Data Processing & EDA

Charlie Beirnaert

Thomas More University of Applied Sciences
Business School
Data Science, Protection & Security

Data Science
Advanced

Outline

Outline

Recap

- Machine learning workflow
- Data format:
  - Matrix format
  - No missing values (some ML algorithms can cope with this)
- Train/Test data
- Decision boundaries

Outline

Exploratory data analysis

- A bit like an interrogation:
  - You ask question and the data answers
  - Ask stupid questions and get stupid answers
  - Ask the right questions and you get deeper to the truth

Exploratory data analysis

- Goal is to get to the core of the data
- The core is dependent on the question we want to solve
- The core is hidden by problems:
  - missing values
  - strange/unknown variables
  - faulty measurements
  - correlated variables (not this lesson)

Outline

Measurement types

Remember the Data Processing & Analysis course?

- Nominal data
- Ordinal data
- Interval data
- Ratio data

Measurement types: processing

Not all measurement/data types can be used 'as is' for machine learning. Some need some processing.

- Nominal data is tricky for most ML algorithms (work internally with numbers). One-hot encoding possible solution.
- Ordinal data can be converted to a number

Interval and Ratio data can often be used to create more data with preprocessing and feature engineering.

Outline

Missing Data

- Different types of missing data:
  - Missing completely at random (MCAR)
  - Missing at random (MAR)
  - Missing not at random (MNAR / NMAR)

Missing Data strategies

Below are some missing data strategies from least to most complex

- Remove the data (complete case analysis)
  - Remove rows
  - remove columns
- Impute with a default value:
  - 0 (really basic)
  - mean/median (better in certain situation)
  - NA
  - 999/-999 (If you ever do this I wil come and kill you)
- Impute with a realistic guess
- Build a machine learning model to impute the values (e.g. missforest)

Outline

That's it!

By next lesson, figure out how to impute data and do complete case analysis in python/pandas!
Also, exercise your python skills :)