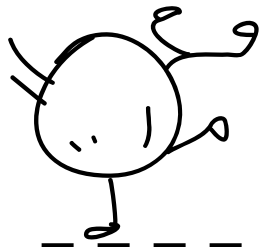


02 Grundlagen

Vorlesung 186.844

29.10.2015

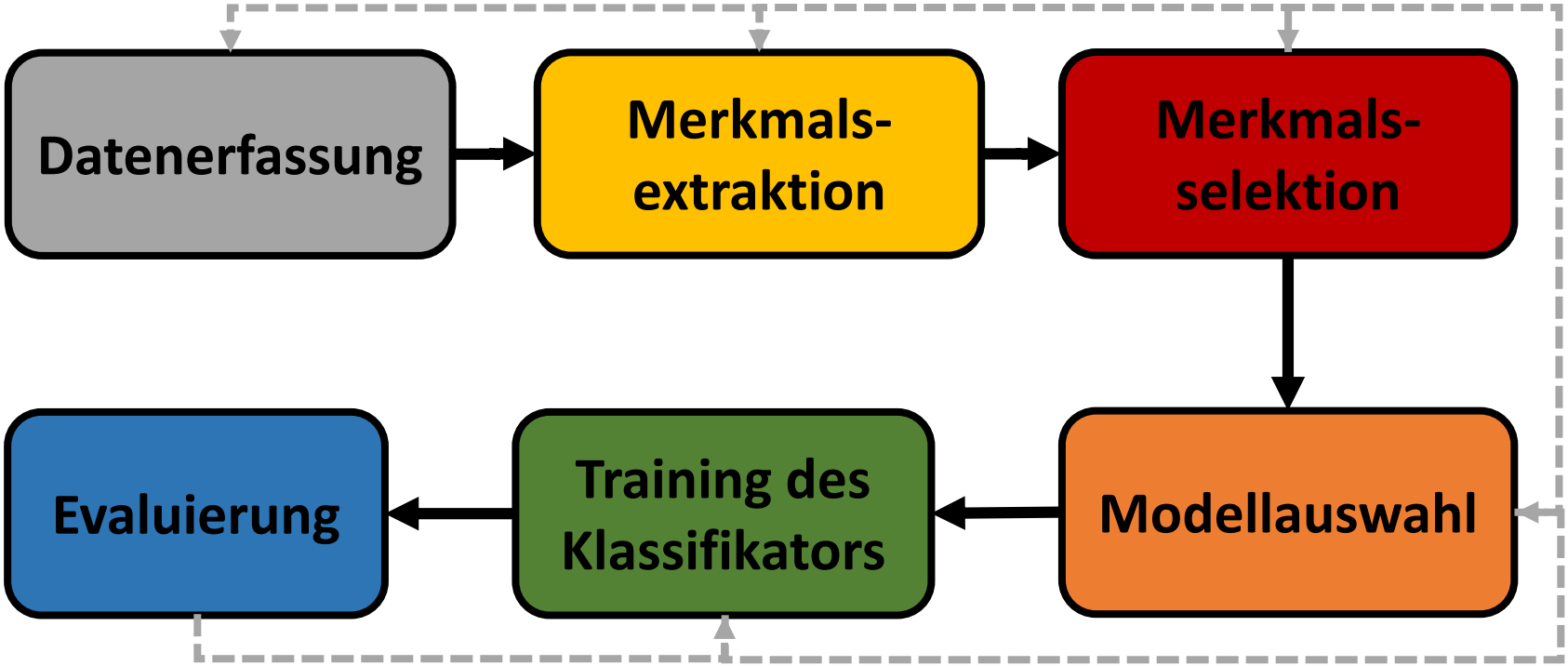
TECHNISCHE
UNIVERSITÄT
WIEN
Vienna University of Technology

