Active Learning

Jonathan Wischhusen

HAW Hamburg, 2018 – GSM

Übersicht

1. Einleitung

2. Active Learning

3. Forschung

Einleitung

Machine Learning

Unsupervised Learning $\rightarrow$ *Verständnis schaffen*

- Keine Labels
- Algorithmus generiert Datenmodell

Supervised Learning $\rightarrow$ *Vorhersagen treffen*

- Jedem Datenpunkt ist ein Label zugeordnet
- Algorithmus „approximiert“ Datenmodell

Motivation

- Verfügbarkeit von Daten und Datenmengen wachsen rasant
- Daten verhelfen zu Erkenntnis

Deep Learning/Neuronale Netze können es doch!!11

$\rightarrow$ Naja, geht so ..

- Daten sind nicht klassifiziert
- Datenklassifikation ist sehr teuer (Training auch)
- Training bedarf i.d.R. sehr vieler klassifierzter Daten

Active Learning

Active Learning

Semi-supervised Ansatz

- benötigt Klassifikation einer kleine Teilmenge der Daten

Bewerten statt stumpfem Lernen

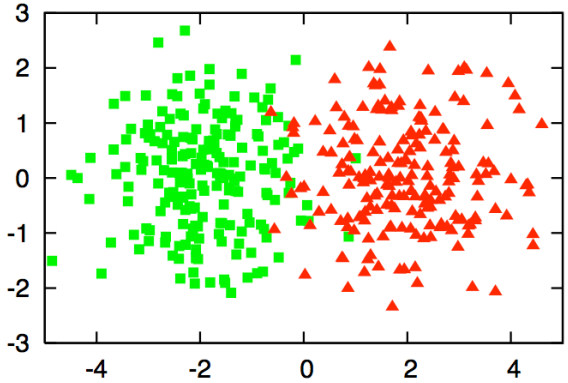
- Algorithmus entscheidet welche Datenpunkte zum Lernen interessant sind

Allwissendes Orakel

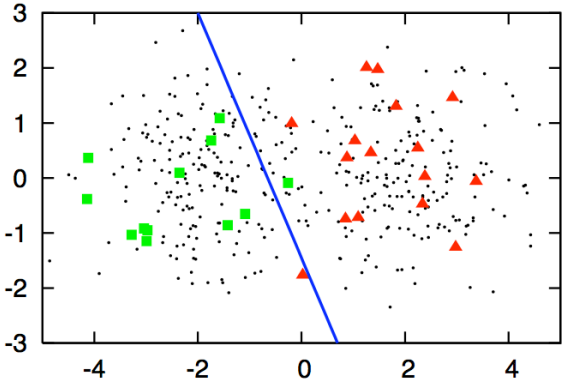
- Algorithmus als Schüler, fragt Orakel nach wichtiger Klassifikation

  $\rightarrow$ ermöglicht *Human in the Loop*, meist Mensch als Orakel

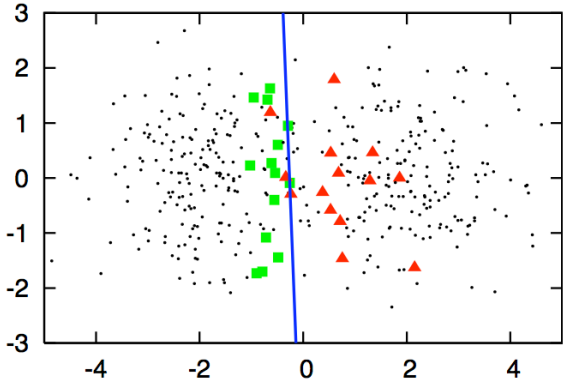
Lineare Regression Beispiel



Verteilung zweier Klassen, eingefärbt nach Klasse.



