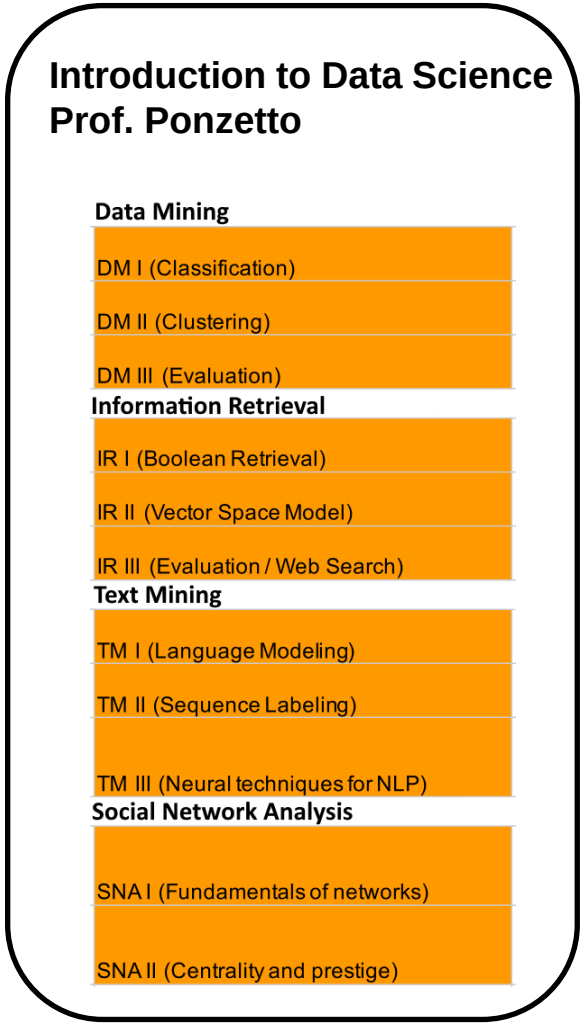
Künstliche Intelligenz

Maschinelles Lernen

Dr. Christian Meilicke
Research Group Data and Web Science
Universität Mannheim

Mehr Maschinelles Lernen

- Maschinelles Lernen zentrales Thema der Künstlichen Intelligenz
- Aber: Mindestens drei Vorlesungen in Introduction to Data Sciene
  - Viele andere Inhalte bauen darauf auf
- Hier daher nur eine Vorlesung
  - … plus MCTS Vorlesung
  - Weniger angewandt



Gliederung

- Maschinelles Lernen: Überblick
- Überwachtes Lernen
  - Beispiel
- Entscheidungsbaum
  - Repräsentation
  - Lernen des Baums
- Evaluationsmethodik
- Ausblick Deep Learning

Was haben wir bisher gemacht

- Beispiel eines klassischen Suchproblems:
  - Platziere 8 Damen auf einem Schachbrett, so dass diese sich nicht schlagen können. Zwei Damen schlagen sich, wenn …
- Suche mit Logik:
  - Gegeben eine logische Minesweeper Modellierung, finde eine Position auf der sicher keine Bombe liegt
- In beiden Fällen handelt es sich um Deduktion, d.h., der Schluss vom Allgemeinen auf besonderes
  - Wann sich zwei Damen schlagen ist eine allgemeine Regel
  - Was z.B. eine 3 auf einem aufgedeckten Feld bedeutet, wird durch eine allgemeine Regel festgelegt
  - Auch Handlungen werden durch allgemeine Regeln beschrieben

Induktion und Deduktion

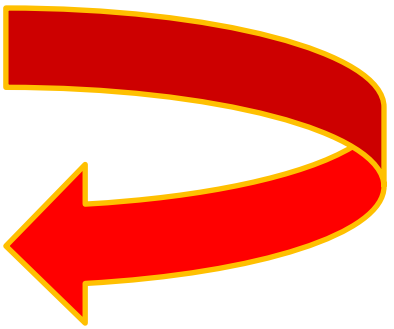
- Induktion = Aus der Betrachtung mehrerer einzelner Fälle wird auf eine allgemeine Regel geschlossen
  - GEGEBEN: Da ist ein weißer Schwan, und noch einer …
    ALSO: Alle Schwäne sind weiß

- Deduktion = Aus einer Menge allgemeiner Regeln wird auf einen einzelnen Fall geschlossen
  - GEGEBEN: Alle Schwäne sind weiss und Susi ist ein Schwan.
    ALSO: Susi ist weiss.

Induktion und Deduktion

- Induktion = Aus der Betrachtung mehrerer einzelner Fälle wird auf eine allgemeine Regel geschlossen
  - GEGEBEN: Da ist ein weißer Schwan, und noch einer …
    ALSO: Alle Schwäne sind weiß

- Deduktion = Aus einer Menge allgemeiner Regeln wird auf einen einzelnen Fall geschlossen
  - GEGEBEN: Partiell aufgedecktes Minenfeld und die Minesweeper Regeln.
    ALSO: Auf Kachel 3,7 ist keine Bombe.
  - **Indirektes Verfahren: Suche nach Modell für Minenfeld, Regel und der Annahme dass auf 3,7 eine Bombe ist**



Definition

- Maschinelles Lernen kann als eine Form induktiven Schließens verstanden werden:

- Eine mögliche Definition:

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

Tom Mitchell "Machine Learning"

Formen Maschinellen Lernens

- Überwachtes Lernen (Supervised Learning)
  - Lernen einer Funktion, die einen Input auf einen Output abbildet
  - E = Ein/Ausgabe Paare, auch Trainings-Beispiele genannt
  - Task T:
    - Klassifikation = Output ist eine (kleine) Menge diskreter Werte
    - Regression = Output ist eine Zahl
- Unüberwachtes Lernen (Unsupervised Learning)
  - Erkennen statistischer Zusammenhänge in E
  - Oft geht es dabei um Clustering (= E in Mengen ähnlicher Beispiele unterteilen)
- B/Vestärkendes Lernen (Reinforcement Learning)
  - Agent erlernt Strategie, um Belohnungen zu maximieren

Formen Maschinellen Lernens

- Überwachtes Lernen (Supervised Learning)
  - Regression:
    - Schätze erreichbaren Verkauspreis eines Hauses basierend auf Lage, Quadratmeterzahl, Baujahr, …
  - Klassifikation:
    - Erkennen von Buchstaben in Handschrift
    - Email: Spam ja oder nein?
- Unüberwachtes Lernen (Unsupervised Learning)
  - Kunden in Gruppen segmentieren um Marketing zielgerichtet einsetzen zu können
- B/Vestärkendes Lernen (Reinforcement Learning)
  - MCTS!