Überblick

I. Regression
II. Klassifikator: Nearest Neighbor
III. Statistische Grundlagen für diskrete Zufallsvariablen
IV. Bayes Theorem für diskrete Merkmale
V. Statistische Grundlagen für stetige Zufallsvariablen
VI. Bayes-Theorem für stetige Merkmale

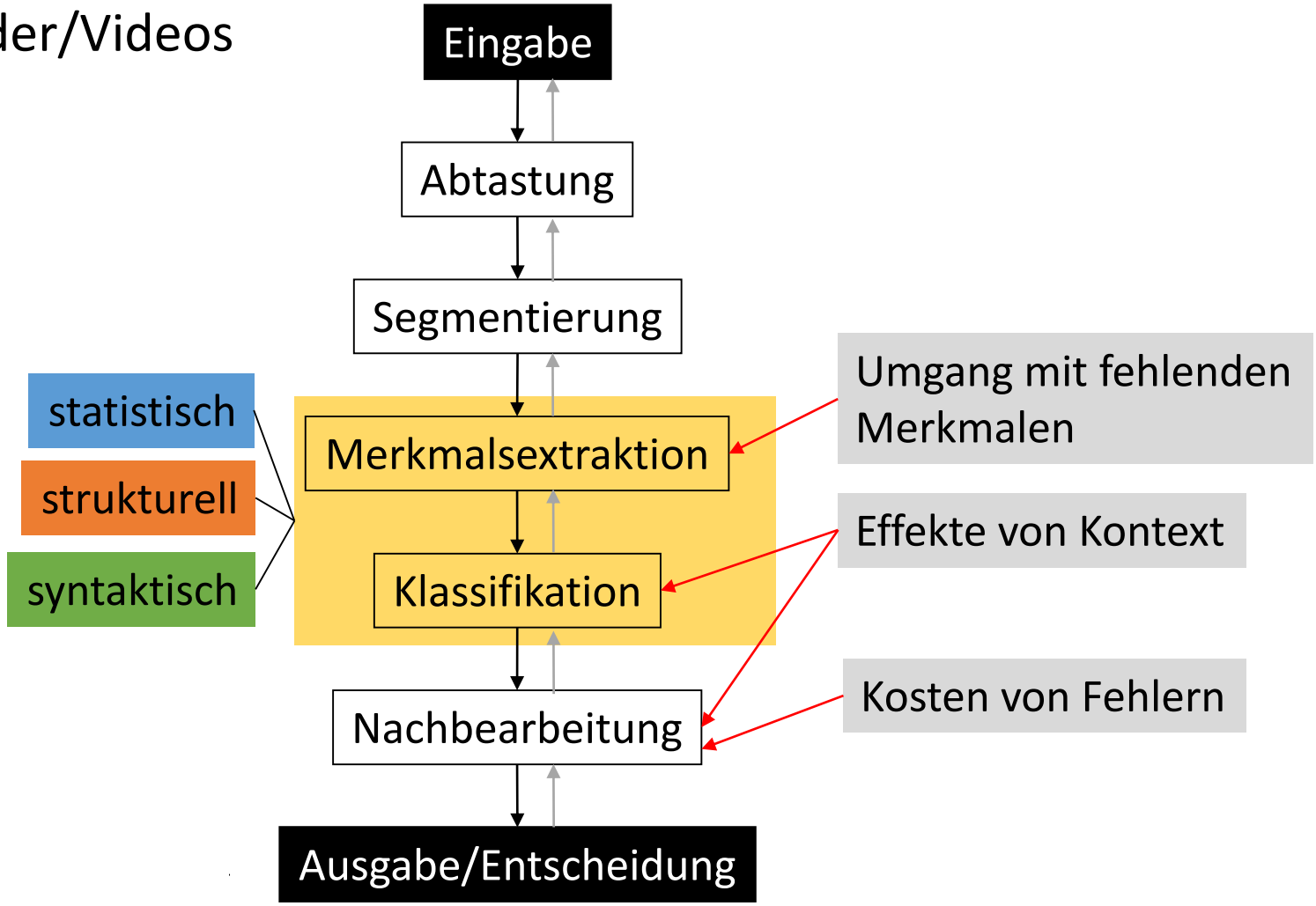
Design eines ME-Systems

… durch folgende Prozesse, die meist wiederholt werden müssen



ME-System

… für Bilder/Videos



Ausgabe

Qualitative Ausgabe [Fokus dieser VO]

