****

人工智能与机器学习

数据集导入和处理实验

实 验 报 告

|  |  |
| --- | --- |
| **学 院** | 网络空间安全学院 |
| **专 业** | 网络空间安全 |
| **班 级** | 22270311 |
| **学 号** | 22200215 |
| **学生姓名** | 王翔宇 |
| **教师姓名** | 唐会军、赵治栋 |
| **完成日期** | 2024.06.13 |
| **成 绩** |  |
| **数据集导入和处理编程实验** | | |
| 1. **实验任务** 2. 熟练掌握scikit-learn库中datasets模块的各种命令； 3. 对任一数据集（除鸢尾花数据集以外的其他数据集），导入该数据集； 4. 将numpy数据转化为DataFrame格式； 5. 对该数据集进行可视化； 6. 使用K折交叉验证法将数据集划分为训练集、验证集、测试集； 7. 实验总结。 | | |
| 1. **实验分析**   任务1：  在scikit-learn库中，datasets模块提供了多种用于加载示例数据集的功能。这些数据集通常用于演示机器学习算法的使用，因为它们具有相对简单的结构和已知的特性。以下是一些datasets模块中常用的命令和功能的概述：   1. 加载数据集    * load\_<dataset\_name>(): 加载并返回一个包含数据（特征）和目标（标签）的字典。这些函数通常返回NumPy数组，以及一些其他元信息（如特征名称、描述等）。例如，load\_iris() 加载鸢尾花数据集。 2. 数据集名称    * datasets.fetch\_names(): 返回一个包含所有可获取数据集名称的列表。 3. 数据集详情    * 加载的数据集通常包含一个名为DESCR的属性，它包含了数据集的详细描述。   任务2：  导入威斯康星州乳腺癌数据集        任务3：  将numpy数据转化为DataFrame格式    将标签添加进可视化数据中    任务4： 绘制散点图矩阵表示数据集，矩阵的对角线是每个特征的直方图，其他位置为通过两两特征绘制的散点图      由于该数据集样本包含特征过多，所以直接筛选了部分特征进行可视化表示  任务5：  使用K折交叉验证法将数据集划分为训练集、验证集、测试集    首先将数据集划分为训练集和测试集，之后对训练集进行K折交叉验证法划分为训练集和验证集  实验总结：  经过上述实验，我成功地掌握了如何使用scikit-learn的datasets模块加载除鸢尾花数据集外的其他数据集，如乳腺癌数据集。我通过将数据集的数据部分转换为pandas DataFrame，不仅方便了后续的数据分析和可视化，还提高了数据处理的灵活性。通过K折交叉验证法，我能够有效地将数据划分为训练集、验证集和测试集，为模型的训练和评估提供了有力的数据支持。整个实验过程不仅加深了我对scikit-learn库的理解，也提升了我处理和分析实际数据集的能力。   1. **实验过程中遇到的问题和解决方案**   问题：  对scikit-learn的datasets模块一知半解，不了解具体的用法以及函数各个参数的具体含义。  解决方案：  在CSDN上搜索了相关的学习教程，明白了Scikit-learn是一个非常知名的Python机器学习库，它广泛地用于统计分析和机器学习建模等数据科学领域。通过对它的datasets模块仔细学习，完成了本次实验 | | |