INT104W04_CW作业要求v0.2

任务

- 使用箱型图（box plot）观察原始数据分别，讨论为减少因原始数据规模不同而导致的影响所应当采取的措施。
- 对数据进行主成分分析（PCA），观察组成成分的分布，找到一组能使程序分类更容易的组成成分。
- 用你自己的方式，提取出易于分类学生所属专业（programme）的特征。
- 可视化并比较原始特征（raw），缩放特征（scaled），PCA特征和你自己的结果特征。

原始数据格式

- 学生索引（1~619）
- 学生性别（1，2）
- 学生专业（1，2，3，4）
- 学生年级（2,3）
- 学生总分（满分100）
- 6道题小分（MCQ=54，Q1=8，Q2=8，Q3=14，Q4=10，Q5=6）

要求

您必须使用Python来执行任务。在实验过程中，将为您指派一名助教（TA）来支持您的工作。当您完成实验时，请确保您生动地展示了您的工作（以确保实验是由您设计和执行的）。但是，请记住，助教没有责任教你Python编程，也没有责任为你设计实验。

在实验室会议之后，您应该写一份实验室报告，记录您进行的实验、获得的结果以及证明您推荐的特征提取方法的讨论。

已经单独提供了指导实验的MATLAB脚本。请提醒学生：

1）用MATLAB实现所提供的MATLAB脚本将导致零分；2） 用Python实现所提供的MATLAB脚本并不能保证获得高分。

在现场演示中，您将被问到不超过三个与以下内容有关的问题：

1）您的代码，2）您使用的算法，以及3）您获得的结果。

您可能还需要根据请求对Python脚本进行一些小的更改，并解释相应的结果。
实验室会议后的实验室报告长度不应超过3页，双栏（参考IEEE格式），不包括参考列表。报告可以简单地命名为“实验室报告”（“lab report“），但学生也可以用自己喜欢的方式命名报告。报告不需要封面，学生应在标题下写下自己的名字，并提供学生id。学生还应以自己的名字命名指定的助教。虽然可以引用文献来支持报告中的观点，但没有必要审查报告中的相关文献，因此报告中不引用论文绝对没有问题。实验室报告应与一个PDF文件一起提交，不附带源代码。

强烈建议使用latex。
ChatGPT仅允许用于校对和头脑风暴。然而，将人工智能生成的解决方案复制到任务中并不能保证你能通过课程。重要的是批判性思维、实验设计和结果分析。你必须完全理解你的代码和你在本课程中设计的实验。

评分标准

实验室报告：
编辑和语言问题（10分）
10分：未发现格式问题。
8分：轻微的语言问题或格式问题。
6分：报告总体不错，有一些语言问题和格式问题。
4分：这份报告几乎不可读。
2分：这份报告很难阅读，但可以理解。
0分：报告不可理解。

任务1、2、3和4（共60分，每个任务值15分）
15分：科学假设已通过所提供的结果得到证实。
12分：对不同实验配置的结果进行比较分析。
9分：将不同实验配置的结果与深度进行比较。
6分：任务目标已全部完成。
3分：通过良好的尝试，任务的目标没有部分实现。
0分：任务的目标没有完全实现，也没有做出合理的尝试。

现场演示：
回答问题（共15分，每题5分）
5分：表现出对概念的充分理解，并提供令人满意的答案。
4分：回答满意。
3分：回答满意，但有轻微误解。
2分：答案勉强令人满意。
1分：答案不正确。
0分：学生无法回答问题。

代码运行（15分）
15分：可以以高效的方式实现代码，并在对算法有良好理解的情况下预测结果。
12分：可以根据需要实现代码并深入讨论结果。

9分：可以在帮助下执行代码，并显示对结果的理解。
6分：可以在一段时间内在帮助下实施代码，并表现出一定的理解的结果。
3分：不能在对结果有合理预期的情况下实施所需的更改。
0分：无法理解所需更改的意图。

注意：对于此处未列出的情况，TA将匹配列表中的条件，并为代码运行会话打上标记。TA也有权标记涉嫌抄袭，并将案件提交给模块负责人。

奖励分数（总分上限为100分，无个人奖励上限）
+10分：在任何任务中证明一个新颖的科学假设。
+5分：以一种其他人可以轻松地重新实施实验的方式呈现实验。
+5分：报告格式可发布。
处罚：

-10分：不当引用

-20分：严重不当引用（多次不当引用或整段重复）-大学学术诚信处罚适用。

提交

只接受PDF格式的提交文件。

Submit your lab report via the dedicated Learning Mall coursework link before the Friday of week 8.

Please name your submission file as ID_FirstName_LastName_C1.pdf (e.g., 1234567_FirstName_Surname_C1.pdf).

Late submission policy of XJTLU applies.

作业技巧

CW1作业目的

提取数据特征，分析数据（特征）的分布与学生专业间的关系

作业任务

0.数据清洗

本次数据非常干净，但我们依然需要提到进行过此步骤，没有发现异常值（缺失，离群，错误）

（可叙述原本计划如何做）

1.观察原始数据特征

探索性数据分析（Exploratory Data Analysis），可视化原始数据特征，找到数据特征/模式（pattern）和你对数据的见解（insights）

看数据是否符合高斯分布（正态分布）（可以使用高斯密度点云图，不是规则的圆形，或偏度）看每一个列的标准差离群值等。高斯用皮尔森，非高斯分布用spearman相关系数。

相关性分析

可以预先去除几个不相关feature（性别，序号）

数据预处理（归一化，编码转换，minmax？zscore？）归一化标准化的应用条件，使用不同

去除数据规模不同对分析可能造成的影响，标准化或归一化，

2.主成分分析

对数据进行PCA，观察新空间中各组成成分的分布，找到一组能使程序分类更容易的特征向量组合。

同时分析你选择的特征向量，表明选择的原因（该方向上的方差比（特征值），数据分布情况）

可尝试选择2个或3个特征向量作图

或也可尝试用NMF降维（矩阵拆解），探索组合模式

也可以在PCA之前做fusion，特征工程

3.特征提取

用你自己的方式，提取出易于分类学生所属专业（programme）的特征。

4.总结

比较原始特征（raw），缩放特征（scaled），PCA特征和你自己的结果特征。

效果最好的降维的结果作为cw2的输入。

经过你的处理和选择，数据特征的变化情况。

你提取的特征是否能成为CW2中分类器的通道（大约5个候选特征）？说明原因

注意：

代码规范，如大小写命名格式，代码格式等，可能是扣分项

每张图片需要有在文章中有引用（如fig.1），图的坐标轴，标题，图例名字等不能漏。

每个公式下方需要标注其中每个变量和部分的意义。

在使用每种方法算法时建议引用提出该方法或相关的论文，数量不用多。

可以使用typst简化latex写作

CW2分类器

介绍

超参数-网格化搜索（穷举最优参数）

解释工作原理（scboost）h2o分类器（穷举各种分类结果）

放分类结果图（可视化）与准确率表格

分类器详细讲两三个（表格中可以展示试验了很多方法，但最终选择效果最好的两个）当然最后还要提到目前的缺点和可以提升之处（critical thinking）

CW3聚类器

主要使用kmeans dbscan