Syntaktische und Statistische Mustererkennung

VO 1.0 840.040
(UE 1.0 840.041)

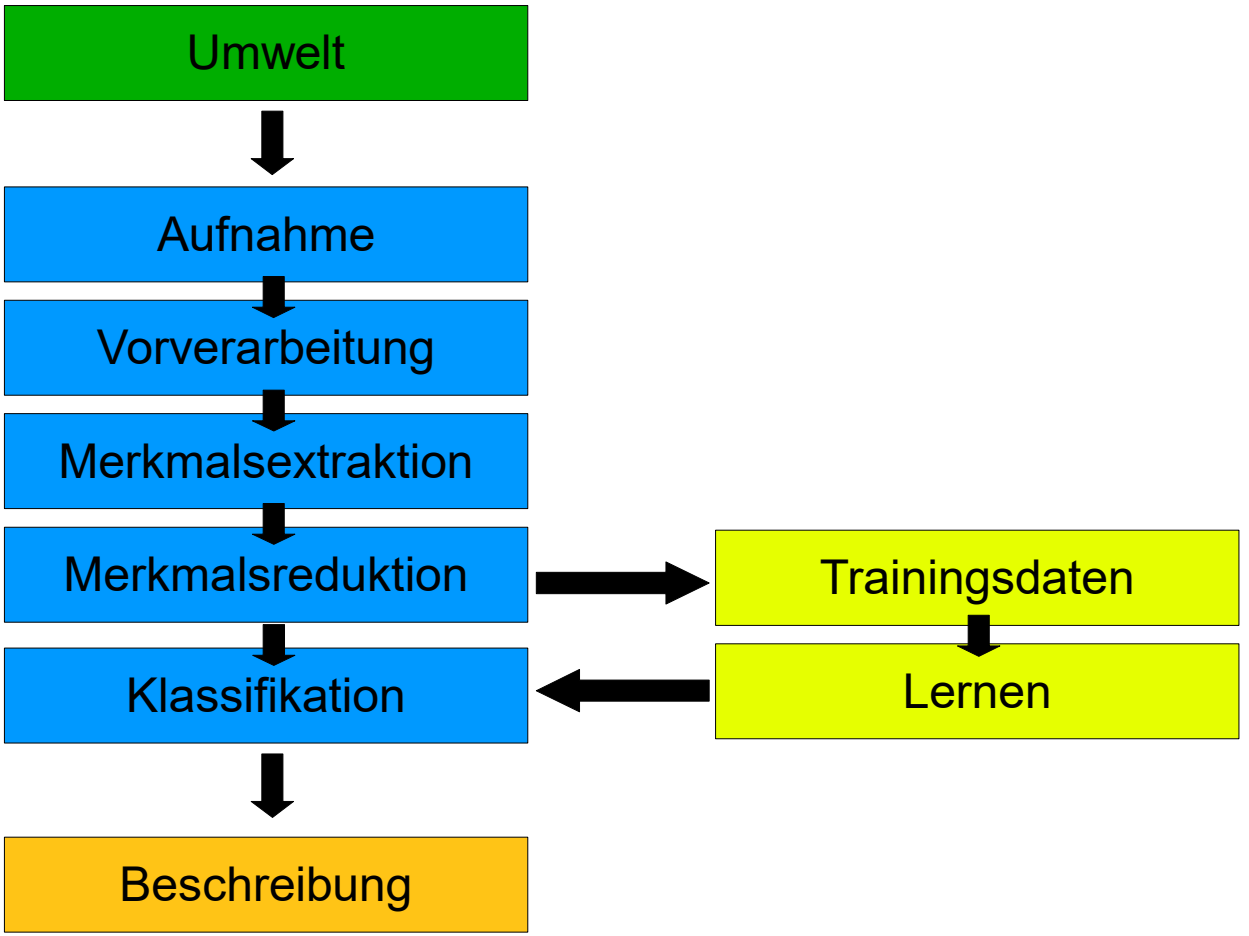
Bernhard Jung

bernhard@jung.name
http://bernhard.jung.name/VUSSME/

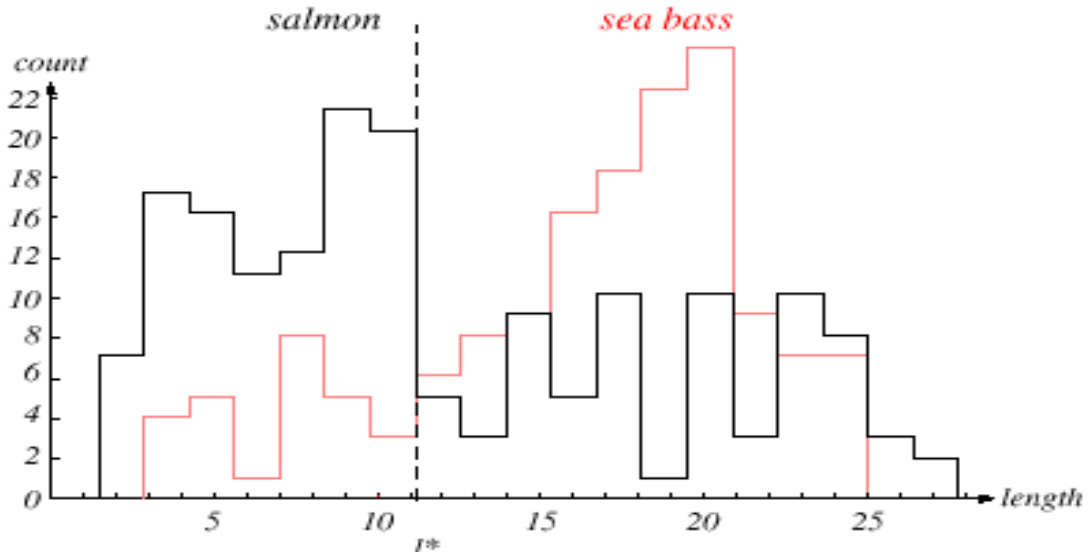
Rückblick

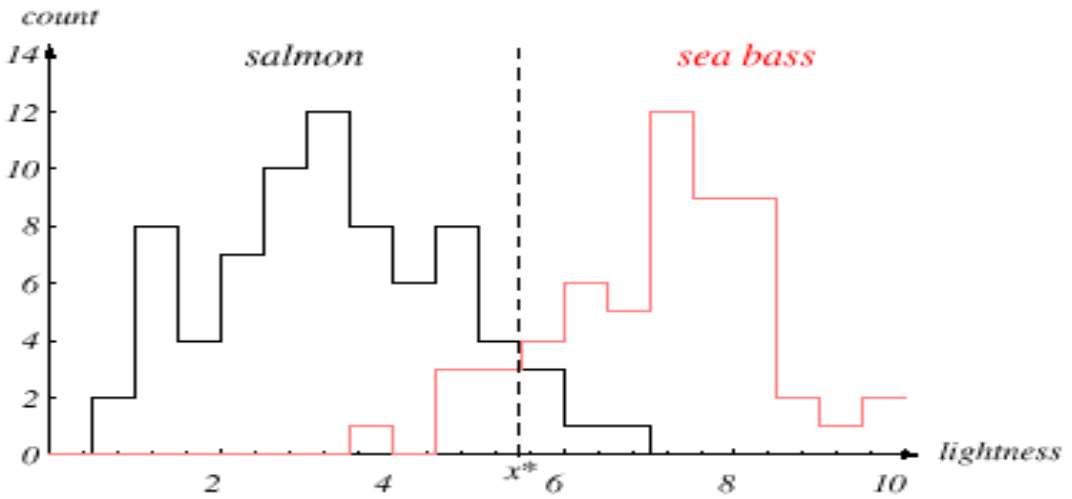
- Was ist ein Muster? Wo treten Muster auf? Fokus auf perzeptive Muster.
- Mustererkennung: Theorie und Praxis, Beispiele
- Lachs oder Wolfsbarsch
- Schritte des Mustererkennungsprozesses
- Klassifikation, Regression, Clustering
- Entscheidungsgrenze:
  Optimalität, Generalisierungsfähigkeit
- Musteranalyse

Mustererkennungsprozess

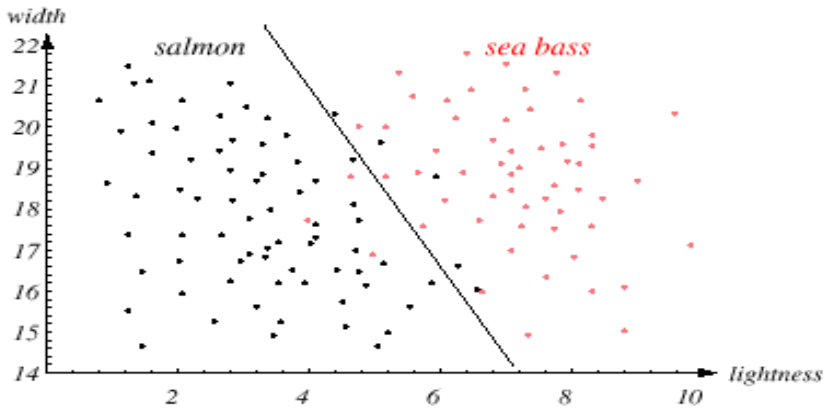


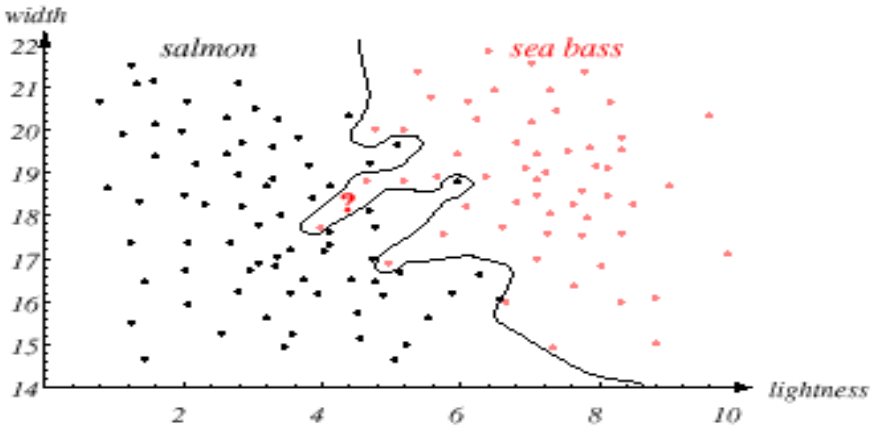
