Ludwig-Maximilians-Universität München
Institut für Informatik
Lehr- und Forschungseinheit für Datenbanksysteme

Skript zur Vorlesung

Knowledge Discovery in Databases

im Sommersemester 2011

Kapitel 3: Klassifikation

Vorlesung+Übungen:
PD Dr. Matthias Schubert, Dr. Arthur Zimek

http://www.dbs.ifi.lmu.de/cms/Knowledge_Discovery_in_Databases_I_(KDD_I)

3. Klassifikation

Inhalt dieses Kapitels

3.1 Grundbegriffe der Klassifikation

Das Klassifikationsproblem

Gegeben: Eine Menge O von Objekten des Formats $(o_1, \ldots, o_d)$
mit *Attributen* A_i, $1 \leq i \leq d$, und Klassenzugehörigkeit c_i, $c_i \in C = \{c_1, \ldots, c_k\}$

Gesucht: die Klassenzugehörigkeit für Objekte aus $DB \setminus O$
ein *Klassifikator* $K : DB \rightarrow C$

Abgrenzung zum Clustering

Klassifikation: Klassen a priori bekannt

Clustering: Klassen werden erst gesucht

Verwandtes Problem: *Vorhersage* (Prediction)
gesucht ist der Wert für ein numerisches Attribut
Methode z.B. Regression (siehe Kapitel 4).

Einfacher Klassifikator

Beispiel

ID	Alter	Autotyp	Risiko
1	23	Familie	hoch
2	17	Sport	hoch
3	43	Sport	hoch
4	68	Familie	niedrig
5	32	LKW	niedrig

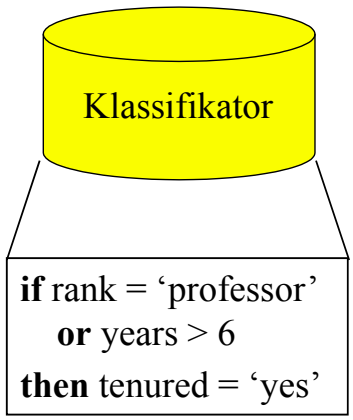
Einfacher Klassifikator

```
if Alter > 50 then Risikoklasse = Niedrig;
if Alter ≤ 50 and Autotyp=LKW then Risikoklasse=Niedrig;
if Alter ≤ 50 and Autotyp ≠ LKW
                        then Risikoklasse = Hoch.
```

Der Prozess der Klassifikation

Konstruktion des Modells

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

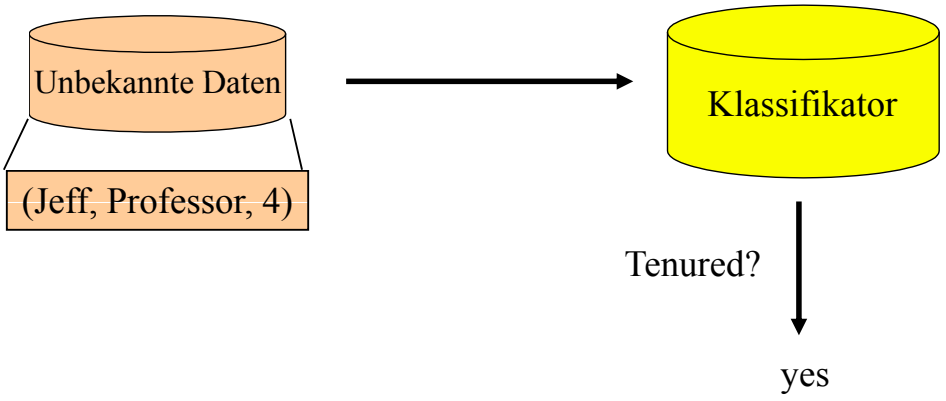


DATABASE SYSTEMS GROUP

Der Prozess der Klassifikation

LMU

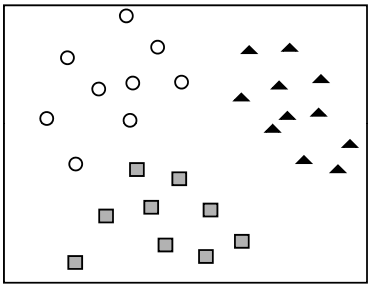
Anwendung des Modells



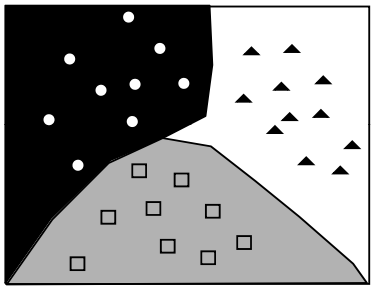
manchmal: keine Klassifikation unbekannter Daten sondern „nur“ besseres Verständnis der Daten

Überblick über Klassifikationsmethoden

Trainingsmenge mit 3 Klassen



3 Klassenbereiche (weiß, grau, schwarz)

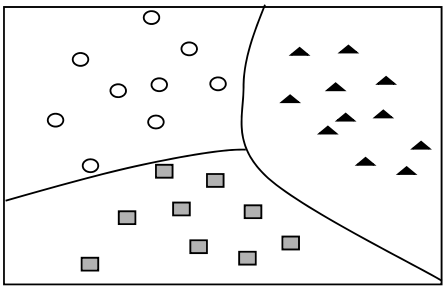


Klassifikatoren legen beim Training im Allgemeinen Klassengrenzen fest.

Aber: Es gibt viele Methoden, Klassengrenzen aus Trainingsdaten abzuleiten.

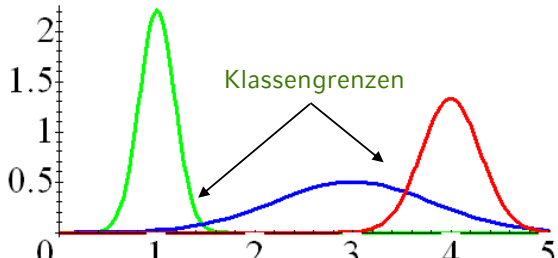
=> Unterschiedliche Klassifikatoren
(statische Kl., Entscheidungsbäume, Support Vektor Maschinen,
kNN-Klassifikatoren, neuronale Netze, ...)

Motivation der Klassifikationsmethoden(1)

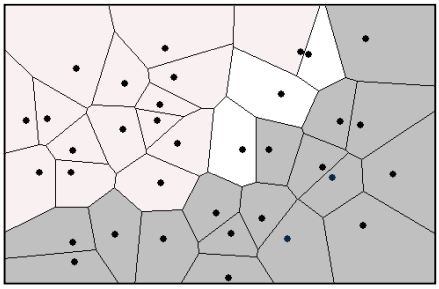


Bayes Klassifikatoren

1-dimensionale Projektion



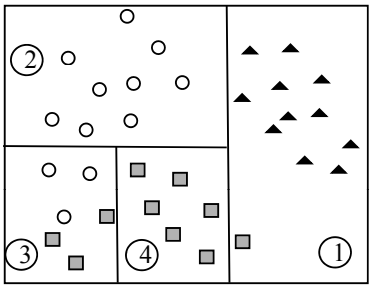
Unterscheidung durch Dichtefunktionen.



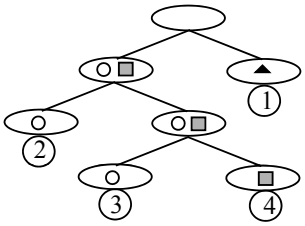
NN-Klassifikator

Unterscheidung durch Voronoi-Zellen
(1 nächster Nachbar Klassifikator)

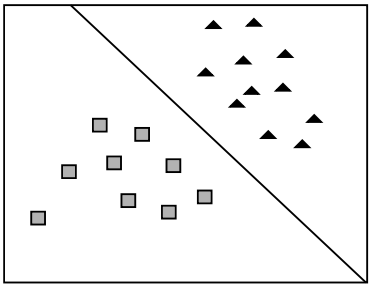
Motivation der Klassifikationsmethoden(2)



Entscheidungsbäume



Festlegen der Grenzen durch rekursive Unterteilung in Einzeldimension.



Support Vektor Maschinen

Grenzen über lineare Separation

Intuition und Grundannahmen

- Es gibt einen (natürlichen, technischen, sozial-dynamischen, …) (Entscheidungs-)Prozess (im statistischen Sinne), der die beobachtbaren Daten O als Teilmenge der möglichen Daten D erzeugt bzw. für ein $x \in D$ eine Klassenentscheidung für eine Klasse $c_i \in C$ trifft.
- Die beobachteten Daten sind Beispiele für die Wirkung des Prozesses.
- Es gibt eine ideale (unbekannte) Funktion, die einen Beispiel-Datensatz auf die zugehörige Klasse abbildet:
  $f: D \rightarrow C$
- Aufgabe des Lernalgorithmus ist es, eine möglichst gute Approximation h an f zu finden, eine Hypothese.

Intuition und Grundannahmen

- Beispiel:

$$f : Sky \times AirTemp \times Humidity \times Wind \times Water \times Forecast \rightarrow \{Yes, No\}$$

Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

- Gibt es ein generelles Konzept, wann Sport getrieben wird?
- Lernen braucht Annahmen (Bias), z.B., das Konzept ist eine Konjunktion ausgewählter Attributwerte.

DATABASE SYSTEMS GROUP

Intuition und Grundannahmen

LMU

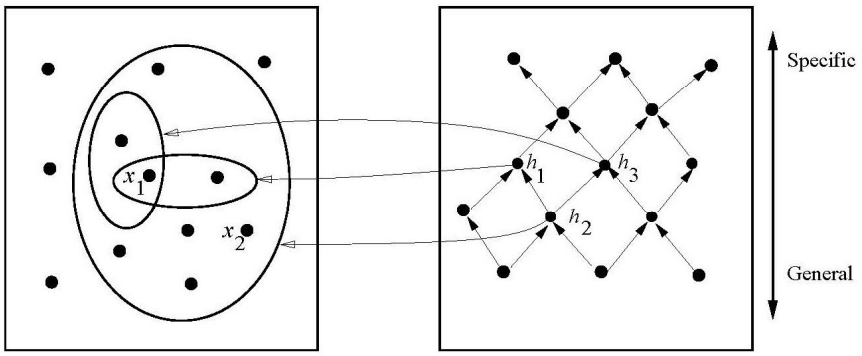
Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

- Mögliche Hypothesen für *Yes* unter dieser Annahme:

<Sunny,?,?,?,?,?>
<?,Warm,?,?,?,?>
<Sunny,Warm,?,?,?,?>
<Sunny,?,?,Strong,?,?>
<?,Warm,?,Strong,?,?>
<Sunny,Warm,?,Strong,?,?>

Intuition und Grundannahmen

- Durch die Annahmen eines Klassifikators wird der Raum möglicher Hypothesen (lernbarer Konzepte) definiert.



x_1 = <Sunny, Warm, High, Strong, Cool, Same>
x_2 = <Sunny, Warm, High, Light, Warm, Same>

h_1 = <Sunny,?,?,Strong,?,?>
h_2 = <Sunny,?,?,?,?,?>
h_3 = <Sunny,?,?,?,Cool,?>

- Sind komplexere Annahmen sinnvoll (erlaubt auch Disjunktionen, Negationen)?

Intuition und Grundannahmen

Konjunktive Hypothesen für *Yes*:

<Sunny,?,?,?,?,?>
<?,Warm,?,?,?,?>
<Sunny,Warm,?,?,?,?>
<Sunny,?,?,Strong,?,?>
<?,Warm,?,Strong,?,?>
<Sunny,Warm,?,Strong,?,?>

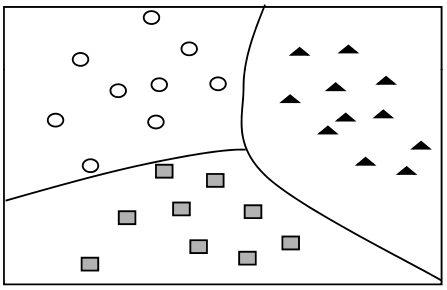
Neue Daten:

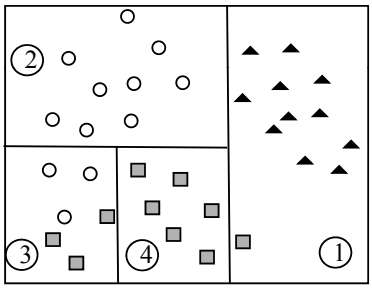
Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
Sunny	Warm	Normal	Strong	Cool	Change	Yes
Cloudy	Warm	Normal	Strong	Cool	Change	Yes
Rainy	Warm	Normal	Strong	Cool	Change	No

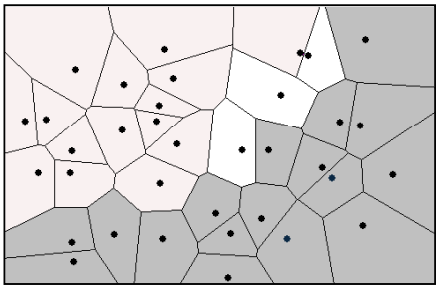
- Disjunktives Konzept: *if Sky=Sunny or Sky = Cloudy, then Yes.*
- Ermöglicht aber Auflistung aller Lernbeispiele als Hypothese.
- Allgemein: Ohne einschränkende Annahmen kann Auswendiglernen nicht mehr ausgeschlossen werden.

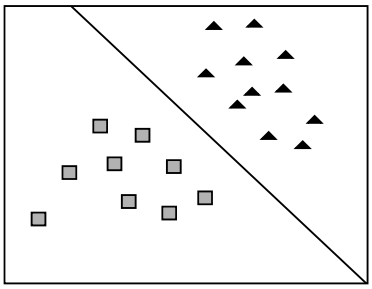
Intuition und Grundannahmen

Verschiedene Klassifikationsalgorithmen basieren auf verschiedenen Annahmen (unterschiedlicher Bias).









3.2 Bewertung von Klassifikatoren

Grundbegriffe

Sei *K* ein Klassifikator und sei $TR \subseteq O$ die Trainingsmenge. $O \subseteq DB$ ist die Menge der Objekte, bei denen die Klassenzugehörigkeit bereits bekannt ist .

Problem der Bewertung:

- gewünscht ist gute Performanz auf ganz *DB*.
- Klassifikator ist für *TR* optimiert.
- Test auf *TR* erzeugt in der Regel viel bessere Ergebnisse, als auf *DB\TR*.
  Daher kein realistisches Bild der Performanz auf *DB*.

