Project

Professor Julien Maitre,
Ph.D. Winter 2022

This project is graded. The grade is 100 points and represents a percentage of 40% in the final grade for this course. You must form groups of 7 or 8 persons to achieve this lab.

这个项目是计分的，满分为100分， 占本课程期末成绩的40%。你必须组成7到8人的小组来完成这个实验。

1. General Description

The general objective of this project is to confirm your knowledge in Python and Data Science at the end of the course.

In this project, you should produce a scientific report, which describes each step you made until the final results. Then, of course, I expect you to analyze and interpret the results. The page ***limit for the scientific report is 15 pages***. ***All student names of the group should appear on the first page.***

本项目的总体目标是在课程结束时确认您在Python和数据科学方面的知识。

在这个项目中，你应该制作一份科学报告，描述你所做的每一步，直到最终结果。当然，我希望你们能分析和解释结果。科学报告的页数限制为15页。小组的所有学生名字都应该出现在第一页上。

2. Formalités

The deadline for submitting your work is March 11, 2022, at 11.59 p.m (China time). After this deadline, there will be a penalty of 10% per day of delay.

You will email me a WeTransfer link with the scientific report and code.

提交作业的截止日期为2022年3月11日晚上11:59（中国时间）。在此截止日期后，每延迟一天将被处以10%的罚分。

您将通过电子邮件向我发送带有科学报告和代码的WetTransfer链接。

3. What is expected?

The scientific report should include:

- the description of your dataset
    - For example:
        - what are the variables?;
        - the meaning of each variable;
        - the number of instances;
        - the number of classes (***it is an obligation that the dataset can be classified***);
        - the values (e.g., min-max interval) that each of the variables can take?.

- 数据集的描述
  - o 例如：
    - ▪ 有哪些变量？；
    - ▪ 每个变量的含义；
    - ▪ 实例数量；
    - ▪ 类别的数量（数据集必须是可分类的）；
    - ▪ 每个变量可以取的值（例如，最小值-最大值区间）？
- the description of each step for the creation of the dataset
  - You should :
    - define a sliding time window length and an overlap (to define as user parameters);
    - extract features (those from the course and others – at least three new features – that you will find on the internet);
    - store the features in a Pandas DataFrame;
    - give a name at each column of the DataFrame (name of the features);
- 创建数据集的每个步骤的描述
  - o你应该：
    - ▪ 定义滑动时间窗口长度和重叠区域（定义为用户参数）；
    - ▪ 提取特征（你在课程中学到的，或者从互联网上找到的特征——至少三个新特征）；
    - ▪ 使用Pandas包中的DataFrame格式存储特征；
    - ▪ 在 DataFrame 的每一列给出一个名称（即，特征的名称）；
    - save the DataFrame in a .pickle file.

- data checking and pre-processing
  - For example:
    - A summary of the number of instances per class after the pre-processing;
      - ➢ **a pre-processing can include:**
        - ✓ **limiting the number of instances per class;**
        - ✓ **a class should have a minimum number of instances to exist;**
        - ✓ **a normalization of the values;**
      - ***We did not see that in the course, but it exists very easy functions to use in scikit-learn.***
    - What are the statistics (e.g., mean, variance, std) for each variable?;
  - ***In this part, do not hesitate to use data visualization tools.***

- 数据检查和预处理
  - o 例如：
    - ▪预处理后每个类的实例数概括；
      - ➢预处理可以包括：
        - ✓限制每个类的实例数量；
        - ✓一个类应该有最少数量的实例；
        - ✓数值的标准化；
      - ▪我们在本课程中没有看到这一点，但在 scikit learn 中存在非常容易使用的函数。
    - ▪每个变量的统计数据（如均值、方差、标准差）是什么？；
  - o 在本部分中，请尽可能地使用数据可视化工具。

- the description of the results obtained after the dimensionality reduction (reduce the number of features)
  - You should:
    - import your saved DataFrame previously;
    - apply a dimensionality reduction;
      - ➢ ***I showed you an algorithm in a live coding session.***
    - cite (with identification if applicable) the left features;
    - use data visualization tools to interpret your results;
    - re-run these steps from point 2 by defining a new window length and a new overlap to create a new dataset;
      - ➢ save all your results.
    - compare the two results from the dimensionality reduction;

• 描述降维后获得的结果（减少特征数量）

o 你应该：

▪ 导入之前保存的 DataFrame；

▪ 使用降维方法；

➢ 我曾展示过一个实时编码过程中的算法。

▪ 引用（如适用，带有标识）左侧特征；

▪ 使用数据可视化工具来解释结果；

▪ 通过定义新的窗口长度和重叠区来创建新的数据集，从第 2 点重新运行以上步骤；

➢保存所有结果。

▪ 比较降维后的两个结果；

- ***a comparative study of classification;***
  - ***I want several tests (train-test splitting 40/60, 50/50, 60/40, 70/30, 10 fold cross-validation with:***
    - ***3 lengths of time window ;***
    - ***2 differents overlaps***
  - ***You also should an analysis of the results and give an interpretation.***

• 对分类进行比较研究；

o 我想要几个测试（训练/测试的比例分为 40/60、50/50、60/40、70/30、10 倍交叉验证）：

▪ 3 个时间窗长度；

▪ 2 个不同的重叠区域

o 你还应该对结果进行分析并给出解释。

- a conclusion of the study
  - summarize the essential information of your work.

• 研究结论

o 总结工作的基本信息。

- a general conclusion
  - summarize what you appreciated, learned, appreciated less in this project.

• 一般结论

o 总结你在这个项目中的收获和感悟。

4. Précisions

Regarding the dataset, there is only one restriction. ***The number of classes should be more than 2.*** Finally, you will search on the Web to find a dataset in a field that interests you for more "fun" (e,g., bioinformatics, marketing, commerce, etc.). Here is a sample of web links that provide access to datasets:

- https://www.data.gov/
- https://www.reddit.com/r/datasets/
- https://www.reddit.com/r/data/
- https://registry.opendata.aws/
- https://rs.io/100-interesting-data-sets-for-statistics/
- https://www.kaggle.com/datasets
- https://archive.ics.uci.edu/ml/datasets.php
- https://datasetsearch.research.google.com/
- etc.

关于数据集，只有一个限制。*类别数量应超过2 个（3 类以上）。*最后，你将在网络上搜索一个你感兴趣的领域（例如，生物信息学、营销、商业等）的数据集。以下是提供数据集访问权限的 web 链接示例。

Annex A

Scientific report	
Content of your report	5/100
Description of your dataset	10/100
Description of the feature extraction process	20/100
Presentation of the results of the dimensionality reduction	20/100
Classification + Interpretations + Conclusion	25/100
Total	**80/100**

Python scripts	
Code	15/100
Comments	5/100
Total	**20/100**