Merkmale

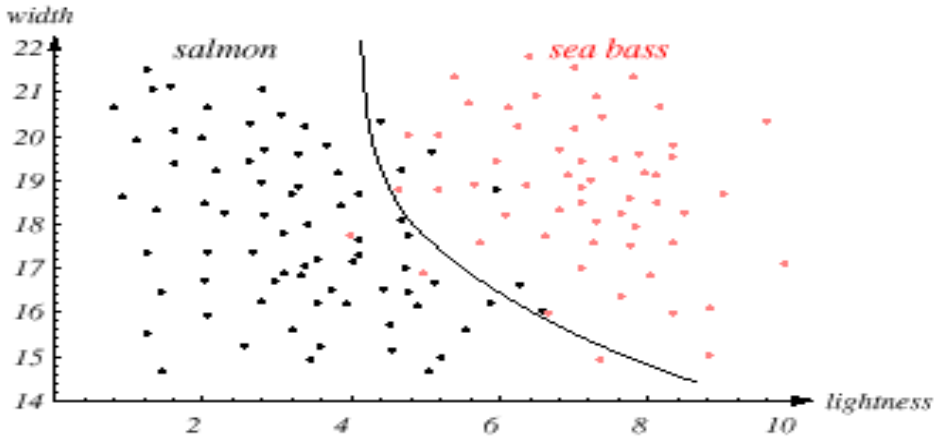

salmon
sea bass
count
22
20
18
16
12
10
8
6
4
2
0
5
10
15
20
25
length
l*


count
14
12
10
8
6
4
2
0
salmon
sea bass
2
4
x*
6
8
10
lightness

Entscheidungsgrenzen







Ähnlichkeiten

Notwendig und bestimmend für Klassifikation

- Abstands-/Distanzmaße
- Pattern Matching
- Matching Score (strukturelles Matchen)
- Komplexe (graphen-basierte) Strukturen
- Syntaktische Analyse (Grammatik)

Metrik (Distanzmaß)

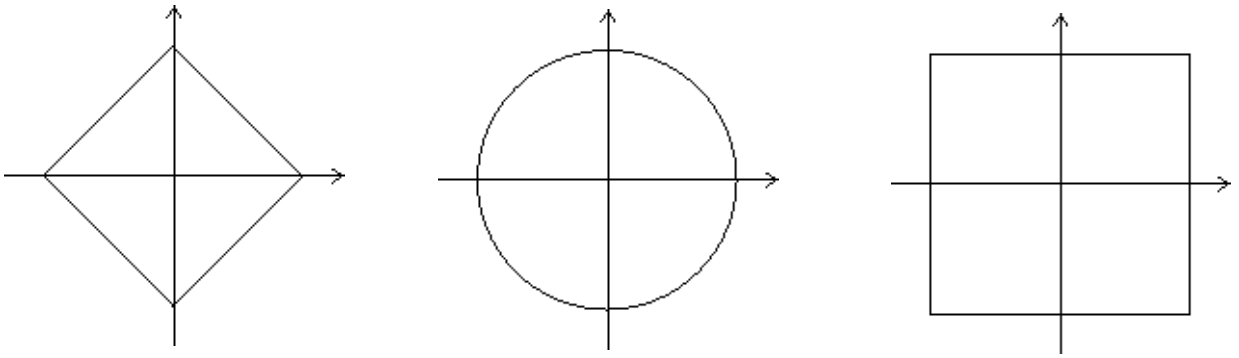
Abbildung:

$$d : \chi \times \chi \rightarrow \Re_{\geqslant 0}$$
$$d(x, x) = 0$$
$$d(x, y) = 0 \Rightarrow x = y$$
$$d(x, y) = d(y, x)$$
$$d(x, y) \leq d(x, z) + d(z, y)$$
$$d(x, y) \geq 0$$