  $\Rightarrow$ *Overfitting*

Bewertung von Klassifikatoren

Train-and-Test

Bewertung ohne *Overfitting* durch Aufteilen von *O* in :

- *Trainingsmenge TR*
  zum Lernen des Klassifikators (Konstruktion des Modells)
- *Testmenge TE*
  zum unabhängigen Bewerten des Klassifikators
- *Ziel: Abschätzen der Erfolg- bzw. Fehlerrate des Klassifikators*
  Daher: Test-Daten müssen unabhängig von Trainingsdaten sein
  Trainings- und Testdaten sollen das Problem angemessen wiederspiegeln
  (z.B. proportionale Anteile der verschiedenen Klassen)

DATABASE SYSTEMS GROUP

Bewertung von Klassifikatoren

LMU

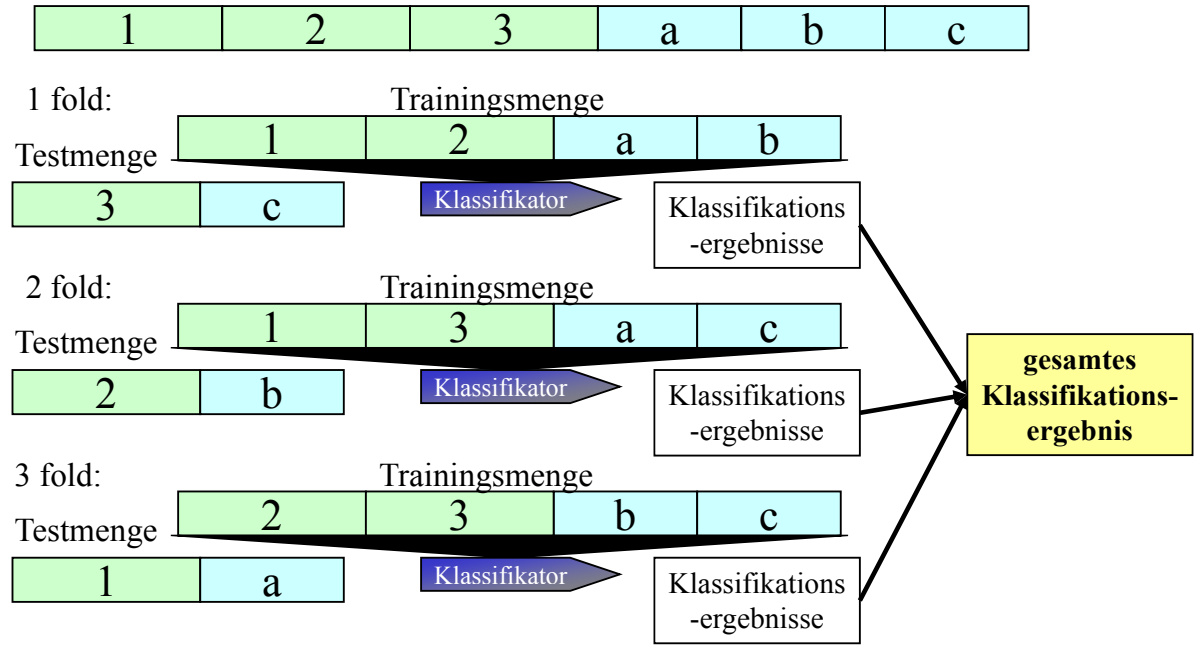
m-fache Überkreuz-Validierung

- sinnvolle manuelle Aufteilung in Trainings- und Testmenge nicht trivial
- Train-and-Test nicht anwendbar, wenn nur wenige Objekte mit bekannter Klassenzugehörigkeit vorhanden sind.
- Stattdessen: *m*-fache *Überkreuz-Validierung* (*m-fold Cross-Validation*)
  - teile die Menge *O* zufällig in *m* gleich große Teilmengen
  - verwende jeweils *m*–1 Teilmengen zum Training
    und die verbleibende Teilmenge zur Bewertung
  - kombiniere die erhaltenen *m* Klassifikationsfehler
  - Wiederhole das Verfahren mehrmals

Bewertung von Klassifikatoren

Ablauf 3-fache Überkreuzvalidierung (3-fold Cross Validation)

Sei n = 3 : Menge aller Daten mit Klasseninformation die zu Verfügung stehen



DATABASE SYSTEMS GROUP

Bewertung von Klassifikatoren

LMU

zusätzliche Anforderung: Stratifikation

- Proportionalität der Klassen erhalten
- zumindest: jede Klasse sollte in Trainingsmenge enthalten sein
- sinnvoll: Verteilung der Klassen sollte in Trainings- und Testmenge die Verteilung auf dem gegebenen Problem wiederspiegeln
- Standard: 10-fache, stratifizierte Kreuzvalidierung, 10-mal wiederholt (Erfahrungswerte)

Bewertung von Klassifikatoren

Alternative: Bootstrap

bilden einer Trainingsmenge aus einer gegebenen Datenmenge durch Ziehen mit Zurücklegen.

- jedes Sample hat die gleiche Größe wie die ursprüngliche Trainingsmenge
- ein Sample enthält durchschnittlich 63% der Ausgangsbeispiele (einige mehrfach, etwa 37% gar nicht):
  - ein einzelnes Beispiel in einem Datensatz mit *n* Beispielen hat bei jedem Ziehen die Chance *1/n* gezogen zu werden, wird also mit Wahrscheinlichkeit *1-1/n* **nicht** gezogen
  - nach *n*-mal Ziehen ist ein bestimmtes Element mit Wahrscheinlichkeit $\left(1-\frac{1}{n}\right)^n$ nicht gezogen worden
  - für große *n* ist $\left(1-\frac{1}{n}\right)^n \approx e^{-1} \approx 0.368$
- daher auch der Name "0.632 bootstrap" für diese Sampling-Methode

Im Allgemeinen eher optimistische Fehlerschätzung

DATABASE SYSTEMS GROUP

Bewertung von Klassifikatoren

LMU

Alternative: Leave-one-out (auch: Jackknife)

- Trainingsmenge wird gebildet durch Weglassen eines einzigen Elementes, dieses wird als Test-Objekt verwendet
- Verfahren wird wiederholt für alle Objekte der gelabelten Daten, Fehler-Abschätzung gemittelt

- Vorteil: kein Zufallselement
- Nachteil: garantiert nicht stratifiziert

- Im Allgemeinen eher pessimistische Fehlerschätzung

Bewertung von Klassifikatoren

Ergebnis des Tests : Konfusionsmatrix (confusion matrix)

klassifiziert als ...

tatsächliche Klasse ...	Klasse1	Klasse 2	Klasse 3	Klasse 4	other
Klasse 1	35	1	1	1	4
Klasse 2	0	31	1	1	5
Klasse 3	3	1	50	1	2
Klasse 4	1	0	1	10	2
other	3	1	9	15	13

korrekt klassifizierte Objekte

Aus der Konfusionsmatrix lassen sich u.a. folgende Kennzahlen berechnen : Accuracy, Classification Error, Precision und Recall.

DATABASE SYSTEMS GROUP

Bewertung von Klassifikatoren

LMU

Gütemaße für Klassifikatoren

Sei *K* ein Klassifikator, $TR \subseteq O$ die Trainingsmenge, $TE \subseteq O$ die Testmenge. Bezeichne *C(o)* die tatsächliche Klasse eines Objekts *o*, *K(o)* die von *K* vorhergesagte.

- *Klassifikationsgenauigkeit* (classification accuracy) von *K* auf *TE*:

$$G_{TE}(K) = \frac{|\{o \in TE \mid K(o) = C(o)\}|}{|TE|}$$

- *Tatsächlicher Klassifikationsfehler* (true classification error)

$$F_{TE}(K) = \frac{|\{o \in TE \mid K(o) \neq C(o)\}|}{|TE|}$$

- *Beobachteter Klassifikationsfehler* (apparent classification error)

$$F_{TR}(K) = \frac{|\{o \in TR \mid K(o) \neq C(o)\}|}{|TR|}$$

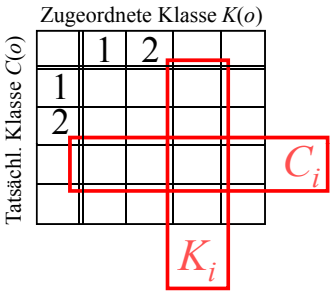
Bewertung von Klassifikatoren

Recall:

Anteil der Testobjekte einer Klasse *i*, die richtig erkannt wurden.

Sei $C_i = \{o \in TE \mid C(o) = i\}$, dann ist

$$\text{Recall}_{TE}(K, i) = \frac{|\{o \in C_i | K(o) = C(o)\}|}{|C_i|}$$



Precision:

Anteil der zu einer Klasse *i* zugeordneten Testobjekte, die richtig erkannt wurden.

Sei $K_i = \{o \in TE \mid K(o) = i\}$, dann ist

$$\text{Precision}_{TE}(K, i) = \frac{|\{o \in K_i | K(o) = C(o)\}|}{|K_i|}$$

DATABASE SYSTEMS GROUP

Bewertung von Klassifikatoren

LMU

weitere Gütekriterien für Klassifikatoren

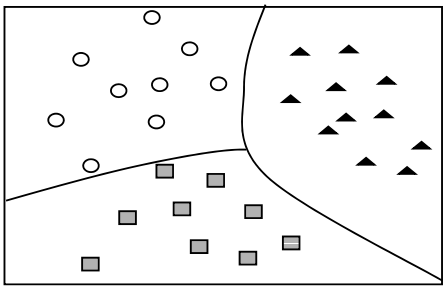
- **Kompaktheit** des Modells
  - z.B. Größe eines Entscheidungsbaums
- **Interpretierbarkeit** des Modells
  - Wieviel Einsichten vermittelt das Modell dem Benutzer?
- **Effizienz**
  - der Konstruktion des Modells
  - der Anwendung des Modells
- **Skalierbarkeit**
  - für große Datenmengen
  - für sekundärspeicherresidente Daten
- **Robustheit** gegenüber Rauschen und fehlenden Werten

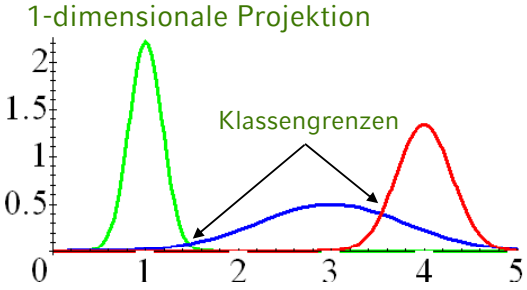
3.3 Bayes-Klassifikatoren

Was sind Bayes-Klassifikatoren?

Statistische Klassifikatoren

- Klassen werden durch statistische Prozesse beschrieben
- Beruht auf dem Satz von Bayes
- Bestimme Wahrscheinlichkeiten mit denen jeder Prozess das Objekt erklärt
  (Class-Membership-Probability)
- Vorhersage der wahrscheinlichsten Klasse
  (Maximum Likelihood Classification)





Überblick Bayes Klassifikatoren

Grundlagen statistischer Klassifikatoren

1. A-priori und A-posteriori Wahrscheinlichkeiten
2. Regel von Bayes
3. „Maximum Likelihood“ Klassifikation

Klassifikatoren und Statistische Prozeße

1. Naive Bayes
2. Bayes Netzwerke
3. LDA
4. multivariate Gauss-Prozesse

Bayes-Klassifikatoren

Grundlagen

Regeln und Fakten zur Klassifikation werden mit Hilfe des Satzes von Bayes als bedingte Wahrscheinlichkeiten formuliert

A-Priori-Wahrscheinlichkeiten modellieren Faktenwissen über die Häufigkeit einer Klasse und das Auftreten von Merkmalen, z.B.

- 20% der Objekte sind Äpfel
- 30% sind Orangen

} A-Priori Wahrsch. f. Klassenzugehörigk.

- 50% der Objekte sind rund
- 40% haben Farbe orange

} A-Priori Merkmalshäufigkeit

Bedingte Wahrscheinlichkeiten („A-Posteriori“) modellieren Zusammenhänge zwischen Klassen und Merkmalen:

- 100% der Orangen sind rund: $P(\text{rund} \mid \text{Orange}) = 100\%$
- 100% der Äpfel sind rund: $P(\text{rund} \mid \text{Apfel}) = 100\%$
- 90% der Orangen sind orange: $P(\text{orange} \mid \text{Orange}) = 90\%$

DATABASE SYSTEMS GROUP

Bayes-Klassifikatoren

LMU

Bei einem gegebenen Merkmals-Vektor M lässt sich die Wahrscheinlichkeit der Klassenzugehörigkeit zu Klasse C_i mit dem Satz von Bayes ermitteln:

$$P(C_i \mid M) = \frac{P(M \mid C_i) \cdot P(C_i)}{P(M)} = \frac{P(M \mid C_i) \cdot P(C_i)}{\sum_{c_j \in C} P(C_j) \cdot P(M \mid C_j)}$$

Im Beispiel: Wahrscheinlichkeit, dass ein oranges Objekt eine Orange ist:

$$P(\text{Orange} \mid \text{orange}) = \frac{P(\text{orange} \mid \text{Orange}) \cdot P(\text{Orange})}{P(\text{orange})} = \frac{0.9 \cdot 0.3}{0.4} = 0.675$$

Die entsprechenden Wahrscheinlichkeiten werden aus den Trainingsdaten geschätzt.

Bayes-Klassifikation

Der Bayes-Klassifikator schätzt die Wahrscheinlichkeit der Klassenzugehörigkeit eines Merkmalsvektors

Zur eindeutigen Zuordnung eines Klassen-Labels geht man meist nach dem Prinzip „Maximum Likelihood“ vor:

$$C = \underset{C_i}{\operatorname{argmax}}\, P(C_i \mid M) = \underset{C_i}{\operatorname{argmax}} \frac{P(M \mid C_i) \cdot P(C_i)}{P(M)} = \underset{C_i}{\operatorname{argmax}}\, P(M \mid C_i) \cdot P(C_i)$$

Da $P(M)$ bei allen C_i gleich ist, ist nur das Produkt zu optimieren. Beispiel:

- P(Apfel | M) = 32%
- P(Orange | M) = 32%
- P(Kiwi | M) = 36%

$\Rightarrow$ C = Kiwi

Schätzung der Wahrscheinlichkeiten

„A-priori“ Wahrscheinlichkeiten

Meistens: relative Häufigkeit in den Trainingsdaten.

Bsp: 7 Orangen , 2 Äpfel , 1 Stein => $P(Orange) = \frac{7}{7+2+1} = 70\%$

„A-Posteriori“ Wahrscheinlichkeiten

- Statistischer Prozess modelliert Zusammenhänge zwischen Merkmalen und einer Klasse
- Unterschiede verwendeter Prozesse:
  - Abhängigkeit der Merkmale (Korrelation oder Unabhängigkeit)
  - Verwendete Verteilungsfunktionen der Merkmalswerte (diskret, Normalverteilung, Multinomialverteilung…)
  - Beschaffenheit der Objekte (Vektor, Sequenz…)

1-dimensionale Verteilungen

Diskrete Merkmale

Auszählen relativer Häufigkeiten

Bsp:

$$P(Form = rund \mid A) = \frac{3}{4} = 75\%$$

$$P(Farbe = grün \mid A) = \frac{2}{4} = \frac{1}{2} = 50\%$$

$$P(Form = oval \mid A) = \frac{0}{4} = 0\%$$

ID	Form	Farbe	Klasse
1	rund	orange	A
2	rund	grün	A
3	rund	gelb	A
4	eckig	grün	A
5	oval	weiß	B

Problem: (Form = oval) => Klasse ≠ A

Man verwendet häufig „Smoothing“, d.h. P(x | Klasse) > ε.

mit $0 < \varepsilon << 1$.