Formen Maschinellen Lernens

- Überwachtes Lernen (Supervised Learning)
  - Regression
    - Schätze erreichbaren Verkauspreis eines Hauses basierend auf Lage, Quadratmeterzahl, Baujahr, …
  - Klassifikation => Decision Tree
    - Erkennen von Buchstaben in Handschrift
    - Email: Spam ja oder nein?
- Unüberwachtes Lernen (Unsupervised Learning)
  - Kunden in Gruppen segmentieren um Marketing zielgerichtet einsetzen zu können
- B/Vestärkendes Lernen (Reinforcement Learning)
  - MCTS!

Gliederung

- Maschinelles Lernen: Überblick
- Überwachtes Lernen
  - Beispiel
- Entscheidungsbaum
  - Repräsentation
  - Lernen des Baums
- Evaluationsmethodik
- Ausblick Deep Learning

Überwachtes Lernen

- Gegeben:
  - Eine Trainingsmenge mit N Beispielen von Ein/Ausgabe Paaren
    - » (x_1, y_1)
    - » (x_2, y_2)
    - » …
    - » (x_N, y_N)
  - Hierbei sei jedes y_i durch eine unbekannte Funktion $y = f(x)$ generiert worden

mit $x_i = (x_i^1, \ldots, x_i^n)$ wobei $x_i^1, \ldots, x_i^n$ die Werte der Attribute $1 \ldots n$ für das i-te Beispiel sind y_i Zielprädikat oder Zielatttribut

- Ziel: Finde möglichst gute Näherung h (Hypothese) der Funktion f
- Methode: **Suche** im Raum der Hypothesen

Beispiel

- Alle Tische besetzt, warten?

- Relevante Attribute können sein:
  - Alt: Gibt es ein alternatives Restaurant in der Nähe? Ja/Nein
  - Bar: Hat das Restaurant eine Bar zum Warten? Ja/Nein
  - Frei: Ist heute Freitag/Samstag oder anderer Tag? Ja/Nein
  - Hun: Sind Ankommenden hungrig? Ja/Nein
  - Gäste: Wieviele sind im Restaurant? Keine/Einige/Voll
  - Preis: Wie sind die Preise? €/€€/€€€
  - Regen: Regnet es? Ja/Nein
  - Res: Wurde eine Reservierung vorgenommen? Ja/Nein
  - Typ: Französisch/Italienisch/Thai/Burger
  - War: Geschätzte Wartezeit des Kellners 0-10/10-30/30-60/>60 Minuten
  - WW: Werden die Ankommenden Warten?

Trainingsmenge

#	Alt	Bar	Frei	Hun	Gäste	Preis	Regen	Res	Typ	War	WW
x_1	Ja	Nein	Nein	Ja	Einige	€€€	Nein	Ja	Franz	0-10	y_1=Ja
x_2	Ja	Nein	Nein	Ja	Voll	€	Nein	Nein	Thai	30-60	y_2=Nein
x_3	Nein	Ja	Nein	Nein	Einige	€	Nein	Nein	Burg	0-10	y_3=Ja
x_4	Ja	Nein	Ja	Ja	Voll	€	Ja	Nein	Thai	10-30	y_4=Ja
x_5	Ja	Nein	Ja	Nein	Voll	€€€	Nein	Ja	Franz	>60	y_5=Nein
x_6	Nein	Ja	Nein	Ja	Einige	€€	Ja	Ja	Ital	0-10	y_6=Ja
x_7	Nein	Ja	Nein	Nein	Keine	€	Ja	Nein	Burg	0-10	y_7=Nein
x_8	Nein	Nein	Nein	Ja	Einige	€€	Ja	Ja	Tha	0-10	y_8=Ja
x_9	Nein	Ja	Ja	Nein	Voll	€	Ja	Nein	Burg	>60	y_9=Nein
x_{10}	Ja	Ja	Ja	Ja	Voll	€€€	Nein	Ja	Ital	10-30	y_{10}=Nein
x_{11}	Nein	Nein	Nein	Nein	Keine	€	Nein	Nein	Thai	0-10	y_{11}=Nein
x_{12}	Ja	Ja	Ja	Ja	Voll	€	Nein	Nein	Burg	30-60	y_{12}=Ja

Trainingsmenge

#	Alt	Bar	Frei	Hun	Gäste	Preis	Regen	Res	Typ	War	WW
x_1	Ja	Nein	Nein	Ja	Einige	€€€	Nein	Ja	Franz	0-10	y_1=Ja
x_2	Ja	Nein	Nein	Ja	Voll	€	Nein	Nein	Thai	30-60	y_2=Nein
x_3	Nein									0-10	y_3=Ja
x_4	Ja	N								10-30	y_4=Ja
x_5	Ja	N								>60	y_5=Nein
x_6	Nein									0-10	y_6=Ja
x_7	Nein									0-10	y_7=Nein
x_8	Nein	N								0-10	y_8=Ja
x_9	Nein									>60	y_9=Nein
x_{10}	Ja									10-30	y_{10}=Nein
x_{11}	Nein	Nein	Nein	Nein	Keine	€	Nein	Nein	Thai	0-10	y_{11}=Nein
x_{12}	Ja	Ja	Ja	Ja	Voll	€	Nein	Nein	Burg	30-60	y_{12}=Ja

- Gegeben:
  - Eine Trainingsmenge mit N Beispielen von Ein/Ausgabe Paaren
    - » (x_1, y_1)
    - » (x_2, y_2)
    - » …
    - » (x_N, y_N)
  - Hierbei sei jedes y_i durch eine unbekannte Funktion $y = f(x)$ generiert worden

Trainingsdaten

- Annotation durch Benutzer oder Experten
  - Spammail
  - Bilderkennung (z.B. Bilder mit Katzen)
  - Positiver oder negativer Review Kommentar

- In der Regel sind vielmehr Beispiele notwendig um etwas sinnvolles Lernen zu können

Komplexität Hypothesenraum

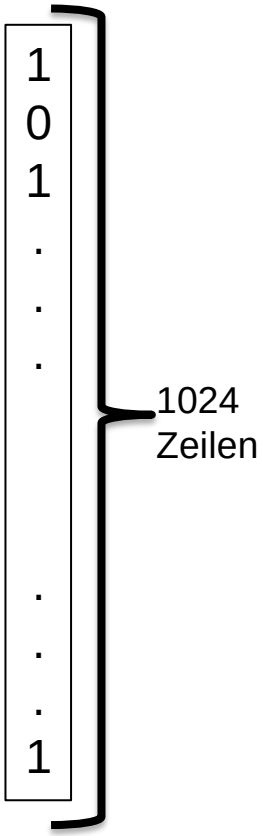
- Wie viele verschiedene Hypothesen h gibt es für unser Beispiel?

