ARISTOTLE UNIVERSITY OF THESSALONIKI

FACULTY OF SCIENCES SCHOOL OF INFORMATICS

MSC IN DATA AND WEB SCIENCE

Advanced Topics in Machine Learning

Team:

- Μωυσιάδης Γιώργος (141)

  gmousiad@csd.auth.gr

- Μικρού Νικόλαος (143)

  nikomikr@csd.auth.gr

Title: ***Handling Missing Values in Time Series Forecasting***

Introduction

Missing values in time series data can significantly affect the accuracy of predictive models. However, there is no single rule that can be applied to all situations to fill in missing values. Hence, there is a need to investigate different strategies to handle missing observations in time series. We aim to investigate various techniques for handling missing values in time series forecasting.

Project Objectives

The primary objective is to compare different methodologies for handling missing values in time series forecasting. Specifically, we aim to achieve the following goals:

1. Investigate standard linear interpolation and more sophisticated techniques for handling missing values in time series data.
2. Analyze the effect of missing values on time series forecasting models.
3. Compare different strategies for handling missing values in time series forecasting.
4. We will try to implement a heuristic method that will reconstruct time series and forecast the missing values, using multiple different models.

Approach:

The proposed methodology involves the following steps:

1. Data Preprocessing: We will select a suitable time series dataset that contains missing values. We will then preprocess the data by identifying missing values and deciding on a suitable strategy for handling them such as:
    a. Forward fill or backward fill
    b. Linear interpolation
2. Time Series Forecasting: We will use different models from the scikit-learn library, for time series forecasting, such as:
    a. Autoregressive Integrated Moving Average (ARIMA)
    b. Exponential Smoothing (ETS)
3. Performance Evaluation: We will compare the accuracy of different time series forecasting models using different methodologies for handling missing values. We will evaluate the performance of the models using classical metrics, such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Dynamic Time Warping (DTW) for the similarity of the final time series.

    The incorporation of DTW will provide a measure of how well the predicted time series aligns with the actual time series data, even if the two time series have different shapes or missing values.

Conclusion:

In summary, we aim to investigate different methodologies for handling missing values in time series forecasting:

1. Identifying the most effective strategies for handling missing values in time series forecasting using the different models that are provided in scikit-learn.
2. An analysis of the effect of missing values on time series forecasting models.
3. A heuristic evaluation of the performance of different time series forecasting models using different methodologies for handling missing values.

We will use scikit-learn library, and any other library we may happen to find online, to carry out our experiments.