非任务导向的时序数据探索 Non-mission-oriented time series data exploration

|  |  |  |  |
| --- | --- | --- | --- |
| Category | Sub-category | Details | Priority |
| Data pre-process | Missing data analysis | 1. Use only valid data, deleting the cases where data is missing. 2. Impute data using values from similar cases or using the mean value. 3. Impute data using model-based methods, in which models are defined to predict the missing values. | P1 |
| Outlier |  | P1 |
| Analyzing target | Descriptive statistics | Classification:   1. Proportion of categories | P0 |
| Regression:  1. Count, Mean, STD, MIN, 25%, 50%, 75%, MAX  2. Histogram (value, percentile)  3. Skewness and Kurtosis |
| Series | Basis | 1. Count, Mean, STD, MIN, 25%, 50%, 75%, MAX 2. Line chart (time, value) 3. Skewness and Kurtosis 4. Rolling Statistics: mean, std. 5. Decomposition into Trend, seasonality and residuals | P0 |
| Stationarity test | 1. ADF (Augmented Dicky Fuller Test) 2. KPSS 3. PP (Phillips-Perron test)   Tips: do some transformations to make the p-value within 5%, then we can assume the stationarity of the series. | P1 |
| AR/MA/ARMA processes test | MA - Next value in the series is a function of the average of the previous n number of values  AR - The errors (difference in mean) of the next value is a function of the errors in the previous n number of values  ARMA - a mixture of both | P1 |
| Feature importance analysis |  | P0 |
| Feature(variable) relationship analysis | Basis | 1. Linear (numeric) 2. Box (category) | P0 |
| Correlation matrix | heatmap style / zoomed heatmap style (top corrcoef variables) | P0 |

Ref:

1. <https://www.kaggle.com/jagangupta/time-series-basics-exploring-traditional-ts/notebook>
2. <https://www.kaggle.com/pmarcelino/comprehensive-data-exploration-with-python>
3. <https://www.quantstart.com/articles/Serial-Correlation-in-Time-Series-Analysis>
4. <https://tsfresh.readthedocs.io/en/latest/text/list_of_features.html>
5. https://github.com/blue-yonder/tsfresh

Schedule of P0 work items

|  |  |  |  |
| --- | --- | --- | --- |
| No. | Work item | Details | ETA (workday) |
| 1 | 理解tsfresh库中的重要feature | 1. 验证tsfresh提供的feature与自己计算得出的差异：为重要feature写注释。 2. 基于测试数据生成样例报告。 | 5 |
| 2 | 扩充tsfresh库的feature | 1. Rolling status 2. Decomposition into Trend, seasonality and residuals | 5 |
| 3 | 定义Series-Basis部分的输入输出：数据格式，文件命名方式，特征命名方式。  （同Analyzing target部分） |  | 2 |
| 4 | 实现3 | Python实现 | 3 |
| 5 | 定义Feature importance analysis部分的输入输出 |  | 2 |
| 6 | 实现5 | Python实现 | 4 |
| 7 | 定义Feature relationship analysis部分的输入输出 |  | 2 |
| 8 | 实现7 | Python实现 | 4 |