Maschinelles Lernen und Datenanalyse in der Soziologie

PD Stefan Bosse

1. Inhalt

2. Überblick

2.1. Motivation

Dieser Online Kurs mit interaktiven Übungen soll:

- Einen **anwendungsorientierten Einstieg** in die Datenanalyse und Interpretation mit Verfahren des **Maschinellen Lernens** bieten;
- Einen **Überblick** über gängige und weniger gängige **Verfahren** geben;

➤ **Interaktive Tutorials und Übungen mit zielgruppenorientierten Fallbeispielen** sollen Verfahren begreifbar und erfahrbar machen!

2.2. Tandemkurs

➤ Dieser Kurs adressiert zwei primäre **Zielgruppen**:

- ❏ FB 4: Produktionstechniker und Materialwissenschaftler (und SysEngs)
- ❏ FB 8: Soziologen (und Psychos)

➤ Dabei gibt es zwei **Inhaltsstränge**:

- ❏ Einen gemeinsamen Strang mit Grundlagen und Verfahren
- ❏ Getrennte Stränge für Anwendungsbeispiele

2.3. Ontologie der Inhalte

➤ Die Ontologie des Kurses besteht aus den Bausteinklassen:

- ❏ **Modelle**
- ❏ **Verfahren** (Training, Test, Inferenz)
- ❏ Überwachtes Training
- ❏ Nicht überwachtes Training

➤ Weiterhin aus den Anwendungs- und **Datenklassen**:

- ❏ Sensorische und experimentelle Daten (Mess- und Prüftechnik)
- ❏ Erhebungs- und Umfragedaten (Soziologie)
- ❏ Metrische und Kategorische Variablen

Die Grenzen der Datenklassen sind fließend! Der Mensch als Sensor!

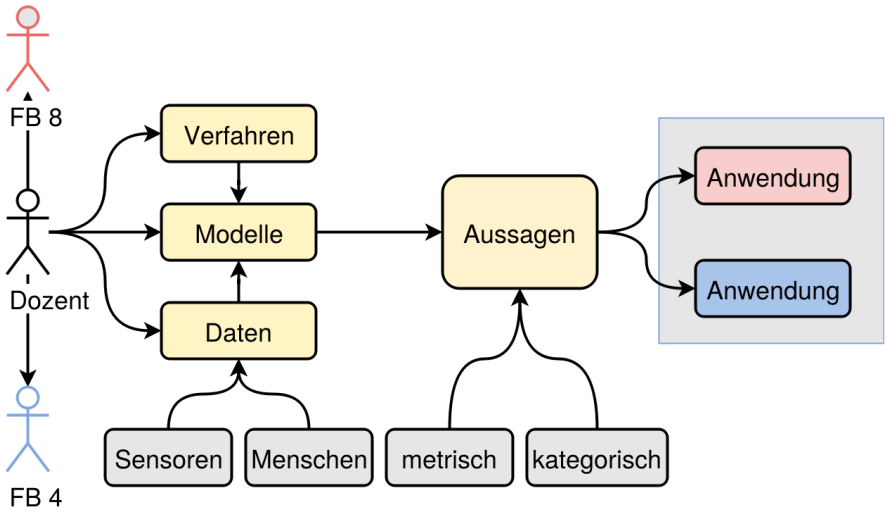


Abb. 1. Gemeinsame Verfahren und Modelle → Unterschiedliche Daten, Aussagen, Anwendungen

2.4. Ontologie der Veranstaltung

1. **Synchrone Vorlesungen mit Livestream** (experimentell!)
   - ➤ Studenten können über einen Chat/Eingabefeld Fragen stellen
   - ➤ Aufzeichnung der Vorlesung → 2.
2. **Asynchrone Video Vorlesungen und Tutorials** (alternativ)
   - ➤ Auch offline seh- und hörbar
3. **Gemeinsame Treffen mit** Videokonferenz (Zoom)
4. **Interaktive Tutorials** und Übungen mit *NoteBook* und ggfs. *WorkBook* im WEB Browser!
   - ➤ Offline ausführbar (evtl. werden Daten von einem Server geladen)
5. Texte und Folien
   - ➤ Vorlesungsskripte (am Anfang: für jedes Modul/jede Einheit) im PDF
   - ➤ Das vorlesungsskript gibt die Folieninhalte 1:1 wieder (nur anderes Layout)

- ➤ Alle Folien im HTML Format (auch offline lesbar)
- ➤ Begleitende Literatur (Bücher im PDF)

2.5. Ontologie der Services

1. WEB Service: Informationen, Dokumente, Folien, Videos:http://edu-9.de/Lehre/ml2k
2. Dokuwiki: **News**, Informationen und Links, **Chats**, **Videostreams**: http://ag-0.de/dokuwiki
   - ➤ Registrierung und Login erforderlich
   - ➤ Interaktiv!
3. SAS: Student Assignment System (TODO): http://edu-9.de/cas
   - ➤ Registrierung und Login erforderlich
4. VIDEO: (Video Opencast Server http://ag-0.de)

2.6. Prüfungsleistungen

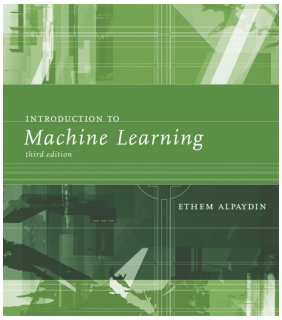
1. Eine mündliche Abschlussprüfung (20 Minuten); **oder alternativ** 2.
2. Eine schriftliche Seminararbeit (Experimentelle Arbeit oder Literaturrecherche)
   - ➤ 15-20 Seiten PDF
3. Bearbeitung und Abgabe der digitalen Übungen (JSON Dateien)

2.7. Literatur

- ➤ Zur Vertiefung!

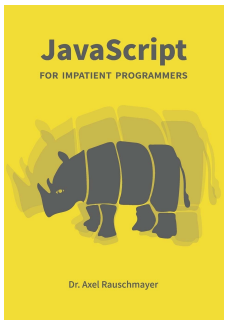
S. Richter, Statistisches und maschinelles Lernen. Springer Spektrum, 2019.

[www.pinterest.com] E. Alpaydın, Introduction to Machine Learning. MIT Press, 2010.

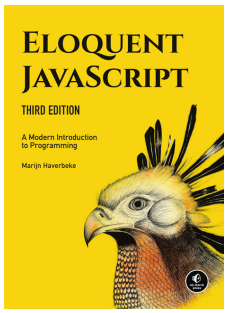


Programmierung

Axel Rauschmayer, JavaScript For Impatient Programmers.

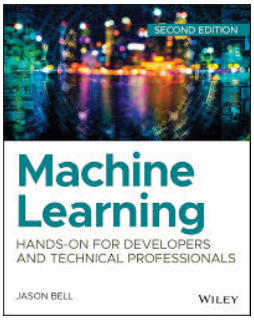


M. Haverbeke, Eloquent JavaScript. 2018.

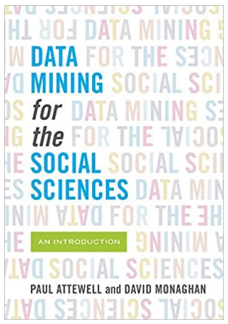


Domainspezifische Literatur

J. Bell, Machine Learning - Hands-On for Developers and Technical Professionals. John Wiley & Sons, Ltd, 2015.



P. Attewell and D. B. Monaghan, Data mining for the social sciences : an introduction. University of California Press, 2015.



2.8. Software

NoteBook

- ➤ Interaktive vorwiegend praktische Übungen werden rein digital im WEB Browser mit den *NoteBooks* durchgeführt
- ➤ Ein digitale Übung (oder Turorial) besteht aus:
  - ❑ Textabschnitten
  - ❑ Informationsblöcken
  - ❑ Aufgaben (mit Lösungen)
  - ❑ Editoren für Programmcode
  - ❑ Ausführungsterminals für Programmcode
  - ❑ uvm.

Abb. 2. Ein NoteBook im WEB Browser

NoteBook Konzept

- ➤ Top-down Bearbeitungsfluß
- ➤ Statische Struktur mit dynamischen Inhalten

- Alle dynamischen Inhalte können in einer JSON Datei gespeichert und wieder geladen werden
- Es können Notizzettel überall im NoteBook angeheftet werden (werden auch gespeichert)
- Musterlösungen (dynamische Inhalte) können eingebettet und mit einem Schlüssel freigeschaltet werden

2.9. WorkBook

- Dynamische Struktur mit dynamischen Inhalten
- Ein WorkBook besteht aus
  - Textabschnitten (MarkDown)
  - Codesnippets mit Editoren und Ausgabekonsolen
  - Speziellen Snippets wie editierbare Tabellen oder allg. Formulare
- Programmierung in JavaScript
- Alle dynamischen Inhalte und Daten können im JSON Format gespeichert und wieder geladen werden

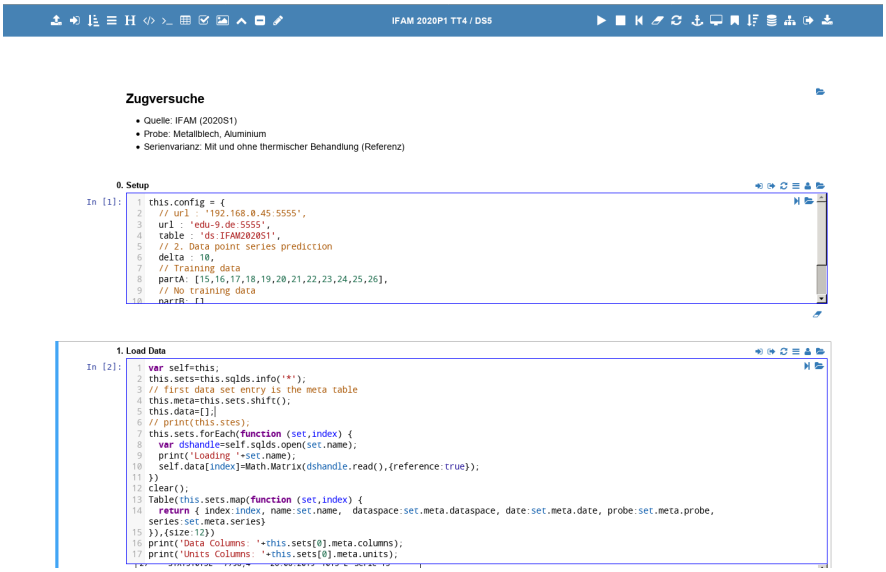


Abb. 3. Ein WorkBook Beispiel

2.10. Machinelles Lernen

Schlüsselwörter und Begriffe

Welche Begriffe werden häufig bei ML genannt:

Anwendungsgebiete

Welche Anwendungsgebiete gibt es:

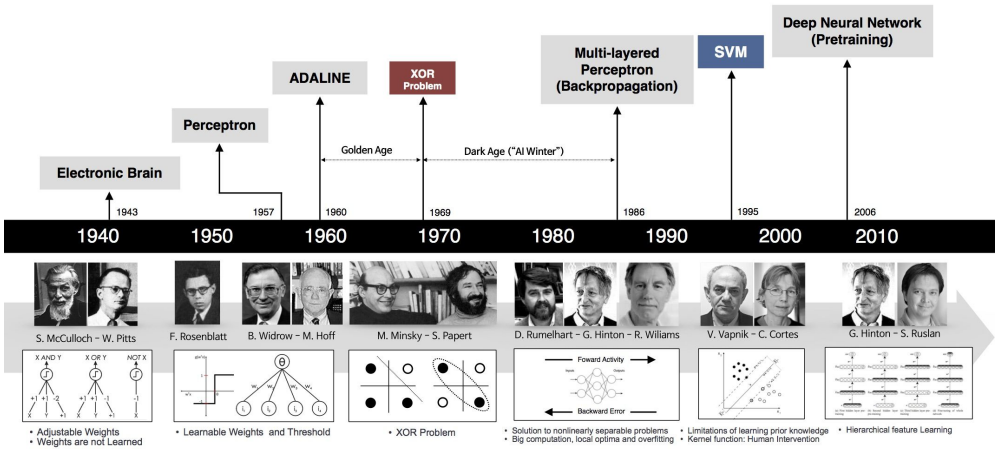
Fragestellungen

Welche Fragestellungen (zu lösende Probleme) gibt es:

2.11. Inhalte

1. **Eingabe** x: Daten (Attribute) und Eigenschaften (Analyse)
2. **Sensoren**: Erfassung von Daten, $S(welt)$: $welt \rightarrow x$
3. **Ausgabe** y: Numerische und kategorische Werte
4. **Metriken und Taxonomie**: Grundlagen des Maschinellen Lernens
5. **Algorithmen und Modelle**: $f(x)$: $x \rightarrow y$
6. **Training, Lernen, Prädiktion, Test** $M(<x,y>)$: $<x,y> \rightarrow f$
7. **Anwendungen**

2.12. Geschichte



[www.pinterest.com]

3. Daten und Sensoren

Metriken von Daten
Metriken von Aussagen
Sensoren als Datenquellen

3.1. Daten

- ➤ Daten sind die Grundlage für die Modellbildung und Modelltestung
- ➤ Daten können aus einer Vielzahl von Quellen stammen
  - ❏ Experiment
  - ❏ Simulation
  - ❏ Feldstudie
  - ❏ Abgeleitet aus anderen Datensätzen: **MapAndReduce**(D): $D \rightarrow D$'
- ➤ Allgemein kann man Daten und deren Werte unterteilen in:
  - ❏ Skalare Werte, wie Temperatur, Alter, usw.

- ❏ Serien von Skalaren Werten, wie Zeitserien
- ❏ Vektorielle Werte wie Bilder
- ❏ Zusammengesetzte Daten, also Datenrecords

➤ Daten haben daher eine Dimensionalität ^{N}, wobei die Wertemenge einer Dimension aus den ganzen , reelen , der Zeit oder kategorischen Wertemengen bestehen kann (oder Untermengen davon).

3.2. Datenreduktion

➤ Ziel der Datenanalyse ist die Reduktion von Eingabedaten bezüglich Größe und Dimensionalität:

$$P(X^N) : X^N \rightarrow Y^M$$
$$|Y| < |X|, M < N$$

```
1: function isStrong(age,weight,length) =
2:   age < 10 ? → false
3:   weight > 200 ? → false
4:   (weight/length) > 30? → false
5:   true
```

Beispiel einer Datenreduktionsfunktion $^{3} \rightarrow$

Datenklassen

Numerische und Metrische Werte

Das sind Werte die abzählbar sind und wo man Relationen (wie kleiner oder größer) sinnvoll definieren kann, also alle reellen und ganzen Zahlen.

➤ Beispiele: Temperatur, Länge, Ort, Zeit

Kategorische Werte

Das sind symbolische Werte für die entweder keine (sinnvolle) Ordnungsrelation existiert oder wo sich wenigstens keine Differenzen bilden lassen.

➤ Beispiele: Staatsangehörigkeit, Farbennamen (rot < gelb???), Schadenstyp

Skalierung der numerischen Werte

Intervallskaliert

Für diese Art von Attributen sind nur Unterschiede (Addition oder Subtraktion) sinnvoll. Beispielsweise wird die in °C oder °F gemessene Temperatur intervallskaliert. Wenn es 20 °C an einem Tag und 10 °C am folgenden Tag ist, ist es sinnvoll, über einen Temperaturabfall von 10 °C zu sprechen, aber es ist nicht sinnvoll zu sagen, dass es doppelt so kalt ist wie am Vortag.

Verhältnisskaliert

Hier kann man sowohl Differenzen als auch Verhältnisse zwischen Werten berechnen. Zum Beispiel kann man für das Alter sagen, dass jemand, der 20 Jahre alt ist, doppelt so alt ist wie jemand, der 10 Jahre alt ist.

Ordnungsrelationen

Nominal

Die Attributwerte in der Domäne sind ungeordnet und somit nur Gleichheitsvergleiche sinnvoll. Das heißt, wir können nur überprüfen, ob der Wert des Attributs für zwei bestimmte Instanzen gleich ist oder nicht. Zum Beispiel ist Geschlecht ein nominales Attribut.

Ordinal

Die Attributwerte sind geordnet und somit Gleichheitsvergleiche (ist ein Wert gleich einem anderen?) und relationale Vergleiche (ist ein Wert kleiner oder größer als ein anderer?) sind erlaubt, obwohl es möglicherweise nicht möglich ist, die Differenz zwischen den Werten zu quantifizieren!

3.3. Daten

Datensätze als Matrizen

➤ Ein Menge von Daten kann in **Matrizenform** als Matrix $\boldsymbol{D}$ dargestellt werden (Analogie zur Tabellenform) [1]:

$$\mathbf{D} = \left(\begin{array}{c|cccc} & X_1 & X_2 & \cdots & X_d \\ \hline \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{array}\right)$$

- Die Zeilen sind Rekords der Variablenmenge $\{X_\mathrm{i}|\mathrm{i}=1,d\}$ und geben als d-stelliges Tupel je nach Anwendung und Zielsetzung einzelne Beispiele, Instanzen, Experimente, Entitäten, Objekte, und Eigenschaftsvektoren wieder

$$\tilde{x_i} = (x_{i1}, x_{i2}, .., x_{id})$$

- Der Vektor $\boldsymbol{X}$ ist die Menge aller Variablen (Sensoren) und die Spalten der Matrix $\boldsymbol{D}$:

$$\tilde{X} = (x_1, x_2, .., x_d)$$

```
type row = { X1:number, X2:number, .., Xd:number }
type table = row array;
```

3.4. Eingabe- und Ausgabevariablen

- Sensoren sind typischerweise Eingabevariablen x
- Aussagen sind Ausgabevariablen y, also Ergebnisse die sich aus den Eingangsvariablen ableiten lassen können (durch eine Funktion F):

$$\tilde{X} = (x_1, x_2, .., x_u, y_1, y_2, .., y_v)$$
$$\tilde{x_i} = (x_{i1}, x_{i2}, .., x_{iu}, y_{i1}, y_{i2}, .., y_{iv})$$
$$F(\tilde{x'}) : \tilde{x'} \rightarrow \tilde{y},$$

mit $u+v=d$.

3.5. Beispiel einer Datenmatrix

- Botanischer Datensatz mit geometrischen (numerischen) Eigenschaften einer Pflanze und kategorischer Klassifikation:

	Sepal length	**Sepal width**	**Petal length**	**Petal width**	**Class**
	X_1	X_2	X_3	X_4	X_5
$\mathbf{x}_1$	5.9	3.0	4.2	1.5	Iris-versicolor
$\mathbf{x}_2$	6.9	3.1	4.9	1.5	Iris-versicolor
$\mathbf{x}_3$	6.6	2.9	4.6	1.3	Iris-versicolor
$\mathbf{x}_4$	4.6	3.2	1.4	0.2	Iris-setosa
$\mathbf{x}_5$	6.0	2.2	4.0	1.0	Iris-versicolor
$\mathbf{x}_6$	4.7	3.2	1.3	0.2	Iris-setosa
$\mathbf{x}_7$	6.5	3.0	5.8	2.2	Iris-virginica
$\mathbf{x}_8$	5.8	2.7	5.1	1.9	Iris-virginica
⋮	⋮	⋮	⋮	⋮	⋮
$\mathbf{x}_{149}$	7.7	3.8	6.7	2.2	Iris-virginica
$\mathbf{x}_{150}$	5.1	3.4	1.5	0.2	Iris-setosa

[1]

Attribute

- Die gemessenen Variablen X_1 bis X_4 sind metrische Datenvariablen, die Variable X_5=y ist eine kategorische Variable!
- Die gemessenen Variablen X_1 bis X_4 (also Sensoren) nennt man **Attribute**, da sie Eigenschaften und beschreibende Variablen der Zielvariablen y sind

3.6. Sensoren

Welche Sensoren und Messdaten kennt ihr:

- Umfragen
  - Umfragevariablen (Antworten auf Fragen) sind Sensoren von einzelnen Menschen
  - Fusionierte Umfragevariablen (z.B. Ensemblemittelwerte) sind Sensoren von Menschengruppen
- Allgemein verfügbare Daten
  - Soziale Netzwerke und soziale Medien

- ❏ Datenbanken von Behörden usw.

3.7. Sensormodell

- ➤ Ein Sensor ist ein Messwandler, auch in der Soziologie (Indikator für eine Eigenschaft die nicht direkt messbar ist)
- ➤ Ein Sensor bildet daher eine i.A. physikalische Größe x auf eine andere Größe y ab:

$$S(x) : x \rightarrow y, K : correct(x \rightarrow y)$$

- ➤ Es gibt i.A. eine Kalibrierungsfunktion $K(f, x, y)$
- ➤ Beispiele: Soziale Vernetzung → Numerischer Radiuswert, Wählerstimmen → Politik, d.h., **Zuordnung von Zahlen zu Objekten oder Ereignissen nach festgelegten Regeln**

3.8. Sensordaten

- ➤ Sensoren S sind Datenquellen d von physikalischen, soziologischen oder sonstigen natürlichen nicht direkt erfassbaren Größen x
- ➤ Die Datenwerte (numerisch) werden in einem definierbaren Intervall liegen
  - ❏ Die Kenntnis des Werteintervalls ist wichtig für spätere Datenverarbeitung, Analyse, und Maschinelles Lernen!
  - ❏ Kategorische Werte werden ebenfalls durch eine Menge definiert

$$S(x) : x \rightarrow d$$
$$d \in [a, b] \Rightarrow \{v_0, v_1, .., v_i\}$$

4. Mess- und Sensorische Systeme

Der Ursprung der Daten für Analyse und Maschinelles Lernen!
Ein Sensor kommt selten allein.

4.1. Sensoraggregation

Sensorklassen

Physische Sensoren

Physische Sensoren messen direkt eine Größe mit einem Messinstrument (kann auch die Auswertung einer Frage in einem Fragebogen sein), Smartphone

Virtuelle Sensoren

Verwenden Daten (von physischen und anderen virtuellen Sensoren) um neue sensorische Werte zu berechnen (kein Messinstrument) → **Aggregatoren**!!

Schichtenmodell von Sensorischen Systemen

In sensorischen Systemen werden Sensordaten in verschiedenen Ebenen verarbeitet:

- **Vertikale Ebenen** repräsentieren die sensorischen Domainen und die Sensorklassen;
- **Horizontale Ebenen** repräsentieren die Datenverarbeitung.

Vertikale Ebenen

Perzeption

Hier findet die Akquisition der rohen Sensordaten statt. Die Sensoren sind räumlich verteilt und werden lokal vorverarbeitet.

Aggregation

Einzelne Sensordaten werden zeitlich und räumlich zusammengeführt und gesammelt (Sensorfusion)

Applikation

Die gesammelten Daten werden nutzbar gemacht: Weitere Datenverarbeitung, Aufbereitung, Eigenschaftsselektion, Informationsgewinnung, Visualisierung

Horizontale Ebenen

- Die horizontalen Ebenen durchziehen alle vertikalen Ebenen:

1. Sicherheit
2. Datenverarbeitung
3. Kommunikation
4. Datenspeicherung
5. Nachrichtenvermittlung
6. Management

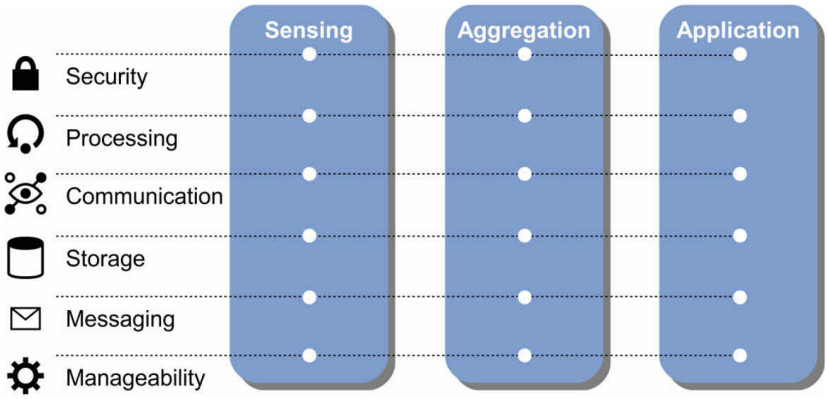


Abb. 4. Grundlegender Zusammenhang der horizontalen und vertikalen Ebenen in Sensorischen Systemen

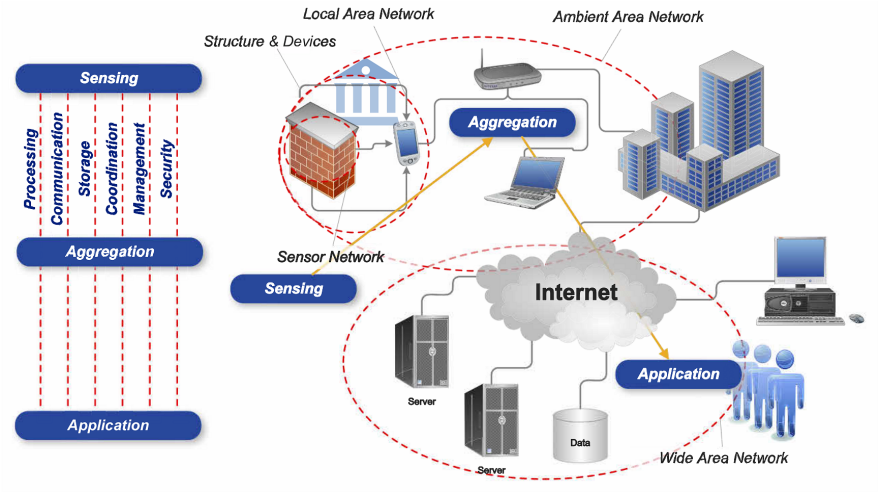


Abb. 5. Räumliche Abbildung der vertikalen Ebenen auf Cloud Computing

4.2. Sensoren in den Ebenen

Perzeption

Vorwiegend physische Sensoren

Aggregation

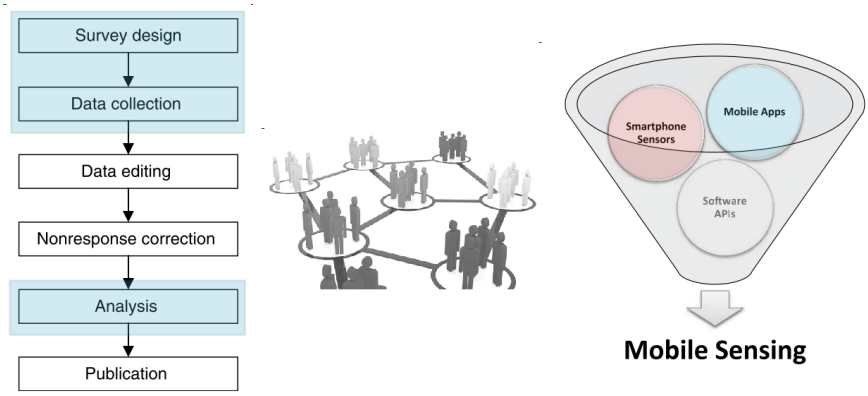
Virtuelle Sensoren, Datenreduktion (Größe und Dimensionalität)

Applikation

Datenanalyse und Modellbildung, Inferenz von Information, Maschinelles Lernen

4.3. Umfragen und Crowd Sensing

➤ Menschen sind Sensoren



[8]

Abb. 6. Von klassischen Umfragen zu mobilen Crowd Sensing mit Smartphones

4.4. Messfehler und Vertrauen

➤ Die Messgrößen können statisch (zeitlich konstant) oder dynamisch (zeitlich veränderlich) sein. Die Wandlung dieser Messgrößen ergeben dann entsprechend Gleich- und Wechselsignale.

➤ Auch eine prinzipiell zeitlich unveränderliche Messgröße (bezogen auf die Messung in einem vorgegeben Zeitinterval) erzeugt kein konstantes Signal. Ursache: Rauschen

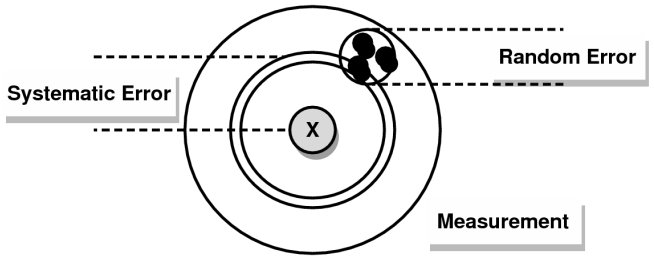
- Wiederholt man daher eine Messung N-mal unter gleichen Bedingungen, so wird man eine Reihe von verschiedenen Messwerten $\{s_1,s_2,...,s_n\}$ erhalten.
- Es gibt systematische und zufällige Fehler bei der Messung, die sich überlagern.