- Umfrage:

  - A: 1 bis 1.000
  - B: 1.001 bis 1.000.000
  - C: 1.000.001 bis 1.000.000.000
  - D: >1.000.000.000

Komplexität Hypothesenraum

- D ist richtig!
  - Vereinfachende Annahme: Alle Attribute sind zweiwertig (Ja/Nein)
  - Es sind $2^{2^{10}} = 2^{1024} = 1.79 * 10^{308}$
- Wie kommt es zu dieser Zahl?
  - 2^10 Zeilen in Wahrheitstabelle mit 10 Variablen
  - Schreibt man eine Abfolge von 0en und 1en in die Zielspalte so entspricht das einer Funktion
  - Man kann die Funktionen durchnummerieren, indem man die Zielspalte als Binär-Zahl liest
- **Ein extrem komplexes Suchproblem!**



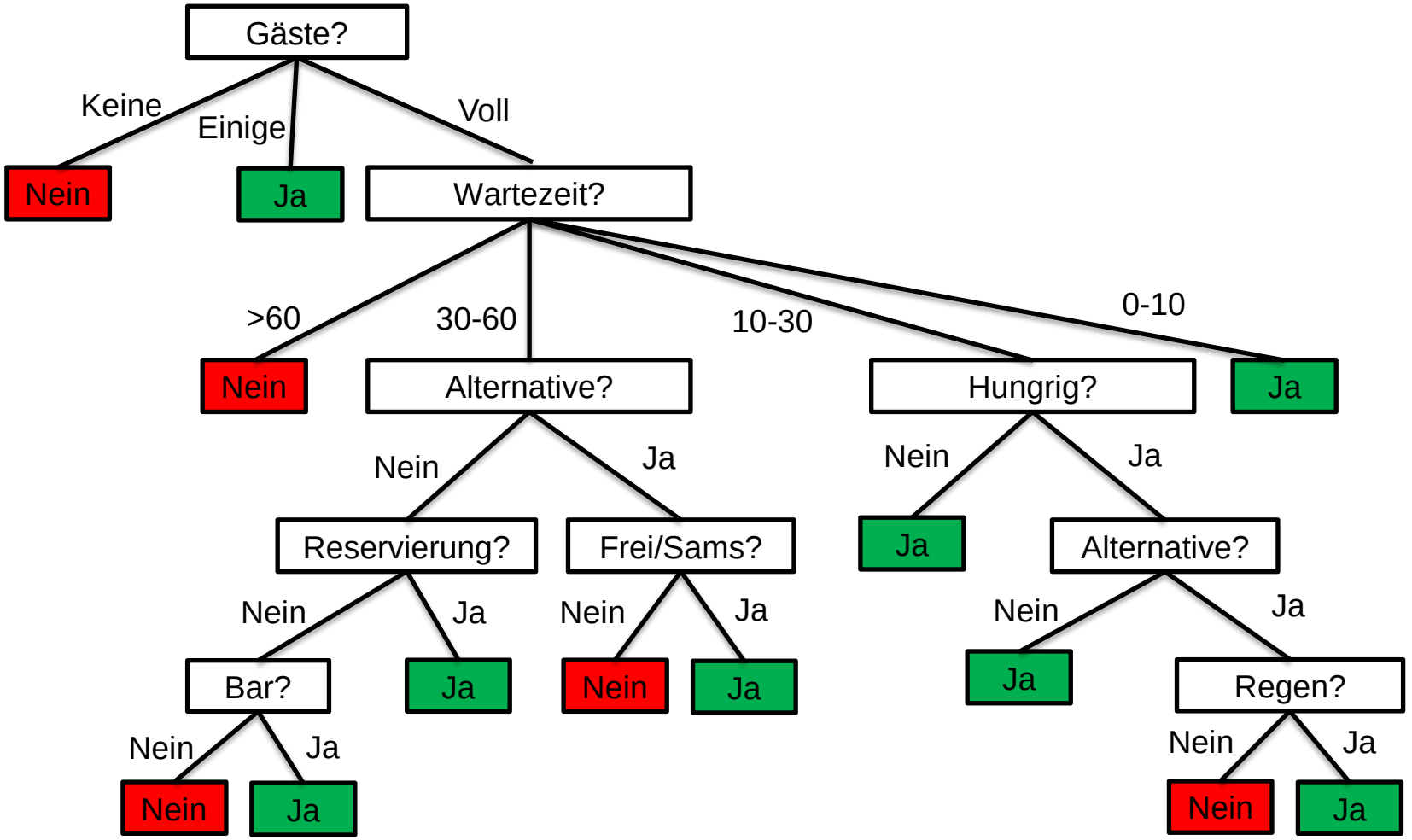
Was ist eine gute Hypothese?