- Pseudometrik:
  - d(x,y) kann 0 sein
- Positivdefinitheit folgt anderen Axiomen

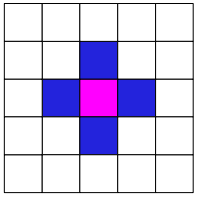
Normen

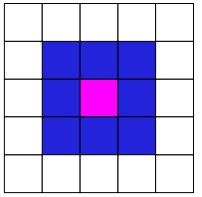
- Norm: $d(x,y)=||x-y||$ auf einem Vektorraum = Metrik
- *p*-Normen $||x||_p := (\sum_{i=1}^{n} |x_i|^p)^{1/p}$
- p=1: Betragssummennorm, Manhattan-Distanz
- p=2: Euklidische Norm $||x||_2 := \sqrt{\sum_{i=1}^{n} |x_i|^2}$
- p=∞: Maximumsnorm $||x||_\infty := \max_{i=1}^{n} |x_i|$



Abstandsmaße in der Bildverarbeitung

Nachbarschaften



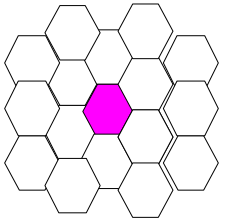


4-connectedness 8-connectedness

Manhatten-Distanz Maximumsnorm

Willkürliche Festlegung auf quadratische Pixel

- Alternative: Hexagonale Strukturen



6-connectedness

Euklidische Distanz

Andere Normen

- Matrixnormen
  - Spaltensummen
  - Zeilensummen
  - Spektralnorm
  - Gesamtnorm
  - Frobeniusnorm
- (Operatornormen)

Andere Metriken (1)

Entfernungstabelle

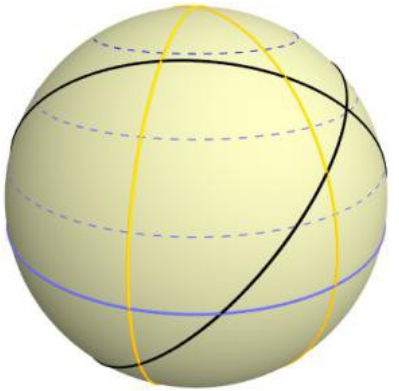
Die Entfernungstabelle Deutschland (gelb) und Europa (blau) können nur wie angegeben miteinander verknüpft werden.

Die rot markierten Städte ermöglichen die Berechnung von Entfernungen zwischen Orten in Deutschland und Europa (siehe Bsp.)	Amsterdam	Athen	Barcelona	Belgrad	Berlin	Bern	Brüssel	Budapest	Bukarest	Dublin	Frankfurt M.	Hamburg	Helsinki	Istanbul	Köln	Kopenhagen	Lissabon	London	Madrid	Minsk	Moskau	München	Oslo	Paris	Prag	Rom	Sofia	Stockholm	Warschau	Wien	Entfernungen Europa Orte aus Deutschland-tabelle in rot
Aachen		2950	1600	1800	650	950	200	1400	2220	950	450	450	1700	2700	260	800	2350	500	1800	1800	2500	850	1300	500	900	1700	2100	1450	1200	1150	Amsterdam
Basel	545		3200	1150	2400	2700	2900	1500	1150	3600	2500	2750	3650	1100	2700	2750	4500	3200	3800	2400	2900	2150	3450	2900	2050	900	750	3450	2200	1750	Athen
Berlin	650	875		2050	1900	950	1350	2000	2600	2000	1350	1750	3050	2950	1400	2100	1300	1500	650	2900	3600	1400	2650	1100	1750	1400	2400	2750	2500	1900	Barcelona
Bremen	370	775	400		1300	1450	1750	400	600	2500	1300	1700	2500	950	1500	1700	3350	2050	2670	1400	2200	950	2300	1800	900	1300	400	2300	1100	650	Belgrad
Dortmund	155	555	495	235		950	750	900	1700	1550	550	300	1250	2200	580	350	2900	1050	2350	1150	1850	590	950	1050	400	1550	1650	950	600	650	Berlin
Dresden	645	745	200	490	515		700	1150	1950	1250	400	950	2050	2300	600	1350	2200	1000	1550	2000	2700	450	1800	550	850	1000	1750	1900	1600	1000	Bern
Düsseldorf	90	550	560	285	70	580		1400	2200	800	450	600	1900	2650	250	950	2150	300	1550	1900	2600	800	1550	300	950	1550	2100	1650	1400	1150	Brüssel
Emden	375	845	520	140	305	620	290		800	2150	1000	1200	2250	1350	1150	1200	3300	1700	2600	1200	1900	700	1900	1550	500	1250	800	1900	700	250	Budapest
Erfurt	440	585	300	340	310	215	375	470		3000	1800	2000	2500	700	2000	2050	3950	2500	3200	1300	1800	1500	2850	2300	1350	1950	400	3150	1200	1050	Bukarest
Flensburg	625	980	450	275	490	660	540	400	520		1200	1350	2750	3500	1000	1700	2750	500	2200	2700	3400	1550	2250	900	1700	2350	2950	2500	2100	1900	Dublin
Frankfurt M.	255	335	550	445	225	460	225	520	260	650		500	1600	2250	200	800	2450	700	1900	1650	2350	400	1350	600	500	1300	1700	1500	1050	700	Frankfurt M.
Frankfurt O.	700	940	105	460	560	180	625	590	370	550	610		1050	2600	430	300	2800	900	2200	1450	2200	780	850	950	700	1750	2050	950	850	1100	Hamburg
Garm.-Patenk.	700	370	675	835	700	550	680	930	510	1020	480	740		3250	1600	850	3950	2100	3400	1250	1150	1850	750	2100	1550	2900	2950	200	1500	1850	Helsinki
Görlitz	750	840	220	575	610	105	680	700	320	690	560	170	650		2450	2650	4250	2950	3500	2000	2500	1900	3500	2700	1850	1600	550	3300	1850	1600	Istanbul
Hamburg	480	825	300	130	350	500	400	255	370	160	500	385	860	530		750	2400	550	1800	1550	2250	580	1150	500	700	1500	1900	1300	1150	900	Köln
Hannover	355	680	290	130	215	385	280	260	220	310	350	350	720	470	155		3100	1250	2500	1550	2250	1100	600	1250	750	2050	2100	650	900	1000	Kopenhagen
Kassel	310	525	385	285	165	350	235	390	150	470	200	450	560	450	310	170		2250	650	4000	4700	2700	3650	1850	3000	2700	3700	3800	3550	3050	Lissabon
Koblenz	165	405	600	410	190	510	145	425	310	670	120	660	550	610	520	390	250		1700	2150	2850	1100	1800	400	1200	1800	2400	1900	1650	1450	London
Köln	75	505	580	315	100	570	40	330	370	570	200	640	650	670	430	300	250	105		3500	4200	2050	3100	1250	2350	2050	3000	3200	3000	2450	Madrid
Leipzig	570	710	190	390	440	115	505	520	140	570	385	255	520	215	440	290	280	440	500		700	1600	2000	2150	1200	2450	1700	1400	550	1300	Minsk
Mannheim	290	270	615	515	300	530	280	600	330	725	85	680	410	630	570	430	270	150	250	460		2300	1900	2850	1900	3150	2150	1300	1250	2000	Moskau
München	650	390	590	750	605	460	610	840	420	930	400	650	90	560	780	630	470	490	580	430	350		1600	850	400	950	1350	1600	1050	450	München
Nürnberg	470	450	440	585	440	315	445	675	270	795	230	510	260	420	610	470	310	340	410	280	240	170		1800	1350	2600	2900	550	1400	1950	Oslo
Passau	690	580	630	800	660	465	655	890	460	980	440	700	280	570	820	680	520	550	620	470	440	190	220		1100	1450	2150	1950	1600	1250	Paris
Rostock	650	1000	230	300	515	440	570	425	490	285	740	325	870	470	180	330	560	690	600	380	810	780	630	820		1350	1300	1350	650	300	Prag
Saarbrücken	310	265	725	550	330	640	280	560	440	810	200	790	500	740	660	530	380	180	250	570	140	430	370	570	920		1050	2750	1850	1150	Rom
Salzburg	800	530	735	910	755	605	770	980	580	1080	540	800	180	540	920	780	625	660	720	570	510	150	320	120	940	600		2750	1450	1000	Sofia
Stuttgart	420	270	630	630	420	510	410	710	350	810	210	700	300	610	660	520	360	280	370	470	135	230	210	400	820	220	390		450	1900	Stockholm
Trier	160	325	715	480	260	635	200	500	430	735	190	780	580	730	590	460	340	140	180	560	180	500	420	620	760	100	700	300		700	Warschau
Wiesbaden	230	350	570	430	210	490	200	480	280	680	40	640	500	590	520	380	220	80	170	410	100	430	260	470	760	160	570	220	150		Wien
Entfernungen Deutschland Orte aus Europatabelle in rot	Aachen	Basel	Berlin	Bremen	Dortmund	Dresden	Düsseldorf	Emden	Erfurt	Flensburg	Frankfurt M.	Frankfurt O.	Garm.-Patenk.	Görlitz	Hamburg	Hannover	Kassel	Koblenz	Köln	Leipzig	Mannheim	München	Nürnberg	Passau	Rostock	Saarbrücken	Salzburg	Stuttgart	Trier	Wiesbaden	