Systematische Abweichung (systematischer Fehler)

- Abweichung wird durch den Sensor verursacht
- z.B.: falsche Eichung, dauernd vorhandene Störungen wie Reibung
- lässt sich nur durch sorgfältiges Untersuchen der Fehlerquelle beseitigen

Zufällige Abweichung (zufälliger oder statistischer Fehler)

- Abweichung wird durch unvermeidbare, regellose Störungen verursacht
- bei wiederholter Messung weichen Einzelergebnisse voneinander ab
- Einzelergebnisse schwanken um einen Mittelwert



[9]

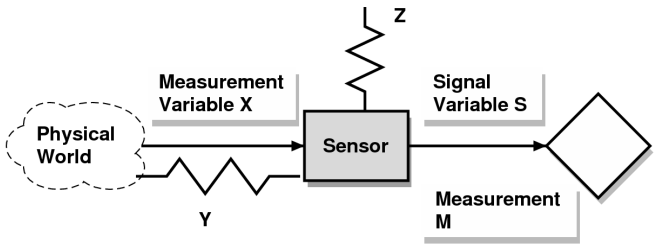
Abb. 7. Offset und Präzision bei der Messung einer Variable X

Systematische Fehler

- Eine Messgröße X ist meistens durch störende Messgrößen Y,Z,... usw. überlagert:

$$K(X,Y,Z) \; : \; X \times Y \times Z \rightarrow S, K(x,y,z) \approx \sum_{n=0}^{m} a_n x^n + \sum_{n=0}^{m} b_n y^n + \sum_{n=0}^{m} c_n z^n$$

➤ So kann z.B. bei einer Messung von sozialpsychologischen Parametern der Wohnort und die Lebensumgebung Einfluss auf den Sensor und dessen "Übertragungsfunktion" und somit auf das "Messsignal" S haben.



[9]

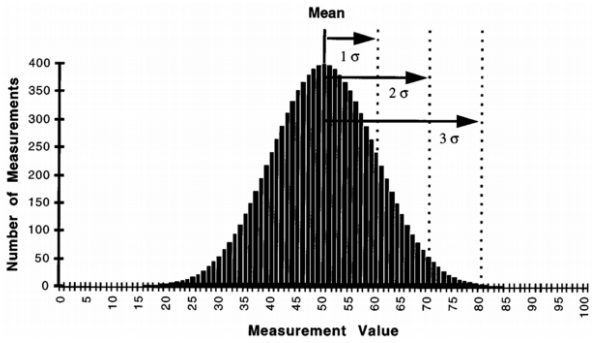
➤ Systematische Fehler verfälschen die Kalibrierungsfunktion (z. B. bei Geraden den Offset und Steigung). Sind sie bekannt, können sie kompensiert (rausgerechnet) werden.

➤ Systematische Fehler können aber auch während der Datenverarbeitung entstehen, so z.B. durch Rundungsfehler oder Verwendung von Funktionsmodellen außerhalb ihres Spezifikationsbereiches.

Zufällige Fehler - Streuung

➤ Zufällige Fehler beeinflussen die Genauigkeit einer Messung (Rauschen).

➤ Wiederholt man eine Messung einer Größe X die durch reine zufälligen Fehler verfälscht wird, so ist die Häufigkeitsverteilung der Messwerte $S=\{s_1,s_2,...,s_n\}$ um einen Mittelwert $\bar{S}$ durch eine Gaussverteilung gegeben (dabei muss die Anzahl der Messungen N groß sein).



[9]

Abb. 8. Häufigkeitsverteilung nach Gauss von Messwerten um einen Mittelwert

➤ Der Mittelwert S repräsentiert die Abschätzung des wahren/wirklichen Wertes Σ der Messgröße X (oder S):

$$\bar{S} = \frac{1}{N} \sum_{i=1}^{N} s_i$$

➤ Die Standardabweichung ist ein Maß für die Zuverlässigkeit (Präzision) der einzelnen Messwerte einer Messreihe $\{s_1, s_2, ..., s_n\}$:

$$\sigma = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (s_i - \bar{S})^2}$$

| Eine Vergrößerung der Anzahl N der Messungen (unter gleichen Bedingungen!) führt zu einer Verbesserung des Mittelwertes $\bar{S}$ (Grenzfall $N \to \infty$), nicht aber zu einer wesentlichen Verkleinerung der Standardabweichung , da die Genauigkeit nicht steigt!

➤ Der wirkliche Mittelwert Σ ist nicht bekannt (nur im Grenzfall $N \to \infty$ ist $\bar{S}$=Σ) - Es gibt aber ein Vertrauensintervall mit einer Wahrscheinlichkeit P dass dieser darin enthalten ist:

Σ [$\bar{S}$- ,$\bar{S}$+] mit 68.3%

Σ [$\bar{S}$-2 ,$\bar{S}$+2] mit 95.4%

Σ [$\bar{S}$-3 ,$\bar{S}$+3] mit 99.73%

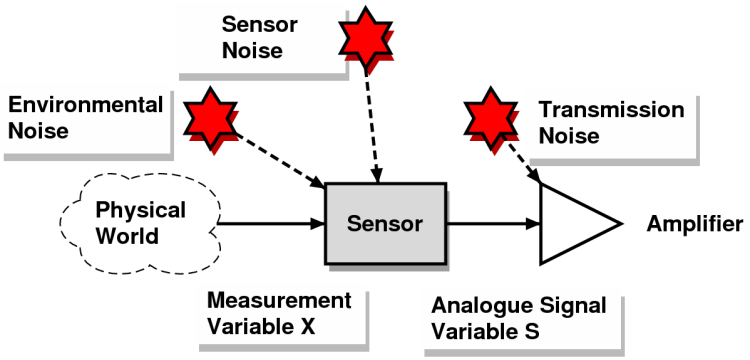


Abb. 9. Rauschquellen bei einer Messung

4.5. Beispiele

/webwork

Plot

```
Plot(Array.random(200),{width:'70%',size:20})
```

Analysis

```
Math.statistics.analysis(Array.random(200))
```

5. Datenanalyse und Eigenschaftsselektion

Häufig sind die rohen sensorischen Daten(variablen) zu hochdimensional und abhängig voneinander
Reduktion auf wesentliche Merkmale kann ML Qualität deutlich verbessern
Häufig besitzen einzelne Sensorvariablen keine oder nur geringe Aussagekraft (geringe Entscheidbarkeitsqualität)

5.1. Datenqualität

➤ Die Daten ***D*** werden durch vier wesentliche Eigenschaften beschrieben, die auch mit statistischer Analyse quantifiziert werden können:

Rauschen. Rauschen ist die Verzerrung der Daten. Diese Verzerrung muss entfernt oder Ihre nachteiligen Auswirkungen vermindert werden, bevor ML Algorithmen ausgeführt werden, da die Leistung und Qualität der Algorithmen beeinträchtigen kwerden ann.

Es gibt eine Vielzahl von Filteralgorithmen um den Effekt von Rauschen zu vermindern.

Ausreißer. Ausreißer sind Instanzen, die sich erheblich von anderen Instanzen im Datensatz unterscheiden. Beispiel: Durchschnittliche Anzahl der Follower von Nutzern auf Twitter. Eine Berühmtheit mit vielen Followern kann die Durchschnittliche Anzahl von Followern pro Person leicht verzerren. Da die prominenten Ausreißer sind, müssen Sie aus der Gruppe der Personen entfernt werden, um die Durchschnittliche Anzahl der Follower genau zu Messen.

Aber: Ausreißer können in besonderen Fällen nützliche Muster darstellen und die Entscheidung, sie zu entfernen, hängt vom Kontext und Fragestellung ab.

Fehlende Werte. Fehlende Werte sind Funktionswerte, die in Instanzen fehlen. Zum Beispiel, Einzelpersonen können es vermeiden, Profilinformationen auf social-media-Websites zu melden, wie Ihr Alter, Standort, oder Hobbys. Um dieses Problem zu lösen, können wir (1) Instanzen mit fehlenden Werten entfernen, (2) fehlende Werte schätzen (Z. B. durch den gängigsten Wert ersetzen) oder (3) fehlende Werte ignorieren, wenn data mining-algorithmen ausgeführt werden.

Duplikate. Doppelte Daten treten auf, wenn mehrere Instanzen mit genau denselben Funktionswerten vorhanden sind. Doppelte blog-posts, doppelte tweets oder Profile auf Social-media-Websites mit doppelten Informationen sind Beispiele für dieses Phänomen. Je nach Kontext können diese Instanzen entweder entfernt oder beibehalten werden. Wenn Instanzen beispielsweise eindeutig sein müssen, sollten doppelte Instanzen entfernt werden.

5.2. Statistische Analyse

- Statistische Analysen von Mess- und Sensordaten können neue Datenvariablen erzeugen und Informationen über die Daten liefern:
  - Eigenschaftsselektion (Feature Selection) für ML und Informationsgewinnung
  - Variablentransformation mit Datenreduktion
- Statistische Analyse liefert eine Reihe von Kennzahlen über Datenvariablen, das können Eigenschaften für die Weiterverabeitung sein:

$$stat(\tilde{x}) : \tilde{x} \rightarrow \tilde{p},$$
$$\tilde{p} = \{mean, \sigma, ..\}$$

Welche statistische Größen gibt es? Was können statistische Größen über Daten aussagen?

5.3. Statistische Funktionen

Peak amplitude (y_{peak})	$y_{peak} = \max \lvert y_i \rvert$
Mean ($\overline{y}$)	$\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$
Mean square ($\overline{y}_{sq}$)	$\overline{y}_{sq} = \frac{1}{n} \sum_{i=1}^{n} (y_i)^2$
Root-mean-square (rms)	$rms = \sqrt{\frac{1}{n} \sum_{i=1}^{n} y_i^2}$
Variance $(\sigma^2)^a$	$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \overline{y})^2$

[2:173]

Standard deviation $(\sigma)^a$	$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2}$
Skewness (dimensionless) $(\gamma)^a$	$\gamma = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^3}{\sigma^3}$
Kurtosis (dimensionless) $(\kappa)^a$	$\kappa = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^4}{\sigma^4}$
Crest factor (X_{cf})	$X_{CF} = y_{peak}/rms$
K-factor (X_k)	$X_{CF} = (y_{peak})(rms)$

[2:173]

5.4. Korrelation von Datenvariablen

- ➤ Variablen **X** sollten möglichst (linear) unabängig sein,
  - ❑ Um eine geeignete, robuste und genaue Modellsynthese (also ML) zu ermöglichen, d.h.
  - ❑ Es sollte möglichst keine Zusammenhänge der Form *correlation*(X_i,X_j) geben!
  - ❑ Um den Modellsyntheseprozess zu beschleunigen (also das Training); Rechenzeit reduzieren
  - ❑ Um Modelle klein und kompakt zu halten
- ➤ Abhängige Variablen sollten identifiziert und in unabhängige "transformiert" werden!

Beispiel Prozessanalyse

- ➤ Eine Datentabelle ***D*** mit experimentellen Messgrößen und Fertigungsparametern (Prozessparameter) von additiv gefertigten Bauteilen hatte

zunächst 7 Variablen (numerisch):

- X_1: Hatchabstand [mm]
- X_2: Scangeschwindigkeit [mm/s]
- X_3: Laserleistung [W]
- X_4: Schichtstärke [mm]
- X_5: Volumenenergiedichte [J/mm^3]
- X_6: Bauplatten Position x
- X_7: Bauplatten Position y
- Y_1: Dichte (%)

➤ $\boldsymbol{D}$ bestand aus 61 Experimenten mit unterschiedlichen Fertigungsserien

➤ Mit einer Principle Component Analysis (PCA) konnte die ganze Tabelle auf die Variablen PC_1 und Y_1reduziert werden!

- Die Genauigkeit der mit ML synthetitisierten Funktion $M(\boldsymbol{X})$: $\boldsymbol{X} \rightarrow$ *Dichte* konnte ohne signifikanten Genauigkeitsverlust nur aus PC_1 abgeleitet werden, d.h. $M(PC1)$: $PC1 \rightarrow$ *Dichte*
- Aber: Für die Inferenz (Applikation) von M muss die PCA für die Eingabedaten $\boldsymbol{X}$ wiederholt werden bzw. die Datentransformation durchgeführt werden!

5.5. Analyse Kategorischer Variablen

➤ Die Analyse von kategorischen Variablen vereint die Konzepte:

- Mengenlehre
- Kodierung/Dekodierung
- Verteilung (Wahrscheinlichkeit des Auftretens)

Attribut	Wertemenge
Aussicht	sonnig, regnerisch, bewölkt
Temperatur	kalt, mild, heiß
Luftfeuchtigkeit	hoch, normal
Windig?	ja, nein

[5]

Messdaten

Beispiel	Aussicht	Temperatur	Luftfeuchtigk.	Windig?	Klasse
1	sonnig	heiß	hoch	nein	N
2	sonnig	heiß	hoch	ja	N
3	bewölkt	heiß	hoch	nein	P
4	regnerisch	mild	hoch	nein	P
5	regnerisch	kalt	normal	nein	P
6	regnerisch	kalt	normal	ja	N
7	bewölkt	kalt	normal	ja	P
8	sonnig	mild	hoch	nein	N
9	sonnig	kalt	normal	nein	P
10	regnerisch	mild	normal	nein	P
11	sonnig	mild	normal	ja	P
12	bewölkt	mild	hoch	ja	P
13	bewölkt	heiß	normal	nein	P
14	regnerisch	mild	hoch	ja	N

[5:53]

Abb. 10. Beispiel einer rein kategorischen Datentabelle ***D***. Die Zielvariable *Klasse* mit den Werten {N,P} ist ebenfalls kategorisch, z.B. Klasse=P Sportliche Aktivität

Gemischte Variablenklassen

Outlook	Temperature	Humidity	Windy	Play-time
Sunny	85	85	False	5
Sunny	80	90	True	0
Overcast	83	86	False	55
Rainy	70	96	False	40
Rainy	68	80	False	65
Rainy	65	70	True	45
Overcast	64	65	True	60
Sunny	72	95	False	0
Sunny	69	70	False	70
Rainy	75	80	False	45
Sunny	75	70	True	50
Overcast	72	90	True	55
Overcast	81	75	False	75
Rainy	71	91	True	10

[7:47]

Abb. 11. Einige Datenvariablen wurden mit numerischen/metrischen Werten ersetzt (Klasse → Play-time)

Kann aus der vorherigen Datentabelle mit numerischen Variablen noch ein Zusammenhang aus $\boldsymbol{X}$ *zu* Y *hergestellt werden?*
Reicht die Anzahl der Experimente im Vergleich zu der rein kategorischen Datentabelle?
Wo liegen die Probleme?

Weiteres Beispiel

Age	Spectacle Prescription	Astigmatism	Tear Production Rate	Recommended Lenses
Young	Myope	No	Reduced	None
Young	Myope	No	Normal	Soft
Young	Myope	Yes	Reduced	None
Young	Myope	Yes	Normal	Hard
Young	Hypermetrope	No	Reduced	None
Young	Hypermetrope	No	Normal	Soft
Young	Hypermetrope	Yes	Reduced	None
Young	Hypermetrope	Yes	Normal	Hard
Prepresbyopic	Myope	No	Reduced	None
Prepresbyopic	Myope	No	Normal	Soft
Prepresbyopic	Myope	Yes	Reduced	None
Prepresbyopic	Myope	Yes	Normal	Hard
Prepresbyopic	Hypermetrope	No	Reduced	None
Prepresbyopic	Hypermetrope	No	Normal	Soft
Prepresbyopic	Hypermetrope	Yes	Reduced	None
Prepresbyopic	Hypermetrope	Yes	Normal	None
Presbyopic	Myope	No	Reduced	None
Presbyopic	Myope	No	Normal	None
Presbyopic	Myope	Yes	Reduced	None
Presbyopic	Myope	Yes	Normal	Hard
Presbyopic	Hypermetrope	No	Reduced	None
Presbyopic	Hypermetrope	No	Normal	Soft
Presbyopic	Hypermetrope	Yes	Reduced	None
Presbyopic	Hypermetrope	Yes	Normal	None

[7:7]

5.6. Kodierung

Kategorische Variablen (sowohl Attribute als auch Zielvariablen) können von einer Vielzahl von numerisch basierten ML Verfahren nicht verarbeitet werden (wie neuronale Netze)

- Eine Lösung kann die Abbildung von kategorischen Werten (also Mengen von Symbolen) auf numerische Werte → **Kodierung**
- Kodierte Werte sind aber i.A. weder intervall- noch verhältnisskalierbar!

Kodierungsformate

- *Linear* und nicht akkummulativ (skalar), d.h. { , , ,..} → { ,2 ,3 ,..}
- *Exponentiell* (z.B. zur Basis B=2) und akkummulativ (skalar), d.h. { , , ,..} → {2^0,2^1,2^2,..}
- *One-hot* und evtl. akkummulativ (vektoriell), d.h. { , , ,..} → {[1,0,0,...],[0,1,0,..],[0,0,1,..],..}

Expontentielle Kodierungen können multiple verschiedene kategorische Werte in einem numerischen Wert darstellen! Z.B. mehrfache kategorische Antworten bei einer Frage einer Umfrage.

Beispiele

```
1: {sonnig,bewölkt,regnerisch} → {3,2,1}
2: {ja,nein} → {1,0}
3: {Schaden A, Schaden B, Schaden C} → { 1,2,3 }
4: {rot,grün,blau,braun,weiß} → {1,2,4,8,16}
5: {Sport, Kino, Theater, Musik} → {1,2,4,8}
6: {heiß,kalt} → {[1,0],[0,1]}
```

➤ Numerische/metrische Werte können auf kategorische durch Intervallkodierung reduziert werden:

$$cat(x) : x \rightarrow \{\alpha_1, \alpha_2, .., \alpha_n\}, x \in \mathbb{R}/\mathbb{N}$$
$$\alpha_i \leftrightarrow x = [x_0 + i\delta, x_0 + (i+1)\delta]$$

➤ Z.B. die Kategorisierung von Schadenspositionen in einer mechanischen Struktur durch räumliche Bereiche (Segmente) {S1,S2,S3,..,S9}

➤ Z.B. Temperaturen durch "gefühlte" Attribute {heiß, warm, moderat, kalt, eiskalt}

➤ Z.B. Zeitangaben durch Epochen {Steinzeit, Bronzezeit, Kohlezeit, .. }

Dekodierung

Die Kodierung ist umkehrbar mit einer Dekodierungsfunktion (unter Kenntnis der Kodierungsvorschrift)

➤ Verwendung der `code=Math.code(val,codes)` und `val=Math.decode(code,codes)` Funktionen

5.7. Entropie und Informationsgehalt

➤ Sensorvariablen können unterschiedlichen *Informationsgehalt* besitzen

❑ Nur auf den Dateninhalt (Werte) der Variable X_i bezogen (inherenter Informationsgehalt)

- ❑ Oder zusätzlich bezogen auf die Zielvariable Y (abhängiger Informationsgehalt)

➤ Der *Informationsgehalt* einer Menge X aus Elementen der Menge C wird durch die *Entropie* $E(X)$ gegeben:

$$E(X) = -\sum_{i=1,k} p_i \log_2(p_i), p_i = \frac{count(c_i, X)}{N}, X = \{c|c \in C\}$$

➤ Dabie ist k die Anzahl der unterscheidbaren Elemente/Klassen $Val(X)$ C in der Datenmenge X (z.B. die Spalte einer Tabelle) und p_i die Häufigkeit des Auftretens eines Elements c_i C in X.

➤ Beispiele:

1: X_1={A, C, B, C, B, C} → Val(X_1)=C={A, B, C},N=6
2:
E(X_1)=-(1/6)log(1/6)-(2/6)log(2/6)-(3/6)log(3/6)=1.46
3: X_2={A, B, C} → Val(X_2)=C={A, B, C},N=3
4:
E(X_2)=-(1/3)log(1/3)-(1/3)log(1/3)-(1/3)log(1/3)=1.58
5:
X_3={A, A, A, A, B, B} → Val(X_3)={A,B} C={A, B, C},N=6
6:
E(X_3)=-(4/6)log(4/6)-(2/6)log(2/6)-(0/6)log(0/6)=0.92
7: X_4={A, A, A, A, B, B} → Val(X_4)=C={A, B},N=6
8: E(X_4)=-(4/6)log(4/6)-(2/6)log(2/6)=0.92

➤ Die Entropie ist Null wenn die Datenmenge X "rein" ist, d.h., nur Elemente einer einzigen Attributklasse c_1 C enthält, z.B. X={A,A,A}.

➤ Die Entropie ist $log_2(|C|)$ wenn alle Werte gleichverteilt vorkommen, wenn nicht dann kleiner (nicht gleichverteilt).

➤ Die Entropie reicht allein zur Bewertung des Informationsgehaltes nicht aus:

X1	**X2**	**Y**
A	C	P
B	C	P
A	D	N
B	D	N

➤ E(X1)=1, E(X2)=1 !! Welche Variable X ist für die Entscheidung der Zielvariable Y geeignet?

5.8. Informationsgewinn (Gain)

➤ Ansatz: Die Datenmenge Y wird nach den möglichen Werten von X partitioniert, also je eine Partition pro $c_i \quad Val(X)$.

$$G(Y|X) = E(Y) - \sum_{v \in Val(X)} \frac{|Y_v|}{|Y|} E(Y_v)$$

➤ Die Menge Y_v enthält nur Werte für die $X{=}v$ ist!

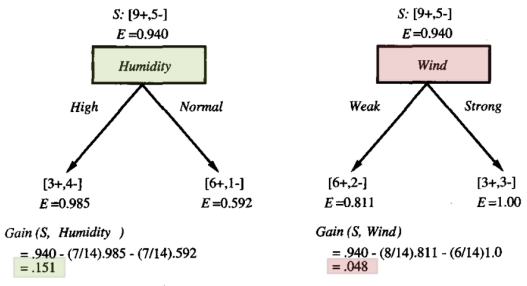
➤ Ein **Verteilungsvektor** ist dann $Dist(X){=}[|v_1|,|v_2|,..]$ und bedeutet wie häufig der bestimmte Wert $v_i \quad Val(X)$ in X auftaucht!

➤ Ein **Verteilungsvektor** ist dann $Dist(Y_v|X){=}[|u_1|,|u_2|,..]$ und bedeutet wie häufig der bestimmte Wert $u \quad Val(Y)$ in Y_v auftaucht!

Beispiele

Beispiel	Aussicht	Temperatur	Luftfeuchtigk.	Windig?	Klasse
1	sonnig	heiß	hoch	nein	N
2	sonnig	heiß	hoch	ja	N
3	bewölkt	heiß	hoch	nein	P
4	regnerisch	mild	hoch	nein	P
5	regnerisch	kalt	normal	nein	P
6	regnerisch	kalt	normal	ja	N
7	bewölkt	kalt	normal	ja	P
8	sonnig	mild	hoch	nein	N
9	sonnig	kalt	normal	nein	P
10	regnerisch	mild	normal	nein	P
11	sonnig	mild	normal	ja	P
12	bewölkt	mild	hoch	ja	P
13	bewölkt	heiß	normal	nein	P
14	regnerisch	mild	hoch	ja	N

[12]



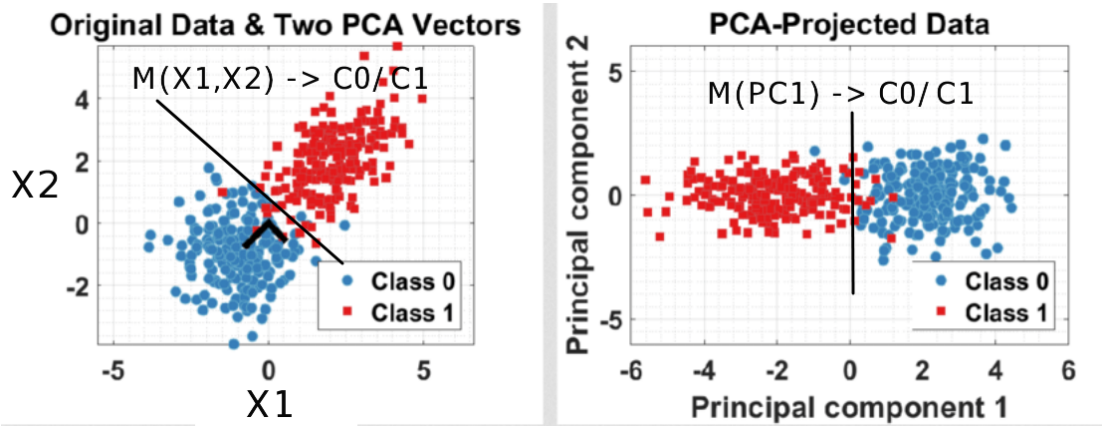
5.9. Principle Component Analysis

➤ PCA: Klassische Methode zur unüberwachten linearen Dimensionsreduktion → Analyse der Hauptkomponenten

- Reduktion von Redundanz in den Attributen $\boldsymbol{X}$
- Bessere Trennung bei der Inferenz von kategorischen (und ggfs. auch numerischen) Zielvariablen
- Weitere Verfahren:
  - Lineare Diskriminanzanalyse (LDA)
  - Singuläre Wertzerlegung (SVD)

Beispiel

- $\boldsymbol{D}$=$[X_1, X_2, Y]$, mit $Val(Y)$={Class1,Class2}
- "Rotation" des zweidimensionalen Attributraums führt zu einer reduzierten Datentabelle $\boldsymbol{D'}$=$[PC_1, Y]$ (PC_2 kann weg gelassen werden)



[Czarnek, RG]

5.10. Merkmalsselektion

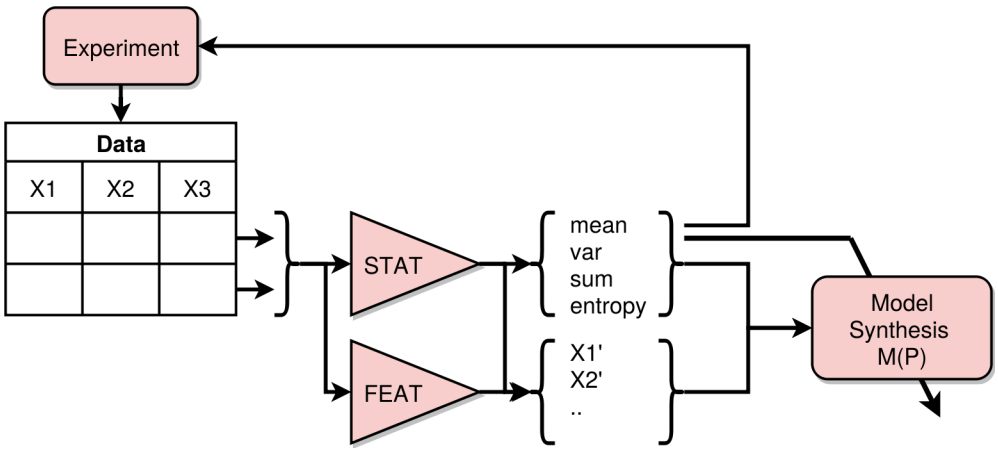


Abb. 12. Die statistische und weitere Analysen können die Eingabe für ML liefern, aber auch die Modellsynthese parametrisieren bzw. beeinflussen

5.11. Zusammenfassung

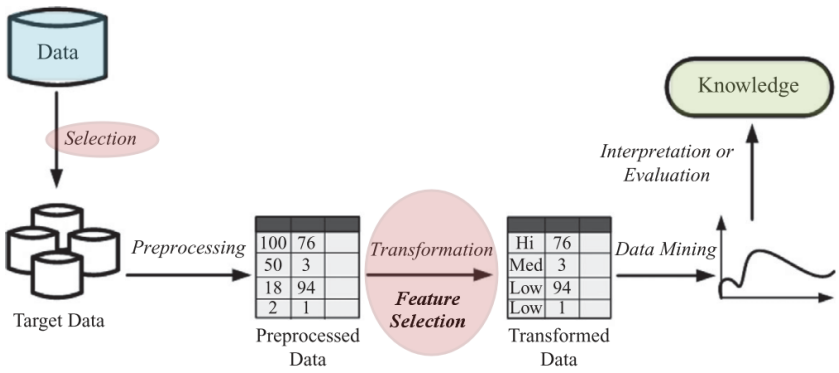
- Die statistische Analyse von Datentabellen liefert wichtige Informationen über die Qualität der Daten
- Die Merkmalsselektion transformiert die Rohdaten auf neue möglichst linear unabhängige Attribute
  - Datenreduktion → Dimensionalität
  - Datenreduktion → Datengröße
  - Datenqualitätserhöhung
- Es werden Verfahren für kategorische und numerische Datenvariablen unterschieden

6. Taxonomie des Maschinellen Lernens

***Zielvariablen**: Kategorische Klassifikation, Numerische Prädiktorfunktionen, Gruppierung*
***Modellfunktionen**: Mit welchen Daten- und Programmarchitekturen können Eingabevariablen auf Zielvariablen abgebildet werden?*
***Training und Algorithmen**: Wie können die Modellfunktionen an das Problem angepasst werden?*
Überwachtes, nicht überwachtes und Agentenlernen

6.1. Datenverarbeitung

➤ Die Daten die als Grundlage für die Induktion (Lernen) und die Deduktion (Applikation/Inferenz der Zielvariablen) müssen i.A. vorverabeitet werden → **Merkmalsselektion**



[6]

Abb. 13. Maschinelles Lernen ist ein Werkzeug der Datenanalyse und des Data Minings

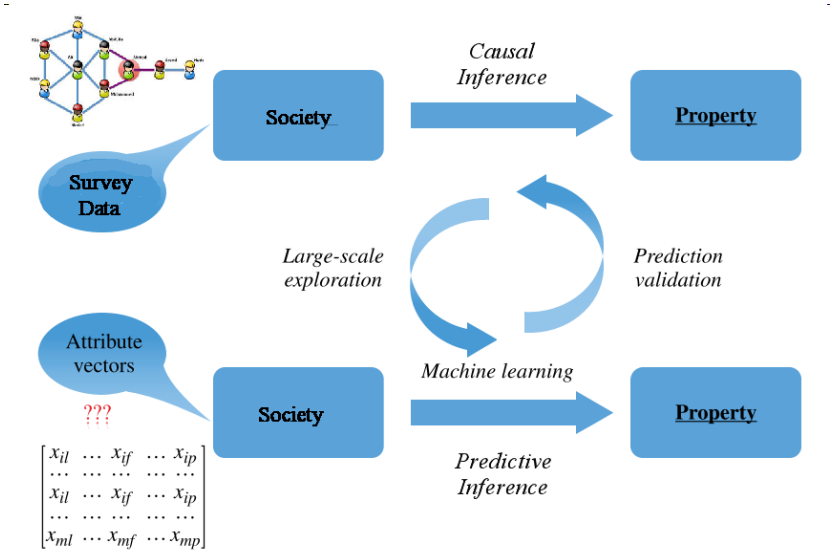
6.2. Die Modellfunktion

➤ Die Modellfunktion soll möglichst genau und effizient die Eingabedaten $\boldsymbol{X}$ auf die Zielvariablen $\boldsymbol{Y}$ abbilden:

$$M(\tilde{X}) : \tilde{X} \rightarrow \tilde{Y},$$
$$X = \begin{cases} \text{diskrete kategorische Werte} \\ \text{numerische Werte } \mathbb{N},\mathbb{R} \end{cases},$$
$$Y = \begin{cases} \text{diskrete kategorische Werte} \\ \text{numerische Werte } \mathbb{N},\mathbb{R} \\ \text{Gruppen(X), Netzwerke} \end{cases}$$

➤ Die Modellfunktion M **approximiert** eine i.A. nicht bekannte Funktion F, d.h. eine axiomatisch oder analytisch abgeleitete Modellfunktion (z.B. phys. Gesetze) $\rightarrow$ M ist **Hypothese** von F!