D.h. $P(Form = oval \mid A) = \max\left(\frac{0}{4}, \varepsilon\right) = \varepsilon$

DATABASE SYSTEMS GROUP

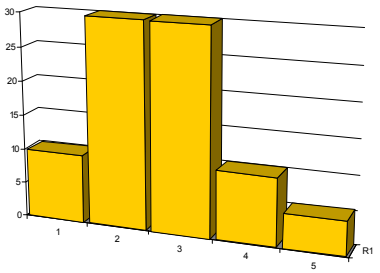
1-dimensionale Verteilungen

LMU

Kontinuierliche metrische Attribute

diskrete Approximation

P (9.0 < Durchmesser ≤ 9.5 | Orange) = 10%
P (9.5 < Durchmesser ≤ 10.0 | Orange) = 30%
P (10.0 < Durchmesser ≤ 10.5 | Orange) = 30%
P (10.5 < Durchmesser ≤ 11.0 | Orange) = 10%
P (11.0 < Durchmesser ≤ 11.5 | Orange) = 5%

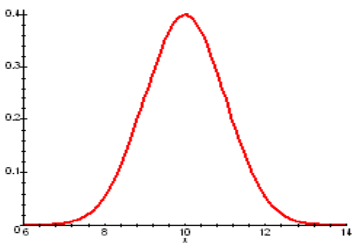


Wahrscheinlichkeits-Dichtefunktionen

z.B. Orangen haben einen Durchmesser von 10±1 cm:

p(Durchmesser | Orange) = *N* (10, 1)

meist Berechnung nach Normalverteilung:



$$P(x \mid C) = \frac{1}{\sqrt{2\pi\sigma\mid}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

wobei $\mu = \frac{\sum_{x \in TR} x}{\mid TR \mid}$ und $\sigma = \sqrt{\frac{\sum_{x \in TR} (x-\mu)^2}{\mid TR \mid}}$

Motivation

Bei hochdimensionalen Merkmalsvektoren schwierige Schätzung der bedingten Wahrscheinlichkeiten $P(M \mid C)$ und damit $P(C \mid M)$:

- M besteht aus vielen einzelnen Komponenten, die UND-verknüpft sind:

$$P(C \mid M_1 \wedge M_2 \wedge \ldots) = \frac{P(M_1 \wedge M_2 \wedge \ldots \mid C) \cdot P(C)}{P(M_1 \wedge M_2 \wedge \ldots)}$$

- Bei d verschiedenen Merkmalen und jeweils r verschiedenen Werten ergeben sich r^d verschiedene Merkmalskombinationen

Probleme:

- Die Wahrscheinlichkeiten lassen sich nicht mehr abspeichern
- Man bräuchte >> r^d Trainingsdatensätze, um die Wahrscheinlichkeit der einzelnen Merkmalskombinationen überhaupt ermitteln zu können

Naive Bayes-Klassifikation

Lösung dieses Problems beim naiven Bayes-Klassifikator:

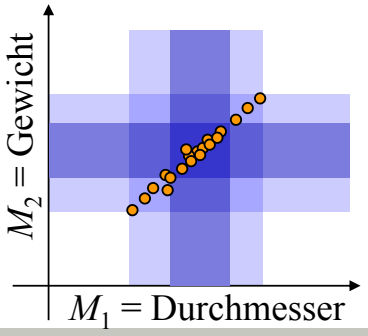
Annahme der bedingten Unabhängigkeit

d.h. bei jeder einzelnen Klasse werden die Merkmale so behandelt als wären sie voneinander statistisch unabhängig:

$$P(M_1 \wedge M_2 \mid C) = P(M_1 \mid C) \cdot P(M_2 \mid C)$$

Was bedeutet dies?

Klasse=Orange:



- Annahme kann falsch sein
- Dies führt *nicht* unbedingt dazu, dass die Klassifikation versagt
- Aber schlechte Leistung, wenn…
  - alle Merkmale bei mehreren Klassen etwa gleich verteilt sind
  - Unterschiede nur in „Relationen" der Merkmale zueinander

Naive Bayes-Klassifikation

Damit ist die Wahrscheinlichkeit der Zugehörigkeit zur Klasse C_i:

$$P(C_i \mid M_1 \wedge M_2 \wedge ...) = \frac{P(C_i) \cdot P(M_1 \wedge M_2 \wedge ... \mid C_i)}{P(M_1 \wedge M_2 \wedge ...)}$$

$$= \frac{P(C_i) \cdot \prod_j P(M_j \mid C_i)}{\sum_k P(C_k) \prod_j P(M_j \mid C_k)}$$

Auch hier ist der Nenner für alle Klassen gleich, so dass nur der Zähler zu maximieren ist:

$$C = \underset{C_i}{\operatorname{argmax}} \{P(C_i) \cdot \prod_j P(M_j \mid C_i)\}$$

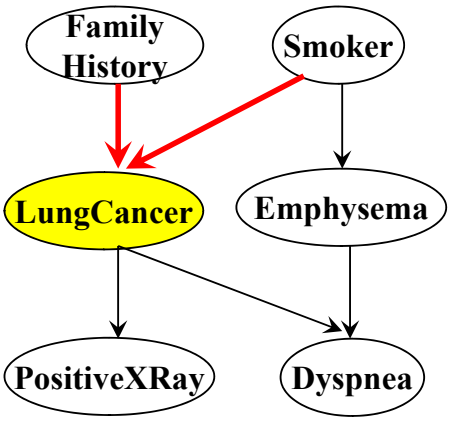
Bayes-Netzwerke

Grundbegriffe

- Graph mit Knoten = *Zufallsvariable* und Kante = *bedingte Abhängigkeit*
- Jede Zufallsvariable ist bei gegebenen Werten für die Vorgänger-Variablen bedingt unabhängig von allen Zufallsvariablen, die keine Nachfolger sind.
- Für jeden Knoten (Zufallsvariable):
  Tabelle der bedingten Wahrscheinlichkeiten
- Trainieren eines Bayes-Netzwerkes
  - bei gegebener Netzwerk-Struktur und allen bekannten Zufallsvariablen
  - bei gegebener Netzwerk-Struktur und teilweise unbekannten Zufallsvariablen
  - bei apriori unbekannter Netzwerk-Struktur

Bayes-Netzwerke

Beispiel



	FH,S	FH, ¬S	¬FH,S	¬FH, ¬S
LC	0.8	0.5	0.7	0.1
~LC	0.2	0.5	0.3	0.9

bedingte Wahrscheinlichkeiten für LungCancer

bei gegebenen Werten für FamilyHistory und Smoker liefert der Wert für Emphysema keine zusätzliche Information über LungCancer.

Lineare Diskriminanz Analyse

- Modelliere alle Klassen als multivariate Normalverteilungen
- Berücksichtigt Korrelationen der Attribute
- Varianzen und Korrelationen für alle Klassen gleich

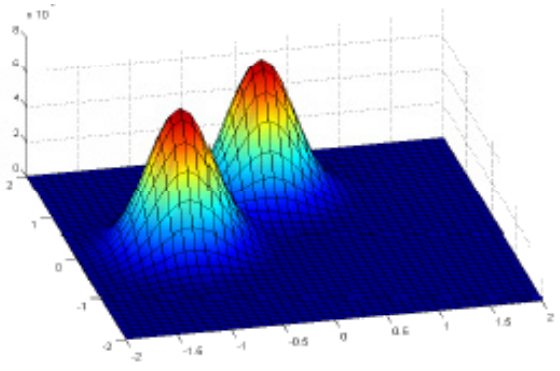
Basis multivariate Normalverteilung (Gauss-Verteilung)

$$P(x \mid C) = \frac{1}{\sqrt{(2\pi)^d \mid \Sigma \mid}} e^{-\frac{1}{2}\cdot(x-\mu)^T \cdot (\Sigma)^{-1} \cdot (x-\mu)}$$

Erwartungsvektor: $\mu = \frac{\sum_{x \in TR} x}{\mid TR \mid}$

Kovarianzmatrix :

$$\Sigma(i,j) = \frac{\sum_{x \in TR} (x_i - \mu_i) \cdot (x_j - \mu_j)}{|TR|}$$



Eigenschaften:

- Korrelation zwischen i und j
- Varianz in der Diagonalen

Lineare Diskriminanz Analyse

Training:

- Bestimme μ_C und Σ_C für alle Klassen C.
- Mittle globale Kovarianzmatrix Σ. (Gewichteter Durchschnitt der Kovarianzmatrizen aller Klassen)

$$\Sigma = \frac{\sum_{C_i \in C} \Sigma_{C_i}}{|C|}$$

Klassifikation:

$$\arg\max_{C_i \in C} P(x \mid C_i) = \arg\max_{C_i \in C} \left(\frac{1}{\sqrt{(2\pi)^d \mid \Sigma \mid}} e^{-\frac{1}{2} \cdot (x-\mu_{C_i})^T \cdot (\Sigma)^{-1} \cdot (x-\mu_{C_i})} \cdot P(C_i) \right)$$

$$= \arg\max_{C_i \in C} \left(-\frac{1}{2} \cdot (x-\mu_{C_i})^T \cdot (\Sigma)^{-1} \cdot (x-\mu_{C_i}) + \log(P(C_i)) \right)$$

$$= \arg\max_{C_i \in C} \left(x^T (\Sigma)^{-1} \mu_{C_i} - \frac{1}{2} \mu_{C_i}^T (\Sigma)^{-1} \mu_{C_i} + \log(P(C_i)) \right) = \sigma_{C_i}(x)$$

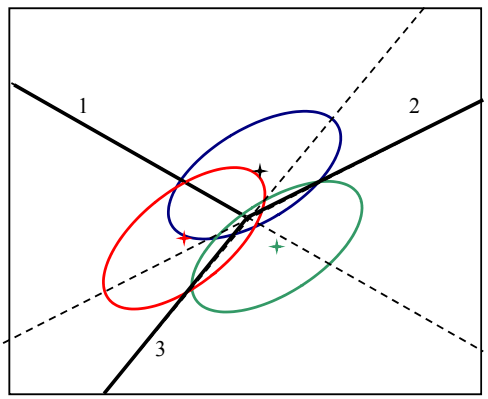
Lineare Diskriminanzfunktion

Lineare Diskriminanz Analyse

- Beobachtung: Da nur Erwartungswerte unterschiedlich
  $\Rightarrow$ Lineare Separation
- Man muss nicht die Wahrscheinlichkeit berechnen.
- Es reicht die Auswertung der folgenden Diskriminanzfunktion:

$$\sigma_{C_i}(x) = x^T \Sigma^{-1} \mu_{C_i} - \frac{1}{2} \mu_{C_i}^T \Sigma^{-1} \mu_{C_i} + \log P(C_i)$$

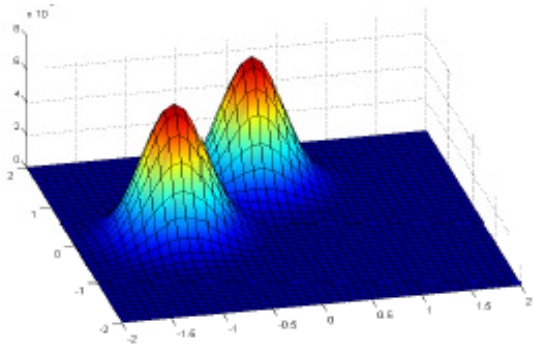
Klasse mit maximalem $\sigma_C(x)$ wird vorhergesagt.



Multivariate Gauss-Prozesse

- Modelliere jede Klasse als multivariate Normalverteilung (Vektoren im R^d)
- Berücksichtigt Korrelationen der Attribute
- Hier: Varianzen und Korrelationen für alle Klassen **individuell**
- Berechnung der Wahrscheinlichkeiten zur Klassifikation (Maximum Likelihood)

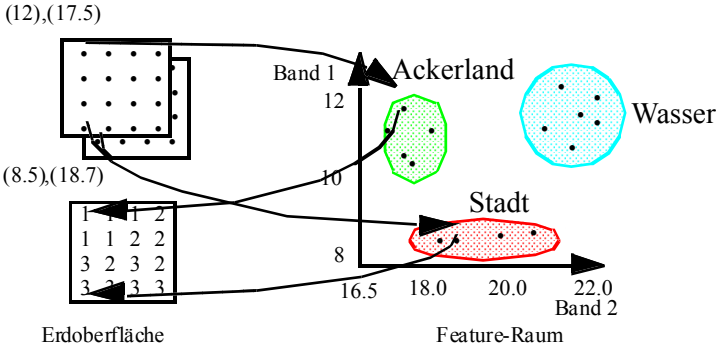
Probleme:
Braucht sehr viel Trainingsobjekte für jede Klasse, um signifikante Korrelationswerte zu bestimmen.