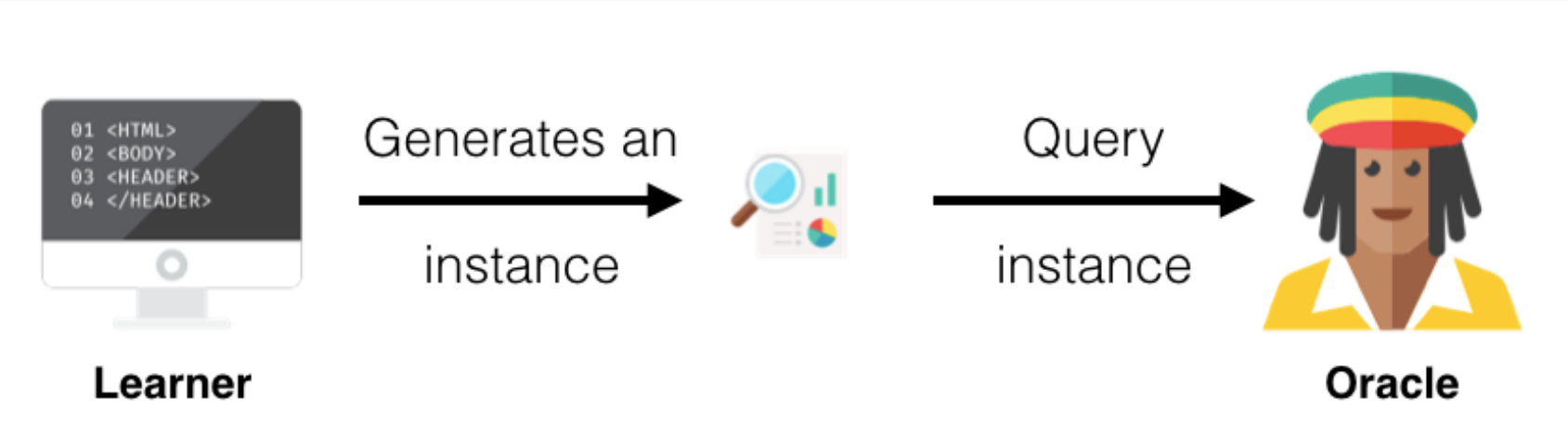
Zufällige Auswahl an interessanten Datenpunkten



Schlaue Auswahl an interessanten Datenpunkten

Abfrage Szenarien 1/3

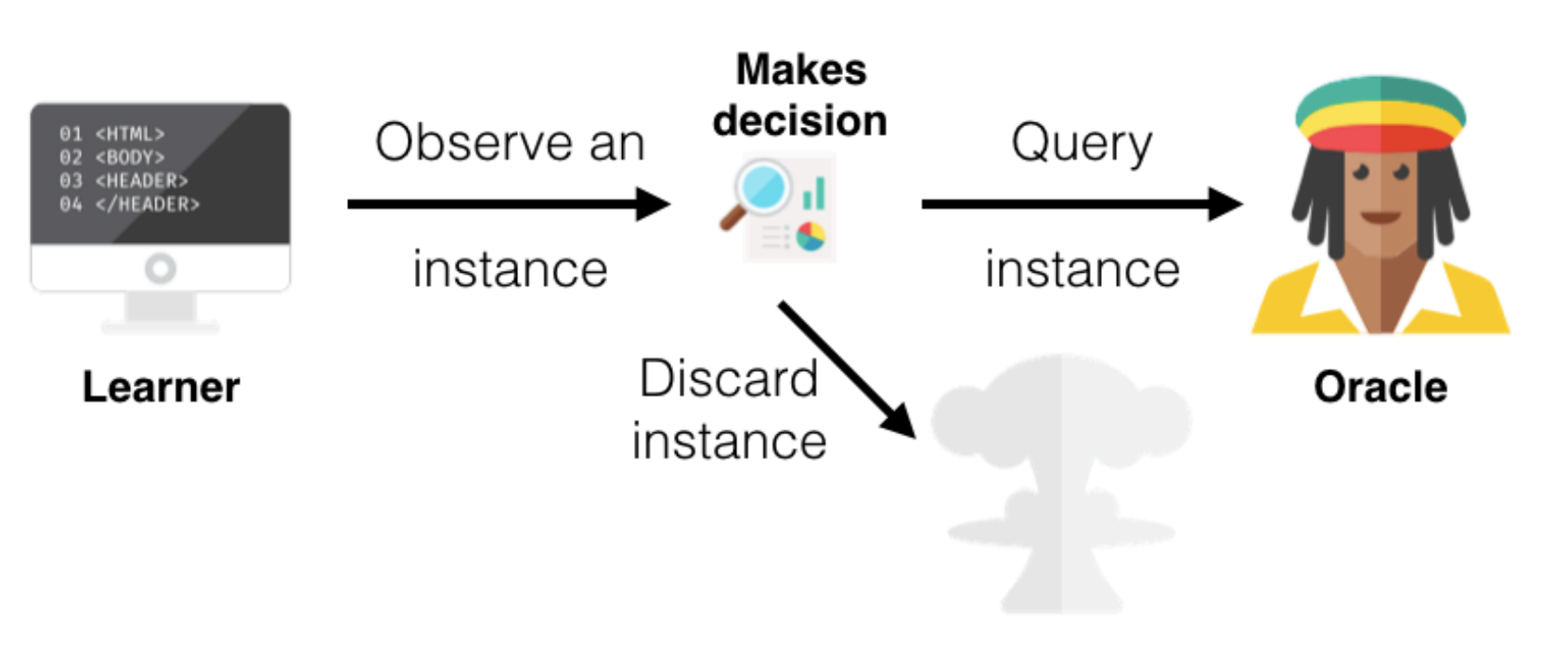
Membership Query Synthesis



Algorithmus generiert auf Grundlage der Daten ein fiktives Beispiel, das wichtige Merkmale zusammenfasst und fragt nach der Antwort.

