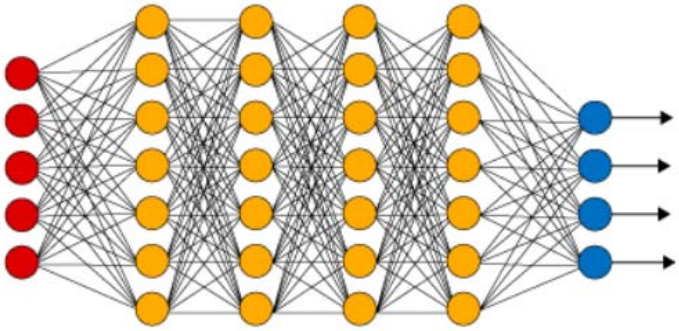
Trends in Machine Learning and Data Analytics

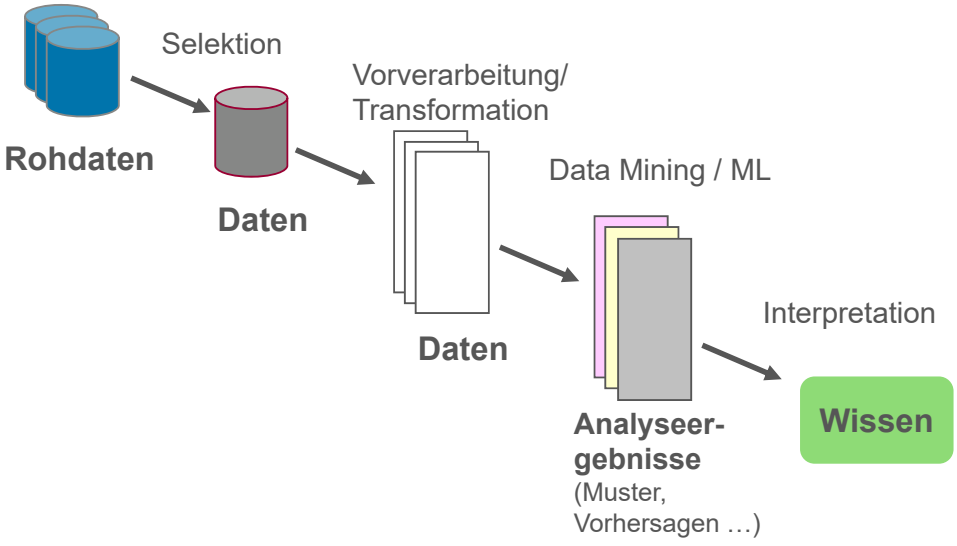
Prof. Dr. E. Rahm
und Mitarbeiter

Seminar, WS 2019/20



KNOWLEDGE DISCOVERY

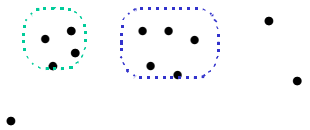
- (semi-)automatische Extraktion von Wissen aus Daten
- Kombination von Verfahren zu Datenbanken, Statistik (Data Mining) und KI (maschinelles Lernen)



TECHNIKEN DES DATA MINING/ML

Clusteranalyse

- Objekte (Kunden, Produkte, ...) werden aufgrund von Ähnlichkeiten in Klassen eingeteilt (Segmentierung)



Assoziationsregeln

- Warenkorbanalyse (z.B. Kunde kauft A und B => Kunde kauft C)
- Nutzung für Kaufvorhersagen / Recommendations, Produkt-Bundling, ...

Klassifikation

- Zuordnung von Objekten zu Gruppen/Klassen mit gemeinsamen Eigenschaften bzw. Vorhersage von Attributwerten
- Verwendung von Stichproben (Trainingsdaten)
- Ansätze: Entscheidungsbaum-Verfahren, neuronale Netze, statistische Auswertungen

weitere Ansätze:

- genetische Algorithmen (multivariate Optimierungsprobleme, z.B. Identifikation der besten Bankkunden)
- Regressionsanalyse zur Vorhersage numerischer Attribute . . .

