Big Data
und
Machine Learning

6. März 2018

Begriffe

- Strukturierte Daten
- Unstrukturierte Daten
- Klassifikation
- Klassifikationsgüte
- Test- und Trainingsdaten
- Regression versus Klassifikation
- Regression and Classification Trees («CARTs»)

Strukturierte versus unstrukturierte Daten

- Daten, die in irgendeiner Weise in **Tabellenform** vorliegen
  - Facebook Likes
  - Spotify-Playlist
  - ….
- Daten, die **nicht** in **Tabellenform** vorliegen
  - Bilder
  - Texte
  - …

Klassifikation versus Regression

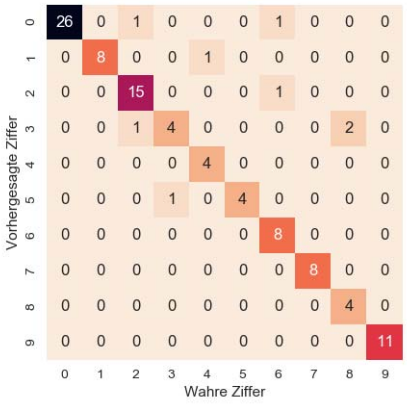
Herangehensweise im EF:

- **Klassifikation** von strukturierten Daten
  - Zeichenerkennung (0,1,…,9) auf Grund Pixelfarben
  - ….
- Andere Fragestellung («**Regression**»): Vorhersage einer stetigen Grösse
  - Einkommen auf Grund Cumulus-Verhalten
  - Occassionsautopreis auf Grund Attributen (Grösse, Türen, Farbe, Hubraum, etc.)
  - ….

Quelle: https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/

Klassifikationsgüte und -matrix

- Die **Klassifikationsgüte** kann als relative Anzahl richtig klassifizierter angegeben werden (siehe Wiki)
- Zusätzlich dazu, kann eine sogenannte **Konfusionsmatrix** angegeben werden:



Test- und Trainingsdaten (Klassifikation)

- Auf den **Trainingsdaten** wird ein Algorithmus «geschult». Diese Daten stehen dem Algorithmus als Datensatz von $[X_i, Y_i]$ zur Verfügung.
- Auf den **Testdaten** stehen dem Algorithmus erstmals nur die **Features** $X_i \in \mathbb{R}^k$ ohne die Klasse zur Verfügung. Zur Überprüfung des Algorithmus werden nachher die doch bekannten Klassen $Y_i \in \mathbb{N}$ mit den Vorhersagen des Algorithmus verglichen.
- Im wahren Leben: Nur Attribute X_i und keine Klassen.

Klassifikation und Regression

Bis jetzt: Nur **Klassifikation**.

Andere Fragestellung: Autopreis auf Grund von Attributen (Farbe, Hubraum, PS, etc.) vorhersagen. Die vorhergesagte Grösse ist keine **Klassifikation** sondern eine **Regression**. Es wird eine **stetige Grösse** (reelle Zahl) vorhergesagt.

«Allerweltswaffe»: CARTs

Sogenannte **C**lassification and **R**egression **T**rees:

Klassifikation: Beispiel Ziffern. Frage: Ziffer auf Grund Pixelbild vorhersagen

«Allerweltswaffe»: CARTs

Sogenannte **C**lassification and **R**egression **T**rees:

Regression: Beispiel Autos. Frage: Preis auf Grund Attributen vorhersagen

Viele Bäume: Zufallswald («random forests»)

- Anstelle eines Baumes werden viele Bäume («Wald», «forest») auf zufälligen («random») Teilmengen der Daten trainiert.
- Die abschliessende Meinung (Klassifikation, Regression) ist dann die Meinung (Mehrheit, Mittelwert) des Waldes.