Abfrage Szenarien 2/3

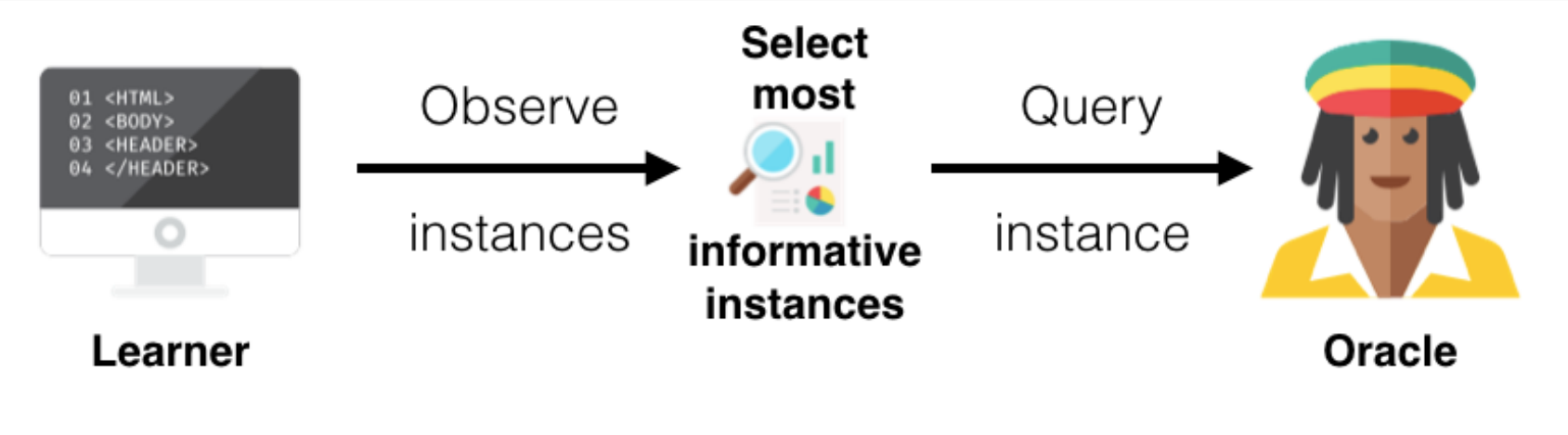
Stream-Based Selective Sampling



Algorithmus iteriert über jeden unklassifizierten Datenpunkt nacheinander und entscheidet, ob es sich lohnt diesen abzufragen und zu lernen.

Abfrage Szenarien 3/3

Pool-Based sampling



Algorithmus bestimmt aus allen Datenpunkten (oder Teilmengen) jene mit dem größten Informationsgehalt zum Lernen.

Abfrage Metriken

Query by committee

Komitee aus verschieden trainierten Modellen bestimmt Abfragen.

Expected model change/error reduction

Datenpunkte die größt möglichst den Fehler reduzieren bzw das Modell beeinflussen werden ausgewählt.

Uncertainty Sampling

Wahrscheinlichkeitsmodelle wählen unklare Datenpunkte.

Balance exploration and exploitation

Active Thomson Sampling

...

Uncertainty Sampling 1/3

Datenpunkt	Label A	Label B	Label C
d1	0.9	0.009	0.01
d2	0.2	0.45	0.35
d3	0.48	0.49	0.01

Least Confidence (LC)

Der Datenpunkt mit der geringsten Wahrscheinlichkeit für das wahrscheinlichste Label wird ausgewählt.

Uncertainty Sampling 2/3

Datenpunkt	Label A	Label B	Label C
d1	0.9	0.009	0.01
d2	0.2	0.45	0.35
d3	0.48	0.49	0.01

Margin Sampling

Der Datenpunkt mit dem geringsten Abstand der beiden wahrscheinlichsten Label wird ausgewählt.

Uncertainty Sampling 3/3

Datenpunkt	Label A	Label B	Label C	H
d1	0.9	0.009	0.01	0.52
d2	0.2	0.45	0.35	1.49
d3	0.48	0.49	0.01	1.08

Entropy Sampling

Für jeden Datenpunkt wird die Entropie (H, der mittleren Informationsgehalt) berechnet, der Datensatz mit dem größten Wert wird ausgewählt.

Forschung

Natural Language Processing

Part-of-Speech Tagging

Zuordnung von Wortarten zu Wörtern und Satzzeichen.