KLASSIFIKATION

Klassifikationsproblem

- gegeben Stichprobe (Trainingsmenge) O von Objekten des Formats $(a_1, \ldots, a_d)$ mit *Attributen* A_i, $1 \leq i \leq d$, und Klassenzugehörigkeit c_i, $c_i \in C = \{c_1, \ldots, c_k\}$
- gesucht: Klassenzugehörigkeit für Objekte aus $D \setminus O$, d.h. *Klassifikator* $K : D \rightarrow C$
- weiteres Ziel: Generierung (Lernen) des expliziten Klassifikationswissens (Klassifikationsmodell, z.B. Klassifikationsregeln oder Entscheidungsbaum)

Abgrenzung zum Clustering

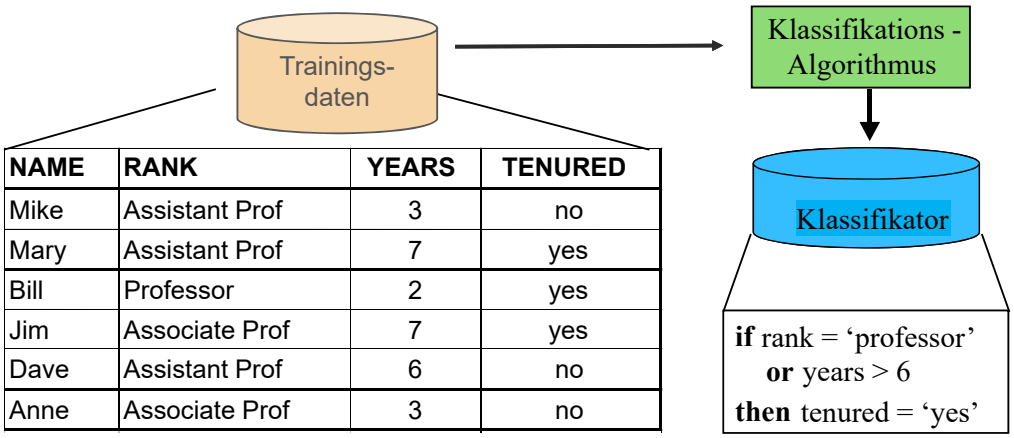
- Klassifikation: Klassen vorab bekannt, Nutzung von Trainingsdaten
- Clustering: Klassen werden erst gesucht, keine Trainingsdaten (unsupervised)

Klassifikationsansätze

- Entscheidungsbaum-Klassifikatoren
- neuronale Netze
- Bayes-Klassifikatoren (Auswertung bedingter Wahrscheinlichkeiten)
- Support Vector Machine (SVM)
- lineare Regression ...