Beispiel



[100]

Abb. 14. Kausale vs. Prädiktive Modellbildung und Soziale Netzwerkmodelle versa algorithmisch bestimmte Modelle (Hypothesen)

6.3. Lernen

Lernen bedeutet die gewünschte Modellfunktion M möglichst genau zu approximieren so dass min error(|Y_0-Y|) für alle (X,Y_0) Paare gilt (Y_0: Referenzdaten).

- I.a. ist M eine parametrisierbare Funktion $f(\boldsymbol{P})$ oder eine parametrisierbare Datenstruktur
  - Der Parametersatz $\boldsymbol{P}$=$\{p_1,p_2,..,p_i\}$ bestimmt sowohl Funktion als auch Struktur (z.B. eines Entscheidungsbaumes)
- Es gibt nicht eine Modellfunktion $\boldsymbol{M}$, sondern eine große Menge möglicher Funktionen, genannt **Hypothesen** .

Lernen bedeutet also die bestmögliche Anpassung des Parametersatzes $\boldsymbol{P}$ *um den Fehler zu minimieren und eine geeignet Hypothesenfunktion zu finden.*

- Man unterscheidet bekannte Referenzwerte der Zielvariablen (und Beziehung zu X) Y_0, auch **Labels** genannt, und prognostische Werte Y die als Ergebnis von $M(X)$ geliefert werden (Inferenzwerte), d.h. bei der Applikation ist der wahre Wert Y_t unbekannt (**Schätzung** von Y_t)

$$H(\tilde{X}) : \tilde{X} \rightarrow \tilde{Y},$$
$$= \{M_1^{P1}, M_2^{P2}, .., M_k^{Pk}\},$$
$$error(X, Y_0, M) = |M(X) - Y_0|$$

Beispiele

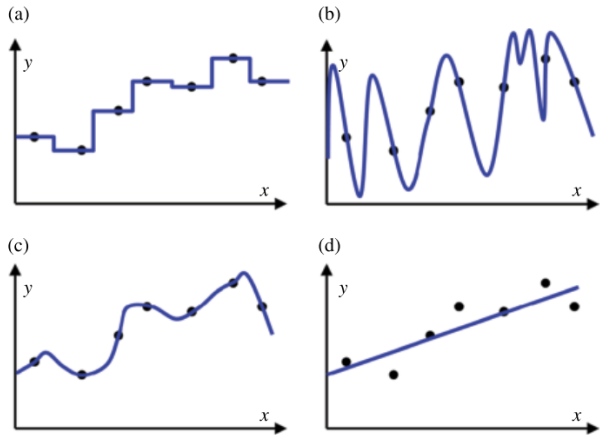
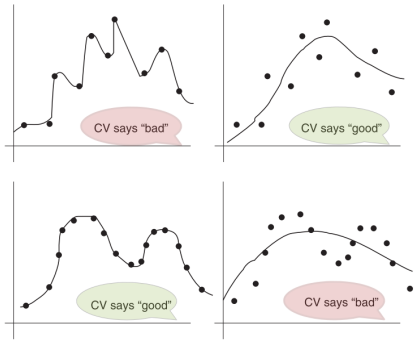


Abb. 15. Verschiedene Modellfunktionen M die die (Trainings) Daten repräsentieren

6.4. Kreuzvalidierung

- Beim Training wird ein Inferenzfehler zunächst aus Trainingsdaten bestimmt → Trugschluß!
- Stattdessen müssen auch unabhängigige Testdaten für eine Kreuzvalidierung herangezogen werden, und dann …



[13]

6.5. Fehler (Verlust)

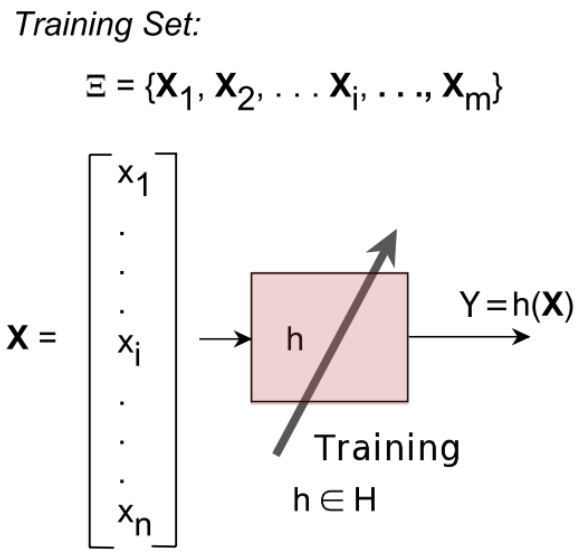
Jede Hypothesenfunktion M führt zu einem Informationsverlust durch Appoximation der tatsächlichen und unbekannten Modellfunktione F.

- Es gilt also:

$$M(x) : x \rightarrow y = F(x) + E(x)$$

mit E als eine Fehlerfunktion (i.A. zufälliger Fehler) und $\hat{E}$ als mittlerer Prädiktionsfehler.

- Die Hypothesenmenge ist also tatsächlich eine Approximation eines unbekannten "exakten" Modells (Modellfunktion) M_{F}, die z.B. mittels physikalischer oder soziologischer Modelle ableitbar wäre.
- Genauso wie eine Sensor eine physikalische Größe nur approximieren kann, der tatsächliche Wert der zu messenden Größe ist nicht bekannt



[11]

Abb. 16. Training als Anpassung von Hypothesen für die Abbildungfunktion $X \rightarrow Y$ mit Trainingsdaten

6.6. Parametrisierung

Die Paramater in dem Parametersatz $\boldsymbol{P}$ bestehen aus zwei Klassen:

Statische Parameter $\boldsymbol{P}_{\mathrm{s}}$

Parameter die die Modellimplementierung (Funktion, Datenstruktur, usw.) festlegen und i.A. während des Trainings und der Applikation unverändert bleiben. (Ausnahme: Evolutionäre Algorithmen) → **Konfiguration**

Dynamische Parameter $\boldsymbol{P}_{\mathrm{d}}$

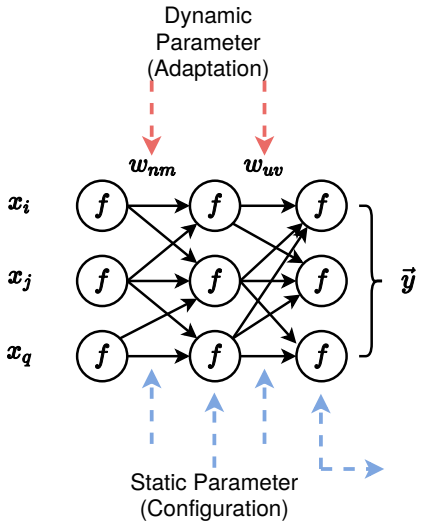
Parameter die während des Trainings verändert (angepasst) werden. Z.B. Funktionsparameter oder Kantengewichte von neuronalen Netzen → **Adaption**

Beispiele

1. Numerische Prädiktorfunktionen (T: Temperatur, S: Satisfaction) → Regression

$$f(T) : T \rightarrow S = a + bT + cT^2 + dT^3,$$
$$P_s = \{degr : 3\}, P_d = \{a, b, c, d\}, S = [0, 1]$$
$$f(T) : T \rightarrow S = a + bT + cT^c + dT^e,$$
$$P_s = \{terms : 4, lin : 2, exp : 2\}, P_d = \{a, b, c, d, e\}, S = [0, 1]$$

2. Künstliches Neuronales Netzwerk



6.7. Daten

Trainingsdaten $\boldsymbol{D}_{\text{train}}$

Datentabellen die aus Zeilen mit einer bekannten Beziehung (X, Y) bestehen und verwendet werden die Modellfunktion M durch Veränderung von $\boldsymbol{P}$ zu approximieren

Testdaten $\boldsymbol{D}_{\text{test}}$

Datentabellen die aus Zeilen mit einer bekannten Beziehung (X, Y) bestehen und verwendet werden die Modellfunktion M auf Genauigkeit und Fehler zu testen. Man spricht auch von einer Kreuzvalidierung da $\boldsymbol{D}_{\text{test}}$ $\boldsymbol{D}_{\text{train}}$ = sein sollte.

Inferenzdaten $\boldsymbol{D}_{\text{inf}}$

Datentabellen die nur aus Zeilen $\boldsymbol{X}$ bestehen (Y ist unbekannt)

Es gilt: $\boldsymbol{D}_{\text{train}}$ $\boldsymbol{D}_{\text{all}}$, $\boldsymbol{D}_{\text{test}}$ $\boldsymbol{D}_{\text{all}}$, $\boldsymbol{D}_{\text{inf}}$ $\boldsymbol{D}_{\text{all}}$ aber $\boldsymbol{D}_{\text{train}}$ $\boldsymbol{D}_{\text{test}}$ = und $\boldsymbol{D}_{\text{train}}$ $\boldsymbol{D}_{\text{test}}$ $\boldsymbol{D}_{\text{inf}}$ (Idealfall!)

Die großen Probleme beim algorthmischen/trainierten Modellieren:

- ➤ Die Trainingsdaten sind nicht repräsentativ (Umfang, Varianz, Qualität)
- ➤ Die Testdaten sind nicht repräsentativ (Umfang, Varianz, Qualität)
- ➤ Die Trainingsdaten enthalten schwache Variablen die nicht entfernt wurden (Inkonsistenz und geringer Informationsgewinn)

***Generalisierung**. Das gelernte Modell M bildet alle drei Datenmengen gleichermaßen gut ab!*

- ➤ Ergänzung:

Bewertungsdaten

Beim Einsatz eines gelernten Modells kann eine Evaluierung bezüglich Qualität / Genauigkeit stattfinden. Diese Daten können dann ggfs. für eine Adaption des Modells und dessen Parametersatz $\boldsymbol{P}$ verwendet werden.

D.h. bei der Anwendung des Modells können somit auch neue Trainingsdaten gewonnen werden, z.B. im Rahmen eines Produktlebenszyklusmanagements!

6.8. Lernverfahren

Überwachtes Lernen

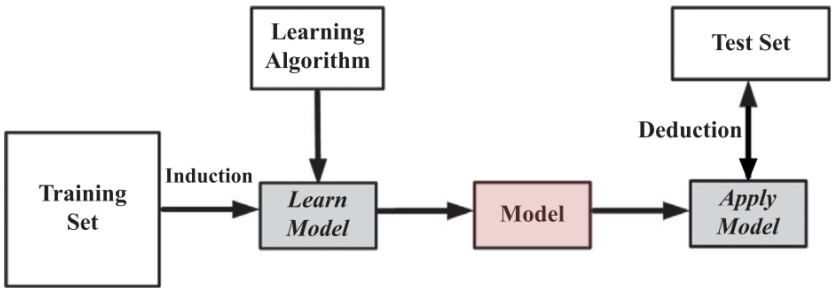
Es gibt Trainingsdaten mit bekannten Beziehungen (X, Y) die verwendet werden um die Modellfunktion mit minimalen Fehler anzupassen. Überwachugn benötigt i.A. einen Experten der die Beziehungen (X, Y) erstellt und analytisch den Fehler bewertet.

Unüberwachtes Lernen

Es gibt Trainingsdaten ohne bekannte beziehung (X, Y), d.h., schon das Lernen führt zu einer automatischen Inferenz der zielvariablen Y, die aber in diesem Fall i.A. nur durch Gruppenmengen bestehen. Eine Gruppenmenge Q=$\{X_i\}$ bringt verschiedene Eingabewerte in Beziehung. D.h. Y .

Belohnungs- und Agentenlernen

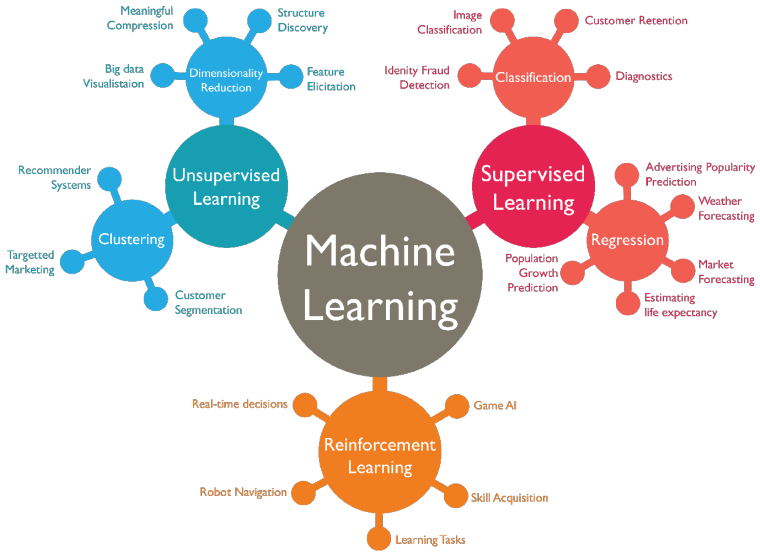
Die Abbildungsfunktion $f(X)$: $X \rightarrow Y$ wird schrittweise durch eine Evaluierung des inferrierten Y mit einem Belohnungswert r=[0,1] gelernt. Training und Inferenz findet gleichzeitig statt.



[6]

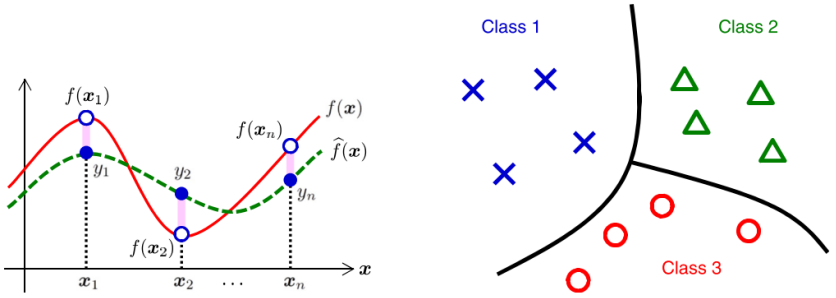
Abb. 17. Ablauf Überwachtes Lernen mit Trainings- (Induktion) und Applikationsphasen (Deduktion)

6.9. Taxonomie der Verfahren



[Abdul Rahid,www.wordstream.com]

6.10. Überwachte Lernverfahren - Unterklassen



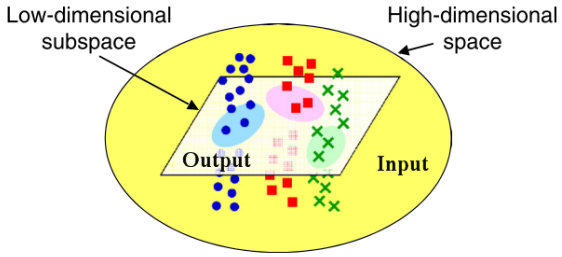
[4]

Abb. 18. Zwei wichtige Unterklassen von überwachten Lernen: Regression (Numerische Zielvariablen) und Klassifikation (Kategorische Zielvariablen)

6.11. Dimensionalitätsreduktion

➤ ML kann auch für die Reduktion von Datendimensionalität eingesetzt werden (Informationen sind reduzierte Daten)

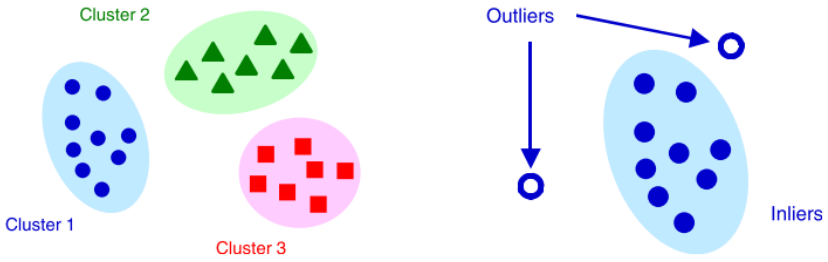
❑ Beispiele: Principle Component Analysis, Single Value Decomposition, ..



[4]

Abb. 19. Abbildung von hochdimensionale Daten X^{n} auf niederdimensionale X^{m} mit $m < n$

6.12. Unüberwachtes Lernen - Unterklassen



(a) Clustering **(b) Ausreißer (Anomalien)**

[4]

Abb. 20. Zwei wichtige Unterklassen von nicht überwachten Lernen: Clustering (Gruppenbildung) und Ausreißerdetektion

6.13. Training

- Das Training einer Modellfunktion M kann
  - **monolitisch** (alle Dateninstanzen werden "parallel" verwendet), oder
  - **stapelbasiert** (d.h. Gruppen von Instanzen werden "parallel" verarbeitet), oder
  - **iterativ** (Dateninstanzen werden "sequenziell" verwendet), und
  - **inkrementell** (iterativ mit neuen Daten).
- Inkrementelle Trainings- und Anpassungsverfahren könnne alte Datensätze verwerfen → Stromldatenlernen!
- Nicht jede Modellimplementierung ist geeignet:
  - Graphen (Bäume) können i.A. nur monolotisch trainiert = erzeugt werden!
  - Regression von math. Funktionen kann monolitisch und/oder iterativ erfolgen;
  - Neuronale Netze können monolitisch, stapelbasiert, iterativ, und vor allem inkrementell trainiert werden.

6.14. Modellimplementierungen

Es gibt im wesentlichen vier verschiedene Architekturen die Modelle M zu implementieren:

Funktionen

Die Struktur einer mathematischen Funktion wird durch ihre Terme gebildet (Berechnungsknoten), z.B. $ax+bx^2$. Zu jedem Term gehört ein dynamischer Parameter der beim Training angepasst wird um den Fehler zu minimieren. Das Ergebnis ist die Zielvariable y.

Gerichtete Graphen

Gerichtete Graphen (oder Entscheidungsbäume) bestehen aus Knoten und Kanten. Die Knoten repräsentieren eine Eingabevariable (Attribute) x $\boldsymbol{X}$. Die Kanten beschreiben die Entwicklung eine Graphens beginnend vom Wurzelknoten hin zu den Blättern. Die Blätter enthalten die Werte der Zielvariable(n) y. Der dynamische Paremetersatz ist der Graph (dessen Struktur).

Funktionale Graphen

Hybrid aus gerichteten Graph und Funktion → Künstliche Neuronale Netze. Die Knoten repräsentieren Berechnungsfunktionen, die Kanten verbinden Ausgänge von Funktionen mit Eingängen. Es gibt Eingangsknoten die mit den Eingabevariablen $\boldsymbol{X}$ verbunden sind, und Ausgangsknoten die mit den Ausgangsvariablen $\boldsymbol{Y}$ verbunden sind.

Ungerichtete Graphen

Hier repräsentieren die Knoten Dateninstanzen X, und die Kanten verbinden die nächsten Nachbarn miteinander. Hier geht es um Gruppenbildung (k nächste Nachbarn/kNN Problem).

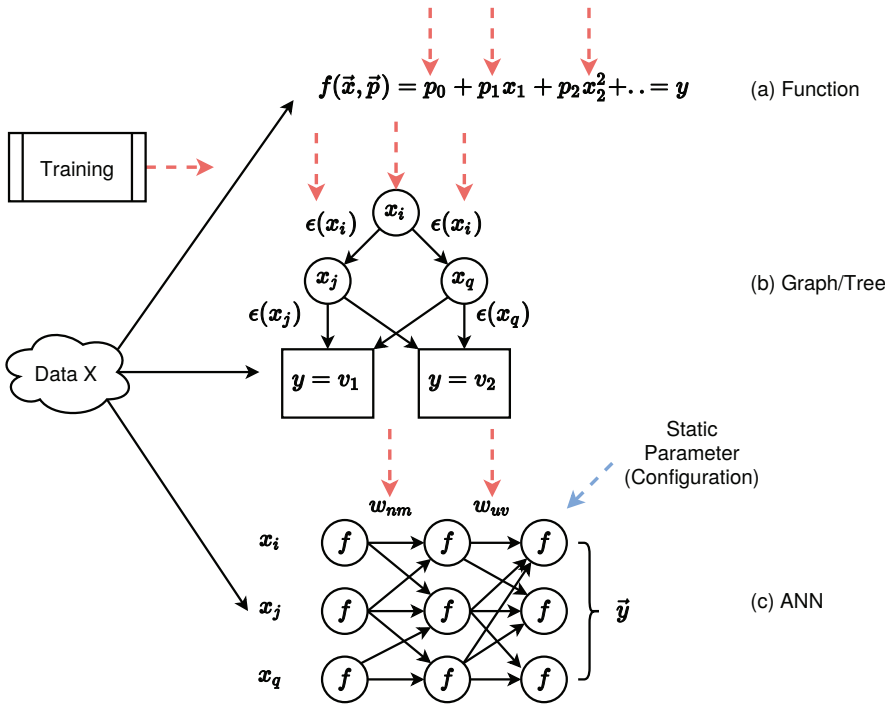


Abb. 21. Verschiedene Modellimplementierungen

6.15. Ablauf und Phasen von ML

0. Statistische Analyse und Bewertung der Daten
1. Merkmalsselektion
2. Aufteilung der Daten in Trainings- und Testdaten (i.A. randomisiert) $\boldsymbol{D}$=$\boldsymbol{D}_{\text{train}}$ $\boldsymbol{D}_{\text{test}}$
3. Training einer Modellfunktion M mit bekannten (gelabelten bei ÜL) Trainingsdaten $\boldsymbol{D}_{\text{train}}$ unter Bewertung des Modellfehlers $E(X)$
4. Test und Bewertung von M mit bekannten Daten $\boldsymbol{D}_{\text{test}}$
5. Applikation (Inferenz) von M auf unbekannten Daten $\boldsymbol{D}$

6.16. ML in der Soziologie

➤ Qualitative und quantitative Sozialwissenschaften wollen aus Daten **erklärbare** Modelle ableiten

- ❏ Die Inferenz von Aussagen mit neuen Daten ist von geringer Bedeutung
- ❏ Das Modell ist das Ziel

➤ Datenwissenschaften wollen aus Daten (ggfs. Black-Box) Modelle ableiten

- ❏ Die Inferenz von Aussagen mit neuen Daten ist Ziel!
- ❏ Das Modell selber ist nur das Werkzeug

6.17. Qualitative Kodierung

Qualitative Kodierung ist eine der wichtigsten Techniken, die in der qualitativen Analyse in den Sozialwissenschaften verwendet werden.
Im Allgemeinen bezieht sich die Kodierung auf den Prozess der Zuweisung beschreibender oder inferentieller Annotierungen zu Datenblöcken, die die Entwicklung von Konzepten oder Theorien unterstützen können.
Kodierung ist in der Regel eine sehr arbeitsintensive und zeitaufwendige Aufgabe.

Einsatz von ML

➤ ML Verfahren können zur Automatisierung der Q. Kodierung eingesetzt werden [101]

ML in der Soziologie findet sich vor allem in den ersten Stufen der "Wertschöpfungskette" → ***Werkzeuge der Datenverarbeitung und Merkmalsselektion***

6.18. Soziale Analysen aus Texten

➤ Rückschlüsse auf soziales Verhalten und Netzwerkbildung können u.A. aus textuellen Quellen gewonnen werden:

- ❏ Soziale Medien (Twitter, Facebook, Blogs, …)
- ❏ Nachrichten
- ❏ Wissensdatenbanken

- ➤ Häufig ist Mustererkennung und Klassifikation zentrale Merkmalsselektion (mit Natural Language Processing NLP)

Einsatz von ML

- ➤ Textklassifikation und Vorhersage

6.19. Soziologische Modellinferenz

- ➤ Neben der kausalen Modellinferenz können auch prädiktive Modellinferenzverfahren - also ML - eingesetzt werden
- ➤ Spannende Frage: Wie ist die Korrelation von kausal und prädiktiv gewonnenen Modellen?
- ➤ Was bedeutet eine Abweichung?

Kernfrage ist die Erklärbarkeit von algorithmisch erzeugten Modellen mit ML Verfahren

6.20. Big Data Analysen

- ➤ Big Data bedeutet nicht groß (wenn auch meistens), sondern die Eingabevariablen sind scheinbar schwach korreliert, gekennzeichnte durch hohes Rauschen und Verzerrung!
- ➤ Aber mit ML kann auch solch schwachen Daten Informationen abgeleitet werden:
  - ❏ Genaue Wahlvorhersage
  - ❏ Demografische Vorhersagen
- ➤ Kritik: Die Datenvoreverarbeitung und ML Datenkette kann (ungewollt) zu Verzerrung und Offset führen.

Daher: Die "Fehler" in der ML Verarbeitungskette bezüglich sozialer Eigenschaften können nicht technisch gelöst und korrigiert werden. Dazu müssen wiederum Modelle der Soziologie verwendet werden. Der "Theorie Rein - Theorie Raus" Ansatz [102]!!