Interpretation von Rasterbildern

Motivation

- *automatische* Interpretation von *d* Rasterbildern eines bestimmten Gebiets für jedes Pixel ein *d*-dimensionaler Grauwertvektor (o_1, . . ., o_d)
- verschiedene Oberflächenbeschaffenheiten der Erde besitzen jeweils ein charakteristisches Reflexions- und Emissionsverhalten

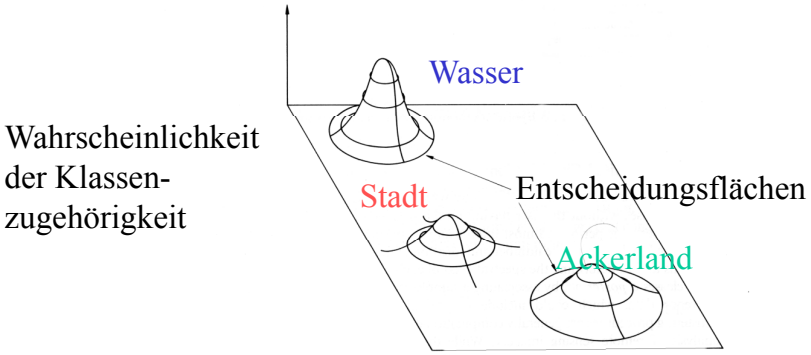


Grundlagen

Anwendung des Bayes-Klassifikators mit Gauss Prozess

Schätzung der $P(o \mid c)$ ohne Annahme der bedingten Unabhängigkeit

Annahme einer d-dimensionalen Normalverteilung für die Grauwertvektoren einer Klasse



DATABASE SYSTEMS GROUP
Interpretation von Rasterbildern
LMU

Methode

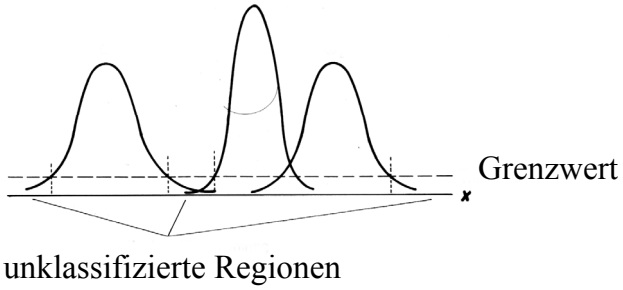
Zu schätzen aus den Trainingsdaten:

μ_i: d-dimensionaler Mittelwertvektor aller Feature-Vektoren der Klasse c_i

Σ_i: $d \times d$ Kovarianzmatrix der Klasse c_i

Probleme der Entscheidungsregel:

- Wahrscheinlichkeit für die gewählte Klasse sehr klein
- Wahrscheinlichkeit für mehrere Klassen ähnlich



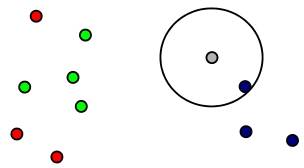
Bayes-Klassifikatoren

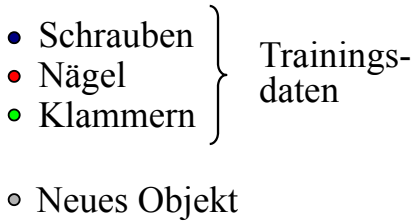
Diskussion

+ hohe Klassifikationsgenauigkeit in vielen Anwendungen
+ Inkrementalität: Klassifikator kann einfach an neue Trainingsobjekte adaptiert werden
+ Einbezug von Anwendungswissen

- Anwendbarkeit (Bayes-Netzwerke): die erforderlichen bedingten Wahrscheinlichkeiten sind oft unbekannt
- Ineffizienz bei sehr vielen Attributen (insbesondere Bayes-Netzwerke)

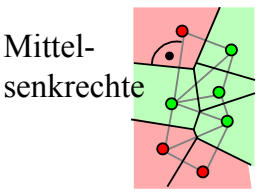
3.4 Nächste-Nachbarn-Klassifikatoren





Instanzbasiertes Lernen (*instance based learning*)

- Einfacher Nächste-Nachbar-Klassifikator: Zuordnung zu der Klasse des nächsten Nachbarpunkts
- Im Beispiel: Nächster Nachbar ist eine Schraube
- Regionen der Klassenzuordnung können als *Voronoi-Diagramme* dargestellt werden:

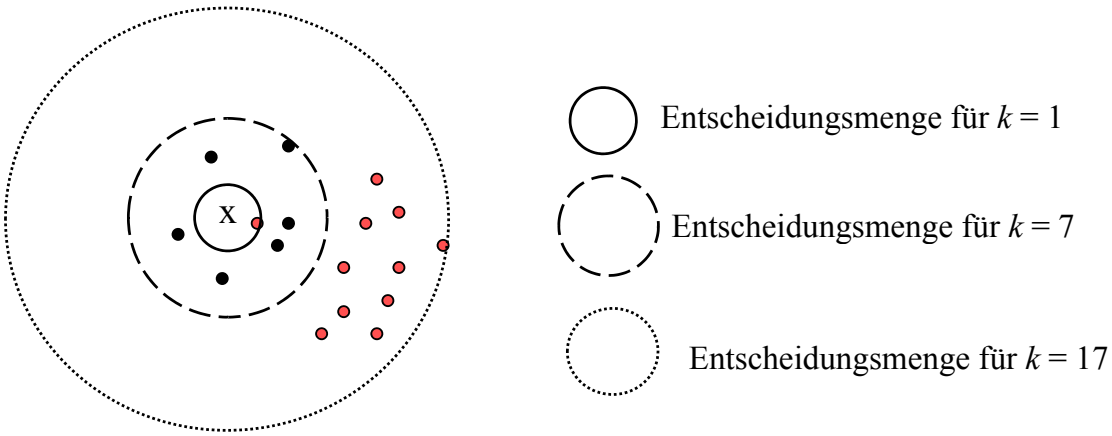


Nächste-Nachbarn-Klassifikatoren

- Problem: Punkt rechts oben wahrscheinlich nur Ausreißer
- Besser: Betrachte mehr als nur einen Nachbarn
  $\Rightarrow$ *k*-Nächste-Nachbarn-Klassifikator
- *Entscheidungsmenge*
  die Menge der zur Klassifikation betrachteten *k*-nächsten Nachbarn
- *Entscheidungsregel*
  Wie bestimmt man aus den Klassen der Entscheidungsmenge die Klasse des zu klassifizierenden Objekts?
  - Interpretiere Häufigkeit einer Klasse in der Entscheidungsmenge als Wahrscheinlichkeit der Klassenzugehörigkeit
  - Maximum-Likelihood-Prinzip: Mehrheitsentscheidung
  - Ggf. Gewichtung

Wahl des Parameters k

- „zu kleines“ *k*: hohe Sensitivität gegenüber Ausreißern
- „zu großes“ *k*: viele Objekte aus anderen Clustern (Klassen) in der Entscheidungsmenge.
- mittleres *k*: höchste Klassifikationsgüte, oft $1 << k < 10$



x: zu klassifizieren

Entscheidungsregel

- Standardregel

  $\Rightarrow$ wähle die Mehrheitsklasse der Entscheidungsmenge

- Gewichtete Entscheidungsregel

  gewichte die Klassen der Entscheidungsmenge

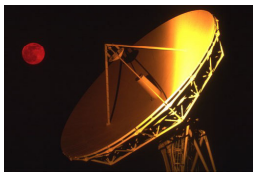
  - nach Distanz, meist invers quadriert: $weight(dist) = 1/dist^2$
  - nach Verteilung der Klassen (oft sehr ungleich!)

    Problem: Klasse mit zu wenig Instanzen ($< k/2$) in der Trainingsmenge bekommt keine Chance, ausgewählt zu werden, selbst bei optimaler Distanzfunktion

    - Klasse A: 95 %, Klasse B 5 %
    - Entscheidungsmenge = {A, A, A, A, B, B, B}
    - Standardregel $\Rightarrow$ A, gewichtete Regel $\Rightarrow$ B

Klassifikation von Sternen

Analyse astronomischer Daten



Entfernen von Rauschen

Bildsegmentierung

Feature-Extraktion

Automatische *Klassifikation* des Sterntyps

Manuelle Analyse der interessanten Sterntypen

Klassifikation des Sterntyps mit Nächste-Nachbarn Klassifikator basierend auf dem Hipparcos-Katalog

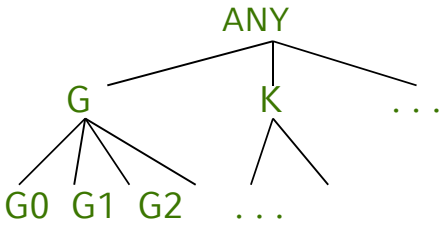
Klassifikation von Sternen

Hipparcos-Katalog [ESA 1998]

- enthält ca. 118 000 Sterne
- mit 78 Attributen (Helligkeit, Entfernung, Farbe,. . .)
- Klassenattribut: Spektraltyp (Attribut H76)

z.B.

H76: G0
H76: G7.2
H76: KIII/IV



- Werte des Spektraltyps sind vage

➡ Hierarchie von Klassen benutze die erste Ebene der Klassenhierarchie

Klassifikation von Sternen

Verteilung der Klassen

Klasse	#Instanzen	Anteil Instanzen	
K	32 036	27.0	häufige Klassen
F	25 607	21.7	
G	22 701	19.3	
A	18 704	15.8	
B	10 421	8.8	
M	4 862	4.1	
O	265	0.22	seltene Klassen
C	165	0.14	
R	89	0.07	
W	75	0.06	
N	63	0.05	
S	25	0.02	
D	27	0.02	

Klassifikation von Sternen

Experimentelle Untersuchung [Poschenrieder 1998]

- Distanzfunktion
  - mit 6 Attributen (Farbe, Helligkeit und Entfernung)
  - mit 5 Attributen (ohne Entfernung)

  ⇒ beste Klassifikationsgenauigkeit mit 6 Attributen
- Anzahl *k* der Nachbarn

  ⇒ beste Klassifikationsgenauigkeit für $k = 15$
- Entscheidungsregel
  - Gewichtung nach Distanz
  - Gewichtung nach Klassenverteilung

  ⇒ beste Klassifikationsgenauigkeit bei Gewichtung nach Distanz aber nicht nach Klassenverteilung

DATABASE SYSTEMS GROUP

Klassifikation von Sternen

LMU

Klasse	Falsch klassifiziert	Korrekt klassifiziert	Klassifikations-genauigkeit
K	408	2338	85.1%
F	350	2110	85.8%
G	784	1405	64.2%
A	312	975	75.8%
B	308	241	43.9%
M	88	349	79.9%
C	4	5	55.6%
R	5	0	0%
W	4	0	0%
O	9	0	0%
N	4	1	20%
D	3	0	0%
S	1	0	0%
Total	2461	7529	75.3%

hohe Klassifikationsgenauigkeit für die häufigen Klassen, schlechte Genauigkeit für die seltenen Klassen

die meisten seltenen Klassen besitzen weniger als $k / 2 = 8$ Instanzen!

Diskussion

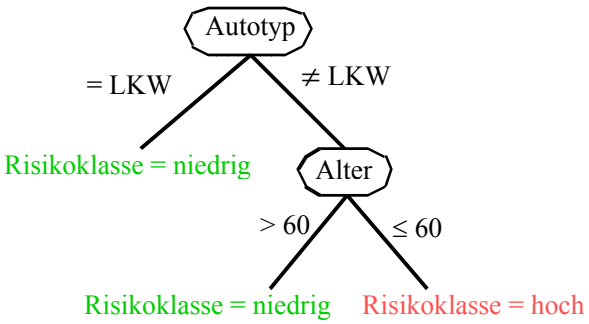
+ Anwendbarkeit erfordert als Eingabe nur die Trainingsdaten
+ hohe Klassifikationsgenauigkeit in vielen Anwendungen
+ inkrementell Klassifikator kann sehr einfach an neue Trainingsobjekte adaptiert werden
+ auch zur Vorhersage einsetzbar

- Ineffizienz bei der Auswertung des "Modells" erfordert *k*-nächste-Nachbarn Anfrage an die Datenbank
- liefert kein explizites Wissen über die Klassen

3.5 Entscheidungsbaum-Klassifikatoren

Motivation

ID	Alter	Autotyp	Risiko
1	23	Familie	hoch
2	17	Sport	hoch
3	43	Sport	hoch
4	68	Familie	niedrig
5	32	LKW	niedrig

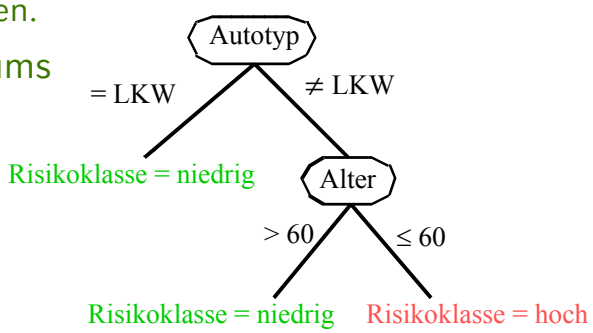


finden explizites Wissen

Entscheidungsbäume sind für die meisten Benutzer verständlich

Grundbegriffe

- Ein *Entscheidungsbaum* ist ein Baum mit folgenden Eigenschaften:
  - ein innerer Knoten repräsentiert ein Attribut,
  - eine Kante repräsentiert einen Test auf dem Attribut des Vaterknotens,
  - ein Blatt repräsentiert eine der Klassen.
- Konstruktion eines Entscheidungsbaums
  - anhand der Trainingsmenge
  - Top-Down



- Anwendung eines Entscheidungsbaums

  Durchlauf des Entscheidungsbaum von der Wurzel zu einem der Blätter

  $\Rightarrow$ eindeutiger Pfad

  Zuordnung des Objekts zur Klasse des erreichten Blatts

Konstruktion eines Entscheidungsbaums

Basis-Algorithmus

- Anfangs gehören alle Trainingsdatensätze zur Wurzel.
- Das nächste Attribut wird ausgewählt (Splitstrategie).
- Die Trainingsdatensätze werden unter Nutzung des Splitattributs partitioniert.
- Das Verfahren wird rekursiv für die Partitionen fortgesetzt.

  $\Rightarrow$ lokal optimierender Algorithmus

Abbruchbedingungen

- keine weiteren Splitattribute
- alle Trainingsdatensätze eines Knotens gehören zur selben Klasse

Entscheidungsbaum-Klassifikatoren

Beispiel

Tag	Aussicht	Temperatur	Feuchtigkeit	Wind	Tennispielen
1	sonnig	heiß	hoch	schwach	nein
2	sonnig	heiß	hoch	stark	nein
3	bedeckt	heiß	hoch	schwach	ja
4	regnerisch	mild	hoch	schwach	ja
5	regnerisch	kühl	normal	schwach	ja
6	regnerisch	kühl	normal	stark	nein
7	bedeckt	kühl	normal	stark	ja
8	sonnig	mild	hoch	schwach	nein
9	sonnig	kühl	normal	schwach	ja
10	regnerisch	mild	normal	schwach	ja
11	sonnig	mild	normal	stark	ja
12	bedeckt	mild	hoch	stark	ja
13	bedeckt	heiß	normal	schwach	ja
14	regnerisch	mild	hoch	stark	nein

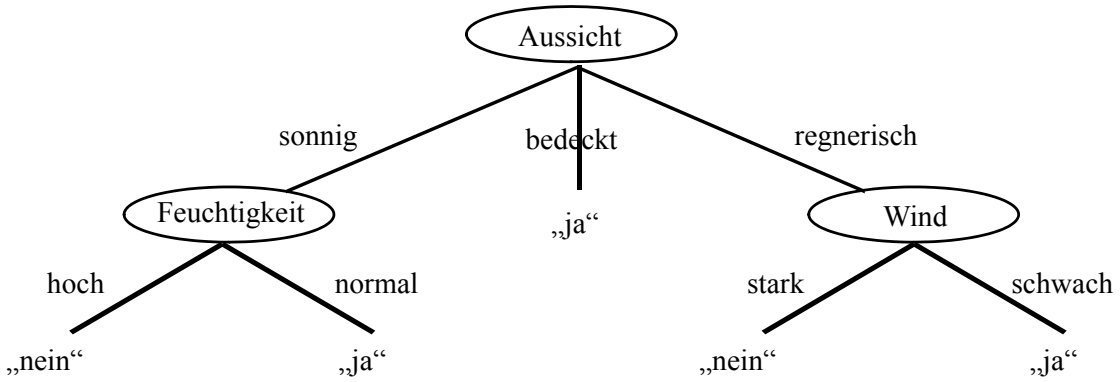
Ist heute ein Tag zum Tennisspielen?