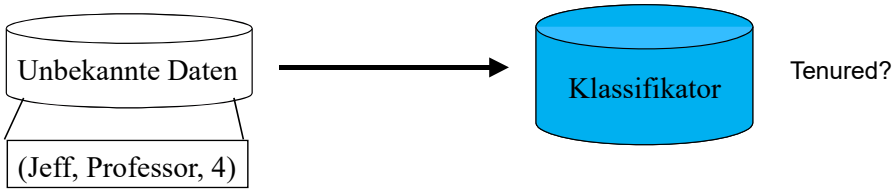
KLASSIFIKATIONSPROZESS

1. Konstruktion des Klassifikationsmodells

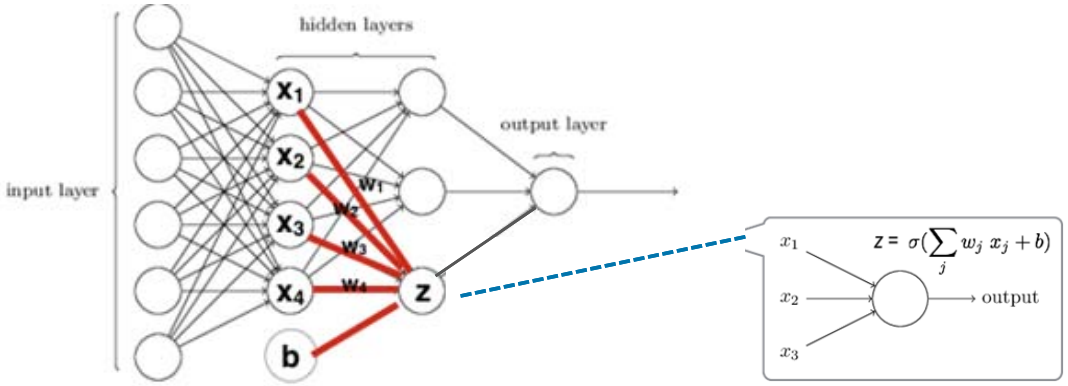


NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

2. Anwendung des Modells zur Vorhersage (Prediction)



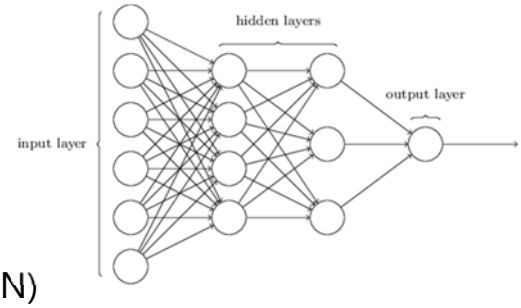
NEURONALE NETZE



- Neuronales Netz (NN) besteht aus mehreren Schichten
  - Eingabe-/Ausgabeschicht
  - mind. einer verdeckten (hidden) Schicht
- jede Schicht besteht aus mehreren Neuronen, welche mit anderen Neuronen verbunden sind
- Verbindungen / Kanten verwenden Zahlen, z.B. Gewichte ($w_i \in \mathbb{R}$)
- Deep Learning: mehrere hidden layers

DEEP LEARNING

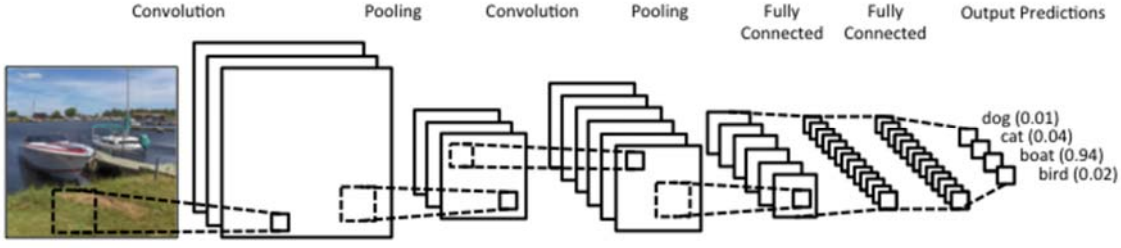
- Nutzung tiefer neuronaler Netze
  - Lernen einer Datenrepräsentation (Embeddings) auf großen Mengen an Trainingsdaten
  - Nutzung des gelernten Wissens für Klassifikation, Vorhersagen …
- zahlreiche Anwendungsfälle
  - Erkennung von Bildern
  - Erkennung von Handschriften
  - Spracherkennung
  - Verarbeitung von Texten …
- verschiedene Varianten von Netzen
  - Convolutional deep neural networks (CNN)
  - Recurrent neural networks (RNN) , u.a. LSTM (Long short-term memory)
  - Autoencoder networks (Erzeugung verbesserter Repräsentationen)