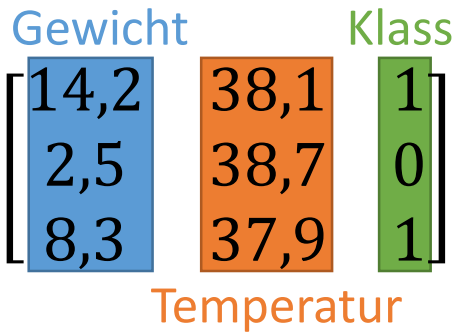
- Werte einer endlichen Menge
  - K = $\{Hund, Katze, Hamster\}$ oder $K = \{a, b, c, \dots, y, z\}$
- keine natürliche Reihenfolge, keine Rechenoperationen möglich
- Klassenzugehörigkeit wird geschätzt
  → Klassifikation

Quantitative Ausgabe

- unendlich viele verschiedene Werte
- Schätzung von quantitativen Werten $f(x) = ?$
  → Regression

Ausgabe

- Qualitative Variablen werden oft durch numerische „Codes“ repräsentiert
  - einfachster Fall: 2 Klassen („ja“ oder „nein“, „success“ oder „failure“, etc.)
    → binäre Repräsentation „0“ oder „1“
  - solche Codes werden auch als „Targets“ bezeichnet

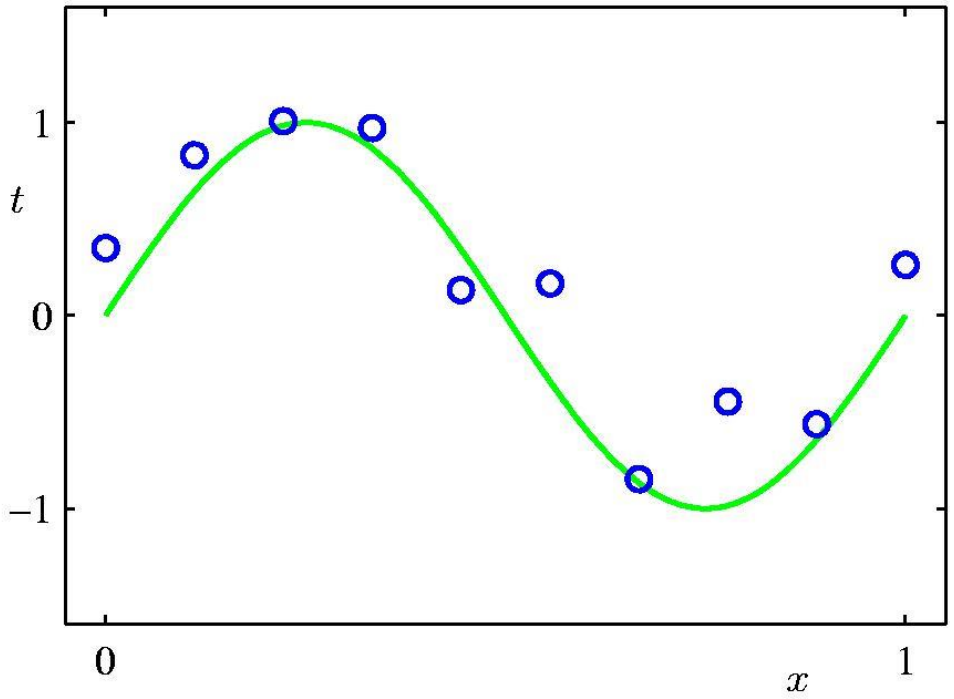


Trainingsdaten:
Merkmale +
Ground Truth (Klasse)