- Die Hypothese soll die gegebenen Daten erklären
  - Passt konsistent zu allen gegebenen Beispielen
- Die Hypothese soll gut generalisieren
  - D.h. sie soll in der Lage sein bisher ungesehene Beispiele korrekt zu klassifizieren
  - Die Hypothese soll möglichst einfach sein
    - Ockhams Raisermesser: Wenn du zwei Möglichkeiten hast etwas zu erklären, dann bevorzuge stets die einfachere
- Die beiden Ziele können unter Umständen nicht vereinbar sein

Gliederung

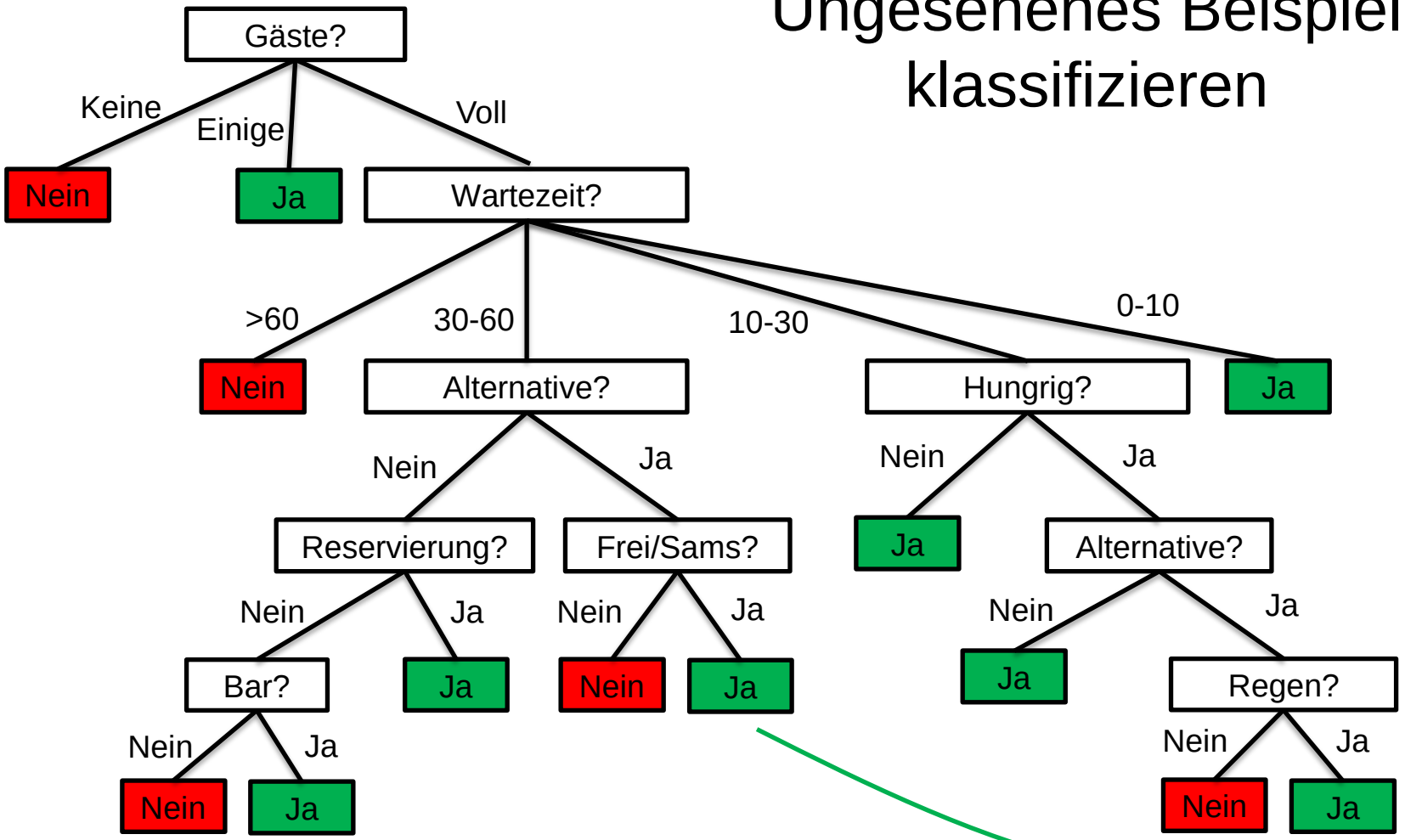
- Maschinelles Lernen: Überblick
- Überwachtes Lernen
  - Beispiel
- Entscheidungsbaum
  - Repräsentation
  - Lernen des Baums
- Evaluationsmethodik
- Ausblick Deep Learning

Beispiel Entscheidungsbaum



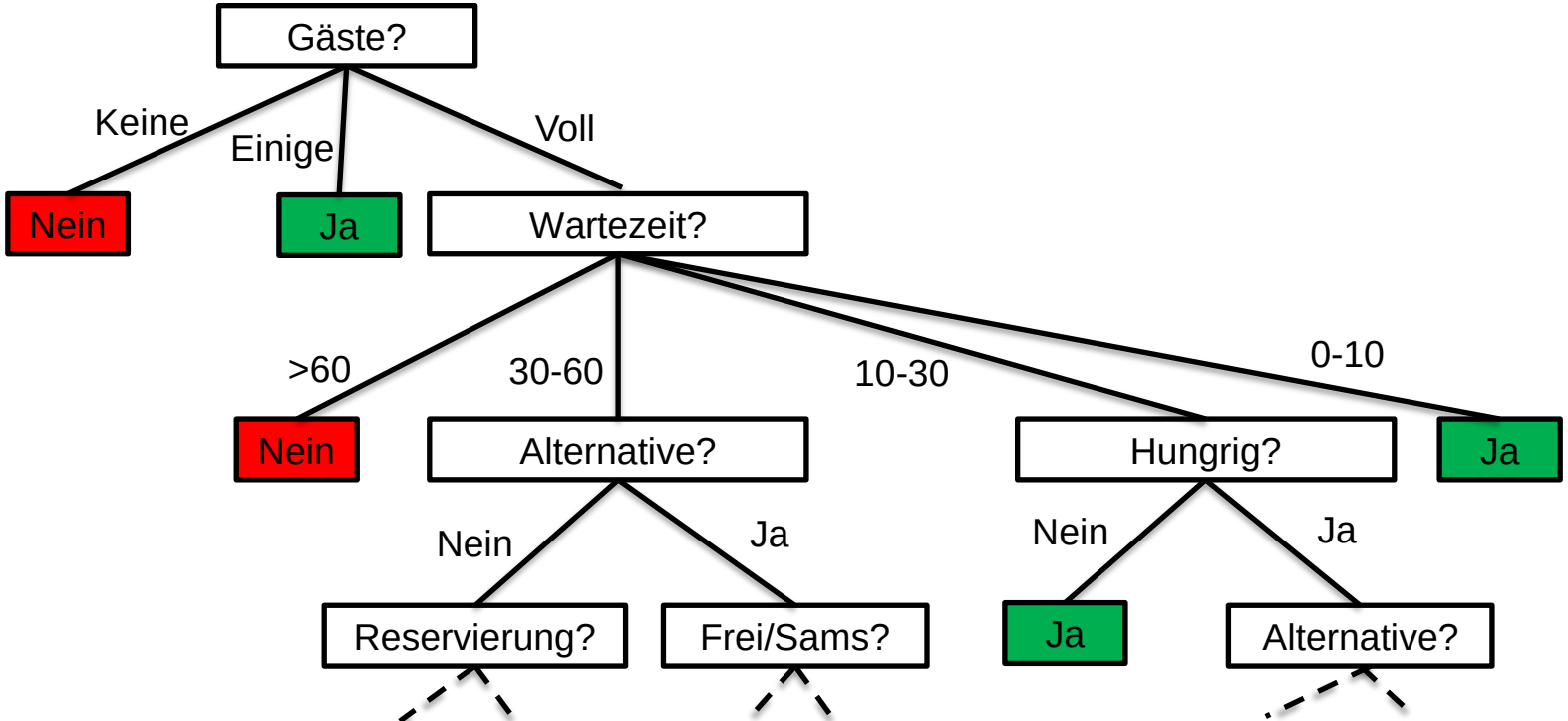
#	Alt	Bar	Frei	Hun	Gäste	Preis	Regen	Res	Typ	War	WW
x_{13}	Ja	Ja	Ja	Ja	Voll	€	Ja	Ja	Franz	30-60	y_{13}=???