DATABASE SYSTEMS GROUP

Entscheidungsbaum-Klassifikatoren

LMU

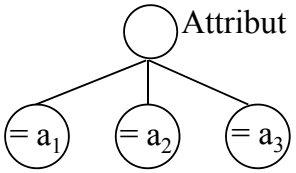
Beispiel

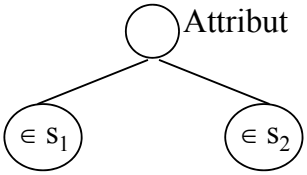


Splitstrategien

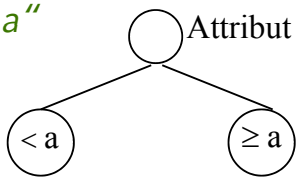
Typen von Splits

- Kategorische Attribute
  Splitbedingungen der Form „*Attribut = a*“ or „*Attribut ∈ set*“
  viele mögliche Teilmengen





- Numerische Attribute
  Splitbedingungen der Form „*Attribut < a*“
  viele mögliche Splitpunkte



Numerische Splitgrenzen

Wo sollen diskrete Attribute gesplittet werden?

=> An den Stellen, die die Qualität maximieren.

Idee: Ordnen der numerischen Attributwerte

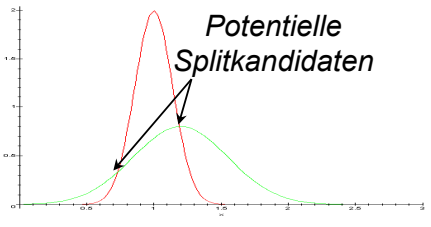
Wert	0.9	0.8	0.65	0.5	0.45	0.3	0.15	0.01
Klasse	A	A	B	B	B	A	A	A

Potentielle Splitkandidaten

Teste die Kombination, die den höchsten Information Gain erzielen.

Schnellere Methode:

- Bilde Gauß-Kurve über alle Klassen
- Wähle Schnittpunkte der Gauß-Kurven als Kandidaten.



Qualitätsmaße für Splits

Gegeben

- eine Menge T von Trainingsobjekten
- eine disjunkte, vollständige Partitionierung $T_1, T_2, \ldots, T_m$ von T
- p_i die relative Häufigkeit der Klasse c_i in T

Gesucht

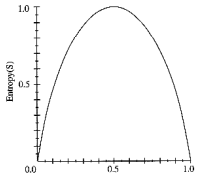
- ein Maß der *Unreinheit* einer Menge S von Trainingsobjekten in Bezug auf die Klassenzugehörigkeit
- ein Split von T in $T_1, T_2, \ldots, T_m$, der dieses Maß der Unreinheit *minimiert*

  $\Rightarrow$ Informationsgewinn, Gini-Index

Informationsgewinn

Entropie: minimale Anzahl von Bits zum Codieren der Nachricht,
mit der man die Klasse eines zufälligen Trainingsobjekts mitteilen möchte.
Die *Entropie* für eine Menge T von Trainingsobjekten ist definiert als

$$entropie(T) = -\sum_{i=1}^{k} p_i \cdot \log p_i$$



$entropie(T) = 0$, falls $p_i = 1$ für ein i
$entropie(T) = 1$ für $k = 2$ Klassen mit $p_i = 1/2$

Das Attribut A habe die Partitionierung $T_1, T_2, \ldots, T_m$ erzeugt.

Der *Informationsgewinn* des Attributs A in Bezug auf T ist definiert als

$$Informationsgewinn(T, A) = entropie(T) - \sum_{i=1}^{m} \frac{|T_i|}{|T|} \cdot entropie(T_i)$$

Splitstrategien

Gini-Index

Gini-Index für eine Menge *T* von Trainingsobjekten

$$gini(T) = 1 - \sum_{j=1}^{k} p_j^2$$

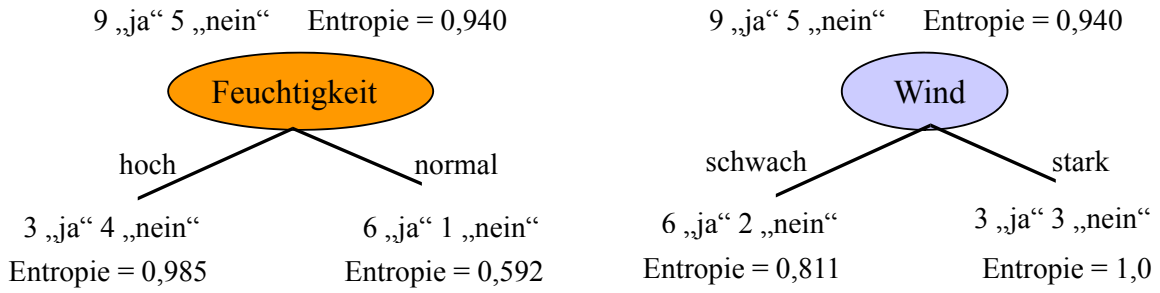
- kleiner Gini-Index ⇔ geringe Unreinheit,
- großer Gini-Index ⇔ hohe Unreinheit

Das Attribut *A* habe die Partitionierung $T_1, T_2, \ldots, T_m$ erzeugt.
Gini-Index des Attributs *A* in Bezug auf *T* ist definiert als

$$gini_A(T) = \sum_{i=1}^{m} \frac{|T_i|}{|T|} \cdot gini(T_i)$$

Splitstrategien

Beispiel

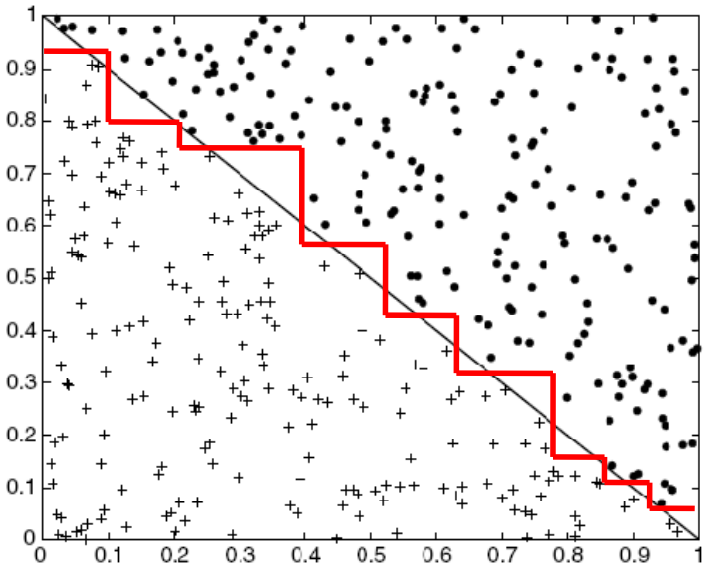


$$Informationsgewinn(T, Feuchtigkeit) = 0{,}94 - \frac{7}{14} \cdot 0{,}985 - \frac{7}{14} \cdot 0{,}592 = 0{,}151$$

$$Informationsgewinn(T, Wind) = 0{,}94 - \frac{8}{14} \cdot 0{,}811 - \frac{6}{14} \cdot 1{,}0 = 0{,}048$$

⇒ Feuchtigkeit liefert den höheren Informationsgewinn

Bias

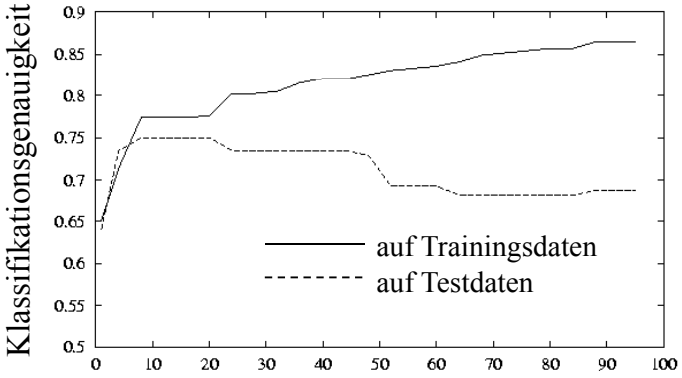


Overfitting

Einführung

Overfitting bei der Konstruktion eines Entscheidungsbaums, wenn es zwei Entscheidungsbäume E und E' gibt mit

- E hat auf der Trainingsmenge eine kleinere Fehlerrate als E',
- E' hat auf der Grundgesamtheit der Daten eine kleinere Fehlerrate als E.



Overfitting

Trainingsdaten

Name	Body Temp	Gives Birth	Four-legged	Hibernates	Mammal
porcupine	warm-blooded	yes	yes	yes	yes
cat	warm-blooded	yes	yes	no	yes
bat	warm-blooded	yes	no	yes	no*
whale	warm-blooded	yes	no	no	no*
salamander	cold-blooded	no	yes	yes	no
komodo dragon	cold-blooded	no	yes	no	no
python	cold-blooded	no	no	yes	no
salmon	cold-blooded	no	no	no	no
eagle	warm-blooded	no	no	no	no
guppy	cold-blooded	yes	no	no	no

*Fehlerhafte Daten

DATABASE SYSTEMS GROUP

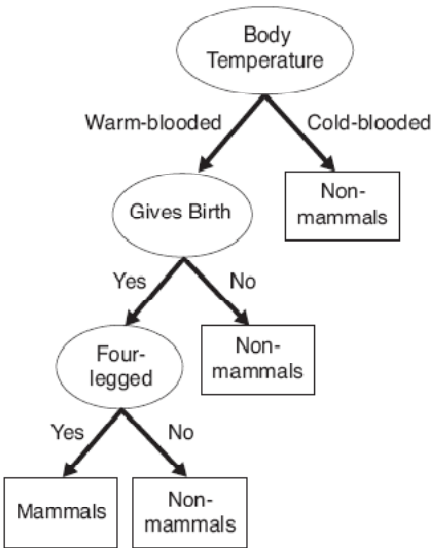
Overfitting

LMU

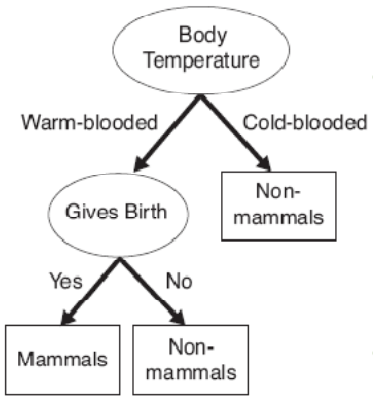
Testdaten

Name	Body Temp	Gives Birth	Four-legged	Hibernates	Mammal
human	warm-blooded	yes	no	no	yes
pigeon	warm-blooded	no	no	no	no
elephant	warm-blooded	yes	yes	no	yes
leopard shark	cold-blooded	yes	no	no	no
turtle	cold-blooded	no	yes	no	no
penguin	warm-blooded	no	no	no	no
eel	cold-blooded	no	no	no	no
dolphin	warm-blooded	yes	no	no	yes
spiny anteater	warm-blooded	no	yes	yes	yes
gila monster	cold-blooded	no	yes	yes	no

Overfitting



(a) Model M1



(b) Model M2

Trainingserror: 0%
Testerror: 30%

Trainingserror: 20%
Testerror: 10%

M1:

- Fehlklassifikation von human und dolphin aufgrund von fehlerhaften Trainingsdaten
- Spiny anteater: ungewöhnlicher Fall

M2:

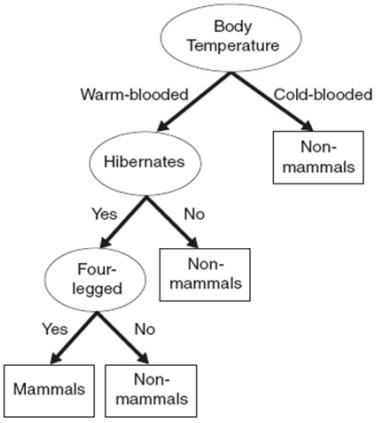
ungewöhnlicher Fall kann nicht vermieden werden

Overfitting

Name	Body Temp	Gives Birth	Four-legged	Hibernates	Mammal
salamander	cold-blooded	no	yes	yes	no
poorwill	warm-blooded	no	no	yes	no
platypus	warm-blooded	no	yes	yes	yes
eagle	warm-blooded	no	no	no	no
guppy	cold-blooded	yes	no	no	no

Problem:

Mangel an repräsentativen Daten



Overfitting

Ansätze zum Vermeiden von Overfitting

- Entfernen von fehlerhaften Trainingsdaten

  insbesondere widersprüchliche Trainingsdaten
- Wahl einer geeigneten Größe der Trainingsmenge

  nicht zu klein, nicht zu groß
- Wahl einer geeigneten Größe des minimum support

  minimum support:

  Anzahl der Datensätze, die mindestens zu einem Blattknoten des Baums gehören müssen

  minimum support >> 1

DATABASE SYSTEMS GROUP

Overfitting

LMU

Ansätze zum Vermeiden von Overfitting

- Wahl einer geeigneten Größe der minimum confidence

  minimum confidence: Anteil, den die Mehrheitsklasse eines Blattknotens mindestens besitzen muß.

  minimum confidence << 100%

  Blätter können auch fehlerhafte Datensätze oder Rauschen „absorbieren“
- nachträgliches Pruning des Entscheidungsbaums

  Abschneiden der überspezialisierten Äste

  kleinere Bäume sind weniger anfällig für Overfitting (Occam’s Razor)

Pruning von Entscheidungsbäumen

Fehlerreduktions-Pruning [Mitchell 1997]

- Aufteilung der klassifizierten Daten in Trainingsmenge und Testmenge
- Konstruktion eines Entscheidungsbaums *E* für die Trainingsmenge
- Pruning von *E* mit Hilfe der Testmenge *T*
  - bestimme denjenigen Teilbaum von *E*, dessen Abschneiden den Klassifikationsfehler auf *T* am stärksten reduziert
  - entferne diesen Teilbaum
  - fertig, falls kein solcher Teilbaum mehr existiert

nur anwendbar, wenn genügend viele klassifizierte Daten

Entscheidungsbaum-Klassifikatoren

Diskussion

+ Interpretation des gefundenen Baumes relativ einfach
+ Implizite Gewichtung der Attribute
+ Leistungsfähiger Klassifikator, häufig in der Praxis verwendet
+ Effiziente Auswertung des gefundenen Modells

- Finden eines optimalen Entscheidungsbaums ist exponentiell
- Heuristische Methoden können nur lokales Optimum finden
- Anfällig für Overfitting (besondere Methoden zur Vermeidung von Overfitting für Entscheidungsbäume)