I. Regression

Regressionsanalyse

- Schätzung der Zusammenhänge zwischen Variablen
  → Suche nach Modell
- Trainingsdaten: $N = 10$, $X \equiv (x_1, \ldots, x_N)^T$ und $\mathrm{T} \equiv (t_1, \ldots, t_N)^T$
- Tatsächliches Modell: $\sin(2\pi x)$, Trainingsdaten beinhalten Störungen (noise)



Polynomial Curve Fitting

… Kurvenanpassung mit Polynomen (freie Übersetzung)

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

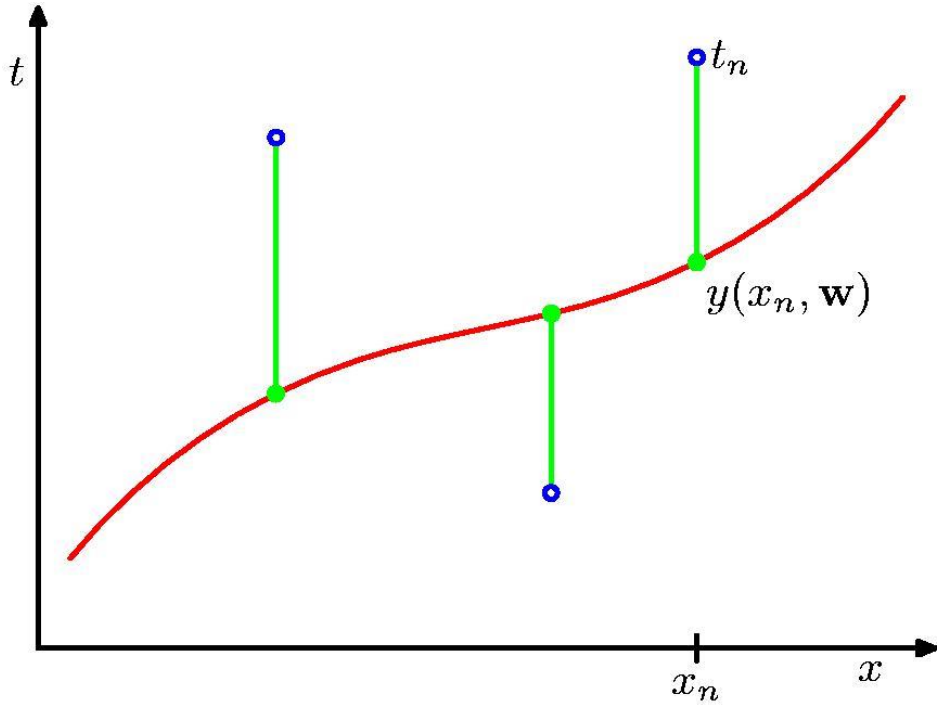
Vektor mit Koeffizienten

Grad des Polynoms

- Werte der Koeffizienten werden durch Anpassung des Polynoms an die Trainingsdaten ermittelt

Fehlerfunktion

- Fehlerfunktion misst Abweichung der Beobachtung t_n von Schätzung $y(x_n, \boldsymbol{w})$



Sum-of-Squares Fehlerfunktion:

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2$$

für verschiedene **w** (Koeffizienten)

