Stream Mining

One-Hot Encoding and DGIM

Zeno Adrian Weil

Data Science 1
Goethe University Frankfurt

7th of June, 2022

One-Hot Encoding

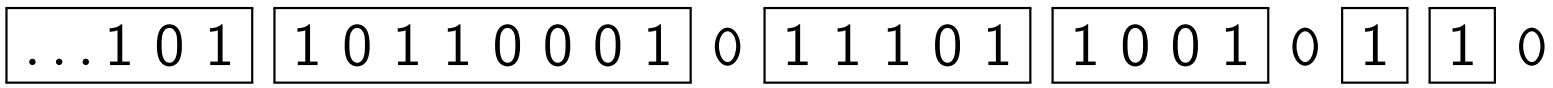
- categorical features
  - nominal (e.g. colours)
  - ordinal (e.g. satisfaction levels)
- need for numbers in algorithms
- naive approach: number serially
  - may introduce arbitrary orders
  - calculating non-sensical differences
- one-hot encoding
  - one binary feature for each possible value

The Datar-Gionis-Indyk-Motwani Algorithm

Objectives

- Estimate the number of ones in a bit stream!
- Be space-efficient!

- window size N
- $\mathcal{O}(\log_2 N)$ buckets
  - timestamp
  - size = number of ones
    - powers of two
    - one or two of each size
    - sizes never decreasing moving back
  - include all ones; end with ones
- estimation: sum of sizes of all included buckets + half the size of partially included bucket (if any)
  - error rate: 50%
- needs only $\mathcal{O}\left((\log_2 N)^2\right)$ many bits



estimation: 16 reality: 14

References

- Project code: `https://github.com/s9770652/DS1-DGIM`
- Mushroom data set: `https://archive-beta.ics.uci.edu/ml/datasets/mushroom`
- Streamlit: `https://streamlit.io/`
- Python package *dgim*: `https://pypi.org/project/dgim/`
- Description of one-hot encoding: `https://sherbold.github.io/intro-to-data-science/04_Data-Analysis-Overview.html#Features`
- Description of the DGIM algorithm: `http://infolab.stanford.edu/~ullman/mmds/ch4.pdf`