Advanced Data Analysis and Machine Learning

Lecture: Sequential Data

Lasse Lensu

2023-11-06

Outline

1 Sequential data

2 Sequential data encoding

Time series

- A time series is a set of data points that are ordered in time.
- Univariate data is typically equally spaced in time, but series in multivariate data may have different spacings in time (sampling rates).
- It is common that the time series does not contain other features than the series itself, but it is possible to:
  - Compute additional features from the time series.
  - Make use of exogenous variables that are presumed to improve model performance.

Text [1]

- Natural languages differ from machine languages designed for machine communication and described by clear formal rules.
- In the case of natural languages, they were first used and the rules were defined later.
- Linguists and engineers have earlier defined complex sets of rules to enable natural language processing (NLP), but the success was limited.
- More recently the focus has turned to finding the language-related rules from large collections of text.

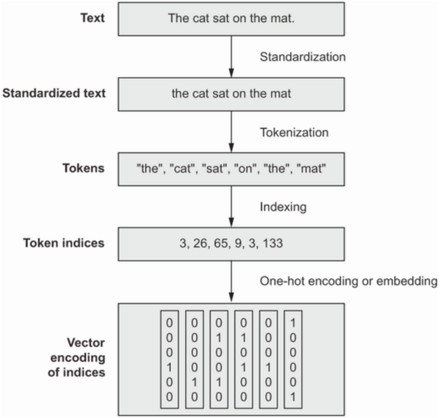
Time series encoding

LUT University

- Univariate or multivariate series
- Single or multiple variable types
- Sampling rate(s) in the case of multivariate series

Text encoding

- To make NLP easier, it is necessary to carry out preprocessing of natural language text [1]:
  - Text standardization
  - Vocabulary indexing
  - Text splitting (tokenization)
  - Text vectorization



Text: The cat sat on the mat.

Standardization

Standardized text: the cat sat on the mat

Tokenization

Tokens: "the", "cat", "sat", "on", "the", "mat"

Indexing

Token indices: 3, 26, 65, 9, 3, 133

One-hot encoding or embedding

Vector encoding of indices:

0 0 0 0 0 1
0 0 1 1 0 0
0 0 0 0 0 0
1 0 0 1 1 0
0 1 1 0 0 0
0 0 0 0 0 1

Figure: From raw text to vectors [1].

Text encoding [1]

- Groups of words represented as sets or sequences.
- Words are categorical features encoded as dimensions in a feature space or as category/word vectors, but how to encode the word order.
- Words in a sentence do not have a canonical order and languages differ from each other.
- The order can be discarded (bag-of-words) or considered as critical (sequences like time series), but also a hybrid approach exists complementing the separate words about their positions.
- Another detail is how the vector representation of the words is encoded.
- One-hot encoding problem: orthogonal vectors.

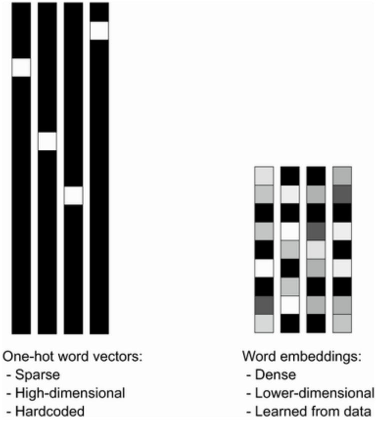


Figure: Word representations [1].

Summary

- A time series is a set of data points that are ordered in time.
- Machine languages are described by clear formal rules, but natural languages differ from them.
- In the case of multivariate time series, the variable types and sampling rate(s) are relevant.
- How to represent groups of words and encode word vectors are important design choices for NLP.

References

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