Lösung

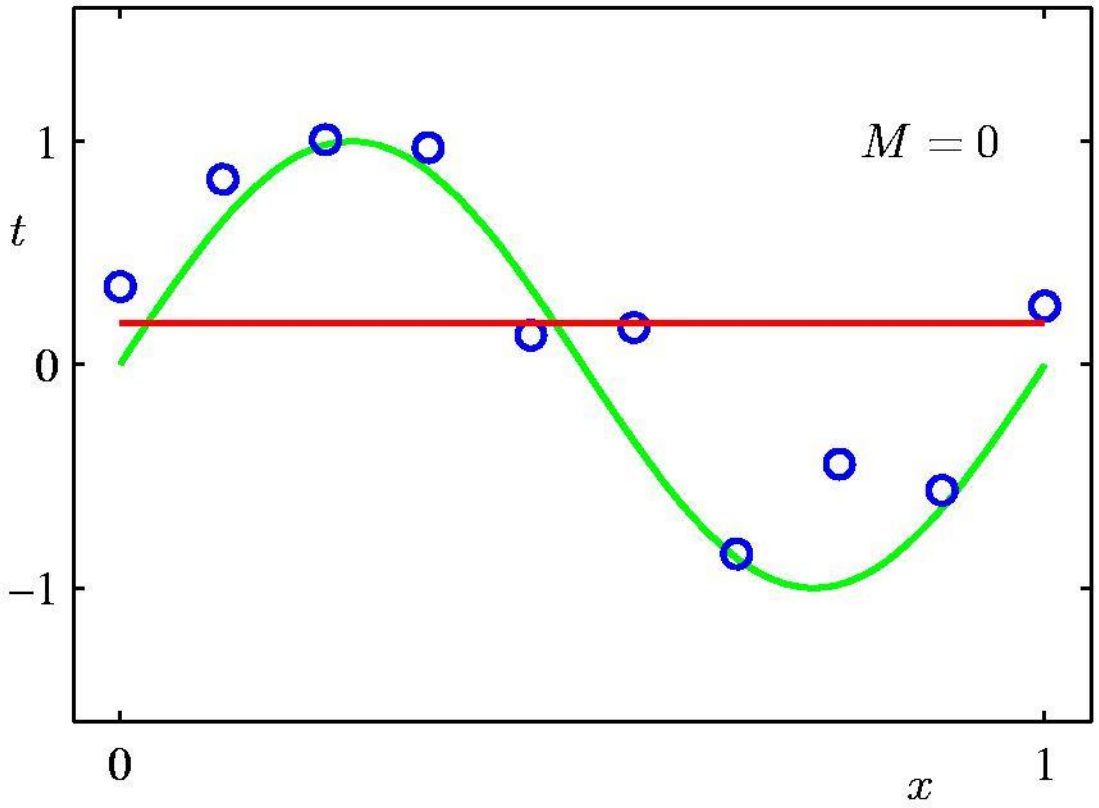
- Kurvenanpassung durch Minimierung der Fehlerfunktion
- Eindeutige Lösung mit Methode der kleinsten Quadrate (Least-Squares Estimation)
- Ergebnis: Koeffizienten $\boldsymbol{w}^*$ und Polynom $\mathrm{y}(\mathrm{x}, \boldsymbol{w}^*)$

