详细设计V1

1. 读取用户输入并储存

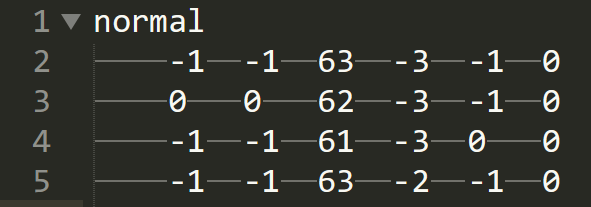
**前后台**：

* 1. 建立用户文件夹，用user\_id（或其它可作为唯一标识符的id）作为文件夹名。
  2. 在用户文件夹下，建立此次上传文件的子文件夹，用当前的time\_stamp作为子文件名用以标识同一用户的多次上传分析行为。

例如：user\_id/time\_stamp\_1/xxx.json

1. 数据预处理

**前后台**：将用户上传数据整理为要求格式，并存至timeseries.csv。完整示例文件见：timeseries.data。



说明：

第1行：class(classification) / value (regression)

第i行：以“\t”开始；各个信号量的值，用“\t”隔开；n个信号量则对应n个”\t”

连续两个”\n”标识开始一个新的sample。

**算法**：处理为可用作数据分析的标准数据格式（Time Series Data Format 参见<https://tsfresh.readthedocs.io/en/latest/text/data_formats.html> ，

暂时选定**Option 1**作为目前的格式），并存储为文件，文件名命名为timeseries.data。

1. 数据分析 PART 1 （Understand the Problem）

**算法**：将分析结果存至part1.csv中，结果说明：

第1行：1（Classification，分类）或者2（Regression，回归），代表问题的类别。

第2-n行：

* 1. 若为分类问题，则数据为不同类别的数量。（不同的类别用“\t”分隔，类别和数量直接与“:”分隔）

**前后台：**根据数据，生成柱形图并显示。

|  |
| --- |
| Class\_1: long\_1 \t Class\_2: long\_2 \t … |

* 1. 若为回归问题，数据包含如下信息：
     1. Descriptive statistics summary

**前后台**：显示数据（考虑用表格的形式展现UI）。

|  |
| --- |
| Descriptive statistics summary # Statistics\_item\_1: double\_1 \t Statistics\_item\_2: double\_2 … |

* + 1. Histogram

**前后台**：调用绘图工具库绘制“直方图”。

|  |
| --- |
| Histogram # Value\_1: count\_1 \t Value\_2: count \_2 … |

* + 1. Skewness and Kurtosis

**前后台**：显示数据

|  |
| --- |
| Skewness # value  Kurtosis # value |

示例文件见附件part1\_1.csv（分类），part1\_2.csv（回归）.

1. 数据分析PART 2 (Series Basic)

**算法**：将分析结果存至part2.csv中，结果说明：

* 1. 分析结果可能包含了多条Series的结果，不同的Series分析结果用“\n\n”分隔；
  2. 对于某一Series，结果包含：
     1. Descriptive statistics summary （数据格式及展示同3-b-i）
     2. Line chart（数据格式同3-b-ii，展示方式为“折线图”）
     3. Skewness and Kurtosis（数据格式及展示同3-b-iii）
     4. Rolling Statistics (mean, std)

**前后台**：显示数据并绘制折线图（同时显示mean和std）

|  |
| --- |
| Rolling mean # time\_1: Value \_1 \t time \_2: Value \_2 …  Rolling std # time\_1: Value \_1 \t time \_2: Value \_2 … |

* + 1. Decomposition into Trend, seasonality and residuals

**前后台**：显示数据并绘制折线图（同时显示Trend, Seasonality和Residuals）

|  |
| --- |
| Trend # time\_1: Value \_1 \t time \_2: Value \_2 …  Seasonality # time\_1: Value \_1 \t time \_2: Value \_2 …  Residuals # time\_1: Value \_1 \t time \_2: Value \_2 … |

示例文件见附件part2.csv.

1. 数据分析PART 3 (Feature relationship analysis)

**算法**：将分析结果存至part3.csv中，分析结果暂包含两类：

* 1. Basic分析了不同的feature与target之间的关系；
  2. Correlation matrix则是feature与feature之间的Correlation关系。
  3. 注意：如果目标为分类问题，则对于每一类别都会有相应的分析结果，用“\n\n”分隔。

数据格式说明（分类问题的某一类别/回归问题）：

1. Basic

|  |
| --- |
| N （表示后面有n行，代表n个feature）  Feature1 # feature1\_value\_1: target \_1 \t feature1\_value\_2: target \_2 …  Feature2 # feature2\_value\_1: target \_1 \t feature2\_value\_2: target \_2 …  … |

**前后台**：显示数据并绘制箱线图（若feature或target为分类特征）或散点图（若feature和target为连续性特征）

1. Correlation matrix

第一行：显示所有用以分析的n个feature，用“\t”分隔；

第2至n+1行，每行显示第i个feature与1-n个feature的correlation value。

|  |
| --- |
| Feature\_1 \t feature\_2 \t … feature\_n  Feature\_1 # corr\_1\_1 \t corr \_1\_2 \t corr\_1\_3 \t … corr\_1\_n  …  Feature\_n # corr\_n\_1 \t corr \_n\_2 \t corr\_n\_3 \t … corr\_n\_n |

示例文件见附件part3.csv.

Document Control

|  |  |  |  |
| --- | --- | --- | --- |
| Version | Details | Modified by | Date |
| 1.0 | 初始化 | Jie Zhou | 10/19/2019 |
|  |  |  |  |