3.6 Neuronale Netze

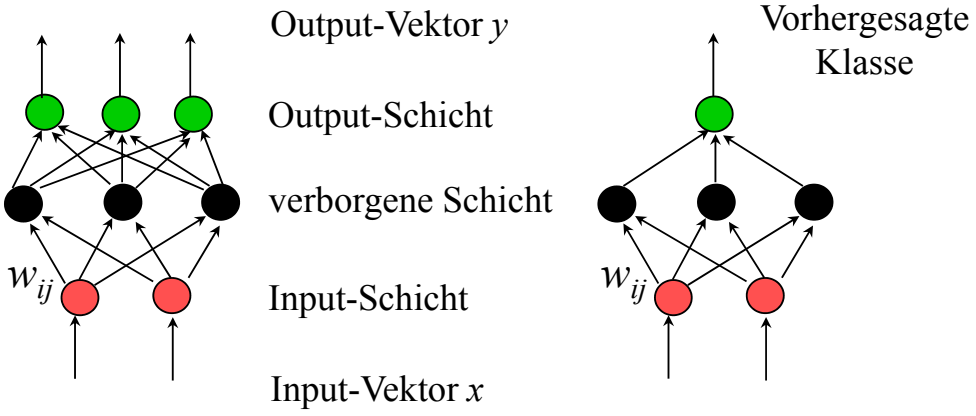
Grundlagen [Bigus 1996], [Bishop 1995]

- Paradigma für ein Maschinen- und Berechnungsmodell
- Funktionsweise ähnlich der von biologischen Gehirnen
  Neuronales Netz: Menge von Neuronen, über Kanten miteinander verbunden
  Neuron: entspricht biologischem Neuron Aktivierung durch Input-Signale an den Synapsen
- Menschliches Gehirn: 10^{11} Neuronen, jedes verbunden mit ~10^4 anderen, schnellste Aktionszeit 10^{-3} Sek. (Computer: 10^{-10}) – aber komplexe Entscheidungen sehr schnell (visuelle Erkennung der eigenen Mutter: 10^{-1} Sek. ≈ 100 Schritte)
- Erzeugung eines Output-Signals, das zu anderen Neuronen weitergeleitet wird.
- Organisation eines neuronalen Netzes
  Input-Schicht, *verborgene Schichten*, *Output-Schicht*
  Knoten einer Schicht mit allen Knoten der vorhergehenden Schicht verbunden

Grundlagen

Kanten besitzen *Gewichte*

Funktion eines neuronalen Netzes

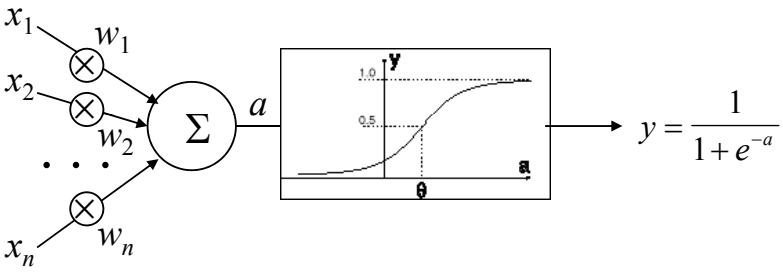


Neuronen

allgemeines Neuron (auch: Perzeptron)

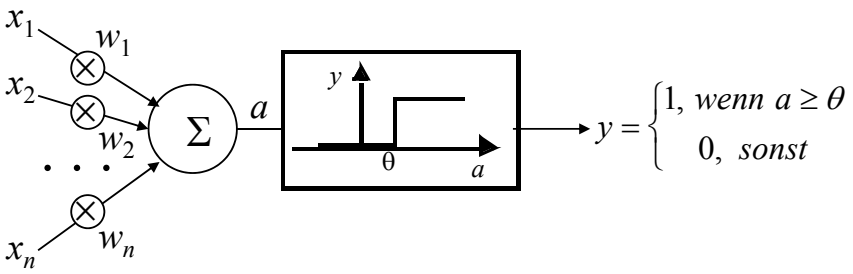
a: *Aktivierungswert*

$$a = \sum_{i=1}^{n} w_i \cdot x_i$$



Threshold Logic Unit

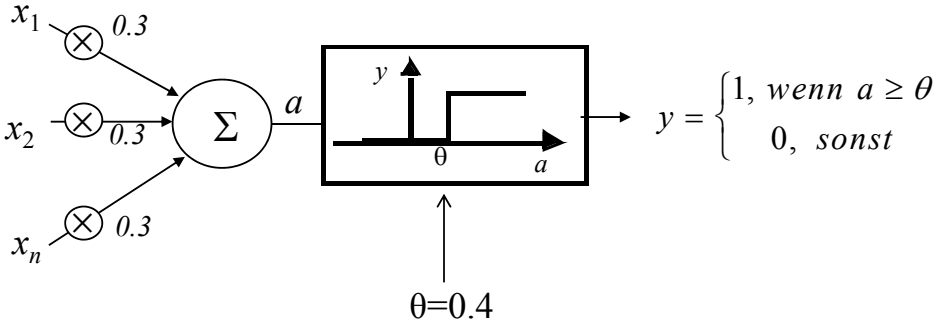
(TLU)



Neuronen

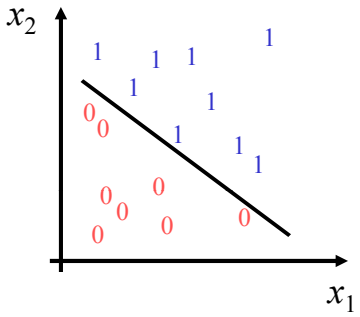
Beispiel: Modellierung einer booleschen Funktion

x_1	x_2	x_3	y
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	1
0	0	1	0
0	1	0	0
0	1	1	1
0	0	0	0



Neuronen

- Klassifikation mit Hilfe einer TLU
  - repräsentiert eine (Hyper-)Ebene $\sum_{i=1}^{n} w_i \cdot x_i = \theta$
  - links von der Ebene: Klasse 0
  - rechts von der Ebene: Klasse 1

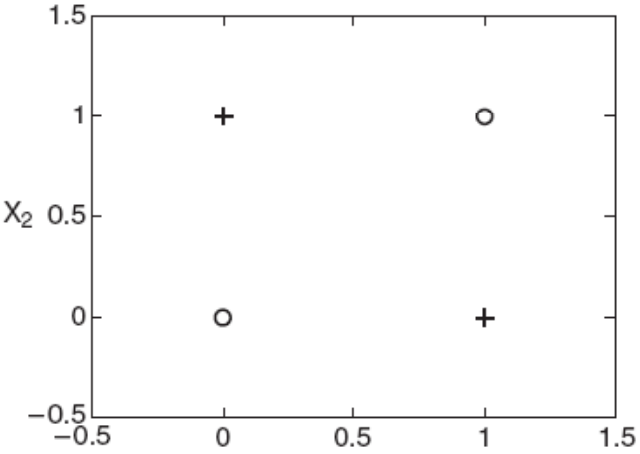


- Trainieren einer TLU
  - Lernen der „richtigen" Gewichte zur Unterscheidung der zwei Klassen
    Iterative Anpassung der Gewichte w_{ij}
  - Rotation der durch w und θ gegebene Hyperebene um einen kleinen Betrag in Richtung v, wenn v noch nicht auf der richtigen Seite der Ebene liegt

XOR-Problem

nicht linear separierbares Problem

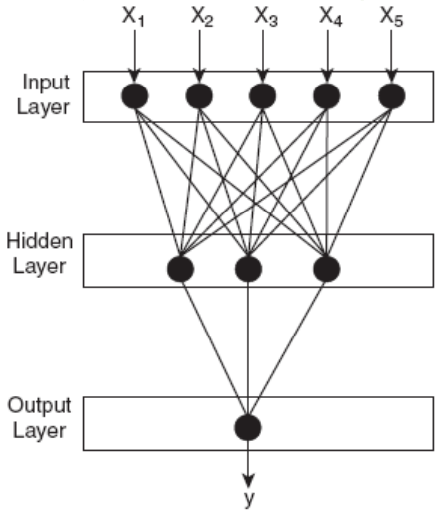
X_1	X_2	y
0	0	−1
1	0	1
0	1	1
1	1	−1



Kombination mehrerer Neuronen

zwei Klassen, die nicht linear separierbar sind:

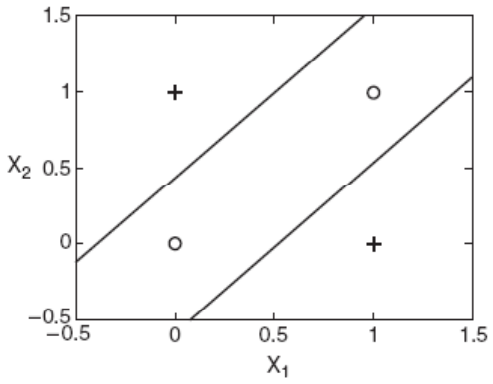
⇒ mehrere innere Knoten (hidden layer) und ein Output-Knoten (zwei-Klassen-Problem, sonst mehr)

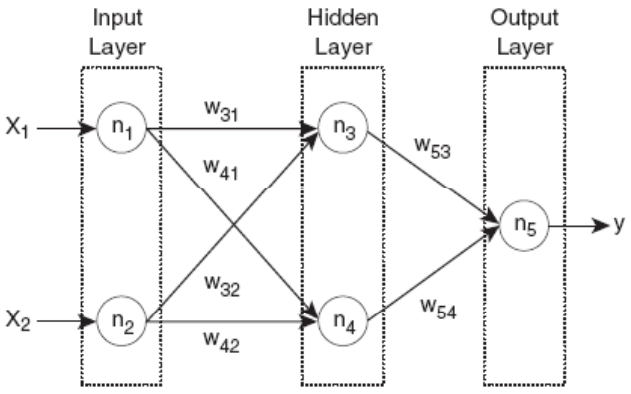


Varianten: feed-forward vs. recurrent (auch Verbindungen auf derselben Ebene oder zu vorherigen Ebenen möglich)

Kombination mehrerer Neuronen

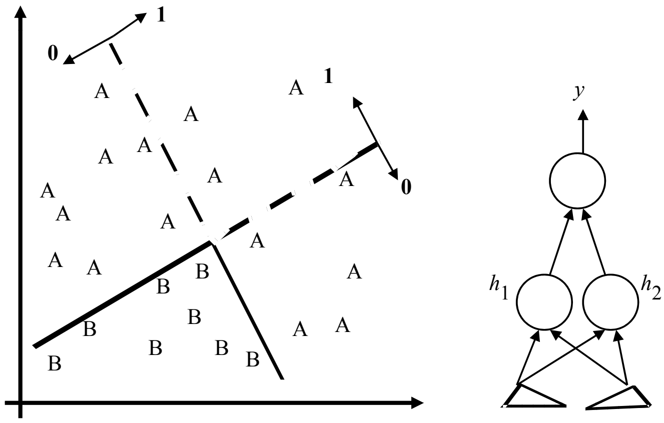
XOR





Kombination mehrerer Neuronen

Beispiel



$h_1 = 0 \wedge h_2 = 0{:}\ y = 0\ (Klasse\ \ B)$

$andernfalls{:}\ y = 1\ (Klasse\ \ A)$

Lernalgorithmus für komplexe Neuronale Netze

Bei Abweichung von vorhergesagter und tatsächlicher Klasse:

Anpassung der Gewichte mehrerer Knoten

Frage

In welchem Maße sind die verschiedenen Knoten an dem Fehler beteiligt?

Anpassung der Gewichte

- durch Gradientenverfahren, das den Gesamtfehler minimiert
- *Gesamtfehler*: Summe der (quadratischen) Abweichungen des tatsächlichen Outputs *y* vom gewünschten Output *t* für die Menge der Inputvektoren. (Least Squares Optimierung)
- *Voraussetzung:* Output *y stetige* Funktion der Aktivierung a

Algorithmus Backpropagation


```
für jedes Paar(v,t) // v = Input,t = gewünschter Output
    „forward pass“:
        Bestimme den tatsächlichen Output y für Eingabe v;
    „backpropagation“:
        Bestimme den Fehler (t - y) der Output-Einheiten
        und passe die Gewichte der Output-Einheiten in die
        Richtung an, die den Fehler minimiert;
        Solange der Input-Layer nicht erreicht ist:
            Propagiere den Fehler auf die nächste Schicht und
            passe auch dort die Gewichte der Einheiten in
            fehlerminimierender Weise an;
```


Design der Netztopologie

Bestimmung von

- Anzahl der Input-Knoten
- Anzahl der inneren Schichten und jeweilige Anzahl der Knoten
- Anzahl der Output-Knoten

starker Einfluss auf die Klassifikationsgüte:

- zu wenige Knoten

  $\Rightarrow$ niedrige Klassifikationsgüte
- zu viele Knoten

  $\Rightarrow$ Overfitting

Bestimmung der Netztopologie

nach [SPSS Clementine 2000]

Statische Topologie

- Topologie wird a priori festgelegt
- eine verborgene Schicht reicht in vielen Anwendungen aus

Dynamische Topologie

- dynamisches Hinzufügen von Neuronen (und verborgenen Schichten) solange Klassifikationsgüte signifikant verbessert wird

Multiple Topologien

- Trainieren mehrerer dynamischer Netze parallel

  z.B. je ein Netz mit 1, 2 und 3 verborgenen Schichten

Bestimmung der Netztopologie

Pruning

- Trainieren eines Netzes mit statischer Topologie
- nachträgliches Entfernen der unwichtigsten Neuronen solange Klassifikationsgüte verbessert wird

Schlussfolgerung

- statische Topologie: niedrige Klassifikationsgüte, aber relativ schnell.
- Pruning: beste Klassifikationsgüte, aber sehr hoher Laufzeitaufwand zum Training.

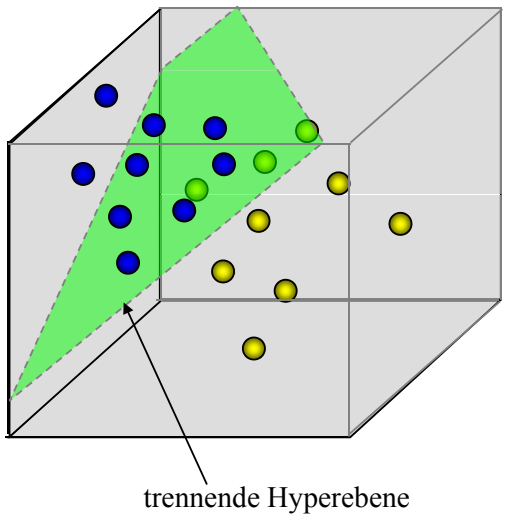
Diskussion

+ im Allgemeinen sehr hohe Klassifikationsgüte
  beliebig komplexe Entscheidungsflächen (möglichst einfache Topologie auswählen, um Overfitting zu vermeiden)
+ redundante Features relativ problemlos, Gewichte werden klein
+ Effizienz der Anwendung

- schlechte Verständlichkeit
  (lernt nur Gewichte, aber keine Klassenbeschreibung)
- Ineffizienz des Lernens (sehr lange Trainingszeiten)
- keine Integration von Hintergrundwissen
- empfindlich gegen Fehler in den Trainingsdaten
- mögliche Konvergenz in lokales Minimum

3.7 Support Vector Machines

Motivation: Lineare Separation



- Vektoren in R^d repräsentieren Objekte.
- Objekte gehören zu genau einer von je 2 Klassen

Klassifikation durch lineare Separation:

- Suche Hyperebene, die beide Klassen „maximal stabil“ voneinander trennt.