Named Entity Recognition

Identifikation von Eigennamen

Speech recognition

Üblicherweise viele nicht gelabelte Daten zu Verfügung.
Transkription erforderlich.

Active Learning: weniger Wordfehler bei weniger Daten

Active Learning Framework

Active Learning als Framework für

- Convolutional Neural Network
- Long short-term memory
- Generative Adversarial Networks
- Reinforcement Learning

Active Learning wird als Framework für verschiedene Bereiche etabliert. Probleme werden vermehrt in den Active Learning Kontext gebracht.

Coactive Learning

Unterschied zu Active Learning

- Algorithmus schlägt Ergebnisse vor, Orakel korrigiert. Korrektur muss nicht optimal sein.
- Preference Feedback

Anwednungen

- Websuche
- Empfehlungsmodelle
- Übersetzungsaufgaben

Konferenzen

ICSE: International Conference on Software Engineering

2018 Gothenburg - Sweden, 2019 Montréal - Canada

SIGKDD Conference on Knowledge Discovery and Data Mining

2018 London - UK, 2019 Alaska - USA

European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning

2018 Belgien, 2019 Belgien

Literatur

Buch:

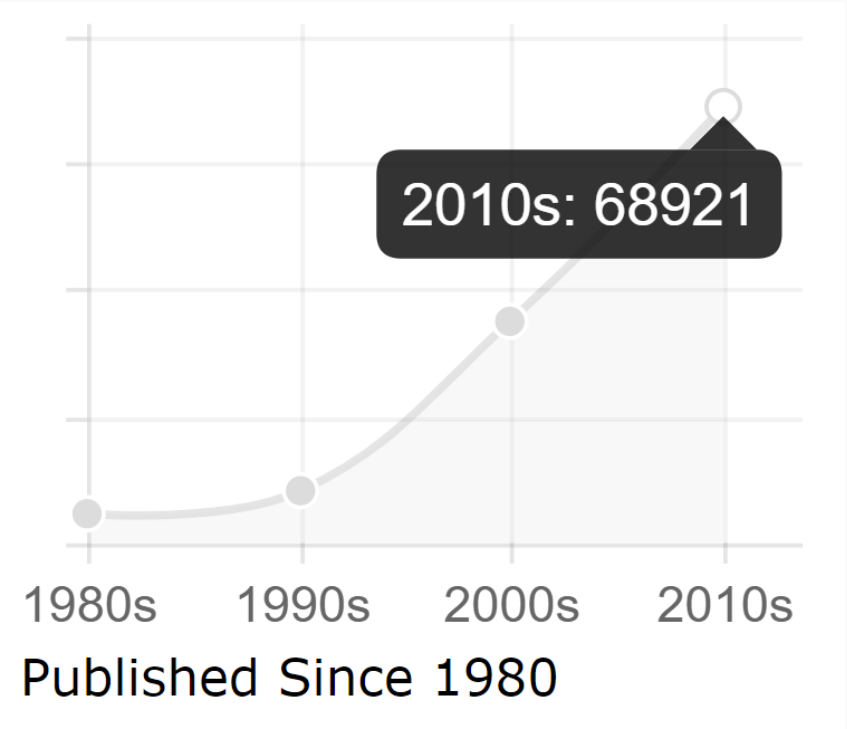
Active Learning

Synthesis Lectures on Artificial Intelligence and Machine Learning

Morgan & Claypool Publishers, June 2012, 114 pages

Author: Burr Settles
http://active-learning.net

Veröffentlichung ACM:



Ausrichtung im Master

Active Learning im Umfeld des Natural Language Processings vertiefen.

Coactive Learning im Bereich Visual Analytics untersuchen.

Möglicher Ansatz

- **Forum 4.0**
  Forum 4.0 will develop new methods based on text analysis, machine learning (with human-in-the-loop) and Empirical Software Engineering to better exploit the constructive and deliberative potential of user comments.

Fragen?

Bildquellen

Datacamp

- http://res.cloudinary.com/dyd911kmh/image/upload/f_auto,q_auto:best/v1518178638/al-eg_pbwzob.png
- http://res.cloudinary.com/dyd911kmh/image/upload/f_auto,q_auto:best/v1518178638/membership_wzptzh.png
- http://res.cloudinary.com/dyd911kmh/image/upload/f_auto,q_auto:best/v1518178638/stream_kdlsz2.png
- http://res.cloudinary.com/dyd911kmh/image/upload/f_auto,q_auto:best/v1518178638/pool_guqwfe.png

ACM

- https://dl.acm.org/results.cfm?within=owners.owner%3DHOSTED&dte=1980&srt=_score&query=active+learning+machine+learning&Go.x=0&Go.y=0