- Die Sozialtheorie hilft bei der Lösung von Problemen, die während des gesamten Aufbaus und der Bewertung von Modellen für maschinelles Lernen für soziale Daten auftreten.

6.21. Zusammenfassung Unterschiede Soziologische Verfahren vs ML

- Soziologische Theorie ist oft hypothesengetrieben, während maschinelles Lernen Daten sind!
- Beim maschinellen Lernen beginnt man mit einem Datensatz, um eine Hypothese aufzustellen, während man in der Soziologie oft mit einer Hypothese beginnt.
- Beide verwenden (oder eher ML, beide sollten zumindest) eine Auswertung außerhalb der Stichprobe, um Ihre Hypothesen zu testen.
- Beim maschinellen Lernen liegt der Fokus im Allgemeinen auf der Vorhersage, in der Soziologie nicht auf der Vorhersage, ohne zu erklären, warum ein Phänomen Auftritt.
- Beim maschinellen Lernen glaubt man nicht, dass das Modell richtig ist, dh. es wird nicht angenommen, dass das Modell der datengenerierende Mechanismus ist.
- Modelle werden nur danach ausgewertet, wie gut Sie anhand von Daten Vorhersagen machen, aus denen Sie selber nicht erstellt wurden, und nicht erklären wie sie zu Stande kommen.
- In der Soziologie betrachtet man allgemein, ob ein Koeffizient eines linearen Modells von null unterscheidbar ist; dies macht starke Annahmen über den datengenerierenden Mechanismus, den maschinelle Lerner nicht für gültig halten würden.
- Der Fokus des maschinellen Lernens lag traditionell nicht auf kausalen Effekten, obwohl Maschinelles lernen bei kausalen inferenzproblemen nützlich sein kann.

6.22. Zusammenfassung

Maschinelles Lernen besteht aus:

1. Modellimplementierungen:
   - Funktionen, Gerichtete Graphen, Funktionalen Graphen, Ungerichtete Graphen, also mit/für

- ➤ Regression, Entscheidungsbäume, Neuronale Netze, Clustering (kNN)

2. Aufgaben

   - ➤ Regression, Klassifikation, Gruppierung (Clustering), Prognostik

3. Methoden und Verfahren

   - ➤ Überwachtes, nicht überwachtes, und rückgekoppeltes Belohnungslernen
   - ➤ Monolitisches, stapelbasiertes, iteratives, und inkrementelles Lernen
   - ➤ Einzel- versa Multiinstanzlernen
   - ➤ Enscheidungsbaumlernen (Konstruktion), Support Vector Machines (Regression), Backpropagation in Neuronalen Netze, usw.

4. ML besteht aus mehreren Phasen:

   - ➤ Datenerhebung, Datenvorverarbeitung, Statistische Bewertung, Merkmalsselektion, Modellertsellung, Training, Test und Analyse (Kreuzvalidierung), Anwendung/Inferenz

5. Daten werden unterteilt in:

   - ➤ Trainingsdaten , Testdaten, Anwendungsdaten
   - ➤ Trainings- und Testdaten bei ÜL mit (x,y) Beziehungen (Labelling)

7. Klassifikation mit Entscheidungsbäumen

***Zielvariablen**: Kategorische Variablen*
***Eigenschaftsvariablen**: Kategorische und Numerische Variablen*
***Modell**: Gerichteter azyklischer Graph (Baumstruktur)*
***Training und Algorithmen**: C4.5, ID3, INN*
***Klasse**: Überwachtes Lernen*

7.1. Entscheidungsbäume

- ➤ Ein Entscheidungsbaum ist ein gerichteter azyklischer Graph bestehend aus einer Menge von Knoten $\boldsymbol{N}$ die mit den Eingabevariablen x verknüpft

sind und Kanten $\boldsymbol{E}$ die die Knoten verbinden

- Die Endknoten sind Blätter und enthalten Werte der Zielvariablen y (daher kann y nur eine kategorische Variable sein, oder eine intervallkategorisierte)
- Die Kanten bestimmen die Evaluierung des Entscheidungsbaum beginnend von dem Wurzelknoten bis zu einem Blattknoten
  - Jede Kante hat eine Evaluierungsbedingung (x) der Variable des ausgehenden Knotens x
- Zusammengefasst ausgedrückt:

$$
\begin{aligned}
&M(X) : X \rightarrow Y, X = \{x_i\}, Y = \{y_j\} \\
&DT = \langle N_x, N_y, E \rangle \\
&N_x = \{n_i : n_i \leftrightarrow x_j\}, N_y = \{n_i : n_i \leftrightarrow val(y_j)\} \\
&E = \{e_{ij} : n_i \mapsto n_j | \epsilon_{ij}\}
\end{aligned}
$$

- Entscheidungsbäume können neben dem Graphen auch funktional dargestellt werden:

$$
M(X) = \left\{ \begin{array}{r} x_i = v_1, \left\{ \begin{array}{c} x_j = v_1, val(y_i) \\ x_j = v_2, val(y_i) \\ x_j = v_3, \{.. \end{array} \right. \\ x_i = v_2, \left\{ \begin{array}{c} x_k = v_1, \{.. \\ x_k = v_2, \{.. \\ x_k = v_3, \{.. \end{array} \right. \\ x_i = v_3, \left\{ \begin{array}{c} x_l = v_1, \{.. \\ x_l = v_2, \{.. \\ x_l = v_3, \{.. \end{array} \right. \end{array} \right.
$$

Baumklassen

Man unterscheidet:

- **Binäre Bäume**. Jeder Knoten hat genau (oder maximal) zwei ausgehende Kanten (Verzweigungen). Der Test der Variable x kann daher nur $x < v$, $x > v$, $x \quad v$, oder $x \quad v$ sein! Wird vor allem bei numerischen Variablen eingesetzt.
- **Bereichs- und Mehrfachbäume**. Jeder Knoten hat $1..k$ ausgehende Kanten (Knotengrad k). Der Test der Variable x kann auf einen bestimmten Wert $x \quad V$ oder auf ein Intervall $[a,b]$ erfolgen! Wird vor allem bei kategorischen Variablen eingesetzt.

Baumstruktur

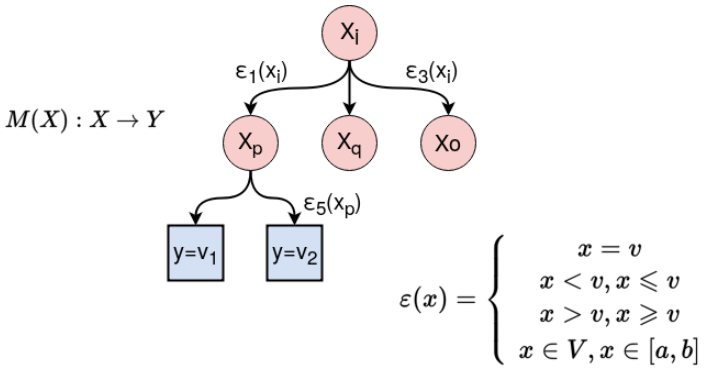


Abb. 22. Grundlegende Struktur eines Entscheidungbaumes

Vorteile

Entscheidungsbäume sind einfach aufgebaut und können mit einfachen Algorithmen erzeugt werden. Entscheidungsbäume als inferriertes Modell erlauben eine **Erklärbarkeit** des Modells, also die Antwort auf die Frage wie sich ein y aus einem x ergibt. Weiterhin ist eine Ableitung eines **inversen Problems** möglich, d.h. welche Werte x für gegebenes y sind möglich?

Nachteile

Entscheidungsbäume können schnell **spezialisieren**, d.h. es fehlt an **Generalisierung**. Theoretisch kann mit einem Entscheidungsbaum jede Trainingsdatentabelle mit einer Trefferquote von 100% abgebildet werden. Der Test mit nicht trainierten Daten ergibt aber Prädiktion in der Größenordnung der Ratewahrscheinlichkeit!

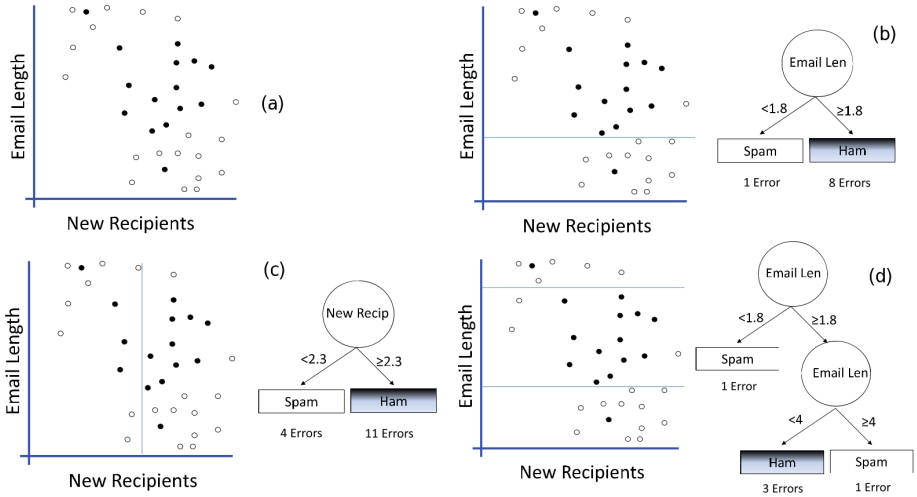
7.2. Training

➤ Das Training mit Trainingsdaten $\boldsymbol{D}_{\text{train}}$ erzeugt den Baum *schrittweise*:

- ❏ Es werden geeignete Variablen x $\boldsymbol{X}$ ausgewählt die einen *Knoten* im Baum erzeugen
- ❏ Jeder hinzugefügte Knoten erzeugt neue Teilbäume (durch Verzweigungen)
- ❏ Die *Verzweigungsbedingungen* (Kanten) werden ebenfalls vom Trainer anhand der Werte der Variable x in Abhängigkeit von der Zielvariablen y gewählt/berechnet.

➤ Die Auswahl der Variablen und die Verzweigungsbedingungen können je nach Algorithmus und Baumklasse variieren!

7.3. Beispiel



[10]

Abb. 23. Schrittweise Erzeugung des Entscheidungsbaums aus den Eingabedaten (a) erst mit einer Variable (b,c), dann mit zwei (d) unter Beachtung des Klassifikationsfehlers

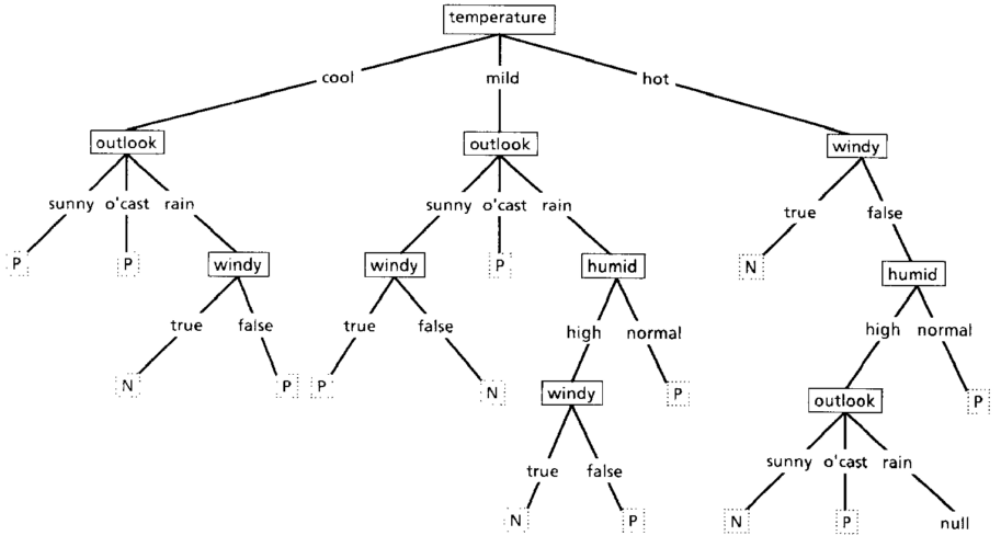
Jeder Knoten in einem binären Baum stellt eine lineare Separation des Eingabedatenraums dar.

Probleme bei Mehrbereichsbäumen

➤ Wenn die Wertemenge $val(x)$ groß ist gibt es entsprechend auch viele Verzweigungen im Baum!

- ❏ Die Größe des Baums wächst an (Speicher)
- ❏ Die Rechenzeit für das Training (Induktion) aber auch die Anwendung (Inferenz, Deduktion) wächst
- ❏ Die Entropie kann als Maß der Varianz der Wertemenge gesehen werden.

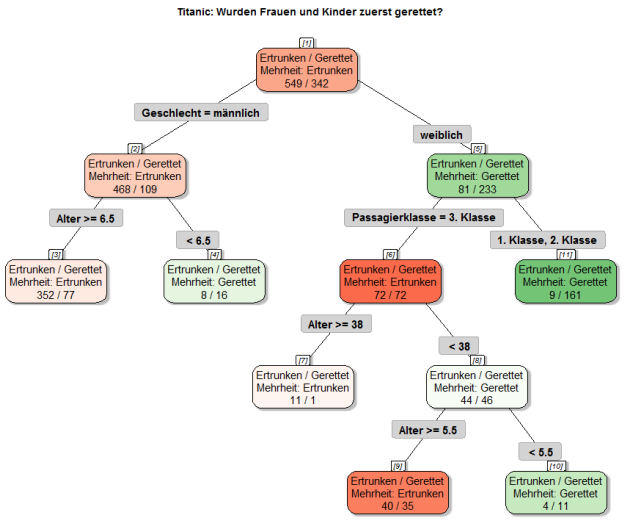
Das "NP" Problembeispiel



[14]

Abb. 24. k-stelliger Entscheidungsbaum für kategorische Variablen

Das Titanic Überlebensbeispiel



[www.statistik-dresden.de]

Abb. 25. Binärer Entscheidungsbaum (Relation und Auswahl) für numerische und kategorische Variablen: Beantowrtung

Trainingsalgorithmen

➤ Es gibt verschiedene Trainingsverfahren (für verschiedene Baumklassen):

- ❑ **ID3**. Der Klassiker (Iterative DiChaudomiser 3, Ross Quinlan, 1975-1986) für kategorische Variablen (k-stelliger Baum)
- ❑ **C4.5**. Der Klassiker (Ross Quinlan 1988-1993) für numerische (und kategorische) Variablen (Binär- und k-stelliger Baum) als Erweiterung des ID3 Verfahrens.
- ❑ **INN**. Die Eigenkreation (ICE, Stefan Bosse, 2016) für numerische Werte mit Intervallarithmetik für unsichere verrauschte Sensorwerte (also im Prinzip mit Intervallkategorisierung und Kantenbedingungen sind $x \quad [a,b]$), basierend auf C4.5 und ID3

7.4. Vergleich ID3 - C4.5

➤ Der ID3-Algorithmus wählt das beste Attribut basierend auf dem Konzept der Entropie und dem Informationsgewinn für die Entwicklung des

Baumes.

➤ Der C4.5-Algorithmus verhält sich ähnlich wie ID3, verbessert jedoch einige ID3-Verhaltensweisen:

- ❏ Möglichkeit, numerische (kont.) Daten zu verarbeiten.
- ❏ Verarbeitung unbekannter (fehlender) Werte
- ❏ Möglichkeit, Attribute mit unterschiedlichen Gewichten zu verwenden.
- ❏ Beschneiden des Baumes nach der Erstellung (**Modellkompaktierung**).
- ❏ Vorhersage der Fehler
- ❏ Hervorhebung und Extraktion von Teilbäumen

7.5. ID3 Verfahren

[1] J. R. Quinlan, "Induction of Decision Trees," in Machine Learning, Kluwer Academic Publishers, Boston, 1986.

➤ Ausgangspunkt für die Konstruktion des Entscheidungsbaums ist die (Shannon) Entropie einer Spalte X der Datentabelle (mit der Variable x):

$$E(X) = -\sum_{i=1,k} p_i \log_2(p_i), p_i = \frac{count(c_i, X)}{N}, X = \{c|c \in C\}$$

➤ Dann der Informationsgewinn einer Spalte X hinsichtlich der Zielvariablenspalte Y:

$$G(Y|X) = E(Y) - \sum_{v \in Val(X)} \frac{|Y_v|}{|Y|} E(Y_v)$$

➤ Der Informationsgewinn, der durch Auswahl des Attributs x und der Spalte X erzielt wird, errechnet sich dann als Differenz der Entropie von Y und der erwarteten/durchschnittlichen Entropie von Y bei Fixierung von x.

7.6. Algorithmus

```
 1:
0. Starte mit leeren Baum, allen Eingansattributen X, der Zielvariablen Y,
 2:    und der vollständigen Datentabelle D(X,Y).
 3:
1. Berechne den Informationsgewinn für jede Attributevariable x   X.
 4:
2. Wenn nicht alle Zeilen zum selben Zielvariablenwert gehören,
 5:
   wird der Datensatz D in Teilmengen D'~xbest,v1~, D'xbest,v(2~~\), usw.
 6:
   aufgeteilt für das Attribut xbest   X mit dem größten Informationsgewinn.
 7:
3. Es wird ein Knoten mit der Attributevariable xbest erstellt.
 8:
4. Wenn alle Zeilen zur selben Klasse gehören, wird ein Blattknoten
 9:    mit dem Wert der Zielvariable erstellt.
10:
5. Wiederholung von 1-4 für die verbleibenden Attribute X'=X / xbest,
11:
   allen Teilbäumen (Verzweigungen von aktuellen Knoten) mit jeweiligen D',
12:    bis alle Attribute verwendet wurden,
13:    oder der Entscheidungsbaum alle Blattknoten enthält.
```

7.7. C4.5 Verfahren

[1] J. R. Quinlan, "C4.5: Programs For Machine Learning". Morgan Kaufmann, 1988.

- Wie ID3 werden die Daten und Attribute an jedem Knoten des Baums bewertet um das beste Teilungsattribut zu bestimmen.
- Aber C4.5 verwendet die Methode der "gain ratio impurity", um das Teilungsattribut zu bewerten (Quinlan, 1993).
- Entscheidungsbäume werden in C4.5 mithilfe eines Satzes von Trainingsdaten oder Datensätzen wie in ID3 erstellt.
- An jedem Knoten des Baums wählt C4.5 ein Attribut der Daten aus, das seinen Satz von Samples am effektivsten in Teilmengen aufteilt, die in der einen oder anderen Klasse verteilt sind.
- Das Kriterium ist der **normalisierte Informationsgewinn**:

- ❏ Verhältnis des Informationsgewinns G (Gain) zu einer sog. Teilungsqualität (Split Info SI), die sich aus der Zielvariable Y zum Aufteilen nach den Y Werten der Daten ergibt.
- ❏ Das Attribut mit dem höchsten Verhältnis GR (Gain Ratio) wird ausgewählt, um die Entscheidung für die Teilung zu treffen.

$$G(Y|X) = E(Y) - \sum_{v \in Val(X)} \frac{|Y_v|}{|Y|} E(Y_v)$$

$$SI(Y) = \sum_{c \in Val(Y)} -\frac{|Y_c|}{|Y|} \log_2 \frac{|Y_c|}{|Y|}$$

$$GR = \frac{G(Y|X)}{SI(Y)}$$

7.8. Teilung von kategorischen und numerischen Variablen

- ➤ Bei kategorischen Variablen bestimmen die Werte $Val(X)$ einer Spalte der Datentabelle einer Variablen x die Aufteilung eines Entscheidungbaums (**Partitionierung**).
- ➤ Bei nuemrischen Variablen muss ein Wert als Teilungspunkt aus der Werteverteilung bestimmt!
  - ❏ Nicht trivial; Welches Kriterium?
  - ❏ Intervallkategorisierung und Wertepartitionierung kann helfen!
  - ❏ D.h. mit intervallkategorisierten diskrete Werter wird die Spalte X entsprechend der Zielvariable Y partitioniert.
  - ❏ Und diese Partitionen werden bewertet und der Teilungspunkt x_{split} X bestimmt (z.B. über Mittelwerte der Intervalle)

Vertiefung

L. Rokach and O. Maimon, Data Mining with Decision Trees - Theory and Applications. World Scientific Publishing, 2015.

7.9. Intervallkodierung

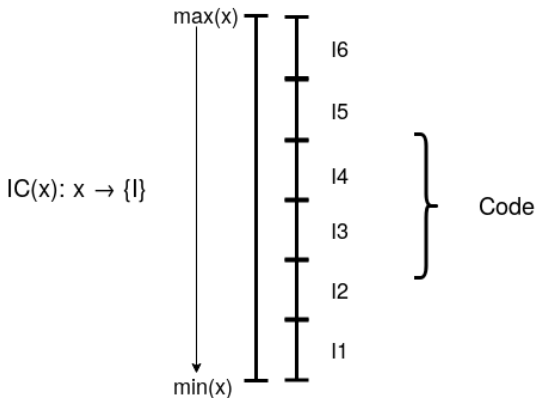


Abb. 26. Einteilung von kontinuierlichen Werteverteilungen in Intervall und Abbildung auf kategorische (diskrete) Werte

7.10. Unvollständige Trainingsdaten

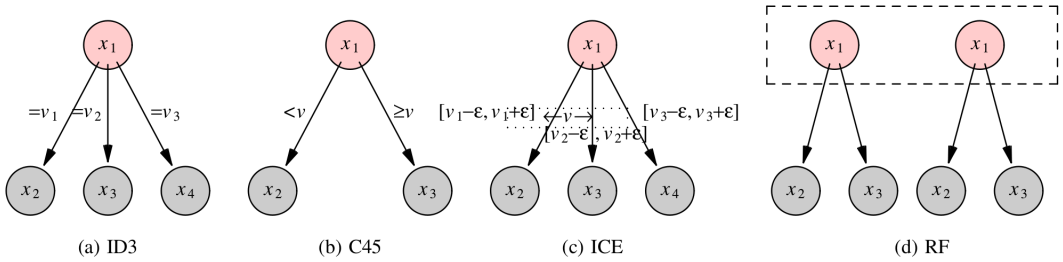
- Es kommt vor allem in der Soziologie aber auch in der Mess- und Prüftechnik vor, dass nicht alle Werte der Attributvariablen $\boldsymbol{X}$ für alle Trainingssätze bekannt sind.
  - Die Behandlung fehlender Attributwerte in den Zeilen der Datentabellen ist schwierig
- Es gibt keine Universallösung für den Umgang mit ? Werten. Möglichkeiten:
  - Ersetzen des fehlenden Werts mit einem Standardwert
  - Ersetzen des fehlenden Werts mit einem probalistisch über Verzeilungshäufigkeiten bestimmten Wert (auch unter Einbeziehung des gesamten Datensamples)
  - Attributevariablen mit fehlenden Werten nicht verwenden

7.11. Intervallkategorisierte Entscheidungsbäume (INN/ICE)

- Bisherige Entscheidungsbäume (C4.5/ID3) wurden entweder mit einer diskreten Anzahl von kategorischen Werten verzweigt oder mittels binärer Relationen!

- Aber Sensoren (sowohl in der Mess-und Prüftechnik als auch ind er Soziologie) sind fehlerbehaftet, d.h. es gibt bei jedem x-Wert ein Unsicherheitsintervall $[x\text{-}, x+] \rightarrow$ **Rauschen**
- Damit können Entscheidngsbäume (anders als Neuronale Netze oder Regressionslerner) nicht umgehen.
  - Wenn der Split mit $x<50$ und $x\ 50$ an einem Knoten mit x erfolgt würde bei Werten um 50 und überlagerten Rauschen ein Entscheidungsproblem entstehen!
- Lösung: k-stellige Knoten mit Intervallverzweigungen, also:

$$M(X) = \begin{cases} x_i \in [v_1 - \varepsilon_i, v_1 + \varepsilon_i], \{\cdots \\ x_i \in [v_1 - \varepsilon_i, v_1 + \varepsilon_i], \{\cdots \\ \cdots \\ x_i \in [v_n - \varepsilon_, v_n + \varepsilon_i], \{\cdots \end{cases}$$



(a) ID3 (b) C45 (c) ICE (d) RF

Abb. 27. Vergleich der verschiedenen Baumarten und Knotenverzweigungen

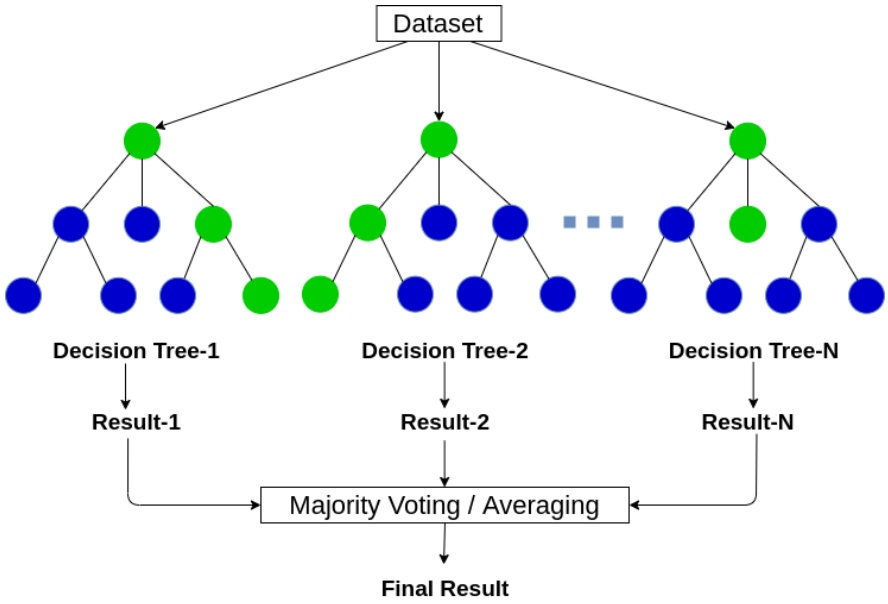
- Bei der Konstruktion des Entscheidungsbaums werden wieder nach Informationsgewinn bzw. Gewinnverhältnis Attributvariablen und Spalten der Datentabelle ausgewählt.
- Die numerischen Werte werden sowohl beim Training als auch bei der Inferenz durch Intervalle ersetzt $\rightarrow$ Ersetzung von diskreter mit **Intervallarithmetik**
- Entropien usw. werden durch kategorisierte Intervalle bestimmt
- Das große Problem: Für jede Variable muss ein abgeschätzt werden $\rightarrow$ Statistisches Modell erforderlich.
- Und was bedeuten jetzt überschneidende Intervalle?
  - Überschneidungen bedeuten Ununterscheidbarkeit!

Inferenz mit NN Suche

- Jeder Knoten x_i hat ausgehende Kanten mit annotierten Intervallen [v_j-,v_j+]
- Bei einem neuen zu testenden Variablenwert v wird einseits auch ein Intervall [v- ,v+] gebildet und mit den Kantenintervallen verglichen, andererseits wird das nächstliegende Intervall gesucht

7.12. Random Forest Trees

- Multiinstanzmodell
  - Es werden m Entscheidungsbäume DT=$\{dt_1,..,dt_m\}$ getrennt gelernt und erzeugt
  - "Random": Die Aufteilung der Daten in Teilungsvariablen erfolgt randomisiert!
  - Eingabedaten werden zur Inferenz an alle Teilbäume dt_i DT gegeben
  - Alle Ausgabevariablen der Teilbäume werden fusioniert
- Fusion:
  - Mittelwert (bei intervallkodierten oder intervallskalierbaren kat. Zielvariablen durch Dekodierung in numerische Werte)
  - Mehrheitsentscheid
  - Konsensfindung (Verhandlung)
- Parametersatz:
  - Stelligkeit eines Knotens (Anzahl der ausgehenden Kanten)
  - **Anzahl der Teilbäume**
  - **Partitionierung** des Eingaberaums (d.h. ein bestimmter Baum verwendet nur eine Teilmenge der Spalten aus $\boldsymbol{D}$)
  - Fusionsmodell und Algorithmus



[Abhishek Sharma, 2020, www.analyticsvidhya.com]

Abb. 28. Grundprinzip von Multibaumklassifikatoren

7.13. Zusammenfassung

- ➤ Entscheidungsbäume sind für die Klassifikation von kategorischen Zielvariablen geeignet
- ➤ Numerische Zielvariablen müssen intervallkodiert werden.
- ➤ ID3/C4.5 Lerner können numerische und kategorische Eingabevariablen (Attribute) verwenden
  - ❏ Ein Attributariable ist ein Teilungspunkt
- ➤ Rauschen auf Sensordaten muss durch "Unsicherheitsintervall" und Intervallarithmetik behandelt werden
- ➤ Vergleich mit anderen Lernverfahren zeigt gute Ergebnisse (je nach Problem)

8. Klassifikation mit Künstlichen Neuronale Netze

***Zielvariablen**: Numerische Variablen*
***Eigenschaftsvariablen**: Numerische Variablen*
***Modell**: Gerichteter Graph (zyklisch oder azyklisch)*
***Training und Algorithmen**: Backpropagation*
***Klasse**: Überwachtes Lernen*

8.1. Künstliche Neuronale Netze

➤ Ein Künstliches Neuronales Netz (KNN) ist ein gerichteter Graph bestehend aus einer Menge von Knoten $\boldsymbol{N}$ und Kanten $\boldsymbol{E}$ die die Knoten verbinden

- ❏ Knoten: Neuron oder Perzeptron mit einem oder mehreren Eingängen $\boldsymbol{I}$ und einem Ausgang o; Berechnungsfunktion $g(\boldsymbol{I})$: $\boldsymbol{I} \rightarrow o$
- ❏ Kanten: Gewichteter Datenfluß vom Ausgang eines Neurons zum Eingang eines anderen (oder des selben) Neurons

Ein KNN ist eine Komposition aus einer Vielzahl von Abbildungsfunktionen $\boldsymbol{G}$=(g_1,g_2,..,g_m). Es gibt Parallelen zu Regressionsverfahren mit Funktionen.