Ungesehenes Beispiel klassifizieren



Disjunktive Normalform (DNF)

WW=Ja ↔ Gäste=Einige ∨ (Gäste=Voll ∧ Wartezeit=0-10)
(Gäste=Voll ∧ Wartezeit=10-30 ∧ Hungrig=Ja) ∨ …



Ausdruckstärke

- Jede aussagenlogische Funktion kann als DNF dargestellt werden
- Zu jedem Entscheidungsbaum kann eine DNF angegeben werden und umgekehrt
- Also:
  - Entscheidungsbaum kann jede beliebige Funktion beschreiben
  - Einschränkung auf Entscheidungsbäume ist keine Einschränkung
  - Lernen eines Entscheidungsbaums = Lernen logischer Formel
- Viele regelhafte Zusammenhänge lassen sich gut und überschaubar mit einem Entscheidungsbaum darstellen
  - Aber es gibt auch Ausnahmen (z.B. Mehrheitsfunktion)

Gliederung

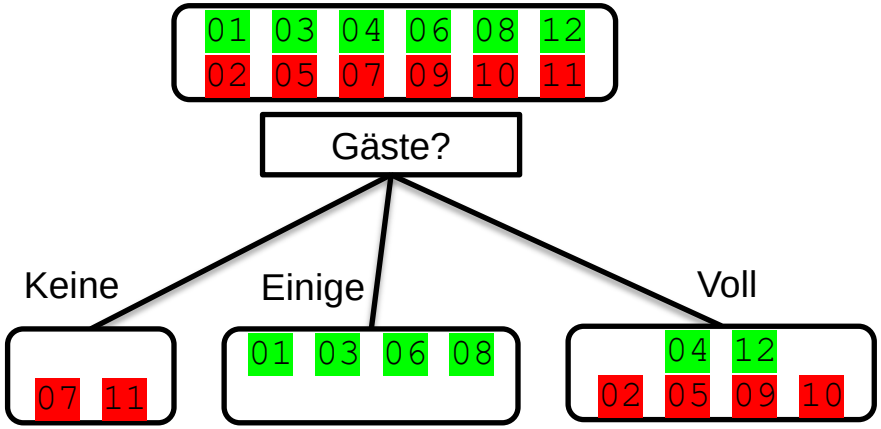
- Maschinelles Lernen: Überblick
- Überwachtes Lernen
  - Beispiel
- Entscheidungsbaum
  - Repräsentation
  - Lernen des Baums
- Evaluationsmethodik
- Ausblick Deep Learning

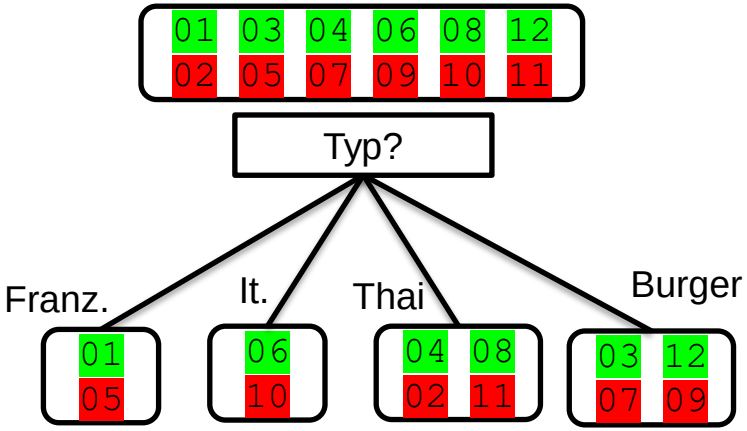
Grundidee

- Gesucht ist ein möglichst kleiner konistenter Baum
- Problem: Es ist nicht möglich die Menge aller Bäume mit einem vollständigen Verfahren zu durchsuchen
- Stattdessen: Divide-and-Conquer
  - Beginne mit einem ungelabelten leeren Wurzelknoten
  - Wähle an jedem ungelabeltem Blatt des Baums das Attribut das den besten Split macht
  - Wende an jedem enstehenden Blatt erneut das Verfahren an, falls notwendig

- Aber was ist mit guten oder bester Split gemeint?