Beispiel:
Berlin bis Dresden =200 km
Berlin bis Prag =400 km
Dresden bis Prag =400 km – 200 km

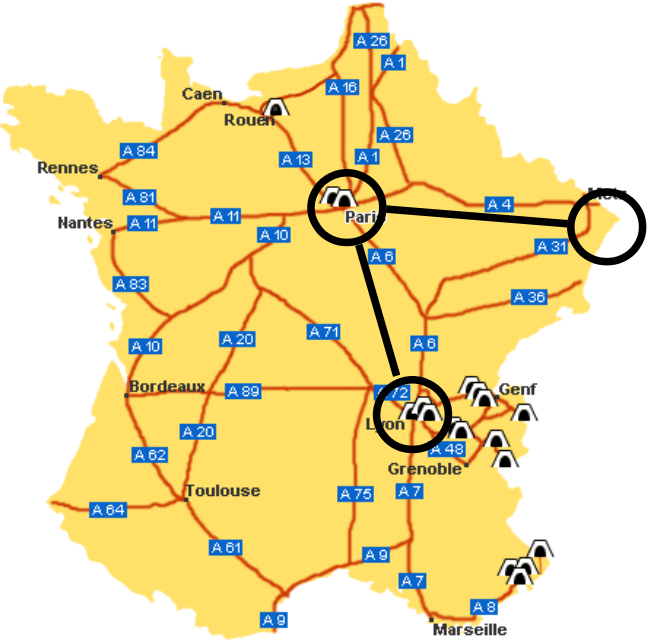
Andere Metriken (2)

- Diskrete Metrik $d(x,x)=0$
  $d(x,y)=1 \quad für\ x \neq y$



- Natürliche Metrik auf einer Kugeloberfläche (Geodäte)

- Französische Eisenbahnmetrik



Andere Metriken (3)

- Hausdorff-Metrik: Abstand zwischen Teilmengen
  - Geringerer Abstand je besser sie einander wechselseitig überdecken

$$D(x, K) := min\{d(x, k) | k \in K\}$$
$$d(A, B) := max\{max\{D(a, B) | a \in A\}, max\{D(b, A) | b \in B\}\}$$

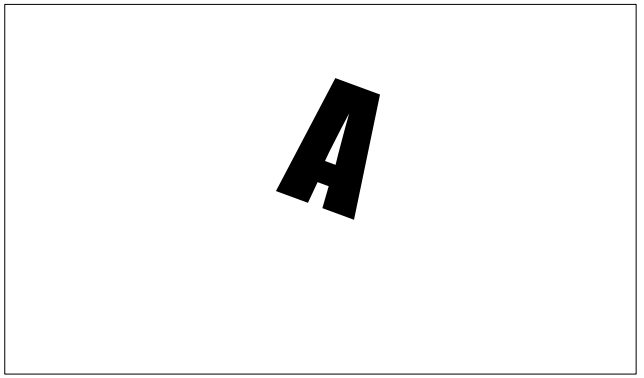
- Hamming Distanz $\Delta(x, y) := \sum_{x \neq y} 1$
- Levenshtein-Distanz (Editing-Distanz)
- Mahalanobis Distanz
  - Berücksichtigt Korrelationen in den Daten
  - Skalierungsinvariant

Metriken für qualitative Merkmale

- Nominalskala
  - Merkmale werden nummeriert, einzige Schlussfolgerung: gleiche Zahl = gleiches Merkmal
- Ordinalskala
  - Rangfolge
- Intervallskala, Verhältnisskala, Absolutskala
  - IS: gemessene Abstände zwischen Messwerten
  - VS: + natürlicher Nullpunkt
  - AS: + Maßeinheit natürlich definiert
- Lageparameter? Umgang mit fehlenden Werten?