BILDKLASSIFIZIERUNG

Nutzung z.B. von *Convolutional Neural Networks*

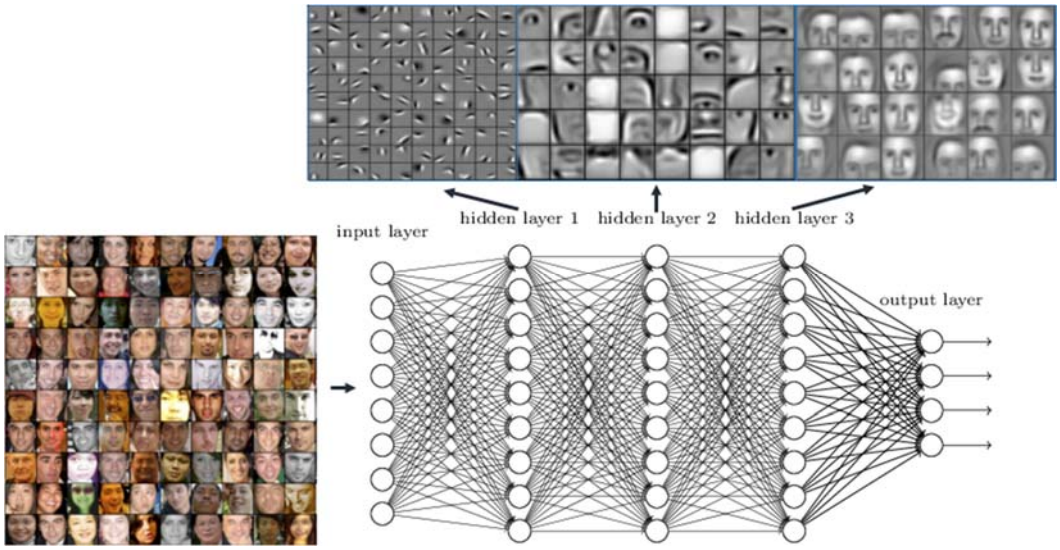
- lokale Filter fassen Pixelaktivität zusammen (convolutional layer)
- nur ausgewählte Informationen daraus werden weitergereicht und somit überflüssige Information verworfen (pool layer)
- dieser Vorgang kann wiederholt Anwendung finden



www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp

GESICHTSERKENNUNG

- neuronale Netze lernen Merkmale der Eingabedaten in Form von aufeinander aufbauenden Konzepten.
- hierarchische Repräsentation der Daten (Farbwerte der Pixel): Kanten -> Teile des Gesichts -> gesamtes Gesicht



https://www.slideshare.net/Tricode/deep-learning-stm-6

TEXT/SPRACHVERARBEITUNG

- Lernen der Nachbarschaft von Wörtern (*word embeddings*) in Text, um deren semantische Ähnlichkeit zu ermitteln
- trainierte Datenrepräsentationen nutzen für weitere ML-Aufgaben, zB
  - Named Entity Recognition
  - Machine Translation
  - Spracherkennung
- häufiger Einsatz von *Recurrent Neural Networks (RNN)*
- vortrainierte Vokabulare
  - Word2vec
  - Glove
  - Fasttext

GENERATIVE ADVERSARIAL NETWORKS

- Zusammenspiel von 2 neuronalen Netzwerken zur Erzeugung neuer Inhalte
  - Generator-Netzwerk zur Erzeugung neuer Kandidaten (z.B. Bilder)
  - Diskriminator-Netzwerk zur Bewertung der Kandidaten (Vergleich mit realen Inhalten)
- vielfältige Anwendungsmöglichkeiten, z.B. Behebung von Bildrauschen, Färben von SW-Fotos etc.

https://blog.codecentric.de/2018/11/eine-kurze-einfuehrung-in-generative-adversarial-networks

ML AUF GRAPHDATEN

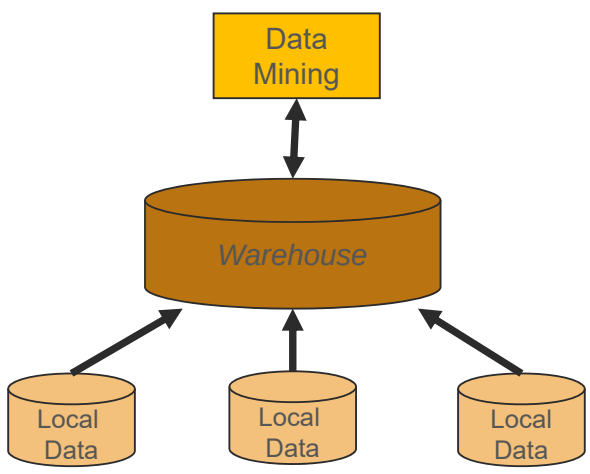
- Vorhersage auf Graphdaten, z.B. Klassifikation von Knoten und Beziehungen (label / link prediction)
- Nutzung des Kontextwissens / Graphnachbarschaft
- Graph-Embeddings statt Word Embeddings
  - mehrere Verfahren (node2vec, DeepWalk, GraphSage, …)
- viele Nutzungsmöglichkeiten: soziale Netzwerke, Wissensgraphen …