Gute und schlechte Aufteilung





Entropie

- Entropie H ist ein Maß für die Unbestimmtheit einer Zufallsvariable
  - Münze die immer Kopf fällt hat Entropie 0
  - Faire Münze hat Entropie 1

$\log_2 0$ ist nicht definiert
gesamter Summand wird
bei P(v_k) = 0 als 0 gewertet

- Sei V Zufallsvariable mit den Werten v_k, dann gilt

$$H(V) = -\sum_k P(v_k) \log_2 P(v_k)$$

- Beispiel: Gezinkte Münze M, die mit 0.99 Kopf anzeigt

$$H(M) = -(0.99 * \log_2 0.99 + 0.01 * \log_2 0.01) = 0.08$$

Informationsgewinn

- Hilfsdefinition: Entropie B einer boolschen Zufallsvariable, die mit Wahrscheinlichkeit q wahr ist

$$B(q) = -(q \log_2 q + (1-q) * \log_2(1-q))$$

- Restliche R Entropie nach einem Split an Attribut A mit d verschiedenen Werten:

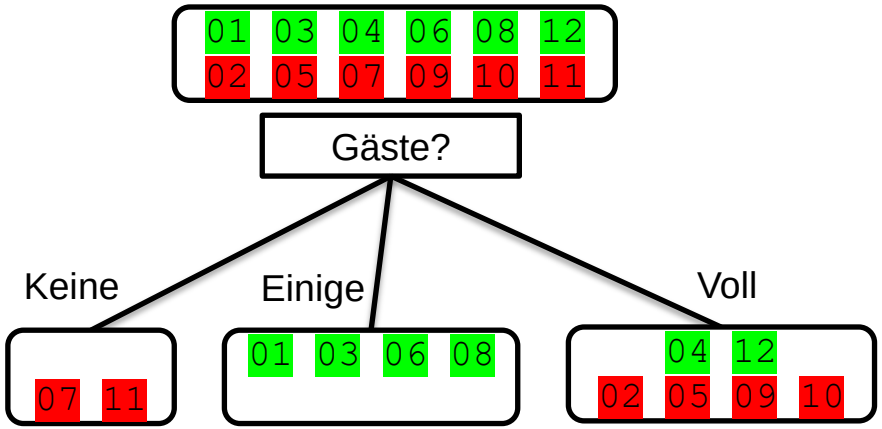
$$R(A) = \sum_{k=1}^{d} \frac{p_k + n_k}{p+n} B\left(\frac{p_k}{p_k + n_k}\right)$$

p = Positive Beispiele
n = Negative Beispiele

- Informationsgewinn eines Attributtests:

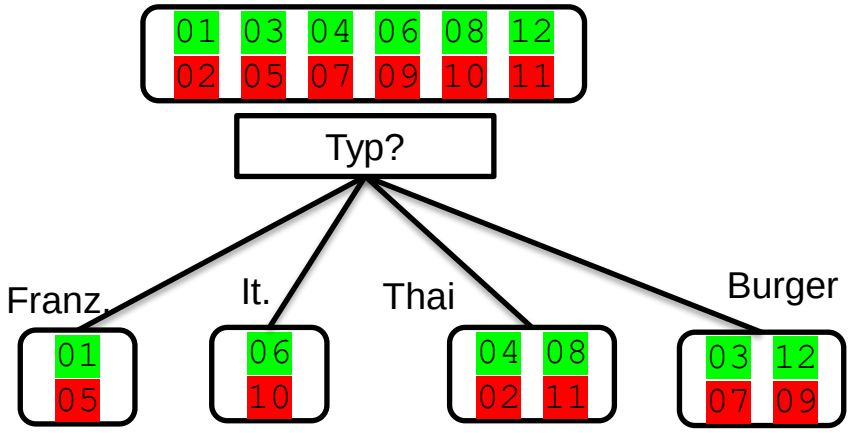
$$\text{Gewinn}(A) = B\left(\frac{p}{p+n}\right) - R(A)$$

Berechnung Informationsgewinn



$$B\left(\frac{6}{12}\right) = 1$$

$$1 - \left[\frac{2}{12}B\left(\frac{0}{2}\right) + \frac{4}{12}B\left(\frac{4}{4}\right) + \frac{6}{12}B\left(\frac{2}{6}\right)\right] \approx 0.541$$



$$1 - \left[\frac{2}{12}B\left(\frac{1}{2}\right) + \frac{2}{12}B\left(\frac{1}{2}\right) + \frac{4}{12}B\left(\frac{2}{4}\right) + \frac{4}{12}B\left(\frac{2}{4}\right)\right] \approx 0$$

Wähle des Attribut mit dem höchsten Informationsgewinn!

Algorithmus: Vier Fälle

(1) Es sind alle Beispiele positiv (oder alle negativ)
- Nicht weiter expandieren: Mit Ja (oder Nein) antworten

(2) Es gibt einige positive und einige negative Fälle und weitere Attribute stehen zur Verfügung
- Informationgewinn für alle Attribute berechnen, bestes Attribut verwenden
- Abbruch, wenn keine Attribut zu einer Verbesserung führt (wie bei (4))

(3) Es gibt an dem Knoten keine Beispiele
- Nicht weiter expandieren: Antwort = Mehrheitsentscheid der Beispiele am Elternknoten

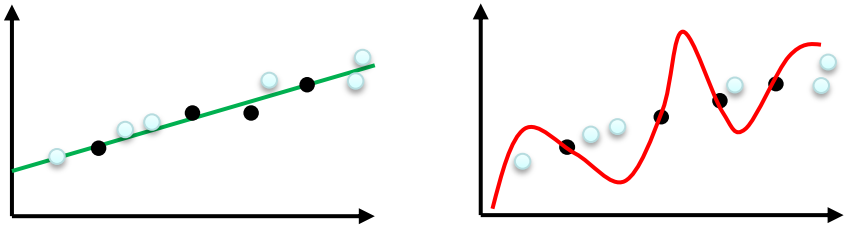
(4) Es gibt einige positive und einige negative Fälle aber keine weiteren Attribute stehen zur Verfügung
- Nicht weiter expandieren: Antwort = Mehrheitsentscheid der verbleibenden Beispiele

Gliederung

- Maschinelles Lernen: Überblick
- Überwachtes Lernen
  - Beispiel
- Entscheidungsbaum
  - Repräsentation
  - Lernen des Baums
- Evaluationsmethodik
- Ausblick Deep Learning

Overfitting (Überanpassung)

- Hypothese h ist überangepasst, wenn es eine Hypothese h^* gibt, so dass h^* einen kleinere Fehler gegenüber h hat in Bezug auf alle Beispiele, während es in Bezug auf die Trainingsdaten genau umgekehrt ist
- Beispiel bei einer Regression



- Bei beiden ital. Restaurantbesuchen zufällig WW => entsprechender Zweig im Baum wird angehängt
  - Kann genausogut für Kombinationen mehrerer Attribute passieren