Problem: Wahl des Grades des Polynoms …

Polynom 0. Grades

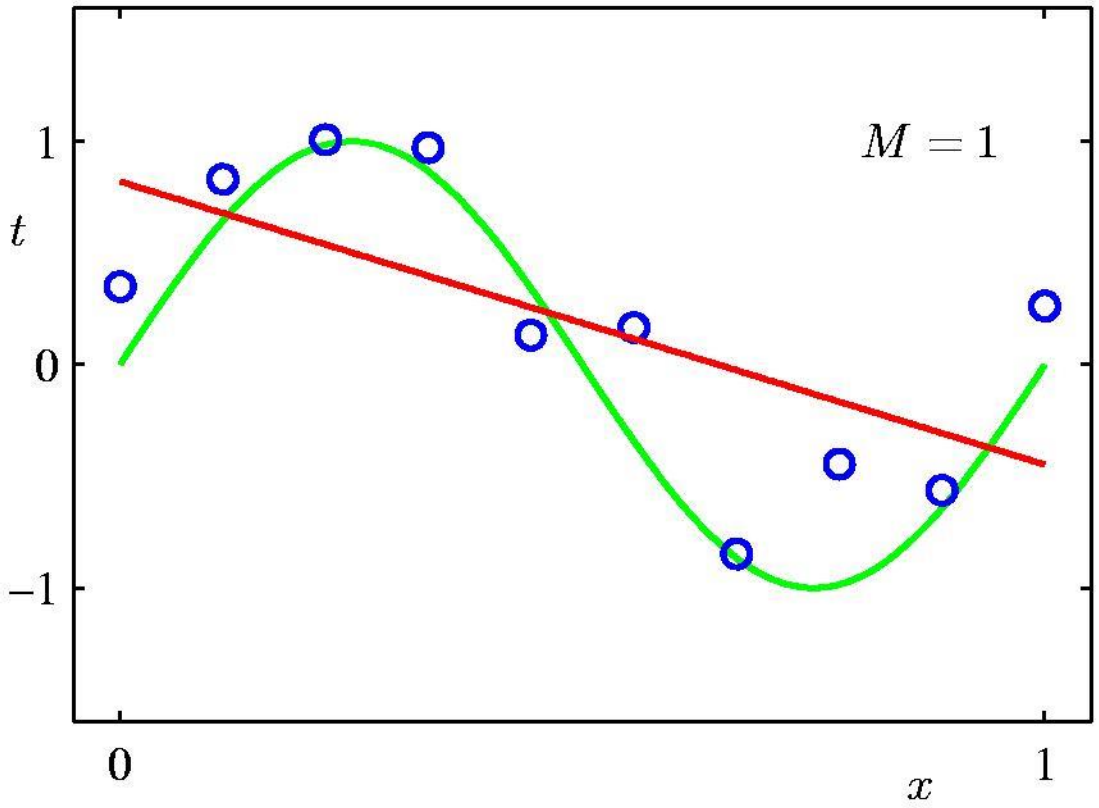
$$y(x, \boldsymbol{w}) = w_0$$

Grad des Polynoms = M
Freiheitsgrad(e) = M+1



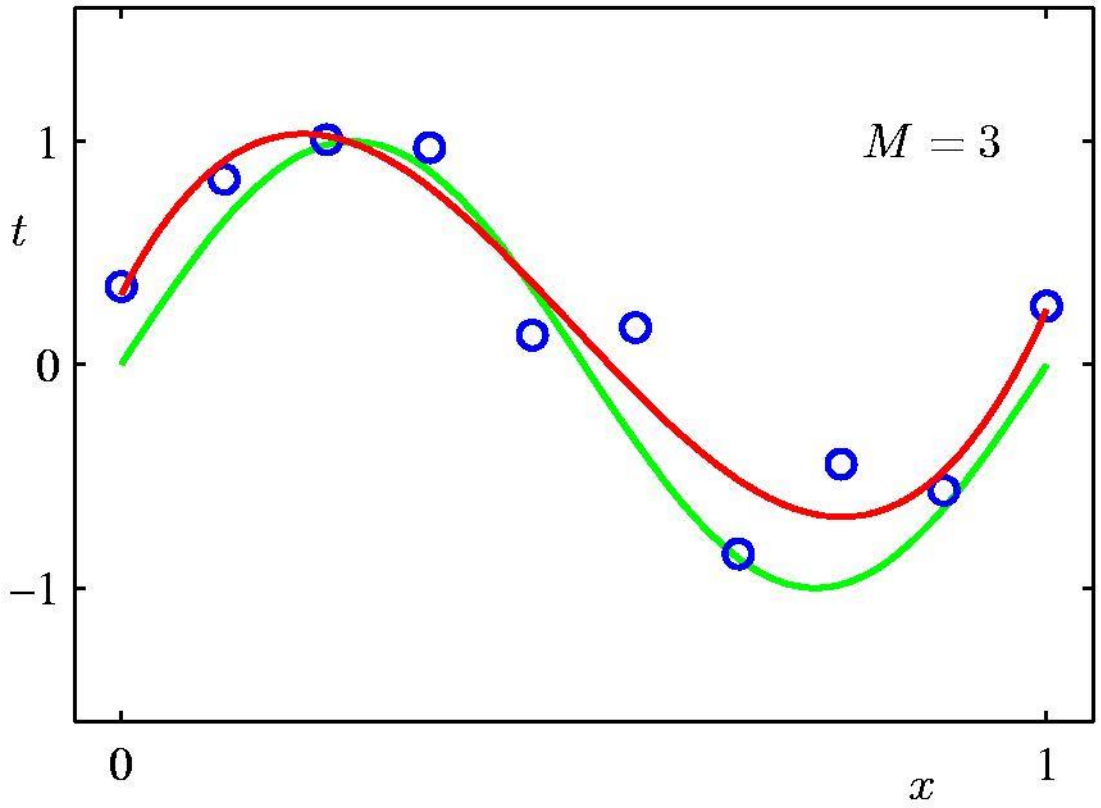