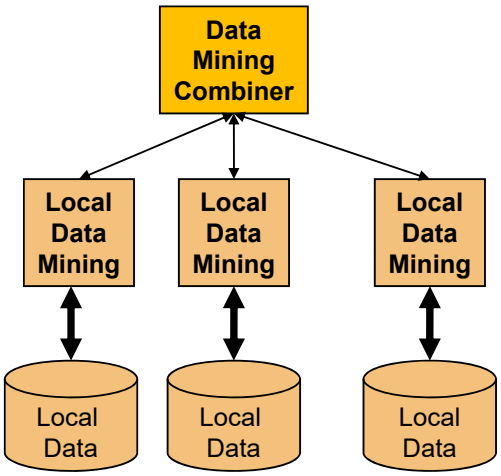
PRIVACY / DATENSCHUTZ

- Privacy-preserving Machine Learning /Data Mining: Analyse personenbezogener Daten unter Wahrung von Datenschutz / Privacy
- Einsatz von Anonymisierungstechniken um Identifizierbarkeit zu verhindern
  - Verallgeminerung kritischer Werte (Geburtsdatum -.> Altersgruppe)
  - gezielte Verfälschung von Werten, z.B. mit differential privacy
- ggf. verteilte Auswertungen statt zentraler Datensammlungen

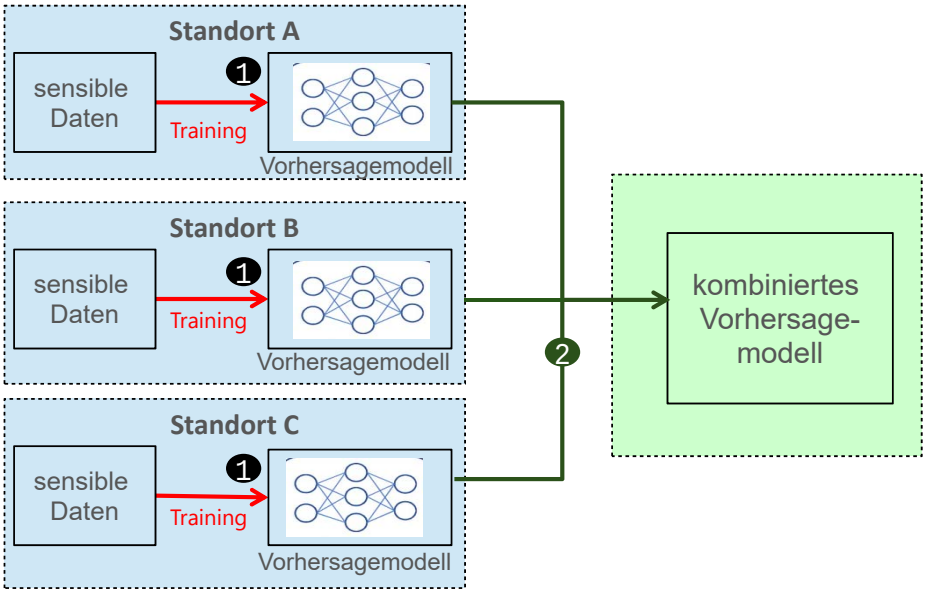
PRIVACY-PRESERVING DATA MINING

- relative hohe Risiken mit Data Warehouses mit integrierten personenbezogenen Daten / Nutzerprofilen
- verteilte Data Mining-Ansätze umgehen Austausch und Fusionierung von personenbezogenen Daten