Overfitting vermeiden

- Sinnvolle Auswahl des Datensatzens
  - Trainingsdatensatz sollte groß sein
    - Deutlich mehr Beispiele als Attribute
  - Trainingsdatensatz sollte repräsentativ sein
    - Zum Beispiel durch zufällige Auswahl
- Absichtliche Einschränkung des Hypothesenraums
  - Bei parametrischen Modellen Anzahl der Parameter einschränken
  - Bei Entscheidungsbaum Tiefe des Baums begrenzen
- Statt Aufteilung in Trainings und Testset, Aufteilung in Training, Validation und Testset

Trainings und Testdatensatz

- Gesamten Datensatz aufteilen in Trainingset und Testset
  - Zum Beispiel 90% zu 10%

- Das Modell lernen auf dem Trainingsdatensatz
- Güte des Modells messen mit dem Testset

- Fehlerrate = Anteil der Klassifikationen des Modells, die nicht stimmen
- Accuracy = 1 – Fehlerrate
  - Anteil richtiger Klassifikationen

k-fold Cross Validation

- Zerteile Datensatz in k ungefähr gleich große Teile
  - Verwende k-1 Teile als Traininingsset
  - Verwende 1 Teil als Testset
  - Ermittle die Qualität der Ergebnisse
- Führe dies für jede mögliche Aufteilung aus und bilde den Mittelwert über die Ergebnisse
  - Verhindert, dass zufällige Schwankungen in den Ergebnissen das Modell zu gut oder schlecht erscheinen lassen
- Oft als 10-fold cross validation angewendet
  - 10-fache Kreuzvalidierung

Hyperparameter

- Auswahl der Parameter, welche den Suchraum bestimmen (Hyperparameter)
  - Z.B. Maximale Tiefe des Baums oder Grad des Polynoms bei einer Regression
- Keine gute Idee:
  - Wähle Hyperparameter, die auf dem Testset am besten funktionieren
- Problem ist ein Overfitting gegen Testset
  - Fehlerrate reflektiert nicht, wie gut das Modell wirklich ist, wenn es auf echte (= andere) Daten angewendet wird

Training, Validation, Test

- Aufteilen in Training, Validation und Testset
- Für alle (oder einige) mögliche(n) Werte der Hyperparameter:
  - Lerne Modell auf Trainingset
  - Messe Qualität gegen Validationset
  - Merke dir das beste Setting
- Wende des beste Setting auf Test an und messe die Qualität der Ergebnisse

Gliederung

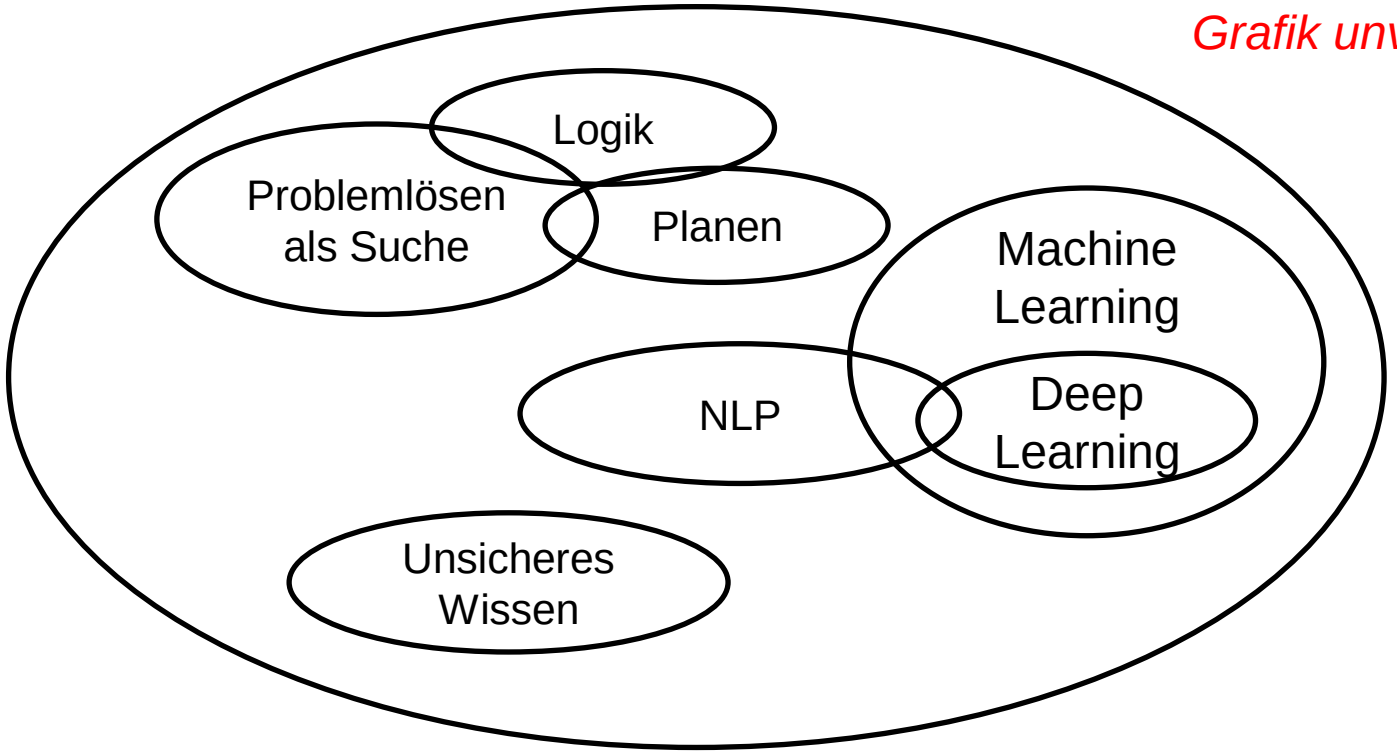
- Maschinelles Lernen: Überblick
- Überwachtes Lernen
  - Beispiel
- Entscheidungsbaum
  - Repräsentation
  - Lernen des Baums
- Evaluationsmethodik
- Ausblick Deep Learning