- Ordne unbekannte Elemente der Seite der Ebene zu, auf der sie sich befinden.

Support Vector Machines

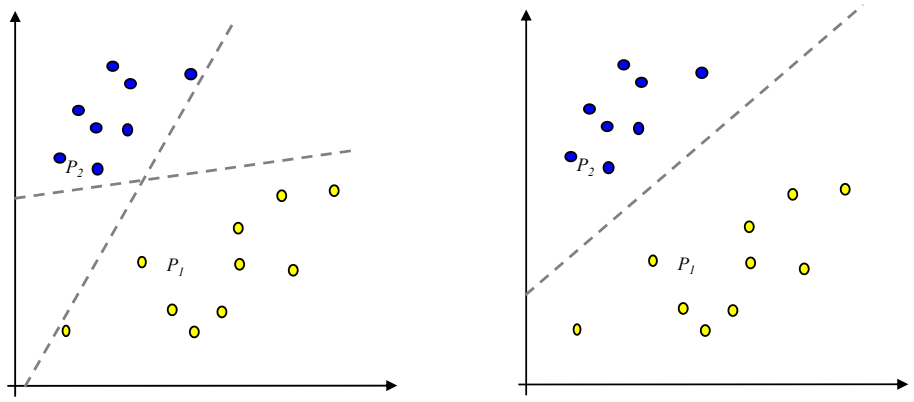
Probleme bei linearer Separation:

- Was ist die „maximal stabile“ Hyperebene und wie berechnet man sie effizient?
- Klassen nicht immer linear trennbar.
- Berechnung von Hyperebenen nach Auswahl sehr aufwendig.
- Einschränkung auf 2 Klassen.
- ...

⇨ *Lösungen dieser Probleme mit Support Vector Machines(SVMs) [Vapnik 1979 u. 1995].*

Maximum Margin Hyperplane

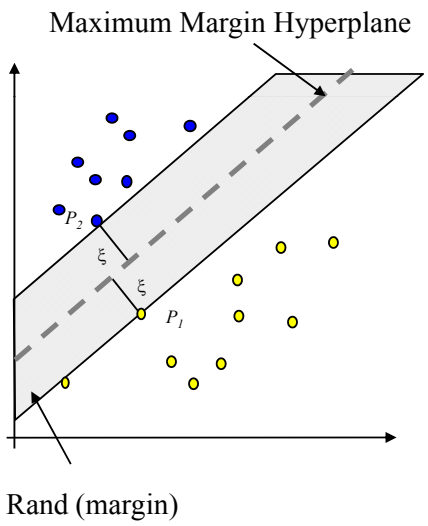
Problem: Hyperebene die P_1 und P_2 trennt ist nicht eindeutig.
⇒ *Welche Hyperebene ist für die Separation die Beste ?*



Kriterien:
- Stabilität beim Einfügen
- Abstand zu den Objekten beider Klassen

Maximum Margin Hyperplane

Lineare Separation mit der „*Maximum Margin Hyperplane*“



- Abstand zu Punkten aus beiden Mengen ist maximal, d.h. mind. ξ.
- Wahrscheinlichkeit, dass beim Einfügen die trennende Hyperebene verschoben werden muss, ist minimal.
- generalisiert am besten.

⇒ Maximum Margin Hyperplane (MMH) ist „maximal stabil“

MMH ist nur von Punkten P_i abhängig, die Abstand ξ zur Ebene aufweisen.

⇒ *P_i heißt Support Vector*

Maximum Margin Hyperplane

Zusammenfassung der Schreibweisen der benötigten algebraischen Konstrukte für Featurespace *FS*:

Skalarprodukt zweier Vektoren: $\langle \vec{x}, \vec{y} \rangle, \vec{x}, \vec{y} \in FS$

z.B. $\langle \vec{x}, \vec{y} \rangle = \sum_{i=1}^{d} (x_i \cdot y_i)$ kanonisches Skalarprodukt

Beschreibung einer Hyperebene: $H(\vec{w}, b) = \left\{ \vec{x} \in FS \,\middle|\, 0 = \langle \vec{w}, \vec{x} \rangle + b \right\}$

Abstand eines Vectors zur Ebene: $dist(\vec{x}, H(\vec{w}, b)) = \left| \frac{1}{\sqrt{\langle \vec{w}, \vec{w} \rangle}} \langle \vec{w}, \vec{x} \rangle + b \right|$

Maximum Margin Hyperplane

Berechnung der Maximum Margin Hyperplane

1. Bedingung: kein Klassifikationsfehler (Klasse1: y_i=1,Klasse 2:y_i=-1)

$$\left.\begin{aligned}(y_i = -1) \Rightarrow [\langle \vec{w}, \vec{x_i} \rangle + b] < 0 \\ (y_i = 1) \Rightarrow [\langle \vec{w}, \vec{x_i} \rangle + b] > 0\end{aligned}\right\} \Leftrightarrow y_i \left[\langle \vec{w}, \vec{x_i} \rangle + b\right] > 0$$

2. Bedingung: Maximaler Rand (Margin)

maximiere: $\xi = \min_{x_i \in TR} \left| \frac{1}{\sqrt{\langle \vec{w}, \vec{w} \rangle}} \left(\langle \vec{w}, \vec{x_i} \rangle + b \right) \right|$ (Abstand von x_i zur Ebene $H(\vec{w},b)$)

oder

maximiere: ξ , so dass $\left[y_i \left(\frac{1}{\sqrt{\langle \vec{w}, \vec{w} \rangle}} \left[\langle \vec{w}, \vec{x_i} \rangle + b \right] \right) \geq \xi \right]$ für ∀i ∈ [1..n]

Maximum Margin Hyperplane

maximiere ξ in $\left[y_i \left(\frac{1}{\sqrt{\langle \vec{w}, \vec{w} \rangle}} \left[\langle \vec{w}, \vec{x_i} \rangle + b \right] \right) \geq \xi \right]$; für ∀i ∈ [1..n]

Setze $\frac{1}{\sqrt{\langle \vec{w}, \vec{w} \rangle}} = \xi$: max. $\frac{1}{\sqrt{\langle \vec{w}, \vec{w} \rangle}}$, mit $\left(y_i \cdot \left(\xi \cdot \left(\langle \vec{w}, \vec{x_i} \rangle + b \right) \right) \geq \xi \right)$ ∀i ∈ [1..n]

$\Rightarrow$ max. $\frac{1}{\sqrt{\langle \vec{w}, \vec{w} \rangle}}$, mit $\left(y_i \left(\langle \vec{w}, \vec{x_i} \rangle + b \right) \geq 1 \right)$ ∀i ∈ [1..n]

Statt $\frac{1}{\sqrt{\langle \vec{w}, \vec{w} \rangle}}$ invertiere, quadriere und minimiere das Ergebnis:

Primäres OP: minimiere $J(\vec{w},b) = \langle \vec{w}, \vec{w} \rangle$

unter Nebenbedingung für ∀i ∈ [1..n] sei $\left(y_i \left(\langle \vec{w}, \vec{x_i} \rangle + b \right) \geq 1 \right)$

Maximum Margin Hyperplane

Zur Berechnung wird das primäre Optimierungsproblem in ein duales OP überführt.
(Umformulierung in Form mit Langrange Multiplikatoren nach Karush-Kuhn-Tucker)

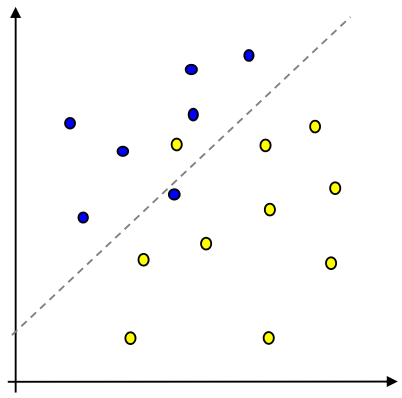
Duales OP: maximiere $L(\vec{\alpha}) = \left(\sum_{i=1}^{n} \alpha_i\right) - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \cdot \alpha_j \cdot y_i \cdot y_j \cdot \left\langle \vec{x}_i, \vec{x}_j \right\rangle$

unter Bedingung $\sum_{i=1}^{n} \alpha_i \cdot y_i = 0$, $0 \leq \alpha_i$ und $\vec{\alpha} \in \Re^n$

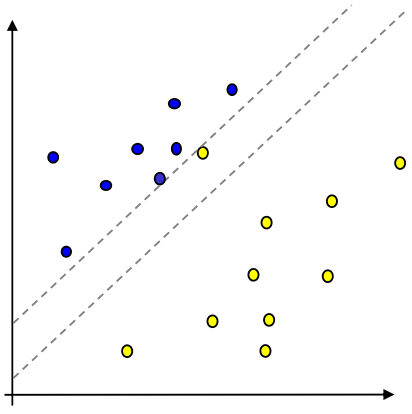
⇒ Lösung des Problems mit Algorithmen aus der Optimierungstheorie
⇒ bis jetzt nur linear separierbarer Fall: Soft Margin Optimierung
⇒ Einführung von Kernelfunktionen zur Steigerung der Kapazität

Soft Margin

Behandlung nicht linear trennbarer Daten:
Soft Margin Optimierung



Daten nicht separierbar

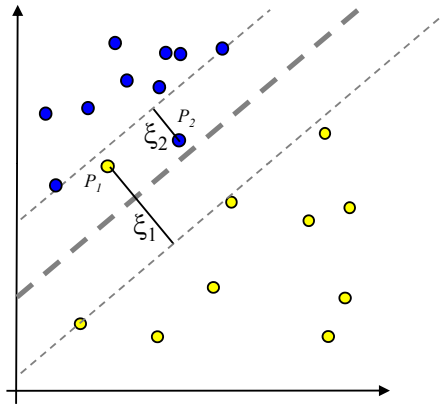


vollständige Separation ist nicht optimal

⇒ Trade-Off zwischen Trainingsfehler und Breite des Randes

Soft Margin

Betrachte beim Optimieren zusätzlich noch die Anzahl der Trainingsfehler.



- *ξ_i ist der Abstand von P_i zum Rand (wird auch Slack-Variable genannt)*
- *C reguliert den Einfluss eines einzelnen Trainingsvektors*

Primäres OP : minimiere $J(\vec{w}, b, \vec{\xi}) = \frac{1}{2}\langle \vec{w}, \vec{w} \rangle + C \cdot \sum_{i=1}^{n} \xi_i$

unter Nebenbedingung für $\forall i \in [1..n]$ sei $y_i\left(\langle \vec{w}, \vec{x_i} \rangle + b\right) \geq 1 - \xi_i$ und $\xi_i \geq 0$

$\Rightarrow$ Primäres Optimierungsproblem unter weichen Grenzen (Soft Margin)

Soft Margin

Das duale OP mit Lagrange Multiplikatoren verändert sich wie folgt:

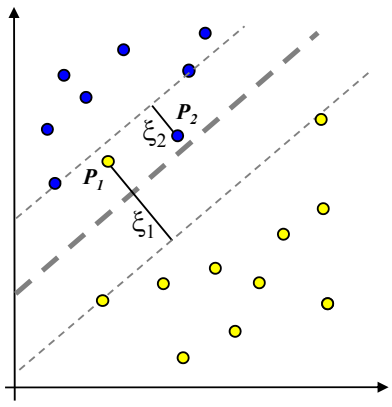
Duales OP: maximiere $L(\vec{\alpha}) = \left(\sum_{i=1}^{n} \alpha_i\right) - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \cdot \alpha_j \cdot y_i \cdot y_j \cdot \langle \vec{x_i}, \vec{x_j} \rangle$

mit Bedingung $\sum_{i=1}^{n} \alpha_i \cdot y_i = 0$ und $0 \leq \alpha_i \leq C$

- *$0 < \alpha_i < C \Leftrightarrow$ Support Vektor mit $\xi_i = 0$*
- *$\alpha_i = C \Leftrightarrow$ Support Vektor mit $\xi_i > 0$*
- *$\alpha_i = 0$ sonst*

Entscheidungsregel:

$$h(\vec{x}) = sign\left(\sum_{x_i \in SV} \alpha_i \cdot y_i \langle \vec{x_i}, \vec{x} \rangle + b\right)$$

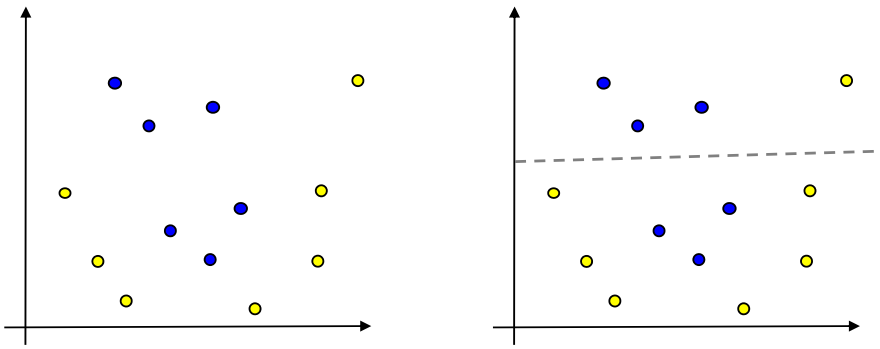


Kernel Machines

Lernen bei nicht linear trennbaren Datenmengen

Problem: Bei realen Problemen ist häufig keine lineare Separation mit hoher Klassifikationsgenauigkeit mehr möglich.