VERTEILTES PPML



- wird u.a. in Google PATE unterstützt

FAIRNESS UND ML

- „Bias“ in Trainingsdaten kann zu „unfairen“ Analyseergebnissen / Vorhersagen führen
  - Benachteiligung von Frauen bei automatisierter Bewerberauswahl (Facebook)
  - geringere Qualität bei automatischer Gesichtserkennung für dunkelhäutige Frauen
  - Vorhersage höherer Rückfallwahrscheinlichkeit für dunkelhäutige Gefängnisinsassen bei Entscheidungen über vorzeitiger Freilassung …
- wie kann Fairness eines ML-Verfahrens bewertet werden?
  - neue Metriken neben precision/recall/accuracy
- wie kann Fairness verbessert werden ?

SEMINAR

SEMINARZIELE

- Beschäftigung mit einem praxis- und wissenschaftlich relevanten Thema
  - kann Grundlage für Abschlussarbeit oder SHK-Tätigkeit sein
- Erarbeitung + Durchführung eines Vortrags unter Verwendung wissenschaftlicher (englischer) Literatur
- Diskussion
- schriftliche Ausarbeitung zum Thema
- Hilfe und Feedback durch zugeteilten Betreuer

SEMINAR: ANRECHNUNGSMÖGLICHKEITEN

- Masterstudium, insbesondere für Schwerpunkt „Big Data“
  - Teil der großen Moduls Moderne Datenbanktechnologien
  - Seminarmodul
- Bachelorstudium
  - Seminarmodul

SCHEINVERGABE / MODULPRÜFUNG

- selbständiger Vortrag mit Diskussion (ca. 45 Minuten)
  - Abnahme der Folien durch Betreuer
- schriftliche Ausarbeitung (15-20 Seiten)
  - Abnahme der Ausarbeitung durch Betreuer
  - Abgabe-Deadline 31.3.2020
- aktive Teilnahme an allen Vortragsterminen
- Modul-Workload 150 h:
  - 30h Präsenzzeit
  - 120 h Selbststudium (Vorbereitung Vortrag, Ausarbeitung)

ScaDS DRESDEN LEIPZIG

SEMINAR (3)

UNIVERSITÄT LEIPZIG

- Themenzuordnung
  - Koordinierungstreffen mit Betreuer innerhalb der nächsten 12 Tage, d.h. bis spätestens 6.11.2019
  - ansonsten verfällt Seminaranmeldung
  - freiwilliger Rücktritt auch bis max. 6.11.2019

- Vortragstermine
  - freitags, Ritterstr. ab 10. 1. 2020
  - max. 2 Doppelstunden ab 13:30 Uhr

Themen	Betreuer	Termin	Studenten
ML Techniken/Trends Neuronale Netze / Architekturen	Täschner Schuchart	10.1.	Busse Meyer Alker
Deep Learning Image classification using CNNs Generative Adversarial Networks Deep learning for entity matching Geometric Deep Learning	Wilke Peukert Saeedi Peukert	10.1 17.1.	Mutzabaupthe Koelble
Machine Learning auf Graphdaten Overview of Graph Representations Random-Walk-based Embeddings Google Deep Mind and GMNs Semi-supervised Learning on Graphs	Obraczka Rost Gomez Christen	17.1.	Akki Reichert
ML-Gefahren und Lösungen Overview Privacy-Preserving ML Privacy-preserving Deep Learning Private Data Sharing Fair ML	Franke Sehili Rohde Rohde	24.1.	Elster / Neu
ML in Medicine ML for biomedical Data Integration Deep Learning for Mortality Prognosis Time Series Classification in Medicine Deep Learning in Radiomics	Franke Lin Christen Martin	24.1.	V. Jehle Blankenburg
Other ML applications ML for intrusion detection Distributed learning Tackling Climate Change with ML AutoML	Grimmer Wilke Wilke Alkhouri	31.1.	Kreibel / Sugcon- winter De Riz