➤ Zusammengefasst ausgedrückt:

$$M(X) : X \rightarrow Y, X = \{x_i\}, Y = \{y_j\}$$
$$KNN = \langle N_x, N_d, N_y, E \rangle$$
$$N_x = \{n_i : n_i \leftrightarrow \{x_j\}\}, N_d = \{n_d\}, N_y = \{n_k : n_k \leftrightarrow y_k\}$$
$$n = g(\tilde{p}, \tilde{w}, b) : \tilde{p} \rightarrow o = f(\sum_i w_i p_i + b)$$
$$E = \{e_{ij} : n_i \mapsto n_j w_{ij}\}$$

➤ f ist eine Transferfunktion die die akkumulierten Eingangswerte auf den Ausgangswert o abbildet, und g ist dann die gewichtete und akkumulative Transferfunktion

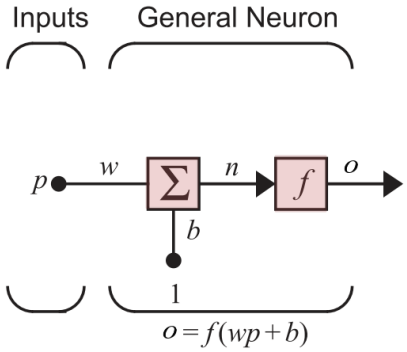
➤ Unterschied (künstliches) Neuron und Perzeptron:

- ❏ Ein Neuron ist immer eine Elementarzelle
- ❏ Ein Perzeptron kann ein einzelnes Neuron oder ein Netzwerk aus Neuronen beschreiben

➤ Daher gibt es:

- ❏ Single Layer Perceptron (SLP) → Nur Eingangs- $\boldsymbol{N}_{\mathrm{x}}$ und Ausgangsneuronen $\boldsymbol{N}_{\mathrm{y}}$
- ❏ Multi Layer Perceptron (MLP) → + Innere Neuronen $\boldsymbol{N}_{\mathrm{d}}$

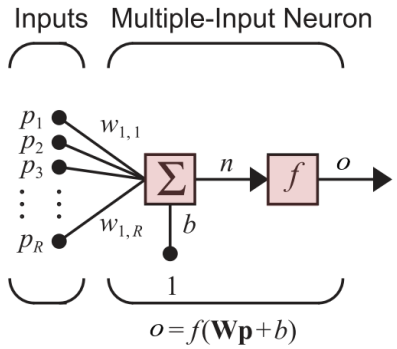
8.2. Das Neuron



[15]

Abb. 29. Ein einzelnes Neuron mit einem einzelnen Eingang p und einem Ausgang o. w ist ein Gewichtungsfaktor (ein Gewicht für eingehendes p) und b ist ein Bias (Offset)

8.3. Das Mehreingangsneuron

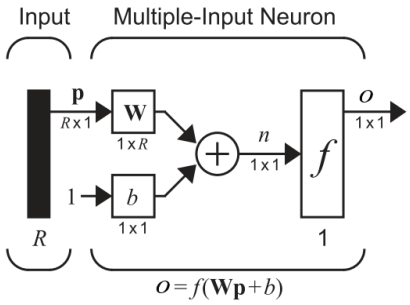


[15]

Abb. 30. Ein einzelnes Neuron mit einem Eingangsvektor $\boldsymbol{p}$ und einem skalaren Ausgang o. $\boldsymbol{w}$ ist ein Gewichtungsfaktorvektor (ein Gewicht für eingehendes p) und b ist ein Bias (Offset)

8.4. Neuronale Netze und Matrizen

➤ Neuronale Netze werden durch eine Graphenstruktur (statische Paremeter) und mathematisch durch Matrizen (dynamische Parameter) beschrieben:



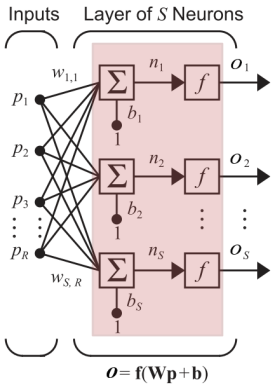
[15]

Abb. 31. Ein einzelnes Neuron mit einem Eingangsvektor $\boldsymbol{p}$ und einem skalaren Ausgang o. $\boldsymbol{w}$ ist ein Gewichtungsfaktorvektor (ein Gewicht für eingehendes p) und b ist ein Bias (Offset); jetzt in Matrizenform (Annotation)

8.5. Schichten von Neuronalen Netzen

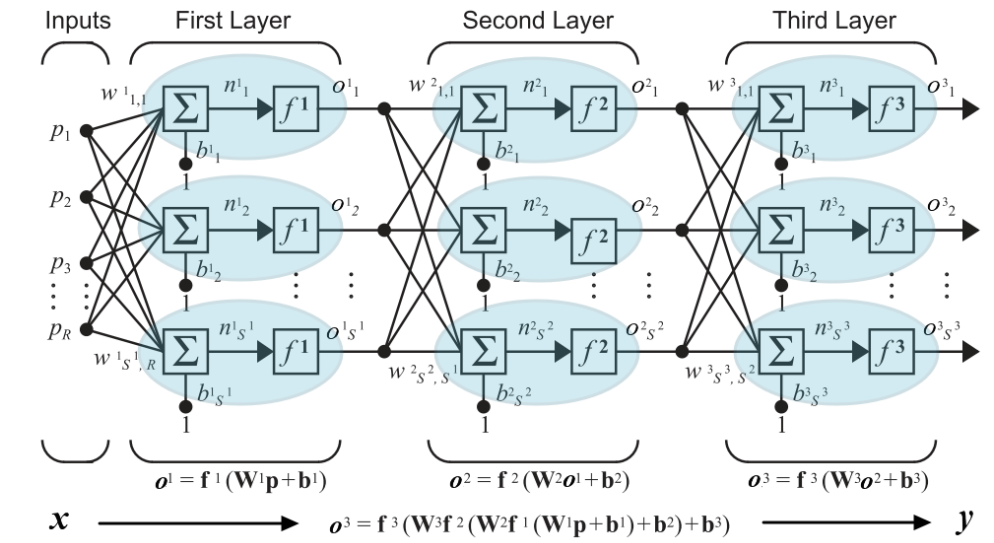
➤ I.A. werden Neuronen von neuronalen Netzen in Schichten (Layern) angeordnet und gruppiert

- ❏ Günstig für Matrixalgebra
- ❏ Aber nicht notwendig!



[15]

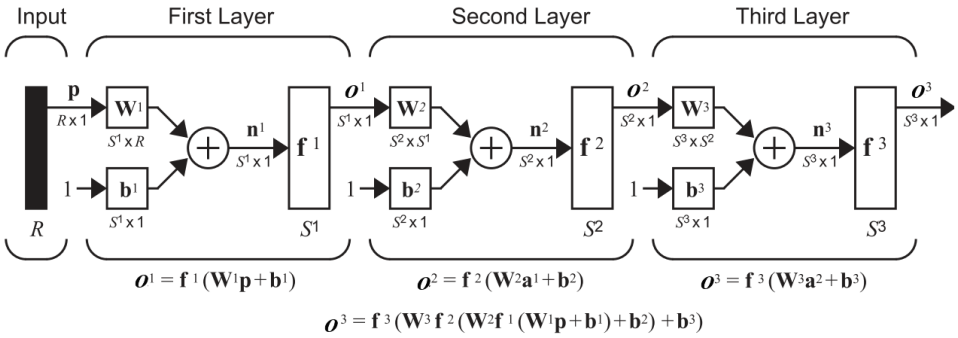
8.6. Struktur eines KNN



[15]

Abb. 32. Grundlegende Struktur eines KNN mit Matrizen (blaue Ellipse=1 Neuron)

8.7. Vereinfachte Form eines KNN



[15]

Abb. 33. Vereinfachte Struktur eines KNN mit Matrizen

8.8. Klassen von KNN

Forwärtsgekoppelte Netzwerke

Azyklischer gerichteter Graph, d.h. es gibt nur eine Vorwärtspropagation von einer Schicht zur nächsten (keine Rückkopplung).

- ➤ Diese Netzwerke können rein funktional beschrieben und berechnet werden.
- ➤ Es gibt keinen Zustand!
- ➤ D.h. die aktuellen Ausgangswerte hängen nur von den aktuellen Eingangswerten ab!

Rückgekoppelte Netzwerke

Zyklischer gerichteter Graph, d.h. es gibt Rückkopplungen (Ausgang eines Neurons geht in Eingänge der aktuellen oder vorherigen Schichten).

- ➤ Diese Netzwerk können nicht rein funktional beschrieben und berechntet werden!
- ➤ Sie besitzen einen Zustand, d.h. der Ausgangswert hängt von der Historie vergangener Eingabewerte ab!

8.9. Transferfunktion

- ➤ Auch Aktivierungsfunktion genannt (in Anlehnung an biologische Vorbild)
  - ❑ Biologisch: Häufig eine Schwellwertfunktion
  - ❑ Künstlich / ML: Auch lineare Übertragunsfunktionen!
- ➤ Es gibt eine Vielzahl verschiedener Funktionen
  - ❑ Die einfachste wäre (wenn auch wenig in Gebrauch): $f(a) = a$

Warum ist eine solche Übertragungsfunktion ungeeignet bzw. problematisch?

- ➤ Welche mathematischen **Eigenschaften** (Übertragungskurve) sollte wohl eine Transferfunktion besitzen?

❑ Zur Erinnerung: Wir nehmen an dass der Wertebereich von einem x [-1,1] ist. Ebenso für ein y [-1,1].

Transferfunktionen besitzen häufig begrenzende Eigenschaften (Sättigungsverhalten), und nicht lineares Übertragungsverhalten

Name	Input/Output Relation	Icon	Function
Hard Limit	$a = 0 \quad n < 0$ $a = 1 \quad n \geq 0$		hardlim
Symmetrical Hard Limit	$a = -1 \quad n < 0$ $a = +1 \quad n \geq 0$		hardlims
Linear	$a = n$		purelin
Saturating Linear	$a = 0 \quad n < 0$ $a = n \quad 0 \leq n \leq 1$ $a = 1 \quad n > 1$		satlin
Symmetric Saturating Linear	$a = -1 \quad n < -1$ $a = n \quad -1 \leq n \leq 1$ $a = 1 \quad n > 1$		satlins
Log-Sigmoid	$a = \frac{1}{1+e^{-n}}$		logsig
Hyperbolic Tangent Sigmoid	$a = \frac{e^n - e^{-n}}{e^n + e^{-n}}$		tansig
Positive Linear	$a = 0 \quad n < 0$ $a = n \quad 0 \leq n$		poslin
Competitive	$a = 1$ neuron with max n $a = 0$ all other neurons	C	compet

[15]

Abb. 34. Verschiedene gebräuchliche Transferfunktionen $f(a)$

8.10. Ein einfaches Neuron - Funktional

$$f_{sigmoid}(a) = \frac{1}{1+e^{-a}}$$

$$g(x_1, x_2, x_3) = f_{sigmoid}(b + \sum w_i x_i)$$

/webwork

Neuron

```
function sigmoid(x) {
  return 1/(1+Math.exp(-x))
}
print(sigmoid(0))
var data = [
  {x1:1,  x2:2,  y:1.0},
  {x1:-1, x2:2,  y:0.0},
  {x1:0,  x2:-1, y:0.0},
]
function neuron(x1,x2,w,b) {
  var accu = x1*w[0]+x2*w[1];
  return sigmoid(accu+b)
}
var w = [1.0,1.0], b=0;
data.forEach(function (row) { print(neuron(row.x1,row.x2,w,b),row.y) });
```

8.11. Parametersatz des KNN

Statische Parameter

- Anzahl der Eingangsneuronen (verbunden mit $\mathbf{x}$), abhängig von der Anzahl der Eingabevariablen $|\mathbf{x}|$ und der Kodierung (numerisch vs. kategorisch)

- ➤ Anzahl der Ausgangsneuronen (abhängig von der Kodierung). Bei numerischen Zielvariablen $\mathbf{y}$ gilt also: $|N_{\mathrm{y}}|=|\mathbf{y}|$
- ➤ Anzahl der inneren verdeckten Neuronen $|N_{\mathrm{d}}|$ und deren Anordnung in Schichten
- ➤ D.h. die **Konfiguration** des Netwerkes ist $[c_1,c_2,..,c_{\mathrm{m}}]$ bei m Schichten und c_{i} Neuronen pro Schicht
- ➤ Bei vollständig verbundenen Schichten ist keine Angabe der Vernetzung notwendig

Dynamische Parameter

- ➤ Im wesentlichen die Gewichtematrix $\mathbf{W}_{\mathrm{i}}$ (Schicht i):

$$W_i = \begin{bmatrix} w_{1,1} & amp; w_{1,2} & amp; \cdots & amp; w_{1,R} \\ w_{2,1} & amp; w_{2,2} & amp; \cdots & amp; w_{2,R} \\ \vdots & amp; \vdots & amp; & amp; \vdots \\ w_{S,1} & amp; w_{S,2} & amp; \cdots & amp; w_{S,R} \end{bmatrix}, B_i = \begin{bmatrix} b_1 \\ \vdots \\ b_S \end{bmatrix}$$

Mit S: Anzahl der Neuronen in der Schicht, R: Anzahl der Eingangsvariablen (oder Neuronen der vorherigen Schicht)

- Der Ausgangswert eines Neurons n_j ist dann gegeben durch einen Wert aus B und die j-te Zeile von $\boldsymbol{W}$:

$$o(\tilde{p}) = f({}_j W^T \tilde{p} + b_i)$$

- Bei mehrschichtigen Netzwerken hat man eine Menge von Gewichtematrizen, die zu einem Tensor zusammengefasst werden können.

8.12. Training von KNN

- Wie bei allen überwachten Lernproblemen gilt es eine Fehlerfunktion zu minimieren:

$$M(\tilde{x}) : \tilde{x} \rightarrow \tilde{y}$$
$$\underset{W}{\operatorname{argmin}}\ err(M) = |y(\tilde{x}) - y_0(\tilde{x})|, \forall (x, y_0) \in D$$

*Ziel ist die Minimierung des Fehlers unserer Modellhypothese M(**x**) durch Anpassung der Gewichtematrix **W** und evtl. (wenn vorhanden) des Offsetvektors **B***
*Es ist leicht zu erkennen dass das Training einen hochdimensionalen Parametersatz anpassen muss. Es ist nicht unmittelbar klar wie ein optimales **W** abgeleitet werden kann!*

Erklärbarkeit

- Der Zusammenhang von y und x ($x \rightarrow y$) ist schon bei einem einschichtigen Netzwerk nur noch schwer nachvollziehbar!
- Eine Invertierung (inverses Problem $y \rightarrow x$) ist ebenso nur schwer möglich
- Eigentlich ist nur ein einzelnes Neuron erklärbar und verständlich
  - Dort ist die Anpassung (des Gewichtevektors $\mathbf{w}$) noch durch multivariate Regression möglich

Beispiel

- "Gradient Descent" Verfahren

- Problem: $\mathbf{x}$=(a,b), y
- Netzwerk: Ein Neuron, Sigmoid Transferfunktion

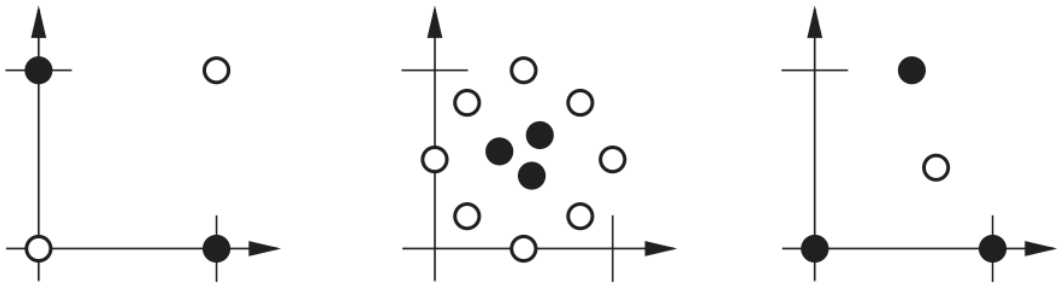
/webwork

Neuron

```
function sigmoid(x) {
  return 1/(1+Math.exp(-x))
}
print(sigmoid(0))
var data = [
  {x1:1,  x2:2,  y:1.0},
  {x1:-1, x2:2,  y:0.0},
  {x1:0,  x2:-1, y:0.0},
]
function neuron(x1,x2,w,b) {
  var accu = x1*w[0]+x2*w[1];
  return sigmoid(accu+b)
}
var w = [1.0,1.0], b=0, sets=[0,1,2], rate=0.1, errors=[];
for(var run=0;run<10000;run++) {
  var set=Math.random.select(sets),
      row=data[set];
  var y=neuron(row.x1,row.x2,w,b), err=y-row.y;
  if ((run % 100)==0) { errors.push(Math.abs(err)); print(Math.abs(err))}
  w[0]=w[0]-rate*err*row.x1;
  w[1]=w[1]-rate*err*row.x2;
}
print(w)
data.forEach(function (row) { print(neuron(row.x1,row.x2,w,b),row.y) });
```

8.13. Nicht lineare Probleme

SLP können nur lineare Probleme separieren.



[15]

Abb. 35. Nicht linear separierbare Probleme - nur mit MLP klassifizierbar

/webwork

Neuron

```
function sigmoid(x) {
  return 1/(1+Math.exp(-x))
}
print(sigmoid(0))
var data = [
  {x1:0,  x2:0,  y:0},
  {x1:1,  x2:0,  y:1},
  {x1:0,  x2:1,  y:1},
  {x1:1,  x2:1,  y:0},
]
function neuron(x1,x2,w,b) {
  var accu = x1*w[0]+x2*w[1];
  return sigmoid(accu+b)
}
var w = [1.0,1.0], b=0, sets=[0,1,2], rate=0.1, errors=[];
for(var run=0;run<10000;run++) {
  var set=Math.random.select(sets),
      row=data[set];
  var y=neuron(row.x1,row.x2,w,b), err=y-row.y;
  if ((run % 100)==0) { errors.push(Math.abs(err)); print(Math.abs(err))}
  w[0]=w[0]-rate*err*row.x1;
  w[1]=w[1]-rate*err*row.x2;
}
print(w)
data.forEach(function (row) { print(neuron(row.x1,row.x2,w,b),row.y) });
```

8.14. Backpropagation Verfahren

➤ Bekanntes und gängiges Verfahren

https://hmkcode.com/ai/backpropagation-step-by-step

Gradientenverfahren

➤ Baut auf dem Minimierungsansatz "Gradient Descent" (GD) auf (Absteigender Gradient)

➤ Beim GD Verfahren wird eine Funktion, z.B. $f(x,w)$: $x \rightarrow y$ derart über den Parameter w angepasst wird dass der Fehler $err=|y\text{-}y_0|$ minimal wird

➤ Es wird nun die Änderung des Fehlers beobachtet Δerr und der (oder später die) Parameter w mit der Ableitung des Fehlerwerts $err/\ w$ zu der Änderung des Parameters korrigiert:

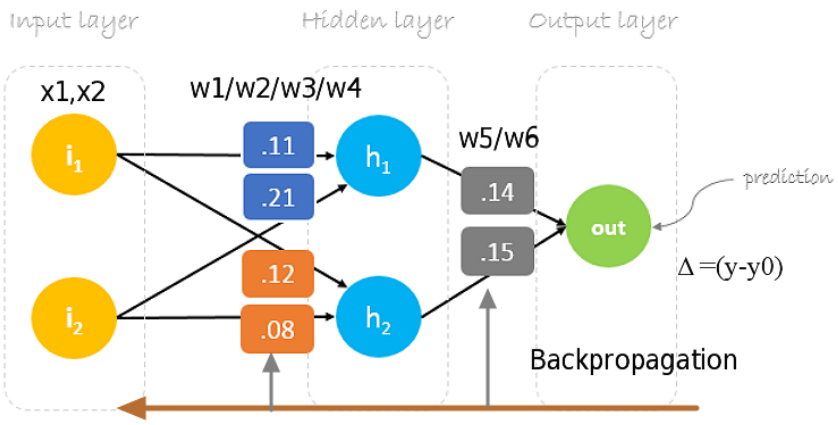
$$w' = w - \alpha \frac{\partial err}{\partial w}$$

➤ Vereinfacht gilt:

$$\frac{\partial err}{\partial w} \sim x(y - y_0)$$

➤ Jetzt wird ein neuronales Netzwerk betrachtet, wo die Neuronen ebenfalls Funktionen mit Eingangsvariablen und Ausgangsvariablen sind

➤ Bei zusammengesetzten Funktionen (z.B. auch Neuronen in inneren Schichten) müssen die Gewichte schrittweise von hinten nach vorne angepasst werden



[hmkcode.com/ai/backpropagation-step-by-step]

Abb. 36. Beispiel eines ANN mit Kantengewichten und dem Ansatz der Backpropagation

➤ Die Gewichte werden nun Schicht für Schicht unter Einbeziehung der gewichteten Fehlerpropagation fleichermaßen angepasst

$$^{*}w_6 = w_6 - \alpha\,(h_2 \,.\, \Delta)$$
$$^{*}w_5 = w_5 - \alpha\,(h_1 \,.\, \Delta)$$
$$^{*}w_4 = w_4 - \alpha\,(i_2 \,.\, \Delta w_6)$$
$$^{*}w_3 = w_3 - \alpha\,(i_1 \,.\, \Delta w_6)$$
$$^{*}w_2 = w_2 - \alpha\,(i_2 \,.\, \Delta w_5)$$
$$^{*}w_1 = w_1 - \alpha\,(i_1 \,.\, \Delta w_5)$$

$$\begin{bmatrix} w_5 \\ w_6 \end{bmatrix} = \begin{bmatrix} w_5 \\ w_6 \end{bmatrix} - \alpha\Delta \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} = \begin{bmatrix} w_5 \\ w_6 \end{bmatrix} - \begin{bmatrix} \alpha h_1 \Delta \\ \alpha h_2 \Delta \end{bmatrix}$$

$$\begin{bmatrix} w_1 & w_3 \\ w_2 & w_4 \end{bmatrix} = \begin{bmatrix} w_1 & w_3 \\ w_2 & w_4 \end{bmatrix} - \alpha\Delta \begin{bmatrix} i_1 \\ i_2 \end{bmatrix} . [w_5 \quad w_6] = \begin{bmatrix} w_1 & w_3 \\ w_2 & w_4 \end{bmatrix} - \begin{bmatrix} \alpha i_1 \Delta w_5 & \alpha i_1 \Delta w_6 \\ \alpha i_2 \Delta w_5 & \alpha i_2 \Delta w_6 \end{bmatrix}$$

[hmkcode.com/ai/backpropagation-step-by-step]

Abb. 37. Backpropagation des Fehlers zu den Eingängen des Beispielnetzwerkes

8.15. Kategorische Multiklassen Probleme

➤ Wenn die Ergbnisvariable vom kategorischen Typ ist dann gibt es zwei Möglichkeiten:

One-Hot Kodierung

Jedes Klassensymbol (also ein diskreter Wert v_i der Zielvariable y) wird durch ein Ausgangsneuron repräsentiert

Multi-level Kodierung

Jedes Klassensymbol wird durch einen Wert aus dem Wertebereich eines Ausgangsneurons repräsentiert

➤ Problem: Nicht lineare Transferfunktion und Sättigungsverhalten

❏ Die gleichen Verfahren sind auch auf kategorische Eingabevariablen anwendbar

8.16. Numerische Prädiktorfunktionen

➤ Neben der Klassfikation lassen sich mit ANN auch numerische (kontinuierliche) Funktionen lernen

➤ Damit wird **Funktionsapproximation** wie bei den Regressionsverfahren möglich

❏ Unterschied: Bei der Regression ist die funktionale Struktur von $f(x)$: $x \rightarrow y$ bereits fest und muss vorgegeben sein

❏ Die Verwendung eines ANN bietet da auch noch indirekt das Lernen der funktionalen Strukturen neben der Anpassung der Parameter

8.17. Literatur zur Vertiefung

[1] M. T. Hagan, howard B. Demuth, M. H. Beale, and O. D. Jesus, *Neural Network Design.* https://hagan.okstate.edu/nnd.html

8.18. Zusammenfassung

➤ Neuronale Netze bestehen aus Neuronen

➤ Neuronen sind Funktionen

➤ Die Kanten verbinden Ausgänge von Neuronen mit den Eingängen nachfolgender Neuronen mit einer Multiplikation eines Gewichtfaktors

➤ Alle Eingänge eines Neurons werden summiert, das Ergebnis einer Transfer/Aktivierungsfunktion übergeben

➤ Training bedeutet Anpassung der Gewichte um den Ausgangsfehler zu minimieren

- ❏ Gängiges Verfahren: Fehlerrückpropagation

9. Ein- und Ausgabeschnittstellen von Prädiktorfunktionen

Kodierung und Dekodierung von Variablen für kont. Prädiktorfunktionen
Normalisierung und Skalierung von Daten

9.1. Kagetorische Variablen

- ➤ Kategorische Variablen müssen für die Verwendung in Prädiktorfunktionen in numerische Werte kodiert werden
  - ❏ Multi-level Kodierung: Abbildung von allen kategorischen Werten auf unterschiedlich numerische Werte (skalare und vektorielle Werte)
  - ❏ One-hot Kodierung: Abbildung von allen kategorischen Werten auf 01 Vektoren

Aber: Anders als die Intervallkodierung von numerischen Variablen (z.B. für Entsch.bäume) muss im umgekehrten Fall der Kodierung einer kat. Variable in ***eine*** *numerische Variable eine Intervall- und Verhältnisskalierung existieren!!*

Beispiele

- ➤ Farben **C**={rot,grün,blau,braun}
  - ❏ Kodierung in skalaren kont. Variable aber mit diskreten Werten
  - ❏ Kodierung in vektorielle Variable mit kont. Werten → RGB

$$z_1(C) = \begin{cases} 1, C = \text{rot} \\ 2, C = \text{gruen} \\ 3, C = \text{blau} \\ 4, C = \text{braun} \end{cases}, z_2(C) = \begin{cases} (1.0, 0.0, 0.0), C = \text{rot} \\ (0.0, 1.0, 0.0), C = \text{gruen} \\ (0.0, 0.0, 1.0), C = \text{blau} \\ (0.5, 0.0, 0.5), C = \text{braun} \end{cases}$$

Welche Probleme ergeben sich bei der Multi-level Kodierung (skalar) von Zielvariablen bei typischen Sigmoid Transferfunktionen von Neuronen? Nichtlinearität an den Rändern [0,1]

9.2. Ein- und Ausgabeschnittstellen

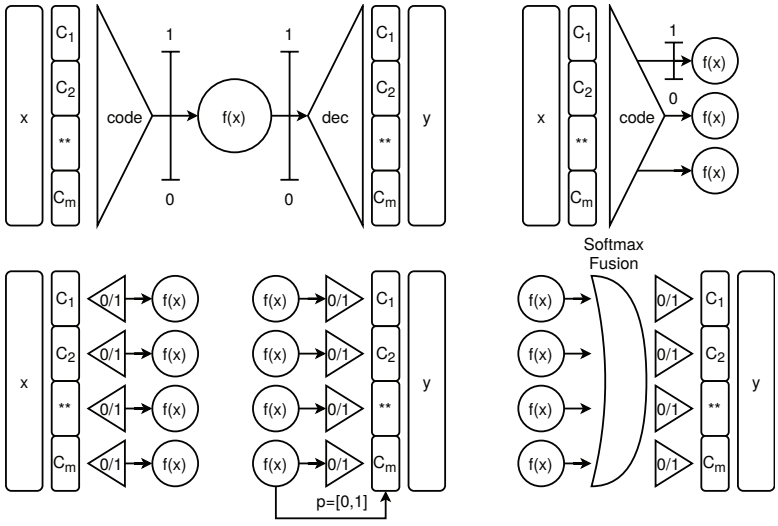


Abb. 38. Verschiedene Kodierungen von ein-/ausg. kat. Variablen

- Die One-hot Kodierung von Eingabevariablen erfordert die Berechnung eines diskretwertigen Vektors
- Aber die One-hot Kodierung von Ausgabevariablen ergibt einen kontinuierlichen Vektor der noch diskretisiert werden muss:
  - Schwellwertfunktion → Es kann zu Mehrdeutigkeiten kommen (Unentscheidbarkeit)
  - Der kont. Ausgangswert y_i (C_i) wenn im Bereich [0,1] kann als ein Maß für die Wahkrscheinlichkeit des Auftretens eines Klassensymbols C_i verwendet werden!
  - Softmax Funktion: Diese bildet einen n-dimensionalen Vektor y mit y_i [0,1] und $max\ (\Sigma\ y_\mathrm{i}) = |\mathrm{y}|$ auf einen n-dimensionalen Vektor ab mit $max\ (\Sigma\ y_\mathrm{i}) = 1$!