Pattern Matching

- Schablonen-basiert
  - Probleme:
    - Varianten: A A A A A A A
    - Kongruenz (Deckungsgleichheit)



Pattern Matching

- Ähnlichkeit eines Bildes f(x,y) und einer Schablone g(x,y)

$$\sum_x \sum_y |f(x,y) - g(x,y)| \quad \text{oder} \quad \sum_x \sum_y [f(x,y) - g(x,y)]^2$$

$$\sum_x \sum_y f(x,y) \times g(x,y)$$

$$sofern \sum_x \sum_y f^2(x,y) \, und \sum_x \sum_y g^2(x,y) \, konstant \, sind$$

  - Kreuzkorrelation (cross covariance, sliding dot-product)

$$R_{fg} = \frac{\sum_x \sum_y f(x,y) \times g(x,y)}{\sqrt{\sum_x \sum_y f^2(x,y) \times \sum_x \sum_y g^2(x,y)}}$$

Pattern Matching

- Translationsinvarianz:

$$R_{fg}(\tau_x, \tau_y) = \frac{\sum_x \sum_y f(x+\tau_x, y+\tau_y) \times g(x, y)}{\sqrt{\sum_x \sum_y f^2(x, y) \times \sum_x \sum_y g^2(x, y)}}$$

- Berechnung mit Hilfe der diskreten Fourier Transformation

$$DFT^{-1}\{DFT(f(x, y)) \times DFT(g(x, y))\}$$

Pattern Matching

- Reguläre Ausdrücke
- Grammatikalische Analyse (Parsing)
  - Grammatik G = <V, T, P, S>
    - V ... Menge von Nichtterminalsymbolen
    - T ... Menge von Terminalsymbolen
    - P ... Produktionsregeln in der Form

      $X_1 X_2 ... X_n \rightarrow Y_1 Y_2 ... Y_m$
    - S ... Startsymbole
  - Überprüfung, ob ein Satz von einer gegebenen Grammatik erzeugt werden kann
  - Beschreibung der Struktur: Parse-Baum, Syntax-Baum

Entscheidungstabelle

- Auflistung aller zu berücksichtigenden Bedingungen
- Auflistung aller möglichen Aktionen
- Regeln
  - Bedingungskombinationen
  - Aktion(en)

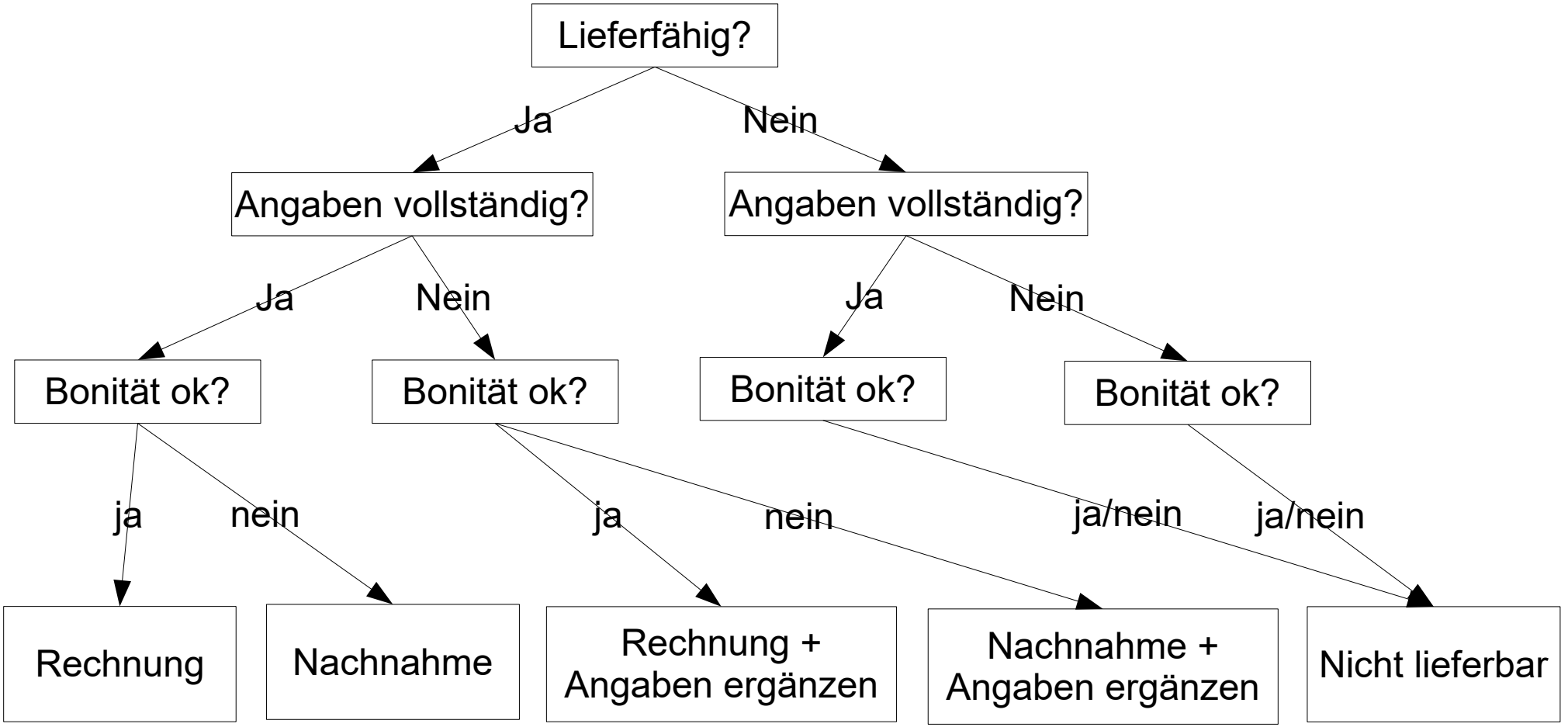
Tabellenbezeichnung	R1	R2	R3	R4	R5	R6	R7	R8
Bedingungen								
Lieferfähig	j	j	j	j	n	n	n	n
Angaben vollständig	j	j	n	n	j	j	n	n
Bonität in Ordnung	j	n	j	n	j	n	j	n
Aktionen								
Lieferung mit Rechnung	x		x					
Lieferung als Nachnahme		x		x				
Angaben vervollständigen			x	x				
Mitteilen: nicht lieferbar					x	x	x	x

Entscheidungstabellen

Eigenschaften:

