰写项目报告，详细记录项目的背景、方法、过程、结果和分析。

项目2：异常检测

SUSTech Southern University of Science and Technology

- 项目简介：运用所学的统计学习方法，设计并实现一个异常检测模型。异常检测模型应用识别时间序列数据中的异常。项目需使用时间序列的数据集，对数据中出现的异常情况进行监测。可以尝试使用不同的机器学习算法，如支持向量机、决策树、孤立森林等，并比较各算法的预测性能。

- 数据集：KDD Cup 99

- 项目提交内容：
  - 一份包含完整Python代码的Jupyter Notebook或.py文件，其中应包括注释以解释代码的作用。
  - 项目报告，包括以下内容：项目背景与目的；数据集描述；数据预处理方法与过程；特征选择依据；选用的算法及原因；模型评估方法与结果；模型优化策略与过程；可视化结果展示；结论与建议

项目2：异常检测

SUSTech Southern University of Science and Technology

- 项目要求：
  - 数据预处理：清洗数据，处理缺失值、异常值、离群值等。对数据进行归一化或标准化处理，使其满足统计学习算法的要求。
  - 特征工程：将原始时间序列数据转换为适合统计学习算法处理的特征表示。
  - 异常检测算法选择：选择合适的异常检测算法，如支持向量机 (SVM)、孤立森林 (Isolation Forest) 等。根据数据集的特点和异常检测目标，可以选择单一算法或结合多个算法进行检测。
  - 模型训练和评估：对数据集进行训练，并评估模型在异常检测任务上的性能。评估指标可以包括准确率、召回率、精确率、F1 分数等，同时还可以绘制混淆矩阵、ROC 曲线等来可视化结果。
  - 结果解释和可视化：解释模型的结果，并使用可视化技术展示异常检测或预测的效果。包括绘制异常分数的分布图、突出显示异常样本等。

与Project 1的不同

SUSTech Southern University of Science and Technology

- 算法创新性：探索新的算法、改进现有算法或结合多个算法以提高性能和鲁棒性。
- 算法比较和分析：进行更全面的算法比较和分析。可以使用多个评价指标、采样不同的数据子集、使用交叉验证等方法来评估算法的性能，并深入分析算法的优缺点。
- 参数调优：进行更深入的参数调优工作。可以使用网格搜索、贝叶斯优化等方法来寻找最佳的参数组合，以最大化算法的性能。
- 可解释性和可视化：关注模型的可解释性和可视化。可以使用特征重要性分析、异常样本的可视化等方法来解释模型的决策过程，并提供直观的结果展示。
- 实际应用和影响：思考方法在实际应用中的潜在影响。可以讨论如何将这些方法应用于现实世界的问题，并提出改进和推广的建议。