Softmax Funktion

$$softmax(Y) = \begin{pmatrix} \frac{e^{y_1}}{weight} \\ ... \\ \frac{e^{y_n}}{weight} \end{pmatrix}, weight = \sum e^{y_k}, y_i \in Y, Y = \begin{pmatrix} y_1 \\ ... \\ y_n \end{pmatrix},$$
$$\sum softmax(Y) \in [0, 1]$$

10. Fehleranalyse und Kostenfunktionen

Bewertung der Qualität eines Klassifikators oder von Prädiktorfunktionen
Aufteilung von Datensätzen
Künstliche Erweiterung von Datensätzen

10.1. Fehlerfunktionen

1. Ziel ist ein aussagekräftiger Fehlerwert um die Qualität des Trainings und des erzeugten Modells $M(x)$:$x \rightarrow y$ bewerten zu können
2. Fehlerwerte bei kategorischen Zielvariablen vergleichen direkt die Übereinstimmung der inferrierten und vorgegebenen Werte der Zielvariablen
   - ➤ Gesamtfehler (falsche Klassifikation)
   - ➤ Falsch-positiver Fehler (binärer Klassifikator)
   - ➤ Falsch-negativer Fehler (binärer Klassifikator)
   - ➤ Falsch-C Fehler (Multiklassifikator)
   - ➤ Klassenspezifischer Fehler (Multiklassifikator)

10.2. Fehlerberechnung

Daten

- ➤ Es sei $\boldsymbol{D}$ die Gesamtmenge der Dateninstanzen $\boldsymbol{D}$
- ➤ Es sei $\boldsymbol{D}_{\text{test}}$ die Testdatenmenge $\boldsymbol{D}_{\text{test}} \quad \boldsymbol{D}$

Klassifikationsfehler

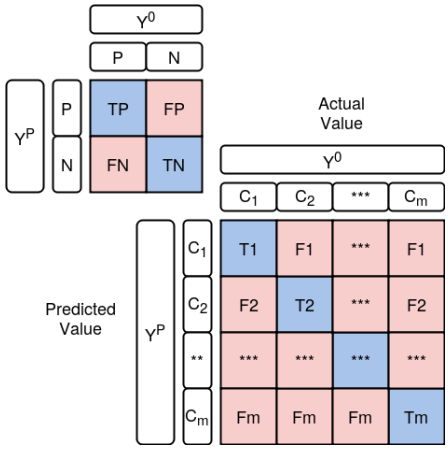
$$err(Y^0, Y^p) = \frac{1}{N} \sum_{i \in D} \begin{cases} 1, Y_i^0 \neq Y_i^p \\ 0 \end{cases}, Y^0 = D_{test}(Y), N = |D_{test}|$$

- Dabei sind Y^0 die vorgegebenen Werte der Zielvariablen und Y^{p} die aus dem trainierten Modell berechneten (inferrierten) Werte
- Der Inferenzfehler liegt dann im Bereich [0,1]
- Bei binärer Klassifikation sollten die Falsch-positiv und Falsch-negativ Fehler zusätzlich getrennt bestimmt werden:

$$err_C(Y^0, Y^p) = \frac{1}{N} \sum_{j \in D(Y=C)} \begin{cases} 1, Y_j^0 \neq Y_j^p \\ 0 \end{cases}, N = |D_{test}(Y = C)|$$

10.3. Konfusionsmatrix

- Bei Multiklassifikation bietet sich die Konfusionsmatrix an um eine Bewertung der Klassifikationsqualität mit falsch-C Bewertungen einer Klasse C und somit err_{C} zu erfassen



10.4. Kreuzentropie

- Bei kontinuierlichen Ausgabevariablen (KNN, SVM) aber evtl. kategorischen Zielvariablen ist die Berechnung der Kreutentropie zwischen Y^0 und Y^{p} aussagekräftiger
- Eine Kreuzentropie $\rightarrow 0$ bedeutet vollständige Übereinstimmung, Werte > 0 bedeuten Abweichungen

- Je größer die Kreuzentropie zwischen zwei Vektoren/Matrizen ist, desto größer ist die Abweichung

Berechnung der Kreuzentropie zweier Matrizen

- Eingabewerte: Matrix U und V mit Zellenwerten im Bereich [0,1]
- Ergebnis: Skalarer Wert
- Problem (siehe unten): log(0) → -∞

$$
\begin{aligned}
&E_{cross}(U,V) = -mean(sumRows(A+B)),\\
&A = (A_{ij}), A_{ij} = U_{ij}\log(\max(V_{ij}, 1^{-9})),\\
&B = (B_{ij}), B_{ij} = (1-U_{ij})\log(1-\min(V_{ij}, 1-1^{-9})),\\
&mean(X) = \sum x_i / |X|,\\
&sumRows(X) = (S_i), S_i = \sum X_{ij}
\end{aligned}
$$

10.5. Beispiele

/webwork

confmat

```
var confmat = [
 [1,0,0],
 [0,0.3,0.2],
 [0,0.7,0.8]
]
Plot(confmat,
     {type:'confusion',
      title:'Machine Learning Results',
      labels:{data:['A','B','C']}})
```

crossEntr

```
U=[[0,1],[1,0],[0,1],[0,1]]
V=[[0.1,0.99],[0.6,0.4],[0.01,0.99],[0.1,0.95]]
var cross=ML.statistics.crossEntropy(U,V)
var error= ML.math.meanMat(
  ML.math.activateTwoMat(U,V,
    function (u,v) { return Math.pow(u-v,2) }))
print('Cross Entropy  = '+cross);
print('Mean Sq. Error = '+(error*100));
```

11. Netzwerkkonfiguration

Konfiguration von ein- und mehrschichtigen neuronalen Netzwerken Festlegung der Anzahl der Schichten, Knoten pro Schicht, und Vernetzung

11.1. Neuronale Netze

- ➤ Ein neuronales Netz ist ein gerichteter Funktionsgraph
- ➤ Einzelne Funktionsknoten können zu Schichten zusammengefasst werden

Eingabeschicht

Die Anzahl der Eingabeneuronen wird durch den Eingangsvektor $\boldsymbol{x}$ und der Kodierung bestimmt! Jedes Element von $\boldsymbol{x}$ ist mit einem Eingang eines Eingangsneurons verbunden

Ausgabeschicht

Die Anzahl der Ausgabeneuronen wird durch den Ausgangsvektor $\boldsymbol{y}$ und der Kodierung bestimmt! Jedes Element von $\boldsymbol{y}$ ist mit einem Ausgang eines Ausgangsneurons verbunden

Innere Schichten

Neuronen die weder mit den Ein- noch den Ausgängen direkt verbunden sind.

Konfiguration

- ➤ Wenn von vollständig verbundenen Schichten ausgegangen wird kann die Konfiguration eines KNN mit einem Vektor/Array angegeben werden.

Jedes Element gibt die Anzahl der Neuronen pro Schicht an.

➤ Beispiele:

```
layers=[1] → SLP
layers=[2,1] → SLP
layers=[2,3,1] → MLP
layers=[1,4,3,1] → MLP
```

Verbindungen

1:1

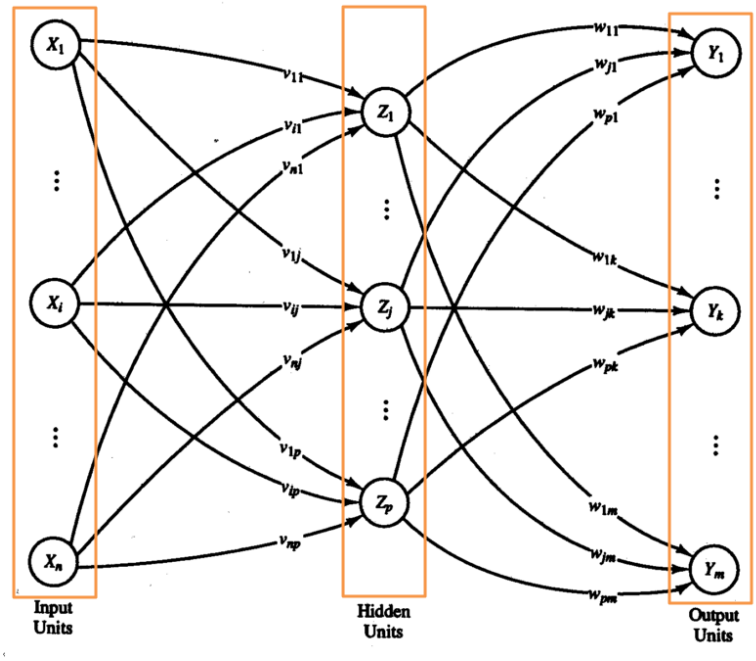
Der Ausgang eines Neurons einer Schicht i (oder einer Eingabevariable x) wird mit dem Eingang eines folgenden Neurons der Schicht j verbunden

1:n

Die Ausgänge aller Neuronen einer Schicht i (oder alle Eingabevariablen $\boldsymbol{x}$) werden mit den Eingängen jedes folgenden Neurons der Schicht j verbunden → **Vollständig verbundene Schicht**

0/1:k

Nur ein Teil der Ausgänge werden mit den Eingängen von folgenden Neuronen verbunden → **Unvollständig und irregulär verbundene Schichten** (eher Sonderfall)



[16]

Abb. 39. Vollständig verbundenes Netzwerk

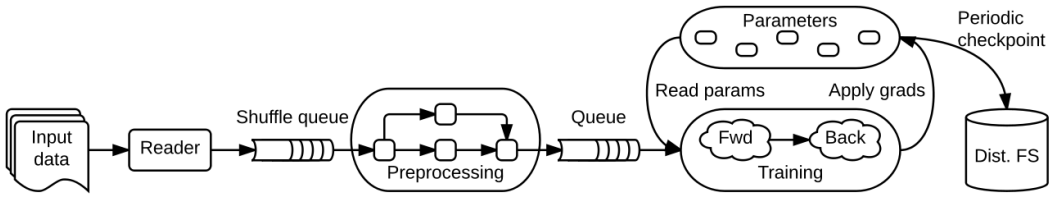
- Die statische Festlegung der Konfiguration der inneren Schichten ist schwierig!
- Es gilt:
  - Linear separierbare Probleme verwenden keine inneren Schichten!
  - Nichtlinear separierbare Probleme können innere Schichten verwenden
  - Beim Training wirken innere Schichten zunächst wie eine “Mauer”, die Wirkung auf den Ausgang wird kleiner
  - Bei inneren Schichten muss die Trainingsdatenvarianz und Anzahl der Instanzen groß sein
  - Je mehr innere Schichten und Neuronen um so mehr Trainingsinstanzen und Anstieg der Rechenzeit!

12. ML Frameworks

Tensorflow
Neataptic
Torch
ML

12.1. Tensorflow

- Datenverarbeitung mit tiefen Künstlichen Neuronalen Netzwerken ("Deep Learning") → "Google Brain"
- Strombasierte Datenverarbeitung mit Pipelines (Queues)
  - Datenverarbeitungsgraph (Datenflussdiagram)



[103]

- TensorFlow verwendet ein einheitliches Datenflussdiagramm, um sowohl die Berechnung in einem Algorithmus als auch den Zustand darzustellen, in dem der Algorithmus arbeitet.
- Im Gegensatz zu herkömmlichen Datenflusssystemen, bei denen Knoten funktionale Berechnungen für unveränderliche Daten darstellen, ermöglicht TensorFlow Kanten mit Berechnungen darzustellen, die einen veränderlichen Zustand besitzen oder aktualisieren [103].
- https://www.tensorflow.org
- Durch die Vereinheitlichung des Berechnungs- und Zustandsmanagements in einem einzigen Programmiermodell ermöglicht TensorFlow Programmierern, mit verschiedenen Parallelisierungsschemata zu experimentieren, die beispielsweise die Berechnung auf Servern oder GPUs erlauben.
- Eine typische TensorFlow-Anwendung hat zwei verschiedene Phasen:

- Die erste Phase definiert das Programm (Z. B. ein zu trainierendes neuronales Netzwerk und die Aktualisierungsregeln) als Symbolisches Datenflussdiagramm mit Platzhaltern für die Eingabedaten und Variablen, die den Status darstellen.
- Die zweite phase führt eine optimierte Version des Programms auf einer Menge verfügbarer Geräte aus (z.B. Server, CPU, GPU).

Verabeitungseinheiten

- Das "Programm" ist zunächst unabhängig von der verwendeten Verarbeitungsarchitektur
- Abstraktion für heterogene Beschleuniger: Neben Allzweckgeräten wie multicore-CPUs und GPUs können spezielle Beschleuniger für Deep Learning signifikante Leistungsverbesserungen und Energieeinsparungen erzielen.

Verabeitungsmodell und Parallelität

TensorFlow unterscheidet sich von batchbasierten Datenflusssystemen in zweierlei Hinsicht:

- Das Modell unterstützt mehrere gleichzeitige Ausführungen auf überlappenden Teilgraphen des Gesamtgraphen.
- Das Modell unterstützt verteilte Berechnung
- Einzelne Knoten des Graphes können einen veränderlichen Zustand haben, der zwischen verschiedenen Ausführungen des Diagramms geteilt werden kann.

Elemente des Datenflussgraphens

- In einem TensorFlow-Diagramm stellt jeder Knoten eine Einheit der lokalen Berechnung dar, und jede Kante repräsentiert die Ausgabe von oder Eingabe in einen Knoten.
  - Die Berechnung findet an Eckpunkten mit Operationen auf Werten statt, die entlang der Kanten als Tensoren fließen.

Tensoren

Alle Daten werden als monosortige Tensoren modelliert (n-dimensionale Arrays), wobei die Elemente vom primitiver Typ wie

int32, float32 oder string sind (wobei string Binärdaten darstellen kann).

Operatoren

Eine Operation nimmt m 0 Tensoren als Eingangswerte und erzeugt n 0 Tensoren als Ausgangswerte.

Zustansbasierte Operationen

Eine Operation kann einen veränderlichen zustand besitzen, der bei jeder Ausführung gelesen und/oder geschrieben wird. D.h. das Ergebnis einer Operation hängt von vorherigen Berechnungen ab!

Queues

TensorFlow unerstützt mehrere Warteschlangenimplementierungen, die Koordination von Berechnungen unterstützen.

12.2. tensforflow.js

- ➤ Tensorflow ist primär in C++ implementiert
  - ❑ Ausführung auf speziellen Verarbeitungseinheiten wie GPUs erfordert eine Kodeübersetzung zur Laufzeit
- ➤ Es gibt Python Anbindungen
- ➤ tensorflow.js ist eine Implementierung von Tensorflow rein in JavaScript und kann in Browsern ausgeführt werden
  - ❑ GPU Nutzung z.B. über Open/WEBGL
- ➤ https://www.tensorflow.org/js

12.3. Nachteile von Tensorflow

- ➤ Durch die Trennung von Daten, Kode, und Verarbeitungseinheiten schlechte Performanz bei “kleinen” Problemen (hohe Initialisierungszeit und Overhead)
  - ❑ Beispiel Lernen des EXOR Problems: Tensorflow.js benötigt ca 3-5 Minuten für das Training, Neataptic.js nur 100ms!
- ➤ Durch primäre Matrixalgebra nicht flexibel anpassbar
  - ❑ Z.B. kaum Möglichkeit der Verwendung von evolutionären Algorithmen (Dynamische Restrukturierung des KNN)

- ➤ Lernkurve ist am Anfang flach

12.4. Neataptic

- ➤ Ebenfalls ein KNN Framework,
  - ❏ Reine JavaScript Implementierung
  - ❏ Basiert NICHT auf Matrixalgebra; die Neuronen werden einzeln berechnet
- ➤ Aber mit (optionalen) evolutionären Algorithmen
  - ❏ Das Training kann neben Kantengewichten auch die Struktur ändern: Knoten und Kanten können getauscht, entfernt, oder hinzugefügt werden!
- ➤ Für kleine Probleme gute Perfomanz
- ➤ Lernkurve ist "steil"
- ➤ Bietet eine Vielzahl von Berechnungsfunktionen und Strukturen
  - ❏ Struktur kann beliebig und irregulär programmiert werden
  - ❏ Es gibt "Architekten" für viele gängige Architekturen
  - ❏ Forwärtsgekoppelte und rückgekoppelte Architekturen
- ➤ https://wagenaartje.github.io/neataptic/

12.5. Torch

- ➤ Torch ist eine open-source-Bibliothek für maschinelles lernen, ein scientific computing framework und eine Skriptsprache, die auf der Programmiersprache Lua basiert.
  - ❏ Daher kann mit einer einfach zu erlernden Programmiersprache direkt auf Torch gearbeitet werden
- ➤ http://torch.ch
- ➤ Torch gibt es auch für *R*!
  - ❏ https://torch.mlverse.org
- ➤ Lernkurve ist "steil"

12.6. ML

- Eigenes Framework (Stefan Bosse) dass eine Vielzahl von Lernalgorithmen und Modellen zusammenfasst:
  - Entscheidungsbäume (C4.5, ID3, ICE, usw.)
  - KNN (ANN → Neataptic, MLP → SMO, usw.)
  - Belohnungslernen (Reinforcement und Agenten L.)
  - kNN
  - Clustering (SOM)
  - Textanalyse
  - ...
- Reine JavaScript Implementierung
- Trennung von Daten und Algorithmen (d.h. prozedurale Progarmmierung)
  - Viele Modelle sind portabel, d.h. können direkt mit `JSON.stringify` und `JSON.parse` serialisiert und deseralisiert werden
- Enthalten in bzw. verfügbar für verschiedene Software Frameworks:
  - WorkBook
  - NoteBook (digitale Übungen)
  - JAM (JavaScript Agent Machine)

12.7. Beispiele

/webwork

data

```
var x = [
  [0,0],
  [0,1],
  [1,0],
  [1,1]
]
var y = [
  [0],
  [1],
  [1],
  [0]
]
ML.log(function () { print(inspect(arguments)) })
```

model

```
model=ML.learner({
  algorithm: ML.ML.MLP,
  layers : [2,4,1],
  verbose: 1,
});
```

train

```
ML.train(model,{
  x:x,
  y:y,
  epochs:1000
});
```

test

```
var results=ML.predict(model,x);
print(results)
print(ML.statistics.crossEntropy(ML.math.vec2Mat(results),y));
```

12.8. Zusammenfassung

➤ Es wird zwischen Klassifikatoren (kategorische Zielvariablen) und Prädiktorfunktionen (numerische Zielvariablen) unterschieden

- ❏ Klassifikation mit Prädiktorfunktionen erfordert Kodierung und Dekodierung!

➤ Die Fehleranalyse während und nach dem Training kann mit den statistischen Größen "Mean Squared Error" und Kreuzentropie erfolgen

➤ Die Konfiguration der inneren Schichten von KNN ist schwierig und häufig ein iterativer Prozess

13. Daten- und Dimensionalitätsreduktion

Datenreduktion ist ein wichtiger Schritt in der Datenvorverarbeitung für ML
Ziel: Reduktion der Datenvariablen (Attribute) → Dimensionalitätsreduktion pro Instanz
Ziel: Reduktion der Dateninstanzen (durch kleine Anzahl von Repräsentanteninstanzen) → Datenvolumenreduktion

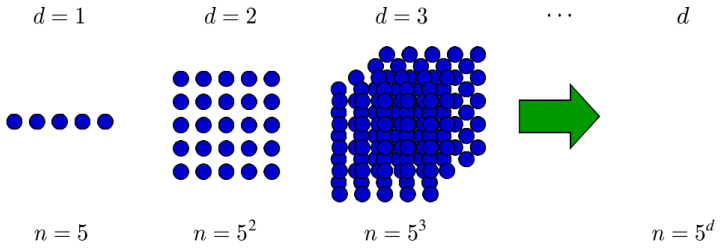
13.1. Motivation für Datenreduktion

1. Die Daten sind sowohl hinsichtlich Dimensionalität des Eingabevektors $\boldsymbol{x}$ als auch hinsichtlich der Anzahl von Dateninstanzen $|\mathbf{D}|$ sehr groß
2. Es gibt Redundanzen
   - ➤ a. Von Datenvariablen (lineare Abhängigkeit)
   - ➤ b. Von Dateninstanzen (Redundanz und Überlappung)
   - ➤ Aber: Reduktion bei b. kann die geforderte Datenvarianz verschlechtern!
3. Trennung von wenig aussagekräftigen (schwachen) von aussagekräftigen (starken) Variablen

Wenn die Dimensionalität der Eingabedaten $\boldsymbol{x}$ zunimmt, wird jedes Lernproblem immer schwieriger und rechenintensiver!

➤ Beispielsweise werden in regelmäßigen Abständen 5 Punkte von [0,1] abgetastet.

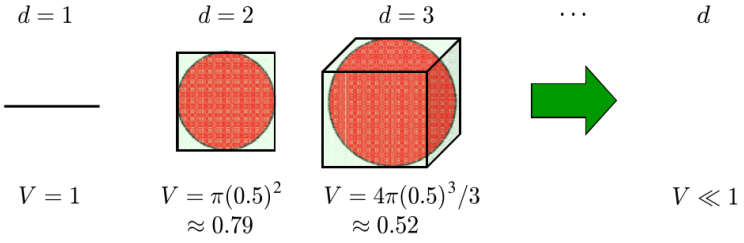
- ❏ Das sammeln von Proben auf die gleiche Weise im d-dimensionalen Raum erfordert $5d$-Punkte, die exponentiell in Bezug auf d wachsen.



[4]

Ein weiteres Problem bei hochdimensionalen Daten ist, dass unsere geometrische Interpretation von Daten irreführend sein kann.

- Betrachtet man zum Beispiel die ausfüllende Hypersphäre des Einheishyperwürfels im d-dimensionalen Raum
  - Wenn $d = 1$ ist, ist das Volumen einer eingepassten Hypersphäre 1, was dem Hyperwürfel entspricht.
  - Wenn d auf 2 und 3 erhöht wird, ist das Volumen des Hyperkubus immer noch 1, aber das Volumen der Hypersphäre ist ungefähr 0,79 bzw. 0,52.



[4]

- Unsere geometrische Intuition ist falsch,
  - da obwohl die eingepasste Hypersphäre kleiner als der Hyperwürfel ist, ist die Hypersphäre nicht extrem klein.
  - Wenn jedoch d weiter erhöht wird, ist das Volumen des Hyperkubus immer noch 1, aber das Volumen der ausfüllenden Hypersphäre neigt zu 0.
  - Dies bedeutet, dass, wenn d groß ist, die Hypersphäre fast vernachlässigbar ist.

13.2. Verfahrem und Methoden

Lineare Algebra

- Bestimmung von Eigenvektoren und Hauptkomponenten mit der **Principle Component Analysis**
  - Abbildung des Eingabedatenraums auf einen niedrigdimensionaleren Datenraum **unter Beibehaltung der Datenrepräsentation**
  - Ziel: Reduktion der Attribute, Beibehaltung der Dateninstanzen

Clustering

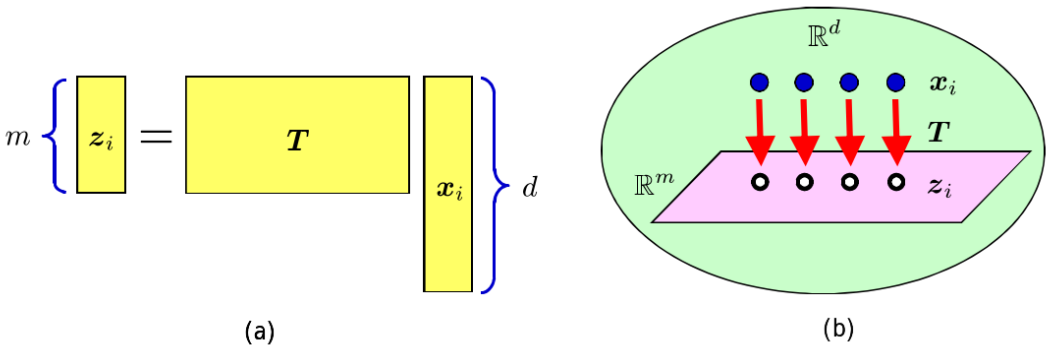
- Bestimmung von Gruppen von Dateninstanzen (mit geometrischer Gemeinsamkeit durch "Nähe") durch dichtebasierte Clusteringverfahren (**DBSCAN**)
  - Ziel: Reduktion der Instanzmenge durch wenige repräsentative Instanzen; Beibehaltung der Datenvariablen
  - Repräsentative Instanzen können (aber müssen nicht) mehrheitlich die Instanzen der Gruppe mit einem bestimmten Wert der Zielvariable (sofern kategortisch oder intervallkodiert) verknüpfen
  - Hohe Zielvariablenwertdiversität in Gruppen zeigt schwache Korrelation von $\boldsymbol{x}$ mit $\boldsymbol{y}$!

13.3. Lineare Dimensionalitätsreduktion

Lineare Dimensionalitätsreduktion transformiert die ursprünglichen d-dimensionalen Dateninstanzen $\{x_\mathrm{i}\}^\mathrm{n}$ in nieder m-dimensionale Ausdrücke $\{z_\mathrm{i}\}^\mathrm{n}$ durch eine lineare Transformation $\boldsymbol{T}^{m\times d}$

$$z_i = \mathbf{T}x_i$$

- $\boldsymbol{T}$ ist die Transformationsmatrix, z der reduzierte Datenvektor (pro Dateninstanz i) und x der ursprüngliche Datenvektor.