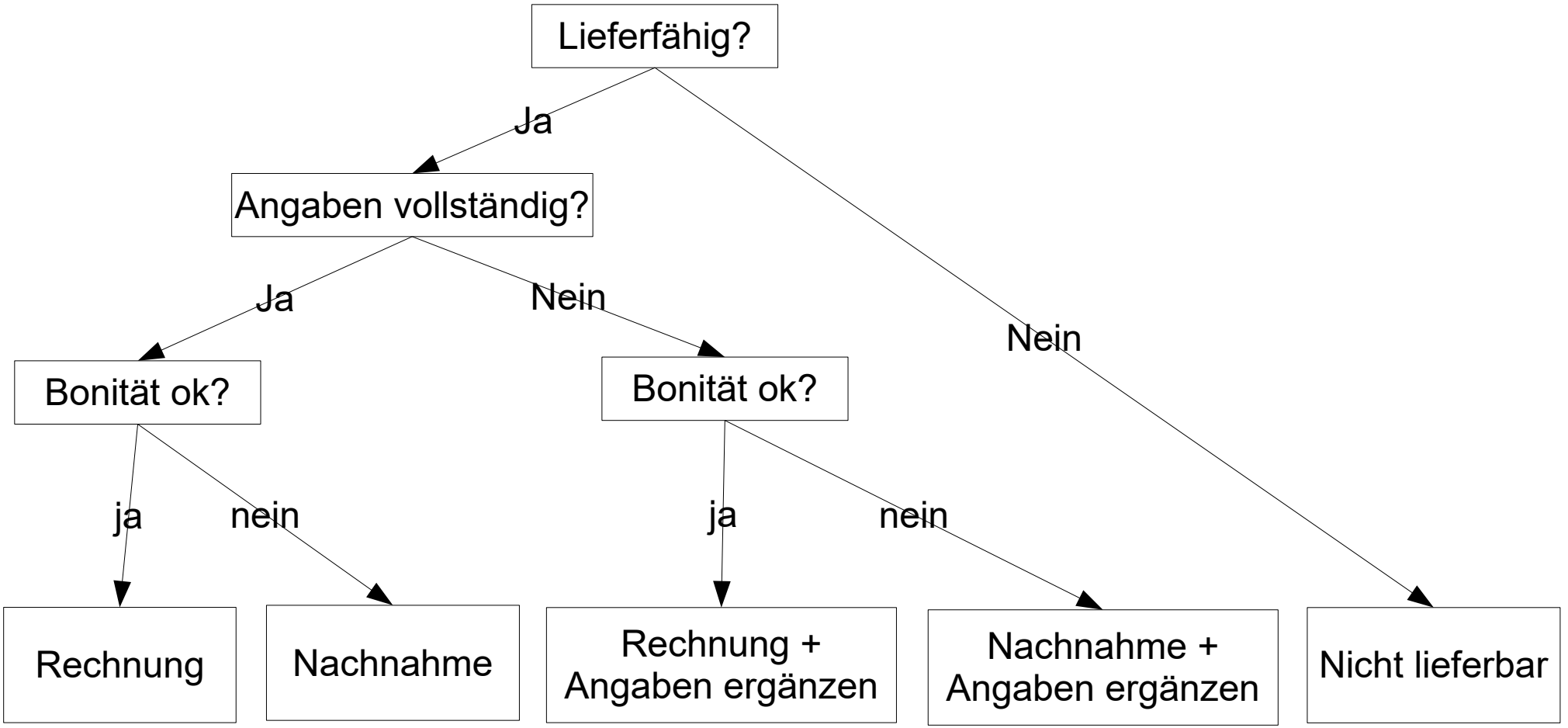
- Vollständigkeit
  - Jede mögliche (2^n) Bedingungskombination ist repräsentiert
- Konsolidierung
  - Regeln können zusammengefasst werden
- Redundanz
  - Mehrere Regeln für identische Fälle
- Konsistenz
  - Widerspruchsfreiheit

Entscheidungsbäume



Entscheidungsbäume

Redundanzelimination/Pruning



Entscheidungsbäume

- Anordnung/Reihenfolge der Merkmale?
  - Information Gain (ID3, C4.5 algorithms)
  - Gini impurity (CART)
- Entscheidungsbäume für numerische Probleme?
  - Kombinationen von einfachen Klassifikatoren
- Erweiterung: Decision Graphs

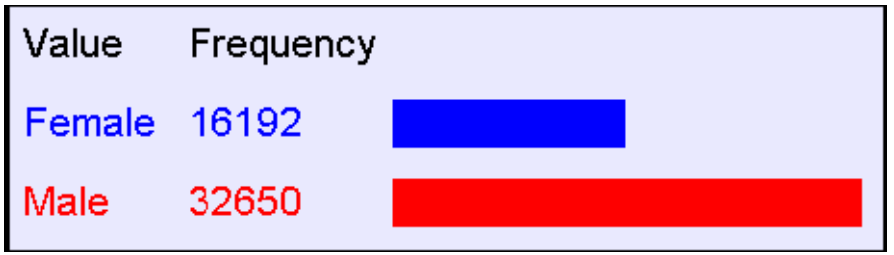
Beispieldaten für Entscheidungsbaumlernen

- US Census (1990) Income Data Set
  - Aus dem UC Irvine Machine Learning Repository http://archive.ics.uci.edu/ml/datasets/Census+Income
  - 48842 records
  - 14 attributes
  - Missing values
  - Numeric & symbolic attributes