Idee: Transformiere Daten in einen nicht linearen Feature-Raum und versuche sie im neuen Raum linear zu separieren.
(Erweiterung des Hypothesenraumes)



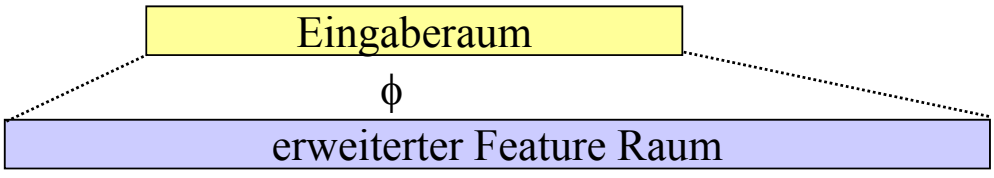
Beispiel: quadratische Transformation

DATABASE SYSTEMS GROUP

Kernel Machines

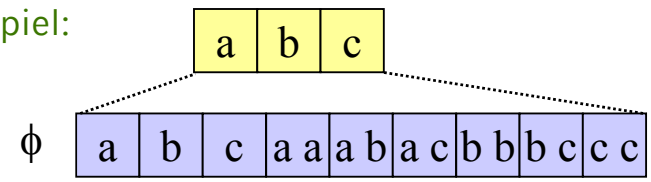
LMU

Erweiterung der Hypothesenraumes



⇒ Versuche jetzt in erweitertem Feature Raum linear zu separieren

Beispiel:

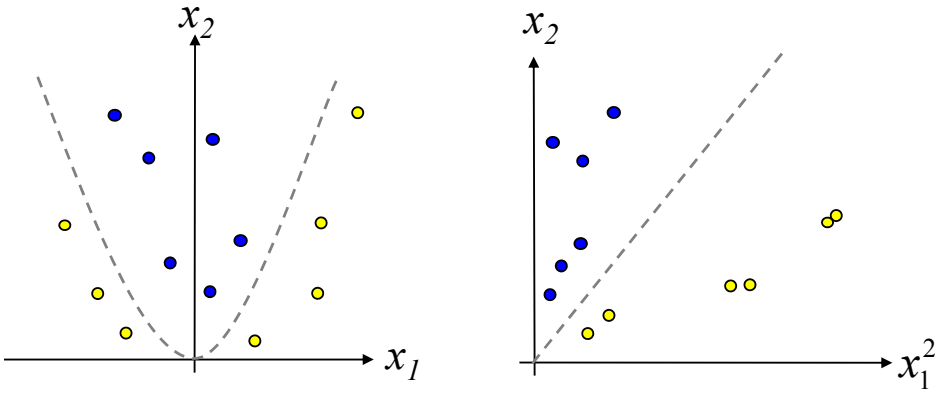


hier: Eine Hyperebene im erweiterten Feature Raum ist ein Polynom 2. Grades im Eingaberaum.

Kernel Machines

Eingaberaum: $\vec{x} = (x_1, x_2)$ (2 Attribute)

im erweiterten Raum: $\phi(\vec{x}) = (x_1^2, x_2^2, \sqrt{2} \cdot x_1, \sqrt{2} \cdot x_2, \sqrt{2} \cdot x_1 \cdot x_2, 1)$
(6 Attribute)



Kernel Machines

Einführung eines Kernels ⇔ (Implizite) Featuretransformation mittels $\phi(\vec{x}): FS_{alt} \longrightarrow FS_{neu}$

Duales OP: maximiere

$$L(\vec{\alpha}) = \left(\sum_{i=1}^{n} \alpha_i \right) - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \cdot \alpha_j \cdot y_i \cdot y_j \cdot \left\langle \phi(\vec{x}_i), \phi(\vec{x}_j) \right\rangle$$

mit Bedingung $\sum_{i=1}^{n} \alpha_i \cdot y_i = 0$ und $0 \le \alpha_i \le C$

Zusätzliche Featuretransformation wirkt sich nur auf das Skalarprodukt der Trainingsvektoren aus. ⇒ Kernel *K* ist eine Funktion mit:

$$K_\phi(\vec{x}_i, \vec{x}_j) = \left\langle \phi(\vec{x}_i), \phi(\vec{x}_j) \right\rangle$$

Wann ist eine Funktion K(x,y) ein Kernel ?

Wenn die **Kernel-Matrix** (Gram Matrix) KM

$$KM(K) = \begin{pmatrix} K(\vec{x}_1, \vec{x}_1) & .. & K(\vec{x}_1, \vec{x}_n) \\ .. & .. & .. \\ K(\vec{x}_n, \vec{x}_1) & .. & K(\vec{x}_n, \vec{x}_n) \end{pmatrix}$$

positiv (semi) definit ist, also keine negativen Eigenwerte besitzt, dann ist *K(x,y)* ein Kernel (siehe Mercer's Theorem)

Notwendige Bedingungen:

- $K_\phi(\vec{x}, \vec{y}) = \langle \phi(\vec{x}), \phi(\vec{y}) \rangle = \langle \phi(\vec{y}), \phi(\vec{x}) \rangle = K_\phi(\vec{y}, \vec{x})$ (Symmetrie)
- $K_\phi(\vec{x}, \vec{y})^2 \leq K_\phi(\vec{x}, \vec{x}) \cdot K_\phi(\vec{y}, \vec{y})$ (Cauchy-Schwarz)

Symmetrie und Cauchy-Schwarz sind keine hinreichenden Bedingungen!

einige Regeln zur Kombination vom Kerneln:

$$K(\vec{x}, \vec{y}) = K_1(\vec{x}, \vec{y}) \cdot K_2(\vec{x}, \vec{y})$$

$$K(\vec{x}, \vec{y}) = K_1(\vec{x}, \vec{y}) + K_2(\vec{x}, \vec{y})$$

$$K(\vec{x}, \vec{y}) = a \cdot K_1(\vec{x}, \vec{y})$$

$$K(\vec{x}, \vec{y}) = x^T \cdot B \cdot y$$

für K_1, K_2 Kernelfunktionen, a eine positive Konstante und B eine symmetrische positiv semi-definite Matrix.

Kernel Machines

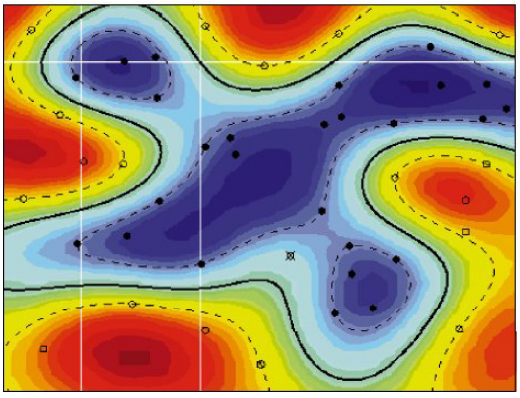
Beispiele für verwendete Kernel-Funktionen:

- linear: $K(\vec{x},\vec{y}) = \langle \vec{x},\vec{y} \rangle$
- polynomiell: $K(\vec{x},\vec{y}) = \left(\langle \vec{x},\vec{y} \rangle + c \right)^d$
- Radiale Basisfunktionen: $K(\vec{x},\vec{y}) = \exp\left(-\gamma \cdot \left| \vec{x} - \vec{y} \right|^2 \right)$
- Gauss Kernel: $K(\vec{x},\vec{y}) = \exp\left(-\frac{\left\| \vec{x} - \vec{y} \right\|^2}{2\sigma^2} \right)$
- Sigmoid: $K(\vec{x},\vec{y}) = \tanh\left(\gamma \cdot \langle \vec{x},\vec{y} \rangle + c \right)$

DATABASE SYSTEMS GROUP

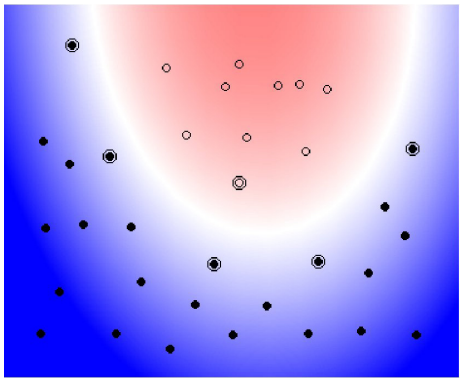
Kernel Machines

LMU



Radial Basis Kernel

Polynomieller Kernel (Grad 2)



Training einer SVM

zu lösen ist folgendes Problem:

Duales OP: maximiere

$$L(\vec{\alpha}) = \left(\sum_{i=1}^{n} \alpha_i\right) - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \cdot \alpha_j \cdot y_i \cdot y_j \cdot K\left(\vec{x}_i, \vec{x}_j\right)$$

mit Bedingung $\sum_{i=1}^{n} \alpha_i \cdot y_i = 0$ und $0 \le \alpha_i \le C$

oder

$$\max \begin{bmatrix} 1 \\ .. \\ 1 \end{bmatrix}^T \cdot \begin{bmatrix} \alpha_1 \\ .. \\ \alpha_n \end{bmatrix} - \frac{1}{2} \begin{bmatrix} \alpha_1 \\ .. \\ \alpha_n \end{bmatrix}^T \cdot \begin{bmatrix} y_1 y_1 K(\vec{x}_1, \vec{x}_1) & .. & y_1 y_n K(\vec{x}_1, \vec{x}_n) \\ .. & .. & .. \\ y_n y_1 K(\vec{x}_n, \vec{x}_1) & .. & y_n y_n K(\vec{x}_n, \vec{x}_n) \end{bmatrix} \cdot \begin{bmatrix} \alpha_1 \\ .. \\ \alpha_n \end{bmatrix}$$

mit Bedingung $\sum_{i=1}^{n} \alpha_i \cdot y_i = 0$ und $0 \le \alpha_i \le C$

Training einer SVM

zur Lösung:

- Standardalgorithmen aus der Optimierungstheorie für *konvexe quadratische Probleme*
- für große Trainingsmengen numerische Algorithmen notwendig
- es existieren einige Spezialalgorithmen für *SVM*-Training:
  - Chunking / Decomposition
  - Sequential Minimal Optimisation (*SMO*)

Multi-Class SVMs

Bisher SVMs nur anwendbar auf 2 Klassen Probleme !!

Idee: Kombination mehrere 2-Klassen SVMs zu Klassifikatoren, die beliebig viele Klassen unterscheiden können.

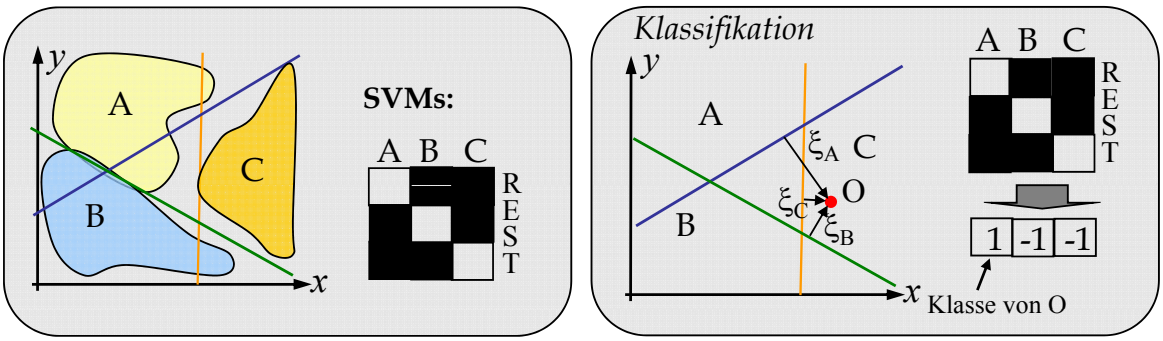
Multi-Class SVMs

2 klassische Ansätze:

⇒ Unterscheide jede Klasse von allen anderen *(1-versus-Rest)*

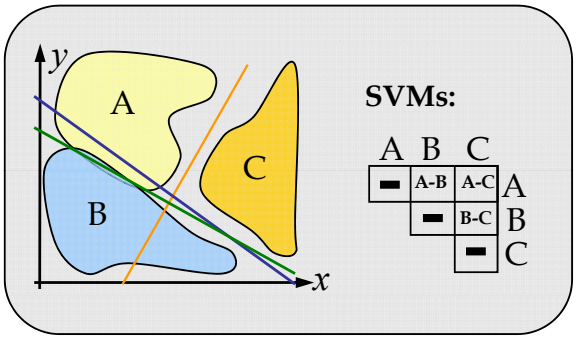
⇒ Unterscheide je 2 Klassen *(1-versus-1)*

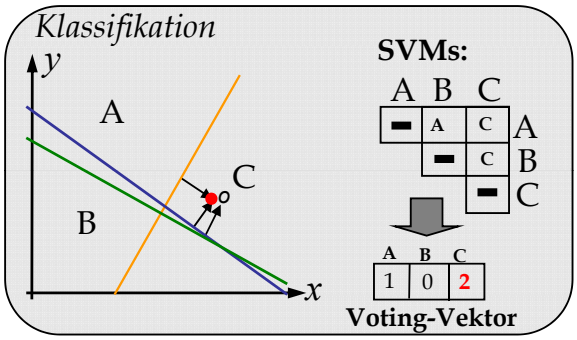
1-versus-Rest Ansatz



- 1-versus-Rest Ansatz : 1 SVM für jede Klasse.
- SVM trennt jedes Klasse von Vereinigung aller anderen Klassen ab
- Klassifiziere O mit allen Basis-SVMs.
  ⇒ **Multiple Klassenzugehörigkeit möglich (Multi-Classification)**
  oder
  Entscheidung für die Klasse, bei der Abstand ξ_i am größten ist.

1-versus-1 Ansatz





- 1-versus-1 Ansatz : 1 SVM für jedes Paar von Klassen.
- Klassifikation von Objekt O:
  1. Klassifiziere O mit allen Basis-SVMs.
  2. Zähle "Stimmen"(Votes) für jede Klasse.
- Maximale Anzahl an Votes => Ergebnis.

Vergleich 1-versus-rest und 1-versus-1

Kriterium	1-versus-rest	1-versus-1
Aufwand Training	linear zur Anzahl der Klassen ($O(\|K\|)$)	Quadratisch zur anzahl der Klassen ($O(\|K\|^2)$)
Aufwand Klassifikation	linear zur Anzahl der Klassen ($O(\|K\|)$)	Quadratisch zur Anzahl der Klassen ($O(\|K\|^2)$) **Verbesserung:** „Decision Directed Acyclic Graphs" Klassifikation in $O(\|K\|)$ [Platt,Christianini 1999]
Genauigkeit	tendenziell schlechter	tendenziell höher, (nutzt wissen über unterschiedliche Klassen besser aus)

Diskussion

+ erzeugt Klassifikatoren mit hoher Genauigkeit
+ verhältnismäßig schwache Tendenz zu Overfitting
  (Begründung durch Generalisierungtheorie)
+ effiziente Klassifikation neuer Objekte
+ kompakte Modelle

- unter Umständen lange Trainingszeiten
- aufwendige Implementierung
- gefundene Modelle schwer zu deuten

Literatur:

- C. Cortes, V. Vapnik: Support-vector networks.
  Machine Learning, 20:273-297, November 1995.
- C.J.C. Burges: A tutorial on support vector machines for pattern recognition.
  Data Mining and Knowledge Discovery, 2(2):121-167,1998.
- T. Joachims: Text categorisation with Support Vector Machines.
  in Proceedings of European Conference on Machine Learning (ECML), 1998.
- N. Cristianini, J Shawne-Taylor: An Introduction to Support Vector Machines and other kernel-based learning methods.
  Cambridge University Press 2000.