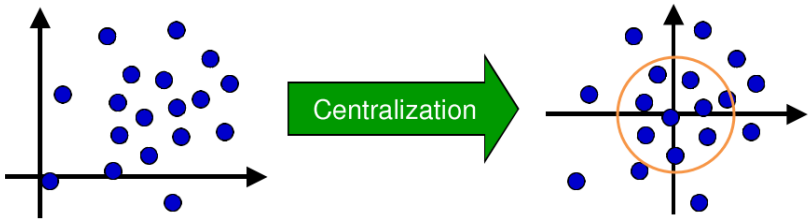
[4]

Abb. 40. (a) Lineare Dimensionalitätsreduktion (b) Projektion von Daten auf einen linearen Subraum

Zentrierung von Daten

➤ Verschiebung der Daten in Richtung "Koordinatenursprung"

$$x_i \leftarrow x_i - 1/n \sum_j x_j$$

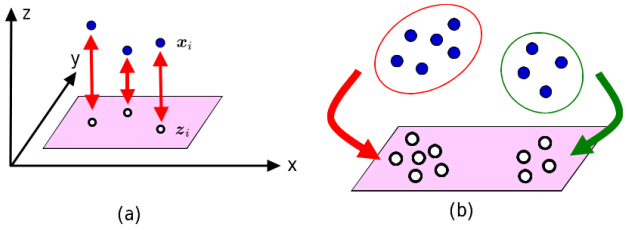


[4]

13.4. Unüberwachte Dimensionsreduktion

Hauptkomponentenanalyse (PCA)

➤ Reduktion der Datendimesionalität unter Beibehaltung der intrinsischen Information der Daten und der Datenstruktur



[4]

Abb. 41. PCA: (a) Reduktion der Dimensionalität 3 → 2 unter Beibehaltung der ursprünglichen geometrischen Eigenschaften der Punkte zueinander (Position) und der Gruppenzugehörigkeit (b)

- Das bedeutet dass $\boldsymbol{z}_\mathrm{i}$ eine **orthogonale Projektion** von $\boldsymbol{x}_\mathrm{i}$ ist
  - D.h. $\boldsymbol{T}\ \boldsymbol{T}^\mathrm{T} = \boldsymbol{I}_\mathrm{m}$ mit $\boldsymbol{I}_\mathrm{m}$ als Identitätsmatrix und a^T die transponierte Matrix
- Die "Distanz" zwischen $\boldsymbol{x}$ und $\boldsymbol{z}$ kann aber (als Fehler) nicht unmittelbar bestimmmt werden
  - Daher wird $\boldsymbol{z}$ wieder in den ursprünglich d-dimensionalen Datenraum durch $\boldsymbol{T}^\mathrm{T}$ zurück transformiert
- Die euklidische Distanz ist dann gegeben durch die totale statistische Streumatrix $\boldsymbol{C}$ und der Spur einer Matrix *tr*:

$$\sum_j ||\mathbf{T}^T\mathbf{T}x_i - x_i||^2 = -tr(\mathbf{TCT}^T) + tr(\mathbf{C})$$

- mit:

$$\mathbf{C} = \sum_j x_j x_j^T$$

- D.h. PCA ist ein Optimierungsproblem mit:

$$\max_{\mathbf{T}} tr(\mathbf{TCT}^T)$$

- Es gibt eine globale Lösung:

$$\mathbf{T} = (\mathbf{e}_1, ..., \mathbf{e}_m)^T$$

- mit e_i als Eigenvektor der Matrix $\boldsymbol{C}$ mit den Eigenwerten $_1 \quad _2 \quad .. \quad _\mathrm{d}$ 0.
- D.h. es gilt:

$$\mathbf{Ce} = \lambda \mathbf{e}$$

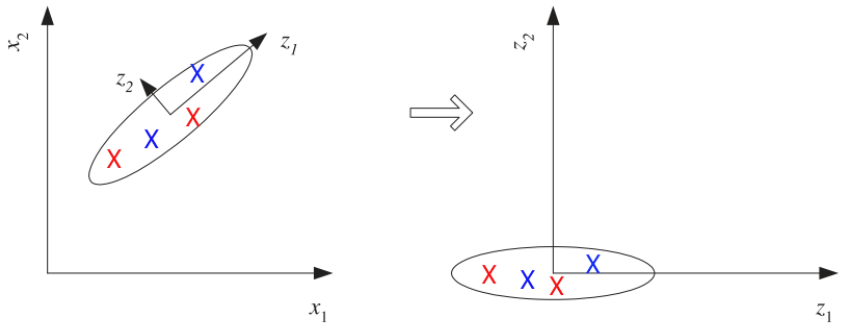
➤ Es wird “kleine” und “große” Eigenvektoren geben

*Die Transformationsmatrix **T** der PCA ist also eine orthogonale Projektion in einen Subraum der durch die “großen” Eigenvektoren aufgespannt wird,*

➤ Die kleinen Eigenvektoren werden daher entfernt

➤ Und: PCA transformierte Variablen sind unkorreliert!

➤ PCA liefert aber i.A. keine Struktureigenschaften wie Cluster



[17]

Abb. 42. Weiteres Beispiel von PCA: Die Analyse der Hauptkomponenten zentriert die Dateninstanzen und dreht dann die Achsen, um Sie mit den Richtungen der höchsten Varianz in Einklang zu bringen. Wenn die Varianz auf z_2 klein ist, kann Sie ignoriert werden und wir haben eine Dimensionalitätsreduktion von zwei auf eins.

13.5. ML.pca

➤ Das ML Framework stellt ein PCA Modul zur Verfügung:

➤ Die Daten ***D*** müssen in Matrixform vorliegen (keine Recordtabellen)

➤ Eigenvektoren berechnen

```
ML.pca.getEigenVectors = function(data:number [][]) → {
  eigenvalue: number, vector: number []
} [];
```

2. Transformation der Datentabelle berechnen

```
ML.pca.computeAdjustedData =
  function (data:number [][],
            eigenVector1:{},eigenVector2?:{},..)
→ {
  adjustedData: number [][],
  formattedAdjustedData: number [][],
  avgData: number [][],
  selectedVectors: number [][]
}
```

➤ Achtung: Das Datenformat von *adjustedData* ist transponiert und muss für eine Tabelle rekonstruiert werden

3. Rekonstruktion der Datentabelle aus der reduzierten Tabelle

```
ML.pca.computeOriginalData = function(
  formattedAdjustedData,
  selectedVectors,
  avgData) → {
  formattedOriginalData,
}
```

13.6. PCA Beispiel

/webwork

data

```
data = [
  {
    "length":5.1,
    "width":3.5,
    "petal_length":1.4,
    "petal_width":0.2,
    "species":"setosa"
  },
  {
    "length":4.7,
    "width":3.2,
    "petal_length":1.3,
    "petal_width":0.2,
    "species":"setosa"
  },
  {
    "length":4.6,
    "width":3.1,
    "petal_length":1.5,
    "petal_width":0.2,
    "species":"setosa"
  },
  {
    "length":5,
    "width":3.6,
    "petal_length":1.4,
    "petal_width":0.2,
    "species":"setosa"
  },
  {
...
```

eigen

```
dataX=ML.preprocess(data,'m',
 {features:
    ['length','width','petal_length','petal_width']
}).data;
eigenX=ML.pca.getEigenVectors(dataX);
```

adjust

```
dataZ=ML.pca.computeAdjustedData(dataX,eigenX[0],eigenX[1]).adjustedData;
```

reconstruct

```
dataXR=ML.pca.computeOriginalData(
  dataZ.formattedAdjustedData,
  dataZ.selectedVectors,
  dataZ.avgData).formattedOriginalData;
```

13.7. Lokalitätsbewahrende Projektion

➤ Ähnlichkeit zwischen Instanzen $\boldsymbol{x}_\mathrm{i}$ und $\boldsymbol{x}_\mathrm{j}$ wird untersucht und mit einem Wert 0 $W_{\mathrm{i,j}}$ 1 ausgedrückt.

Ähnlichkeitsfunktionen

Gaußfunktion

$$W_{i,j} = exp(-\frac{||x_i - x_j||^2}{2t^2})$$

(t ist ein Anpassungsparameter)

k-Nächster Nachbar (kNN)

$$W_{ij} = \begin{cases} 1, (x_i \in N_k(x_j) \vee x_j \in N_k(x_i)) \\ 0 \end{cases}$$

➤ $N_\mathrm{k}(\boldsymbol{x})$ ist die Menge aller Nachbarschaftsinstanzen (k ist die Anzahl der Menge)

➤ k ist ein Anpassungsparameter der die Lokalität einstellt (Reichweite)

Transformation

$$\min_T \sum_{i,j} W_{ij} ||\mathbf{T}x_i - \mathbf{T}x_j||^2$$

➤ Problem: $\boldsymbol{T}$=0 ist eine Lösung, aber nutzlos!

➤ Daher weitere Randbedingung für die Minimierung (Anpassung von $\boldsymbol{T}$):

$$\mathbf{TXDX}^T\mathbf{T}^T = \mathbf{I}_m$$

$$\mathbf{X} = (x_1, .., x_n), \mathbf{D}_{ij} = \begin{cases} \sum_k W_{ik}, (i = j) \\ 0, (i \neq j) \end{cases}$$

13.8. Dichtebasiertes Clustering

- ➤ kNN kann ebenso als ein ML Clusterer mit Inferenz auf unbekannten eingesetzt werden
- ➤ Dichtebasierte Clusteringmethoden können primär für die Datenreduktion eingesetzt werden
  - ❑ Reduktion von einer Menge von Dateninstanzen auf Gruppen (wobei die Instanzen erhalten bleiben, aber in jeder Gruppe reduziert werden können)

DBSCAN

Dichtebasiertes Räumliches Clustering mit Rauschen (DBSCAN) ist ein Basisalgorithmus. Es kann Cluster unterschiedlicher Formen und Größen aus einer großen Datenmenge entdecken, die Rauschen und Ausreißer enthält.

[www.kdnuggets.com/2020/04/dbscan-clustering-algorithm-machine-learning.html]

Algorithmische Schritte für DBSCAN Clustering

- ➤ Der Algorithmus iteriert über alle Punkte indem er willkürlich einen Punkt im Datensatz aufnimmt (bis alle Punkte besucht wurden)
- ➤ Wenn es mindestens *minPoint* - Punkte in einem radius von bis zu dem Punkt gibt, werden alle diese Punkte als Teil desselben Clusters betrachet.
- ➤ Die Cluster werden dann erweitert, indem die Nachbarschaftsberechnung für jeden Nachbarpunkt rekursiv wiederholt wird.
- ➤ Eventuell gibt es Beschränkungen bezüglich der Dimension der Dateninstanzen (typisch 2: Clustering von Bildpunkten)

Wahl der Parameter

Jede Data Mining Aufgabe hat das Problem der Parameterwahl.

➤ Jeder Parameter beeinflusst den Algorithmus auf bestimmte Weise. Für DBSCAN werden die Parameter und *minPts* benötigt.

minPts

Das Minimum von *minPts* kann aus der Dimension des Datensatzes ***D** abgeleitet werden, so dass minPts D + 1 gilt.

Der Wert für kann dann unter Verwendung eines k-Entfernungsgraphen ausgewählt werden, der den Abstand zum $k = minPts$-1 nächsten Nachbarn zeichnet, der vom größten zum kleinsten Wert geordnet ist.

➤ Bei guten Werten von gibt es eine ausgewogene Verteilung der Dateninstanzen in Gruppen

- ❏ Wenn viel zu klein gewählt wird, wird ein großer Teil der Daten nicht gruppiert, während für
- ❏ einen zu hohen Wert von Cluster zusammengeführt werden und sich die Mehrheit der Objekte im selben Cluster befinden.

Aufgrund minPts Randbedingung werden u.U. nicht alle Instanzen gruppiert!

Beispiel DBSCAN

/webwork

data

```
data = [
  {
    "length":5.1,
    "width":3.5,
    "petal_length":1.4,
    "petal_width":0.2,
    "species":"setosa"
  },
  {
    "length":4.7,
    "width":3.2,
    "petal_length":1.3,
    "petal_width":0.2,
    "species":"setosa"
  },
  {
    "length":4.6,
    "width":3.1,
    "petal_length":1.5,
    "petal_width":0.2,
    "species":"setosa"
  },
  {
    "length":5,
    "width":3.6,
    "petal_length":1.4,
    "petal_width":0.2,
    "species":"setosa"
  },
  {
...
```

pca1

```
dataX=ML.preprocess(data,'m',
 {features:
   ['length','width','petal_length','petal_width']
}).data;
eigenX=ML.pca.getEigenVectors(dataX);
```

pca2

```
dataZ=ML.pca.computeAdjustedData(dataX,eigenX[0],eigenX[1]).adjustedData;
```

clust1

```
dataZ2=dataZ[0].merge(dataZ[1],'c')
dbscan=new ML.DBCLUST.DBSCAN();
```

clust2

```
eps=0.3;
clusters=dbscan.run(dataZ2,eps,10);
```

13.9. Zusammenfassung

- ➤ Datenreduktion ist ein wichtiger Schritt in der Datenvorverarbeitung
  - ❑ Datenreduktion bedeutet die Reduktion der Datenmenge und/oder ihrer Dimensionalität
  - ❑ Datenreduktion ist bereits eine Merkmalsselektion (Reduktion auf wesentliche Attribute und Werte)
- ➤ PCA ist geeignet um numerische Eingabedaten in ihrer Dimensionalitä zu reduzieren (Reduktion der Datenvariablen und Ersatz mit transformierten)
- ➤ DBSCAN ist geeignet um Gruppen von Dateninstanzen zu finden um schliesslich repräsentative Instanzen daraus zu ermitteln

14. Probalistisches Lernen

14.1. Wahrscheinlichkeiten und Bayes Regel

[Witten, DMPMLTaT, pp. 335]

- ➤ In einem probalistischen Ansatz sind die Dateninstanzen gemessene Ereignesse oder Beobachtungen.
  - ❑ Die Dateninstanzen in $\boldsymbol{D}$ bilden Zufallsvariablen ab!

Es sei A eine Zufallsvariable mit diskreten Werten $\{a_\mathrm{i}\}$. Dann ist $P(A)$ oder kurz $P(a)$ die Wahrscheinlichkeitsfunktion für das Auftreten eines a_i A!

Es sei x eine Zufallsvariable mit kontinuierlichen Werten $[v_0,v_1]$. Dann ist $p(x)$ die Wahrscheinlichkeitsverteilung der Werte $x \quad [v_0,v_1]$.

Dann ist $p(x{=}x_\mathrm{i})$ die Wahrscheinlichkeit des Auftretens des Wertes x_i von x.

➤ Besondere Rolle nehmen binäre Ereignisse ein (also $A{=}\{0,1\}$). Etwas tritt ein oder ist wahr oder nicht.

Wenn A und B diskrete Zufallsvariablen sind, dann kann man über eine Produktregel die gemeinsame (vereinte) Wahrscheinlichkeit für das Auftreten von A und B bestimmen:

$$P(A,B) = P(A|B)P(B)$$

Die gemeinsame Wahrscheinlichkeit ist ein statistisches Maß, das die Wahrscheinlichkeit berechnet, dass zwei Ereignisse zusammen und gleichzeitig auftreten. Gemeinsame Wahrscheinlichkeit ist die Wahrscheinlichkeit, dass Ereignis B gleichzeitig mit Ereignis A Auftritt.

➤ $P(A)$ ist die Wahrscheinlichkeit eines Ereignisses A

➤ $P(A|B)$ ist die bedingte Wahrscheinlichkeit von A mit dem bedingten Ereignis B (d.h. B muss eintreten damit auch A eintritt): $B \rightarrow A$

➤ $P(B|A)$ ist dann das "inverse Problem": $A \rightarrow B$

Bayes Regel (Umkehr der Schlussfolgerung)

➤ Interessant wäre es zu wissen wenn $P(A|B)$ bekannt wie es mit $P(B|A)$ aussieht:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

14.2. Ein Beispiel: Der Mythos des Infektionstests

➤ Es gibt einen Test auf eine Virusinfektion

- ❑ Dieser wird im Labor getestet: 100 Proben Klasse negativ (kein Virus, $\neg V$), 100 Proben Klasse positiv (mit Virus, V)
- ❑ Der Test zeigt T (positiv) an, ansonsten $\neg T$ (negativ)

❑ Die Analyse der Testexperimente zeigt: TP=99, FP=2, FN=1, TN=98

Sensitivität

$$P(T|V) = TP/(TP+FN) = 99/(99+1) = 0.99 (CV19 : 0.5, 0.7-0.9, Good20)$$

Spezifizität

$$P(\neg T|\neg V) = TN/(TN+FP) = 98/(98+2) = 0.98 (CV19 : 0.99, 0.999, Good20)$$

Genauigkeit

$$Accuracy = (TP + TN)/N = (99 + 98)/200) = 0.985$$

Präzision

$$Precision = TP/(TP + FP) = 99/(98 + 2) = 0.99$$

➤ Es gibt eine Vorbedingung (Vorwahrsch.) bei einer Testanwendung: Der Wahrscheinlichkeit einer Infektion P(V) wenn eine Stichprobe gemacht wird (also n=1). Diese wird mit P(V)=0.001 angenommen.

Anwendung der Bayeschen Regel

$$P(V|T) = \frac{P(T|V)P(V)}{P(T)},$$
$$P(T) = P(T, V) + P(T, \neg V) =$$
$$P(T|V)P(V) + (1 - P(\neg T|\neg V))(1 - P(V))$$

➤ Bei P(V)=0.001 (zufällige Stichprobe ohne Anlass und Differentialdiagnose) ergibt sich:

$$P(V|T) = \frac{0.99 \times 0.001}{0.99 \times 0.001 + (1 - \mathbf{0.98})(1 - 0.001)} = 0.047 \approx P(V)$$
$$= \frac{0.70 \times 0.001}{0.70 \times 0.001 + (1 - \mathbf{0.999})(1 - 0.001)} = 0.41$$
$$= \frac{0.50 \times 0.001}{0.50 \times 0.001 + (1 - \mathbf{0.99})(1 - 0.001)} = 0.047$$

$$P(\neg V|\neg T) = \frac{P(\neg T|\neg V)P(\neg V)}{P(\neg T)},$$
$$P(\neg T) = 1 - P(T) =$$
$$P(\neg T|\neg V)P(\neg V) + (1 - P(T|V))P(V)$$

➤ Bei P(V)=0.001 (zufällige Stichprobe ohne Anlass und Differentialdiagnose) ergibt sich:

$$P(\neg V|\neg T) = \frac{0.98 \times 0.999}{0.98 \times 0.999 + (1 - 0.99)0.001} = 0.999$$

14.3. Naiver Bayes Klassifikator

➤ Zurück zum Golfspielproblem!

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

[Witten]

Abb. 43. Die Ein- und Ausgabevariablen mit bedingter Verteilung (also Y|X)

Annahme: Alle Eingabevariblen $\boldsymbol{x}$*={Outlook, Temperature, Humidity, Windy} sind unabhängig.*

➤ Wenn es nun eine Messung E=$\boldsymbol{x}$ gibt (Evidenz) mit einer Hypothese vom Ergebnis H=y mit C_y=$\{yes|no\}$, dann gilt quasi als Training (wir kennen ja den Zusammenhang $E \quad H$): $P(E|H)$

➤ Es wird angenommen dass alle Variablen gleichwertig sind → unbekannte Modellannahme oder a-priori Wissen!

Nun gibt es ein weiteres unbekanntes Beispiel:

Outlook	Temperature	Humidity	Windy	Play
Sunny	Cool	High	True	?

$$P(y = yes) = 2/9 \times 3/9 \times 3/9 \times 3/9 \times 9/14 = 0.0053$$
$$P(y = no) = 3/5 \times 1/5 \times 4/5 \times 3/5 \times 5/14 = 0.026$$

➤ Es gilt also nach der Bayes Regel:

$$P(C = yes|E) = \frac{\Pi_i P(E_i|C = yes) \times P(C = yes)}{P(E)}$$
$$P(C = no|E) = \frac{\Pi_i P(E_i|C = no) \times P(C = no)}{P(E)}$$

➤ Das ist ein einfacher Klassifikator

Funktionale Beschreibung

Es gilt:

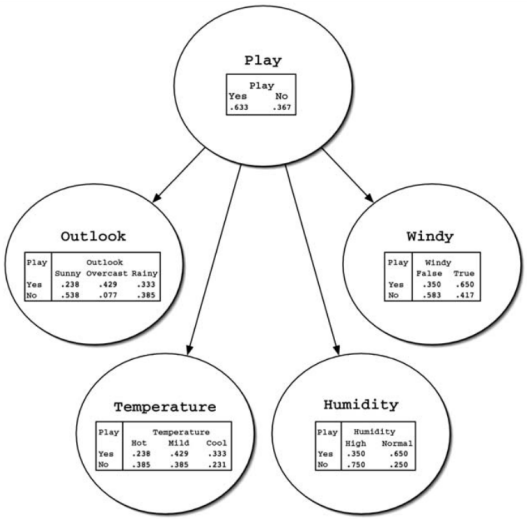
$$c \in C = h_{bayes}(x) = \arg\max_{j=1,m} P(c_j) \prod_{i=1,n} P(x_i|c_j)$$

mit c als eine Klasse aus allen möglichen Klassenwerten C der Zielvariable y und x als eine Dateninstanz.

14.4. Bayes Netzwerke

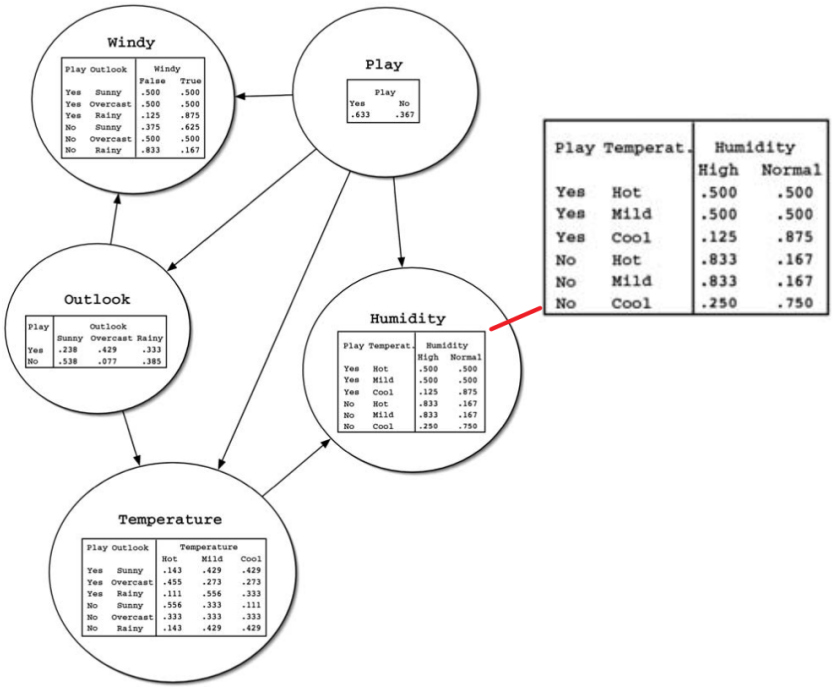
Naive Bayes Netzwerke bieten einen einfachen Ansatz mit klarer Semantik, um probabilistisches Wissen darzustellen, zu verwenden und zu lernen. Naiv Alle x_i sind unabhängig!

Ein bayesisches Netzwerk ist eine Form eines grafischen Modells zur Darstellung multivariater Wahrscheinlichkeitsverteilungen.



[Witten]

Abb. 44. Bayes Netzwerke sind eine Kombination aus Bäumen (gerichteten Graphen) und Wahrscheinlichkeitstabellen (Look-up Tabelle) die dann durch bedingte Wahrscheinlichkeiten das Ergebnis (der Hypothese von y) abschätzen.



[Witten]

Abb. 45. Ein anderes Bayes Netzwerk für das gleiche Problem!

➤ Für das gleiche Problem können verschiedene Bayes Netzwerke aufgebaut werden, die genau die gleiche Wahrscheinlichkeitsverteilung darstellen.

❑ Dies geschieht durch Änderung der Art und Weise, wie die gemeinsame Wahrscheinlichkeitsverteilung faktorisiert wird, um bedingte Unabhängigkeiten auszunutzen.

[Barber, BRaML, 2011]

Belief Netzwerke

Ein "Belief" Netzwerk ist eine Verteilung der Form:

$$p(x_1, x_2, .., x_D) = \Pi_{i=1}^{D} p(x_i|pa(x_i))$$

mit $pa(x_\mathrm{i})$ als die Elternvariable der Variablen x_i.

➤ Dargestellt als gerichteter Graph, wobei ein Pfeil von einer Elternvariablen auf eine untergeordnete Variable zeigt.

➤ Ein Belief Netzwerk ist ein gerichteter azyklischer Graph (DAG), wobei der i-te Knoten im Graph dem Faktor $p(x_\mathrm{i}|pa(x_\mathrm{i}))$ entspricht.

Übergang zu Markovschen Entscheidungsprozessen und Graphen!

14.5. Bayes Entscheidungslerner

Bayes Lerner

Es werden zwei Funktionen zum Erlernen der Netzwerke benötigt (Struktur und Berechnung):

1. Eine Funktion um ein gegebenes Netzwerk zu evaluieren
2. Eine Funktion um geeignete Netzwerke aus dem Raum aller möglichen Netzwerke zu finden (Suche)

[Barber, BRaML, 2011, pp. 288] [Witten, DMPMLTaT, pp. 335]

Ein einfacher und sehr schneller Lernalgorithmus ist **K2**

➤ Er beginnt mit einer bestimmten Reihenfolge der Attribute (d.h. Knoten).

➤ Dann verarbeitet er jeden Knoten der Reihe nach und wird Kanten von zuvor verarbeiteten Knoten zu dem aktuellen Knoten hinzuzufügen.

➤ In jedem Schritt wird die Kante hinzugefügt, die die Trefferwahrscheinlichekit (Score) des Netzwerks maximiert. Wenn es keine weitere Verbesserung gibt, wird der nächste Knoten bearbeitet.

➤ Als zusätzlicher Mechanismus zur Vermeidung von Überanpassungen kann die Anzahl der Eltern für jeden Knoten auf ein vordefiniertes Maximum beschränkt werden.

➤ Da nur Kanten von zuvor verarbeiteten Knoten berücksichtigt werden und es eine feste Reihenfolge gibt, kann dieses Verfahren keine Zyklen einführen.

Randomisiertes Sampling

➤ Das Ergebnis hängt jedoch von der anfänglichen Reihenfolge ab, daher ist es sinnvoll, den Algorithmus mehrmals mit unterschiedlichen zufälligen Ordnungen auszuführen.

Klassifikation und Bewertung

- ➤ Bei Entscheidungsbäumen gibt es keine Information wie "sicher" eine Klassifikation ist (Vorhersage- oder Vertrauenwahrscheinlichkeit), die aber häufig sehr wicht ist!
- ➤ Bei ANN ist der Ausgangswert eines Neurons zwar ein Indikator für die Aktivierungsstärke, ist aber modellbasiert keine Wahrscheinlichkeit der Vorhersage (höchstens grobe Näherung)
- ➤ Bei NB Klassifizieren gibt es als Ergebnis immer eine (statistisch) modellbasierte Wahrscheinlichkeit!

14.6. Anwendungen von Naiven Bayes-Algorithmen

Echtzeit-Vorhersage

Naive Bayesfunktionen sind ein einfacher und schneller Klassifikator und geben Wahrscheinlichkeiten aus für die Beurteilung. Somit könnte es für Vorhersagen in Echtzeit verwendet werden.

Multi-class-Vorhersage

Dieser Algorithmus ist auch für multi-class-Vorhersage-Funktion verwendbar. Hier können die Wahrscheinlichkeiten mehrerer Klassen von Zielvariablen vorhergesagt werden.

Textklassifizierung / Spam-Filterung / Stimmungsanalyse

Naive Bayes-Klassifikatoren, die hauptsächlich in der Textklassifizierung verwendet werden (aufgrund eines besseren Ergebnisses bei Problemen mit mehreren Klassen und der Unabhängigkeitsregel), weisen im Vergleich zu anderen Algorithmen eine höhere Erfolgsrate auf. Infolgedessen werden sie häufig in der Spam-Filterung (Identifizierung von spam-e-Mails) und in der Stimmungsanalyse (in der social-media-Analyse) verwendet, um positive und negative Kundenstimmungen zu identifizieren).

Empfehlungssystem

Der Naive Bayes-Klassifikator und die "kollaborative Filterung" erstellen zusammen ein Empfehlungssystem, das maschinelles Lernen und Data Mining Techniken verwendet, um versteckte Informationen zu filtern und vorherzusagen, ob ein Benutzer eine bestimmte Ressource möchte oder nicht.

14.7. Zusammenfassung

- ➤ Bayes Klassifikation erfolgt über die Berechnung bedingter und abhängiger Wahrscheinlichkeitsverteilungen

- ➤ Das "Training" liefert $P(x|y)$, die Inferenz benötigt die inverse Abhängigkeit und verwendet $P(y|x)$
- ➤ Ein naiver Bayes Klassifikator nimmt unabhänge Varibalen x_i an → Entspricht nicht immer "physikalischen Modellen und Kausalitäten"!
- ➤ Ein Bayes Klassifikator muss Tabellen mit bedingten Verteilungen/Wahrscheinlichkeiten verwenden.