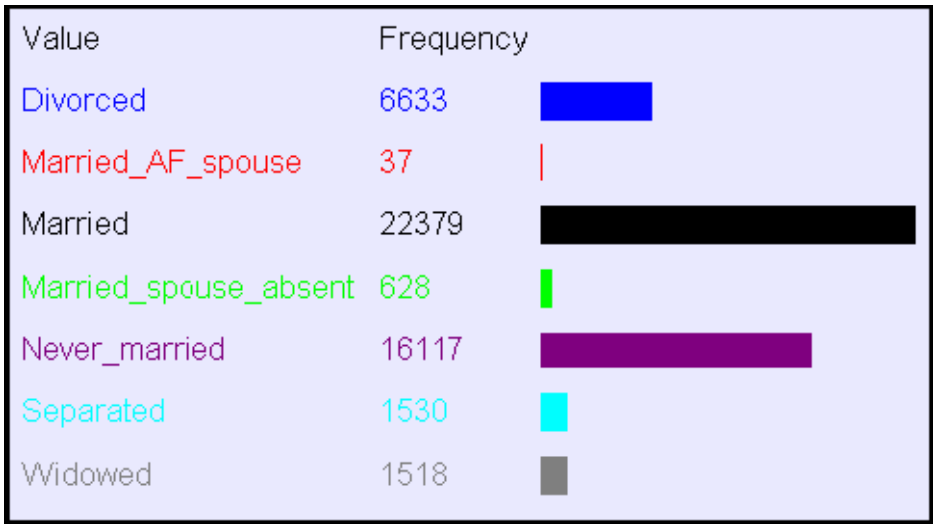
age	workclass	fnlwgt	Edu	occupation	relationship	race	sex	Hou	Native-country	wealth
39	State-gov	77516	13	Adm-clerical	Not-in-family	White	Male	40	United-States	<=50K
50	Self-emp-not-inc	83311	13	Exec-managerial	Husband	White	Male	13	United-States	<=50K
38	Private	215646	9	Handlers-cleaners	Not-in-family	White	Male	40	United-States	<=50K
53	Private	234721	7	Handlers-cleaners	Husband	Black	Male	40	United-States	<=50K
28	Private	338409	13	Prof-specialty	Wife	Black	Female	40	Cuba	<=50K
37	Private	284582	14	Exec-managerial	Wife	White	Female	40	United-States	<=50K
49	Private	160187	5	Other-service	Not-in-family	Black	Female	16	Jamaica	<=50K
52	Self-emp-not-inc	209642	9	Exec-managerial	Husband	White	Male	45	United-States	>50K
31	Private	45781	14	Prof-specialty	Not-in-family	White	Female	50	United-States	>50K
42	Private	159449	13	Exec-managerial	Husband	White	Male	40	United-States	>50K
37	Private	280464	10	Exec-managerial	Husband	Black	Male	80	United-States	>50K
30	State-gov	141297	13	Prof-specialty	Husband	Asian-Pac-Islander	Male	40	India	>50K
23	Private	122272	13	Adm-clerical	Own-child	White	Female	30	United-States	<=50K
32	Private	205019	12	Sales	Not-in-family	Black	Male	50	United-States	<=50K
40	Private	121772	11	Craft-repair	Husband	Asian-Pac-Islander	Male	40	?	>50K
34	Private	245487	4	Transport-moving	Husband	Amer-Indian-Eskimo	Male	45	Mexico	<=50K
25	Self-emp-not-inc	176756	9	Farming-fishing	Own-child	White	Male	35	United-States	<=50K
32	Private	186824	9	Machine-op-inspct	Unmarried	White	Male	40	United-States	<=50K
38	Private	28887	7	Sales	Husband	White	Male	50	United-States	<=50K
43	Self-emp-not-inc	292175	14	Exec-managerial	Unmarried	White	Female	45	United-States	>50K
40	Private	193524	16	Prof-specialty	Husband	White	Male	60	United-States	>50K
54	Private	302146	9	Other-service	Unmarried	Black	Female	20	United-States	<=50K
35	Federal-gov	76845	5	Farming-fishing	Husband	Black	Male	40	United-States	<=50K
43	Private	117037	7	Transport-moving	Husband	White	Male	40	United-States	<=50K
59	Private	109015	9	Tech-support	Unmarried	White	Female	40	United-States	<=50K
56	Local-gov	216851	13	Tech-support	Husband	White	Male	40	United-States	>50K
19	Private	168294	9	Craft-repair	Own-child	White	Male	40	United-States	<=50K
54	?	180211	10	?	Husband	Asian-Pac-Islander	Male	60	South	>50K
39	Private	367260	9	Exec-managerial	Not-in-family	White	Male	80	United-States	<=50K
49	Private	193366	9	Craft-repair	Husband	White	Male	40	United-States	<=50K
23	Local-gov	190709	12	Protective-serv	Not-in-family	White	Male	52	United-States	<=50K
20	Private	266015	10	Sales	Own-child	Black	Male	44	United-States	<=50K
45	Private	386940	13	Exec-managerial	Own-child	White	Male	40	United-States	<=50K
30	Federal-gov	59951	10	Adm-clerical	Own-child	White	Male	40	United-States	<=50K
22	State-gov	311512	10	Other-service	Husband	Black	Male	15	United-States	<=50K
48	Private	242406	7	Machine-op-inspct	Unmarried	White	Male	40	Puerto-Rico	<=50K
21	Private	197200	10	Machine-op-inspct	Own-child	White	Male	40	United-States	<=50K
19	Private	544091	9	Adm-clerical	Wife	White	Female	25	United-States	<=50K
31	Private	84154	10	Sales	Husband	White	Male	38	?	>50K
48	Self-emp-not-inc	265477	12	Prof-specialty	Husband	White	Male	40	United-States	<=50K

Visualisierung

- Histogramme
  - Gender



  - Martial Status



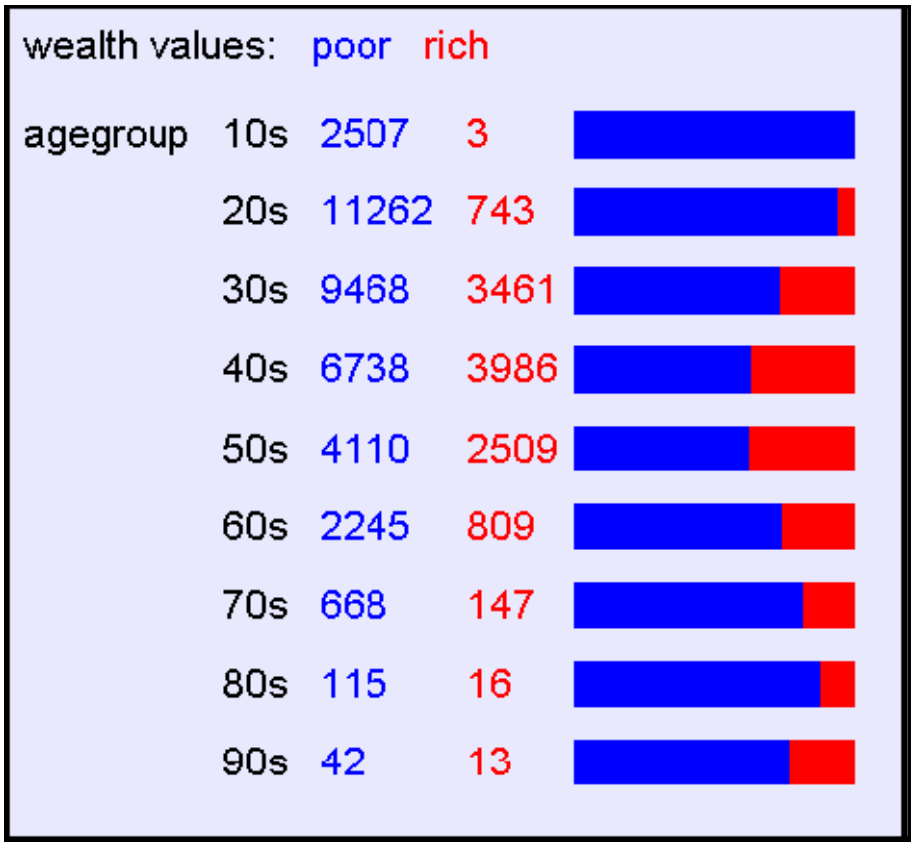
Vom Histogramm zur Kontingenztabelle

1-dimensional -> n-dimensional

- Frequenz der Kombination von n beliebigen Merkmalen

2-dimensionale Kontingenztabelle

wealth values:		poor	rich
agegroup	10s	2507	3
	20s	11262	743
	30s	9468	3461
	40s	6738	3986
	50s	4110	2509
	60s	2245	809
	70s	668	147
	80s	115	16
	90s	42	13