Deep Learning

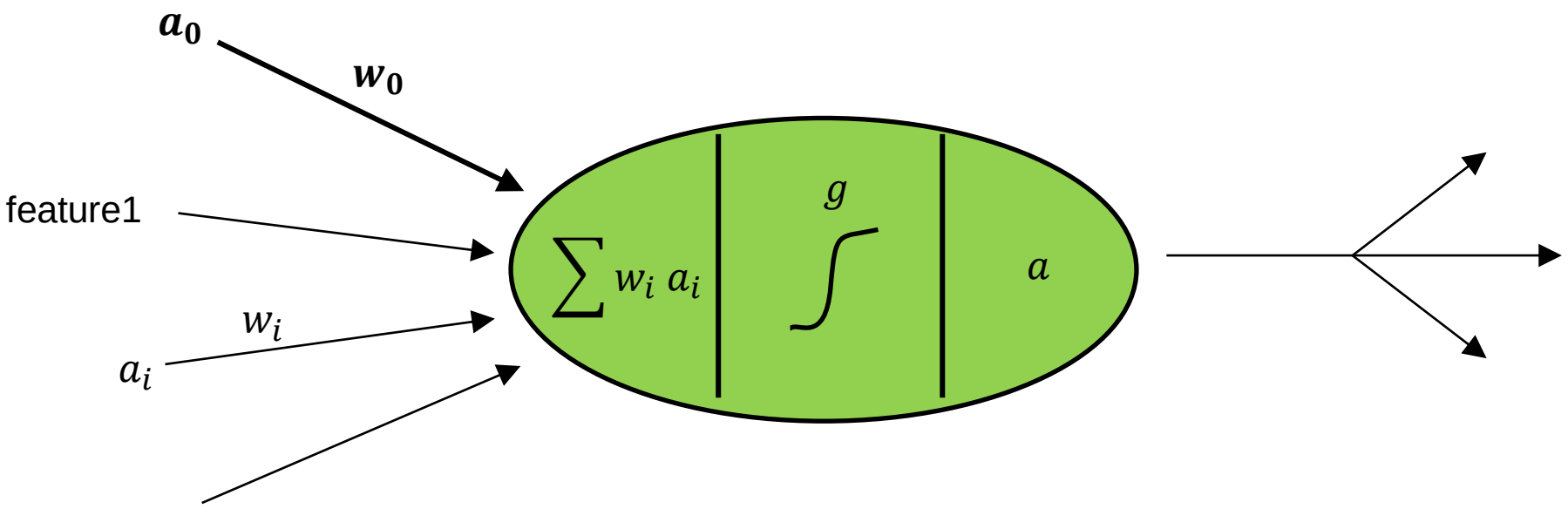
- Deep Learning = Lernen tiefer neuronaler Netze
- Vor über 50 Jahren wurden Neuronale Netze erforscht
- Erst „heute“ (seit 10-20 Jahre) erfolgreich
  - Bild Klassifikation (autonomes Fahren)
  - Natural Language Processing (Übersetzen)
  - Gesprochene Sprache verstehen
  - …
- Gründe für die (späten) positiven Ergebnisse
  - Extrem viele Eingabedaten notwendig
  - Extrem rechenaufwendig
  - Neues Wissen über Verfahren die in tiefen Netzen funktionieren

Deep Learning

- Heute: Deep Learning = KI ?



Grundbaustein: Perzeptron

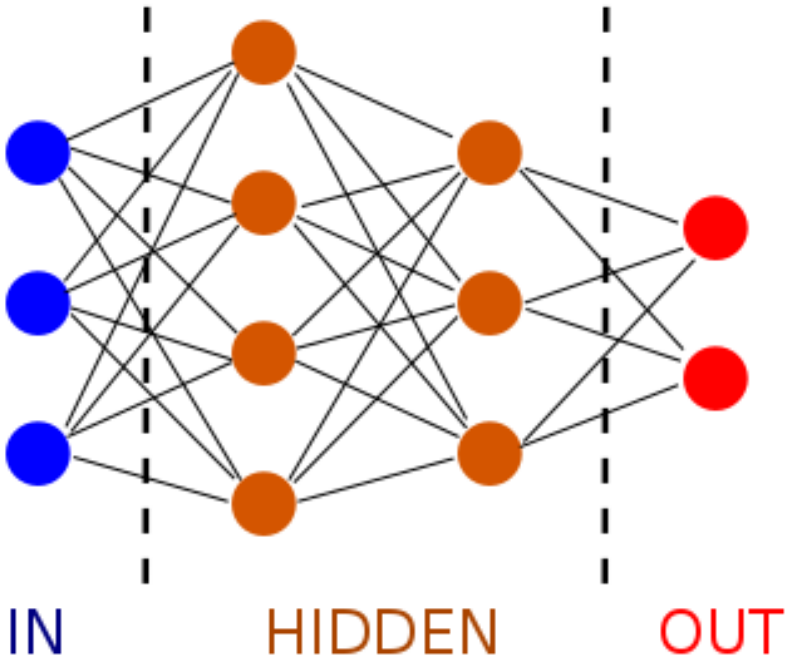


$$a = g(\sum w_i\, a_i)$$

Differenzierbar!

Künstliches Neuronales Netz

- Mehr „hidden layer“ möglich
- Rekurrente Netz: Rückgerichtete Kanten
- …



Künstliches Neuronales Netz

- Erlaubt beliebige Funktionen zu approximieren, wenn genug Layer vorhanden sind
- Dadurch
  - große Gefahr des overfitting
  - extrem mächtig

- Lernen ist Suche nach den Parametern
  - Parameter sind die Gewichte an den Kanten
  - In Prinzip als diskrete lokale Suche denkbar
  - Wird als Gradientenabstiegsverfahren realisiert

Deep Learning

- Konstruktion von tiefen Netzen mit vielen versteckten Ebenen und spezielle Konstrukten
- Kann Zusammenhänge erkennen, die andere Verfahren vielleicht nicht erfassen können
- Benötigt viele Trainingsbeispiele
- Feature müssen nicht explizit modelliert werden

- Ist ein Black-Box Verfahren
  - Im Gegensatz zu z.B. einem Entscheidungsbaum

- Aktuelle Forschungsthemen:
  - Erklärbarkeit von Deep Learning Modellen
  - Kombination von symbolischen Verfahren und Deep-Learning

Nächste Woche / Klausur

- Fragestunde
  - Bitte Fragen im Forum vorab im entsprechenden Thread nennen
- Evaluationsergebnisse
  - Leider nur 13 Teilnehmer
- Klausur am Montag 13.12. vor Ort!
  - 08:30-10:00 in SN163
  - Achtung: Hörsaal Pass benötigt!

Ende