15. Textanalyse

Natürliche Sprachverarbeitung ist ein komplexer Prozess!
Häufig ist man an einer Informationsgewinnung und Klassifikation aus Texten interessiert
Chat Bots benötigen Datenreduktion Text → Klasse/Merkmale/Topics/Keywords
In der Soziologie ist NLP+ML ein Werkzeug für die Textanalyse

15.1. Symbolische vs. Subsymbolische KI

- ➤ Symbolische KI ist regelbasiert!
  - ❑ Z.B. Textpassagen anhand von Musterphrasen identifizieren `"where is * next #Noun"` und Musteranpassung (quasi regulärer Sprachausdrücke aus einer Tabelle suchen)
- ➤ Subsymbolische KI ist Lernerbasiert!
  - ❑ Trainierbare ML Verfahren werden verwendet um aus Rohtexten wichtige Informationen abzuleiten

15.2. Beispiel für ein regelbasiertes NLP Dialog System

Eliza - Der digitale Psychotherapist

In den 1960er Jahren entwickelte Joseph Weizenbaum ein Programm namens ELIZA, welches es *Menschen ermöglichte mit einer Maschine in schriftlicher Form zu kommunizieren*, also einen **Chatbot** (Eliza Doolittle).

- ➤ Eliza kann verschiedene **Gesprächspartner simulieren**, unter anderem auch einen Psychologen.
  - ❏ Die "Einschränkung" ermöglicht die Eingrenzung von Frage-Antwort Mustern (also keine universellen Dialoge) → kein Wissen über die Welt erforderlich!
  - ❏ Konzept der Bottom-Up Kommunikation und "Superklassifizierung" (Oberbegriffe): Nutzer spricht vom Vater, Generalisierung auf Familie
  - ❏ Konzept der **Schlüsselwortsuche** und **Wörterbüchern** (Synonymsuche)!

ELIZA Demo

15.3. Natürliche Sprachverarbeitung und Verstehen

- ➤ Digitalisierung von natürlicher Sprache, Erkennung von Semantik, ..

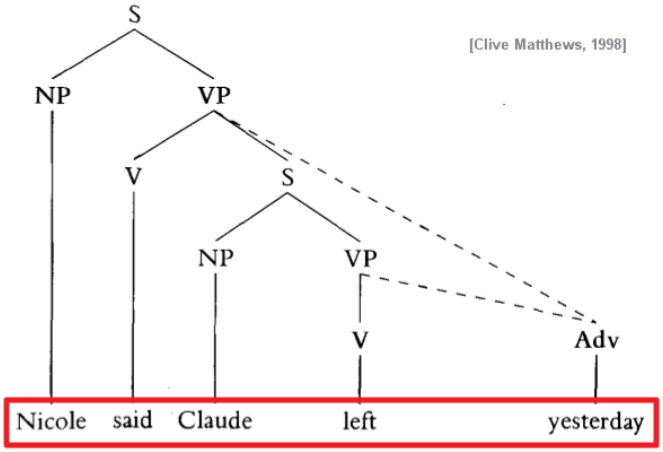
Mustervergleich

```
if (similarity('Die Vorlesung ist interessant',
               'Vorlesungen sind interessant')>0.5) then ..
```

Schlüsselwortsuche

```
if ('Die Vorlesung ist interessant'.contains('interessant')) then ..
```

Syntaxbäume



Neuronale Netze

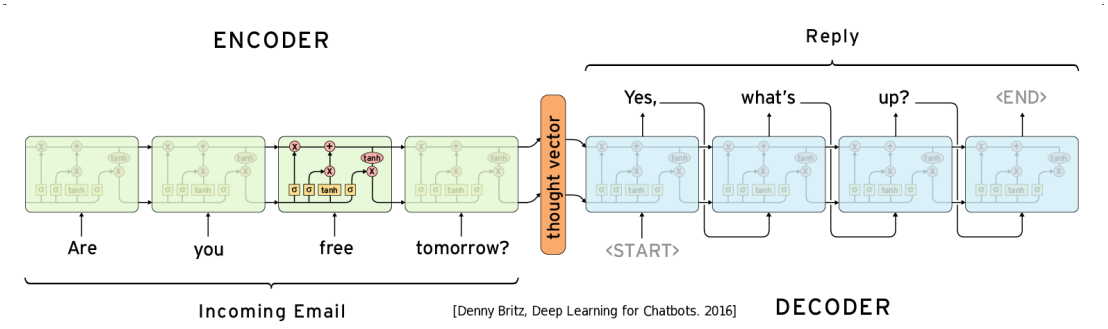
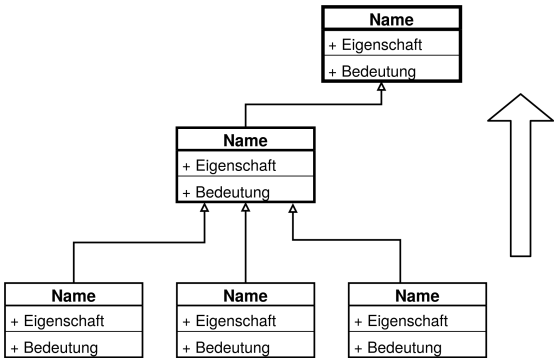


Abb. 46. Verwendung von zustandsbasierten rückgekoppelten Long-short-term Memory ANN für die Sprachanalyse und Synthese

Klassifizierung und "Superklassifizierung"

➤ Suchen von Schlüsselwörten und Ableiten von Themenbereichen (Topics)



15.4. NLP und ML

Menschliches Vokabular kommt im freier Textform. Damit ein maschinelles Lernmodell die natürliche Sprache versteht und verarbeiten kann, müssen die Freitextwörter in numerische Werte umgewandelt werden.

Einer der einfachsten Transformationsansätze besteht darin, eine One-Hot-Codierung durchzuführen, bei der jedes einzelne Wort für eine Dimension des resultierenden Vektors steht und ein Binärwert angibt, ob das Wort (1) darstellt oder nicht (0).

- ➤ Die One-Hot-Codierung ist jedoch rechnerisch unpraktisch, wenn es um das gesamte Vokabular geht, da die Darstellung Hunderttausende von Dimensionen erfordert.
- ➤ **Worteinbettung** stellt Wörter und Phrasen in Vektoren von (nicht binären) numerischen Werten mit viel niedrigeren und damit dichteren Dimensionen dar.

  - ❑ Eine intuitive Annahme für eine gute Worteinbettung ist, dass sie die Ähnlichkeit zwischen Wörtern nähern können.

  https://lilianweng.github.io/lil-log/2017/10/15/learning-word-embedding.html

15.5. Verfahren der Merkmalsselektion

- ➤ Count-Based Vector Space Model
- ➤ Context-Based: Skip-Gram Model
- ➤ Context-Based: Continuous Bag-of-Words (CBOW)

Skip-Gram Verfahren

- Das Skip-Gram Modell liefert Abschätzungen für Wort,Kontext Tupelpaare
- D.h. es gibt ein Wörterbuch mit diesen Tupeln für das Training für probalistische Wahrscheinlichkeitsfunktionen $P(V_\mathrm{c}|V_\mathrm{t})$ (c:Kontext, t:Text, Satz).
- In der Regel werden Satzfenster (Ausschnitte, z.B. jeweils 5 Worte) analysiert

Training des Skip-Gram Modells

https://blog.cambridgespark.com/tutorial-build-your-own-embedding-and-use-it-in-a-neural-network-e9cde4a81296

- Minimierung der folgenden Zielfunktion über eine Softmaxnormalisierung:

$$\sum_{(V_t,V_c)\in D} log P(V_c|V_t) = \sum_{(V_t,V_c)\in D} log \frac{e^{u_c}}{\sum_k e^{u_k}}$$

- Dabei quantifiziert u_c die Nähe eines Wortes zu seinem Kontext:

$$u_c = E_t \cdot O_c$$

mit E_t ist eine Matrix die Eingabeworteinbettungen enthält, und O_c entsprechend die Matrix für Ausgabeworteinbettungen.

15.6. Worteinbettung

[www.r-craft.org/r-news/get-busy-with-word-embeddings-an-introduction]

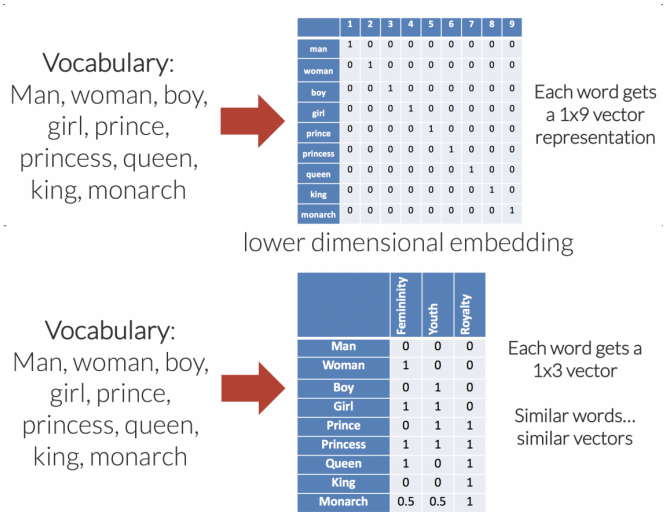
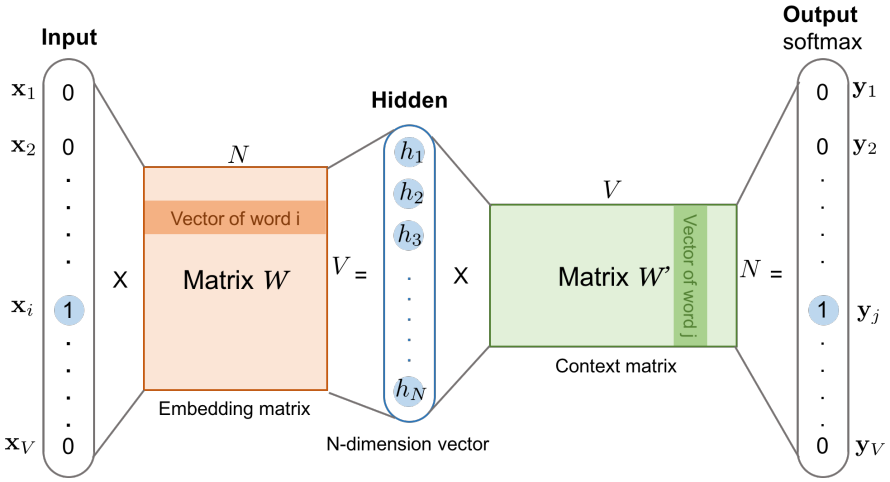


Abb. 47. Hoch- und niedrigdimensionale Worteinbettungsmatrizen

15.7. Word2Vec

➤ Ein neuronales Netzwerk welches sowohl Skip-Gram als auch Continuous Word-of-Bag Modelle lernt



Einfache Schlüsselwortkodierung

➤ Es gibt ein Wörterbuch mit Schlüsselwörtern $W=\{w_1,..,w_n\}$

- Jedes Schlüsselwort wird durche eine Stelle in einem n-dimensionalen Vektor repräsentiert
  - One-hot: Auftreten des Wortes im Satz=1, sonst 0
  - Count: Anzahl des Auftreten des Wortest im Satz
  - Semantik: Zuordnung zu semantischen Begriffen
- Der Kontext und die Stellung von Wörtern im Satz/Text werden nicht berücksichtigt

15.8. Zusammenfassung

- Textanalyse ist ein Werkzeug für qualitative und quantitative Methoden in der Soziologie
- Wichtig ist der Kontext und die Semantik von Wörtern!
- ML Verfahren könnnen helfen die Zuordnung Text-Kontext und Text-Semantik andhand von Beispielen zu erlernen
- Natürtliche Sprachverarbeitung ist komplex und wird ebenso von Dialogrobotern benötigt

16. Inverse Modellierung

Meistens kann ein Modell M(X):X → Y empirisch bestimmt werden
Häufig ist das inverse Modell von Bedeutung: M^{-1} Y: Y → X!
ML bietet Möglichkeit "Prädiktives Modellieren"
Aber wie kann man M^{-1} aus M ableiten?

16.1. Inverse Funktionen: Analytische und numerische Ableitung

- Gegeben sei eine Funktion $f(x)$: $\rightarrow$: $x \rightarrow y$, z.B.
  - $f(x)_1$: $y = x+a$
  - $f(x)_2$: $y = x^2+a$

- ❏ $f(x)_3$: $y = \sin(x)$

➤ Die Bestimmung der inversen Funktion kann häutig durch einfache algebraische Umformung generell und exakt berechntet werden:

- ❏ $f^{-1}(y)_1$: $x = y\text{-}a$
- ❏ $f^{-1}(y)_2$: $x = \{\ \sqrt{}(y\text{-}a), \text{-}\sqrt{}(y\text{-}a)\ \}$
- ❏ $f^{-1}(y)_3$: $x = \arcsin(y)$

➤ Schon bei der zweiten Funktion gibt es mehr als eine Lösung, und die inverse Sinusfunktion kann nicht exakt analytisch berechnet werden sondern benötigt eine **Approximation** durch eine **geometrische Reihe**:

$$arcsin(z) = \sum_{n=0}^{\infty} \frac{(2n-1)!!z^{2n+1}}{(2n)!!(2n+1)}$$

Inverse Probleme sind nicht trivial (zu lösen)!
Wie sieht es bei multivariaten Funktionen aus?

16.2. Multivariate Funktionen

➤ Eine Funktion f: $^{n} \rightarrow$ stellte ein Informationskompression dar;

- ❏ Aber i.A. als irreversible Reduktion (Informationsverlust)!

➤ Die Inversion einer Funktion f: $^{n} \rightarrow$ ergibt eine große Menge an Lösungen da:

- ❏ $f^{-1}(y)$: $\rightarrow$ ^{n} (Informationsexpansion bzw. Dekompression)
- ❏ Die Lösungsmenge kann unendlich groß sein!

➤ Beispiel: $f(x_1,x_2)$: y=x_1+x_2

- ❏ (Unendlich viele) Lösungen für y=0: **x**={ (0,0),(-1,1),(-2,2),(-3,3),… }

Einschränkung des Eingabe- und Lösungsraums

1. **Intervallarithmetik** → D.h. eine Variable x wird nur in einem Intervall $[a,b]$ betrachtet (und f)

2. **Diskretisierung** des Intervalls; $[a,b] \rightarrow \{\ a,\ a+\ ,\ a+2\ ,..,\ b\}$

Randbedingungslösen

➤ Wenn für alle Eingabevariablen und ebenso für die Zielvariable diskrete Werte in einem endlichen Bereich liegen könnten man das Inversionsproblem durch einen Randbedingungslöser (Constraint Solving Problem) lösen

❑ Aber auch dieser Ansatz liefert entweder viele oder nur eine Auswahl an Lösungen

➤ Beispiel eines Erfüllbarkeitsproblems über relationale Ausdrücke:

```
 1: Problem
 2: f(x1,x2) = y = x1+x2
 3: x1 = { 1,2,3 }
 4: x2 = { 1,2,3 }
 5: y  = { 2,3,..,6 }
 6: Randbedingungen
 7:   x1   1   x1   3
 8:   x2   1   x2   3
 9:   x1 = 1   x1 = 2   x1 = 3
10:   x2 = 1   x2 = 2   x2 = 3
11:   y = x1+x2
```

16.3. Das Single Layer Perceptron

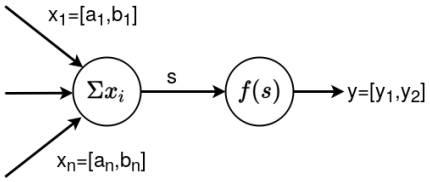


Abb. 48. Ein SLP (künstliches Neuron) besteht aus zwei verbundenen Funktionsblöcken (Summierer und Aktiverungsfunktion)

16.4. Inverses Problem ML: Naiver Lösungsansatz

Datentabelle

- ➤ Problem: Die Reversion der Datentabelle liefert einzelne Dateninstanzen
- ➤ Ziel: Das Modell M soll repräsentativ und generalisierbar sein
- ➤ Daher: Auch M^{-1} sollte möglich nur repräsentative Eingabevektoren liefern
  - ❏ Mittelwertbildung der Dateninstanzen und deren Variablen wenig hilfreich!
  - ❏ Zwei Instanzen: x=Sonnig, x=Regen x=(Sonnig+Regen)/2=wolkig???
  - ❏ Majoritäten könnten repräsentative Variablenwerte ergeben!?

16.5. Inverses Problem ML: Entscheidungsbaum

- ➤ Ein empririsch gelernter Entscheidungsbaum kann ein generalisiertes Modell sein $M(x)$: $x \rightarrow y$
- ➤ Die Invertierung geschieht durch Rückwärtsiteration startend bei allen Endknoten (Blättern) mit y_i=y
- ➤ Es wird i.A. mehr als eine Lösung geben (Repräsentanz?)
- ➤ Die Frage ist die Ableitung des resultierenden Variable x aus den Knoten
  - ❏ Bei kategorischen Variablen triviales und eindeutiges Problem
  - ❏ Bei numerischen Variablen und einem relationen Baum mit $N(x)$={ $x <$, x } ist gerade der Teilungswert nicht repräsentativ (Rand!!)

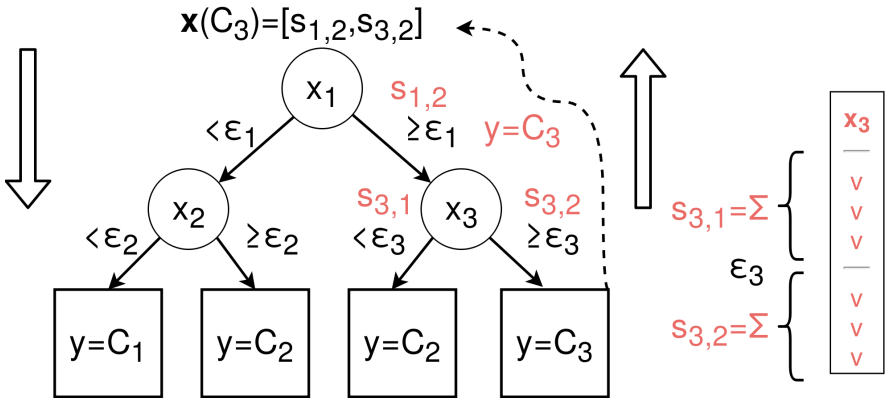
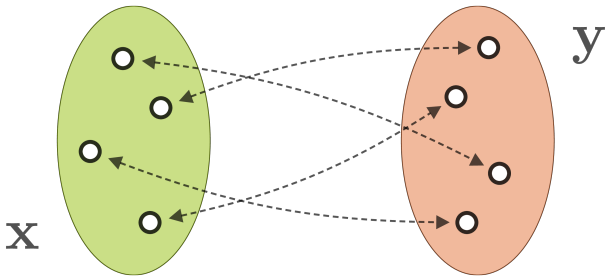


Abb. 49. Invertierung eines relationalen und annotierten Entscheidungbaumes (s: Mittelwert der Partition der Variabele)

16.6. Invertierbare ANN

https://hci.iwr.uni-heidelberg.de/vislearn/inverse-problems-invertible-neural-networks/

Ausgangspunkt: Ein- und Ausgabedaten besitzen die gleiche Dimension!



Invertierbare Netzwerkstruktur

➤ Es wird ein übergeordnetes reversibles Berechnunsnetzwerk eingeführt (Affine Kopplungsschicht)

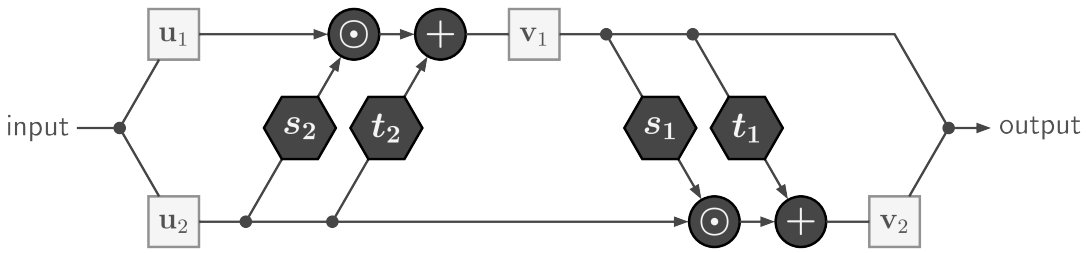


Abb. 50. Die Eingabedaten werden augespalten in [u1,u2] und durch die gelernten Funktionen s_i und t_i transformiert und in Wechselanordnung gekoppelt. Die Ausgabe ist die Verkettung der resultierenden Teile [v1,v2]. - Elementweise Multiplikation

Invertierung

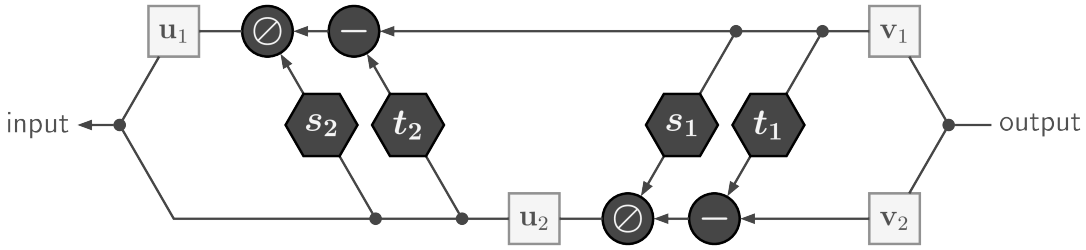


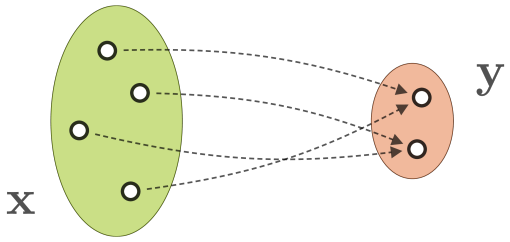
Abb. 51. Invertierung des Netzwerks. Mit einer Umschaltung können [u1,u2] aus [v1,v2] wiederhergestellt werden, um die Umkehrung der gesamten affinen Kopplungsschicht zu berechnen. - Elementweise Division

Entscheidend ist, dass die Transformationen s_i und t_i selbst ***nicht invertierbar*** *sein müssen und durch beliebige neuronale Netze dargestellt werden können, die durch standardmäßige Backpropagation entlang des Berechnungsgraphen trainiert werden.*

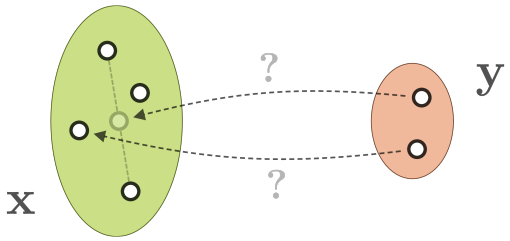
Mehrdeutige Abbildungen

D.h. die Eingabedimension ist wie üblich viel größer als die Ausgabedimension!

➤ Die Inversion erzeugt Mehrdeutigkeit bei der Abbildung $y \rightarrow x$.



Dimensionsreduzierende Abbildung $x \rightarrow y$



Mehrdeutigkeit zwischen y und x

Transformation in Bijektive Abbildungen

Eine zusätzliche latente Variable z wird eingeführt, die die Information erfasst, die sonst im Forward-Prozess verloren gehen würde. Folglich, x [y,z] wird eine bijektive Zuordnung.

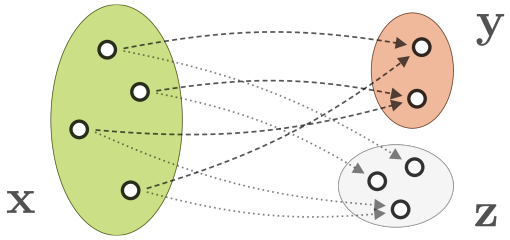


Abb. 52. Durch zusätzliche latente Variable z wird die inverse Abbildung

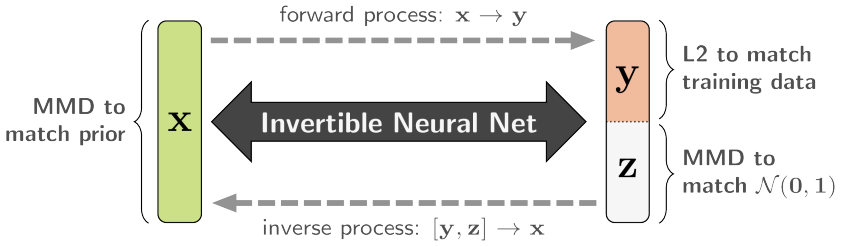


Abb. 53. z muss unabhängig von y sein und muss einer einfachen Stichprobenverteilung $N(0,1)$ folgen.

➤ Beide Bedingungen können mit einem maximalen mittleren Diskrepanzverlust (MMD) erreicht werden, die mit zwei Verteilungen übereinstimmt durch Vergleich von Stichproben.

Die Verteilung $p(x|y)$ kann angenähert werden, indem einfach wiederholt z abgetastet wird und die rückgerichtete Berechnung des Netzwerks durchgeführt wird, d.h. $[y,z] \rightarrow x$.

➤ Aus $p(x|y)$ wird in eine deterministische Funktion $x{=}f(y,z)$ mit der "verrauschten" Variable z.

16.7. Zusammenfassung

➤ Das Training von Vorwärtsmodellen ist ein Standardverfahren

➤ Häufig - gerade in der Soziologie - ist man an Rückwärtsmodellen interessiert, d.h. die Invertierung der aus empirischen Daten algorithmisch gelernten Modelle

➤ Die Inversion ist schwierig durch Mehrdeutigkeit der Abbildung

➤ Variablenintervalle und Wertdiskretisierung können das Inversionsproblem auf Randbedingsungslösen reduzieren und lösbar machen

➤ Inversion von ANN benötigt eine übgeordnete bidirektionale und umschaltbare Netzwerkstruktur

17. Referenzen

17.1. Bücher

1. M. J. Zaki and W. Meira, Data Mining and Machine Learning - Fundamental Concepts and Algorithms. Cambridge University Press, 2020.
2. C. R. Farrar and K. Worden, Structural Health Monitoring: A Machine Learning Perspective. Wiley-Interscience, 2013.
3. X.-S. Yang, Introduction to Algorithms for Data Mining and Machine Learning. Elsevier, 2019.
4. M. Sugiyama, Introduction to Statistical Machine Learning. 2016.
5. J. Herrmann, Maschinelles Lernen und Wissensbasierte Systeme. Springer, 1997.
6. R. Zafarani, M. A. Abbasi, and H. Liu, Social Data Mining. 2014.
7. I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining Practical Machine Learning Tools and Techniques. Morgan Kaufmann.
8. C. Borcea, M. Talasila, and R. Curtmola, Mobile Crowdsensing. CRC Press, 2017.
9. J. G. Webster, Measurement, Instrumentation and Sensors Handbook, no. 0. 1999.
10. L. Rokach and O. Maimon, Data Mining with Decision Trees - Theory and Applications. World Scientific Publishing, 2015.
11. N. J. Nilsson, Introduction To Machine Learning. 1996.
12. T. M. Mitchel, Machine Learning. McGraw Hill, 1997.
13. P. Attewell and D. B. Monaghan, Data mining for the social sciences : an introduction. University of California Press, 2015.
14. J. R. Quinlan, “Induction of Decision Trees,” in Machine Learning, Kluwer Academic Publishers, Boston, 1986
15. M. T. Hagan, howard B. Demuth, M. H. Beale, and O. D. Jesus, Neural Network Design.
16. L. Fausett, Fundamentals of Neural Networks. 1994.
17. E. Alpaydın, Introduction to Machine Learning. MIT Press, 2010.

17.2. Artikel

100. T. Mueller, A. G. Kusne, and R. Ramprasad, “Machine learning in materials science: Recent progress and emerging applications,” Reviews in Computational Chemistry, Volume 29, First Edition., 2016.

101. N.-C. Chen et al., “Challenges of Applying Machine Learning to Qualitative Coding.”

102. J. Radford, K. Joseph, “Theory In, Theory Out: The Uses of Social Theory in Machine Learning for Social Science”, Front. Big Data

103. Abadi et al., “TensorFlow: A system for large-scale machine learning”, 2th USENIX Symposium on Operating Systems Design and Implementation, USENIX Association