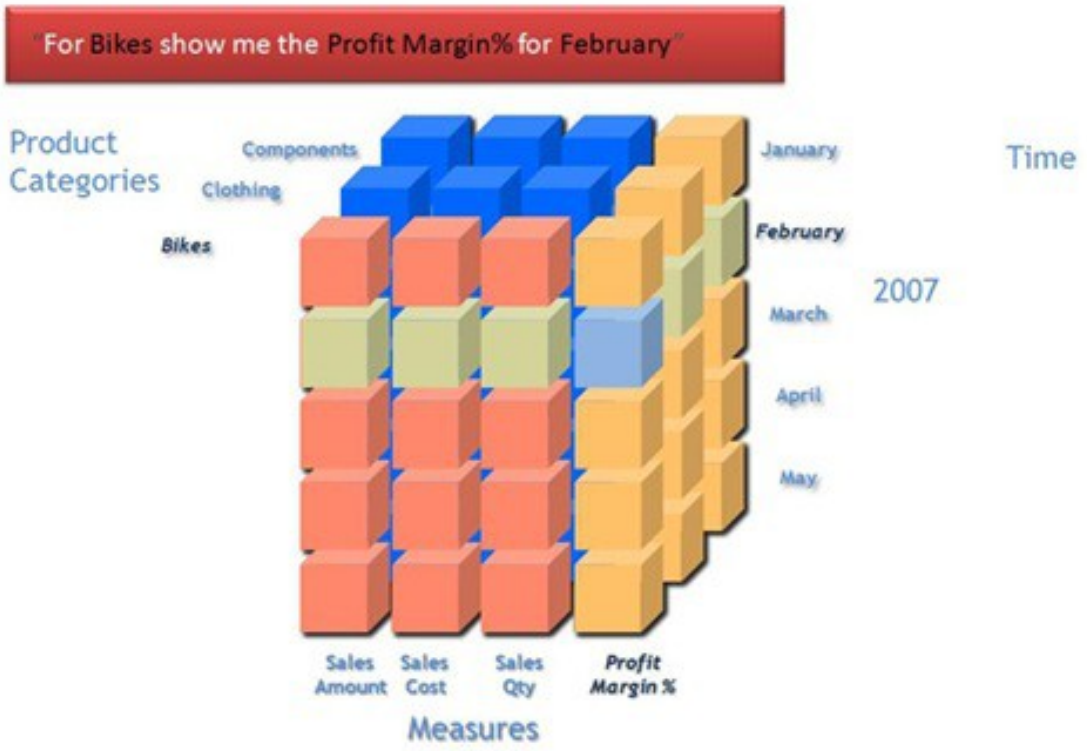


Große 2-D Kontingenztabelle

job values:	Adm_clerical	Craft_repair		Farming_fishing			Machine_op_inspct			Priv_house_serv		Protective_serv	Tech_support	
MissingValue	Armed_Forces	Exec_manageria		Handlers_cleaners			Other_service			Prof_specialty		Sales	Transport_moving	
marital	Divorced	270	1192	0	679	890	90	197	434	762	46	795	121	664	239	254
	Married_AF_spouse	5	6	0	4	3	1	1	1	5	0	4	1	5	0	1
	Married	928	1495	7	3818	3600	869	724	1469	1088	27	3182	583	2491	609	1489
	Married_spouse_absent	45	84	0	77	52	35	32	37	92	9	64	7	55	9	30
	Never_married	1242	2000	0	1301	1200	434	1029	072	2442	99	1049	207	1992	500	400
	Separated	97	224	0	160	126	23	63	123	275	21	145	23	146	48	56
	Widowed	222	250	0	73	155	38	26	86	259	40	133	11	151	35	39

3-D Datenwürfel

OLAP (on-line analytical processing):
Point & click Navigation, Histogramm Visualisierung



Kontingenztabellen?

Angenommen 16 Attribute:

Wieviele???

1-d Histogramme	16
2-d Kontingenztabellen	16*15/(2*1) = 120
3-d Würfel	16*15*14/(3*2*1) = 560

Bei 100 Attributen gibt es 161,700 mögliche
3-d Datenwürfel!!!

Suche nach Mustern

Wann ist ein Muster interessant?

- Informationsgewinn!

Informationstheorie (Shannon)

Übertragung von n unabhängigen möglichen Werten
P(X=A) = ¼ P(X=B) = ¼ P(X=C) = ¼ P(X=D) = ¼

über eine binäre Leitung: Kodierung?

A = 00 B = 01 C = 10 D = 11

2 bits pro Zeichen

Informationstheorie

Andere Kodierung bei anderer Wahrscheinlichkeitsverteilung?
P(X=A) = ½ P(X=B) = ¼ P(X=C) = 1/8 P(X=D) = 1/8

1.75 bits pro Zeichen

A = 0

B = 10

C = 110

D = 111

Informationstheorie

3 gleichwahrscheinliche Werte:
P(X=A) = 1/3 P(X=B) = 1/3 P(X=C) = 1/3

naive Kodierung (2 bits):

A = 00

B = 01

C = 10

A = 1

B = 01

C = 00

(Huffman Codierung)

1.6 bits möglich?

Entropie

Allgemeiner Fall: X kann einen von m verschiedenen Werten V_1, ..., V_n annehmen mit $P(X=V_i)=p_i$

Welches ist die kleinste durchschnittliche Anzahl an bits die zur Übertragung einer Sequenz von Symbolen aus der Verteilung X benötigt wird?

$$H(X) = -p_1 \log_2 p_1 - p_2 \log_2 p_2 - ... - p_m \log_2 p_m = -\sum_{j=1}^{m} p_j \log_2 p_j$$

H(X) = Entropie

hohe Entropie ... gleichförmige „langweilige“ Verteilung von X

niedrige Entropie ... stark variierende Verteilung von X

Entropie

Beispiel: 3 gleichwahrscheinliche Werte
P(X=A) = 1/3 P(X=B) = 1/3 P(X=C) = 1/3

A = 1 1/3 * 1 +

B = 01 1/3 * 2 +

C = 00 1/3 * 2 = 1.66666

$H(X) = - 1/3 \log_2 1/3 - 1/3 \log_2 1/3 - 1/3 \log_2 1/3 =$

$= 1.5849625007211563$

Entropie

- Hohe Entropie
  - Flaches Histogramm der Werte aus X
  - Zufallswerte stammen von irgendwo
- Niedrige Entropie
  - „Zackiges“ Histogramm der Werte aus X
  - Zufallswerte sind vorhersagbarer

Bedingte Entropie

Vorhersage von Variable Y aufgrund des Inputs X

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

Bedingte Wahrscheinlichkeiten:

$P(Y = Yes) = 0.5$

$P(Y = Yes \mid X = History) = 0$

$H(X) = 1.5$

$H(Y) = 1$

Bedingte spezifische Entropie

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

H(Y | X=v) = Die Entropie von Y für jene Einträge, bei denen X den Wert v hat

H(Y|X=Math) = 1

H(Y|X=History) = 0

Durchschnittliche bedingte spezifische Entropie

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

H(Y|X) = die durchschnittliche bedingte spezifische Entropie von Y

$$H(Y|X) = \sum_j P(X = v_j) H(Y|X = v_j)$$

= Wie hoch ist die Entropie von Y nach Auswahl eines beliebenen Records und dem Wissen des Wertes von X?

= Wie viele bits werden benötigt um Y zu übertragen, gegeben dass beide Seiten den Wert von X kennen?

Durchschnittliche bedingte spezifische Entropie H(Y|X)

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

v_j	Prob(X=v_j)	H(Y|X=v_j)
Math	0,50	1,00
History	0,25	0,00
CS	0,25	0,00

$$H(Y|X) = 0.5 * 1.0 + 0.25 * 0 + 0.25 * 0 = 0.5$$

Ausblick

- Nächste Termine:

Donnerstag, 3.11.2016 13-15 (c.t.)

3. Entscheidungstheorie, Lineare Klassifikation