Advanced Data Mining

Piotr Lipiński

List of Assignments 2 – introduction to time series

Assignment 1. *(4 points)*
Weibull distribution is often applied in modeling survivability. Please find the definition of the distribution, see the probability density function and the cumulative distribution function. In Python, you may use the scipy.stats.weibull_min class.

Consider an online service and its users. Each user u visits the service for the first time at time t_u and remains active by T_u minutes (i.e. during T_u minutes he is visiting the service, more or less frequently, but after T_u minutes, he definitely stops to visit the service and never returns).
a) A group of N=1000 users visits the service for the first time at the time t_u=0 and remains active T_u minutes, where T_u is generated randomly with the Weibull distribution W(k=0.5, λ=60) for each user separately. How many users will be active in the successive 150 minutes after the time 0? How many users will be active after 1 day, after 2 days and after 5 days?
b) The same for the W(k=1.0, λ=60), W(k=1.5, λ=60) and W(k=2.5, λ=60). Try to explain the difference between these models from the practical point of view.
c) Try to replace the Weibull distribution with the Gaussian distribution and set its parameters in such a way that the numbers of active users after 1 day and after 2 days are approximately the same than in the case of the Weibull distribution. How many user will be active after 5 or more days? Try to explain the difference in results between the Weibull and the Gaussian distribution.
d) Simulate a group of N=1000000 users that visits the service for the first time at the time t_u generated randomly with the uniform distribution over one year, starting at time 0, for each user separately and remains active T_u minutes, where T_u is generated randomly with the Weibull distribution W(k=0.5, λ=60) for each user separately. Plot the number of active users in the successive days of one year starting at time 0.

Assignment 2. *(4 points)*
Please read the notebook *Introduction to Time Series Prediction* attached to the lecture notes.
a) Improve the notebook by splitting the dataset into the train (the first 65% of recorded data) and the test part (the last 35% of recorded data) and repeating the experiments with learning the model on the train dataset and testing on the test dataset.
b) Is it necessary to remove the trend and the seasonality before the regression? Try to check if it really affects the prediction accuracy.
c) In order to remove the trend, the original data are divided by the trend. Please try to replace it with subtracting the trend from the original data. How does it change the characteristic of the detrended data? Does it improve the prediction accuracy?
d) Similarly, in order to remove the seasonality, the seasonality is subtracted from the detrended data. Please try to replace it with dividing the detrended data by the seasonality. How does it change the characteristic of the preprocessed data? Does it improve the prediction accuracy?

e) Please look at the preprocessed data (i.e. the original data after removing the trend and the seasonality). What is the variance/standard deviation of the data? Is it approximately constant or it changes with time? If it changes with time, try to stabilize it and check whether it improves the prediction accuracy or not.
f) Please find at least 2 other benchmark time series, e.g. in the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets.php - select Time-Series in Data Type at the left panel), and try to predict them in the similar manner.

Assignment 3. *(5 points)*
Recall the general brute-force algorithm for finding characteristic time series shapelets described in our lecture or its detailed version described in [1].
a) Write a function for finding characteristic time series shapelets for a given dataset of time series.
b) Construct a few simple example datasets and verify the results of your function with some illustrations of the shapelets discovered and their location in some selected time series from the dataset.
c) Evaluate computation time of your function. Try to optimize it with your own improvements or the improvements proposed in [1].

Assignment 4. *(4 points)*
Recall the shapelet learning algorithm described in our lecture or in [2]. Use the algorithm to discover characteristic shapelets in water consumption data as well as the Gun Point Dataset and the Arrow Head Dataset. You may use either your own implementation or the implementation available in the pyts [3] package. Try to illustrate the results - present the shapelets discovered and their location in some selected time series from the dataset.

Assignment 5. *(non-obligatory, 4 bonus points)*
Please read the notebook *Bid and Ask Reconstruction* attached to the lecture notes. Please propose an approach to reconstruct bid and ask data from transaction data and validate the approach on some selected periods.

HINT: you may start with inventing *moving k-means*

References:

[1] L. Ye, E. Keogh, "Time Series Shapelets: A New Primitive for Data Mining". International Conference on Knowledge Discovery and Data Mining, 2009, pp.947-956.
[2] J. Grabocka, N. Schilling, M. Wistuba, L. Schmidt-Thieme, "Learning time-series shapelets". International Conference on Knowledge Discovery and Data Mining, 2014, pp.392-401.
[3] J. Faouzi, H. Janati, "pyts: A python package for time series classification". Journal of Machine Learning Research, 21(46), 2020, pp.1-6.
[4] N. Hug, " Surprise: A Python library for recommender systems", Journal of Open Source Software, 5(52), 2020, pp.2174.