Advanced Data Analysis and Machine Learning

Lecture: Sequential Data Characteristics and Issues

Lasse Lensu

2023-11-06

Outline

Time series characteristics

- Trend: a deterministic component that has no periodicity.
- Seasonality: a deterministic seasonal component with periodicity.
- Long-term period that is not connected with seasonality.
- Outliers: observations whose value lie outside the expected set of values.
- Noise representing a stochastic component in the series.
- Volatility representing abrupt changes in the series.

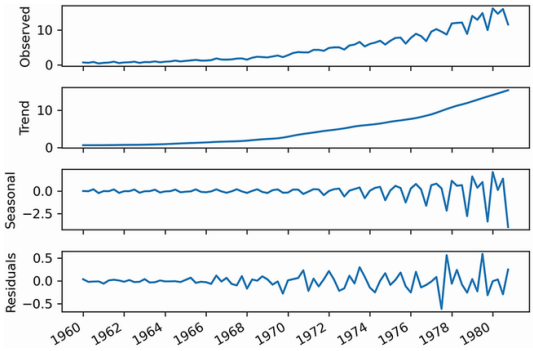

Figure: A typical time series [2].

Time series identification

- A time series is stationary if its statistical properties do not change with time, that is, it has a constant mean, variance, and autocorrelation.
- If a time series is not stationary, it can be transformed into one.
- Differencing can be used for stabilising the mean and applying a logarithm can be used to stabilise the variance.
- Autocorrelation function representing the relationship between lagged values of a series can be used to understand the process that generated the sequence.

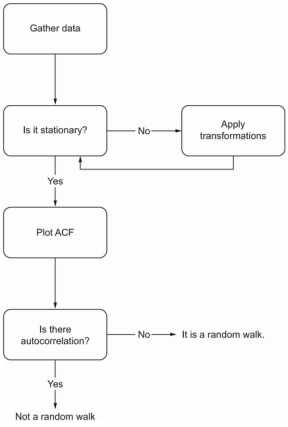

Figure: Identifying the type of time series [2].

Text characteristics

- Recently the focus has turned to finding the language-related rules from large collections of text.
- Modern natural language processing (NLP) relying on machine learning (ML) has enabled predicting the following [1]:
  - What's the topic of this text? (text classification)
  - Does this text contain abuse? (content filtering)
  - Does this text sound positive or negative? (sentiment analysis)
  - What should be the next word in this incomplete sentence? (language modeling)
  - How would you say this in German? (translation)
  - How would you summarize this article in one paragraph? (summarization)

Noise

- A stochastic component in the series that cannot be predicted and completely removed.
- The series can be either homoscedastic (constant variance) or heteroscedastic (non-constant variance).
- Noise can be reduced using different filtering methods.

Missing values

- In addition to errors, a time series may contain missing values causing gaps in data.
- Exclusion of these points/sequences in time would be wasteful.
- Data imputation as a potential remedy:
  - Last observation is used to fill in the gap.
  - Next observation is used to fill in the gap.
  - Statistics of the existing values in the form of a filtering.
  - Interpolation.

Outliers

- Outlier, *n. Statistics.* An observation whose value lies outside the set of values considered likely according to some hypothesis (usually one based on other observations); an isolated point. (OED)
  $\Rightarrow$ Unexpected, non-typical or out-of-range samples.
- Resulting from noise, human error or malfunctioning measurement/data processing equipment.
- The problem of defining outliers in a generally acceptable manner is nontrivial.
- Note: according to information theory [3], the most improbable events carry most of the information!

Outlier detection

- For low-dimensional data, outliers can be easy to identify using visualisation.
- Possible task formulation: Given a set of n data points and the expected number of outliers k, find the top k samples that are considerably dissimilar, exceptional or inconsistent with the rest of the data.
- Modelling can be used to represent the "valid" data, but the representativeness of the data and the validity of the model can significantly affect the outlier detection result.

Outlier detection

- Existing outlier detection methods:
  - Statistical distribution-based methods:
    Assumed distribution and its parameters; working/alternative hypothesis; statistical significance.
  - Distance-based methods:
    At least a fraction of data points lie at a distance greater than a threshold.
  - Density-based methods:
    Non-uniform distributions; local outliers and reachability; degree of being an outlier (not binary).
  - Deviation-based methods:
    Subsets of data points; largest reduction of a dissimilarity metric, or the statistical method within cubes.
  - Regression methods:
    Robust regression; residual error.

Summary

- A time series can be identified using standard techniques such as testing its stationarity and autocorrelation function.
- Typical issues with time series data include noise, missing values and outliers.
- Modern NLP has focused on finding the relevant rules of the language from data.

References

François Chollet.
Deep learning with Python.
Manning, Shelter Island, NY, second edition edition, 2021.

Marco Peixeiro.
Time series forecasting in Python.
Manning Publications Co, Shelter Island, NY, 2022.

Claude E. Shannon.
A mathematical theory of communication.
Bell System Technical Journal, 27(3):379–423, 1948.