Scikit-Learn + Feature-Engineering

Estimator-Api, Pipelines etc.

„There's the joke that 80 percent of data science is cleaning the data and 20 percent is complaining about cleaning the data”

– Anthony Goldbloom

Feature Engineering

„More data beats clever algorithms, but better data beats more data.“

–Peter Norvig

Kategoriale Features

- Onehot-Encoding für lineare Modelle

Country	country_NL	country_BR	country_US
NL	1	0	0
BR	0	1	0
US	0	0	1

- Label-Encoding für xgboost und co
- Feature Hashing
- Fehlende Daten sind ein Problem

TF-IDF

- Term Frequency - verringert Bias zugunsten langer Dokumente

- Inverse Document Frequency - verringert den Bias zugunsten häufiger Tokens

- Multipliziert trennt es ganz gut wichtige von unwichtigen Tokens

Kategorie/Wort-Embeddings

- Verwandt mit Dimensionsreduktion / Autoencodern
- Word2vec
- GloVe

Polynomiales Encoding

- Abbilden von Interaktionen zwischen kategorialen Variablen

- Lineare Modelle können XOR nicht lernen, wenn die Interaktion nicht über die Features abgebildet wird

Aus einer Spalte kann man oft mehrere machen (Expansion encoding)

- Gegeben eine Spalte mit Datum
    - Wochentag
    - Feiertag
    - Quartal
    - Urlaubssaison
- Useragent
    - Ist ein Mobilgerät
    - Operating System
    - OS ist aktuell
    - Etc..

Aus mehreren Spalten eine machen (Consolidation encoding)

- Unterschiedliche kategoriale Features auf eines Mappen
- Stemming
- Spellchecking
- Bezeichnungsnormalisierungen
- Kaputte Abkürzungen

Numerische Features

- Sind generell einfacher

- Fehlende Daten sind einfacher zu behandeln (mean, median, etc)

Typische Verfahren

- Einfach Runden - oft ist zu genau bloß Rauschen
- Gerundete Variablen kategorial behandeln
- Man kann auch den Logarithmus davor berechnen - oder den Logarithmus vom Logarithmus, und dann nochmal runden :)
- Binning, beispielsweise für den Preis, wenn man Angebote kategorisieren möchte und ein lineares Modell hat

Skalieren

- Floating Point hat mehr Auflösung um 0
- Standard (Z) Scaling: mean 0, std 1
- MinMax Scaling $$x_s = \frac{(x - x_{min})}{(x_{max} - x_{min})}$$
- Log Scaling

Interaktionen

- Addition / Subtraktion / Multiplikation / Division
- Hinterher Feature Selection
- Manchmal helfen sehr seltsam aussehende Berechnungen

Statistische Features

- Anzahl NaN / Nullen / negativer Werte
- Oft fügt man einer numerischen Spalte einfach noch mean, median, std hinzu vielleicht auch noch min, max
- Anzahl Spaces, Tabs, Newlines, Punkte

Scikit-Learn

Scikit-learn API

- BaseEstimator
  - get_params
  - set_params
- Transformer
  - fit_transform
- fit
- predict

Pipelines

- Bündeln eine Liste von Transformern, die fit und transform implementieren müssen

- Kann am Schluss einen Estimator haben, der nur fit implementieren muss

- Kann das Ergebnis von Transformern cachen

Feature Union

- Kombiniert eine Liste von Transformern horizontal

- Kann mit Pipeline zusammen verwendet werden, um komplexe Modelle zu bauen

Gridsearch

- Benutzt get_param und set_param, um über ein Parameter-Grid zu iterieren

- Fitted am Schluss den besten Estimator