Polynom 1. Grades

$$y(x, \boldsymbol{w}) = w_0 + w_1 x$$



Polynom 3. Grades

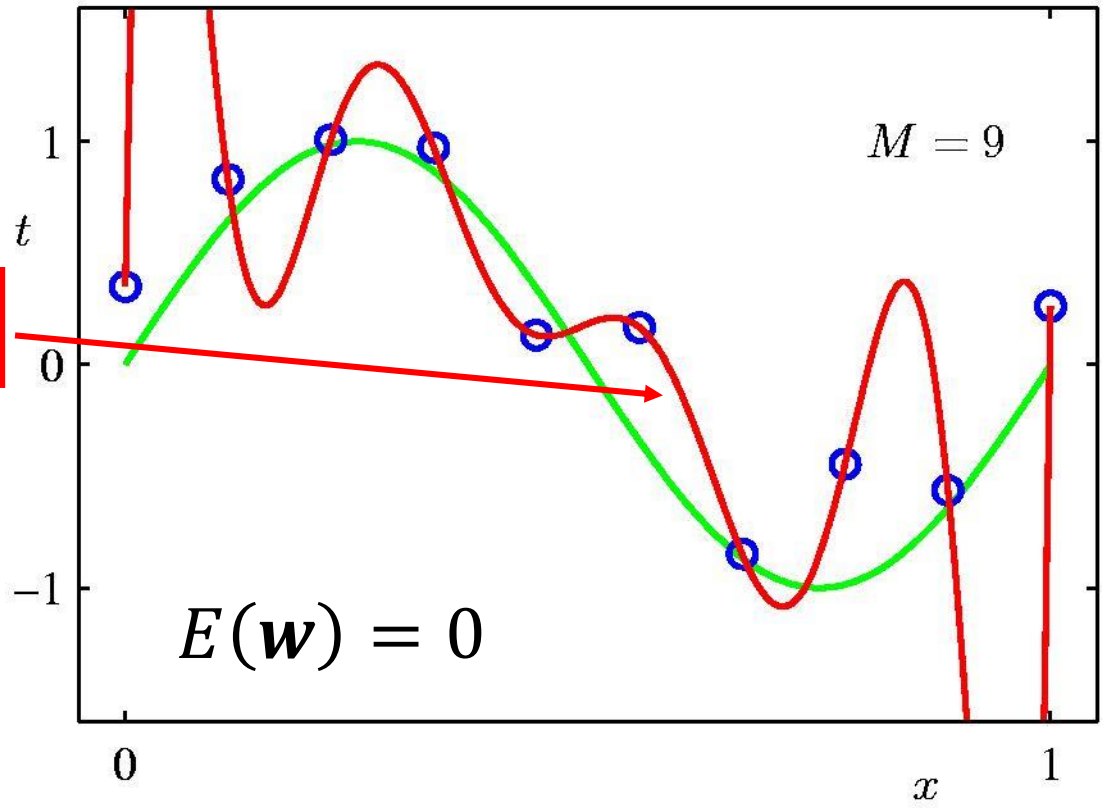
$$y(x, \boldsymbol{w}) = w_0 + w_1 x + w_2 x^2 + w_3 x^3$$



Polynom 9. Grades

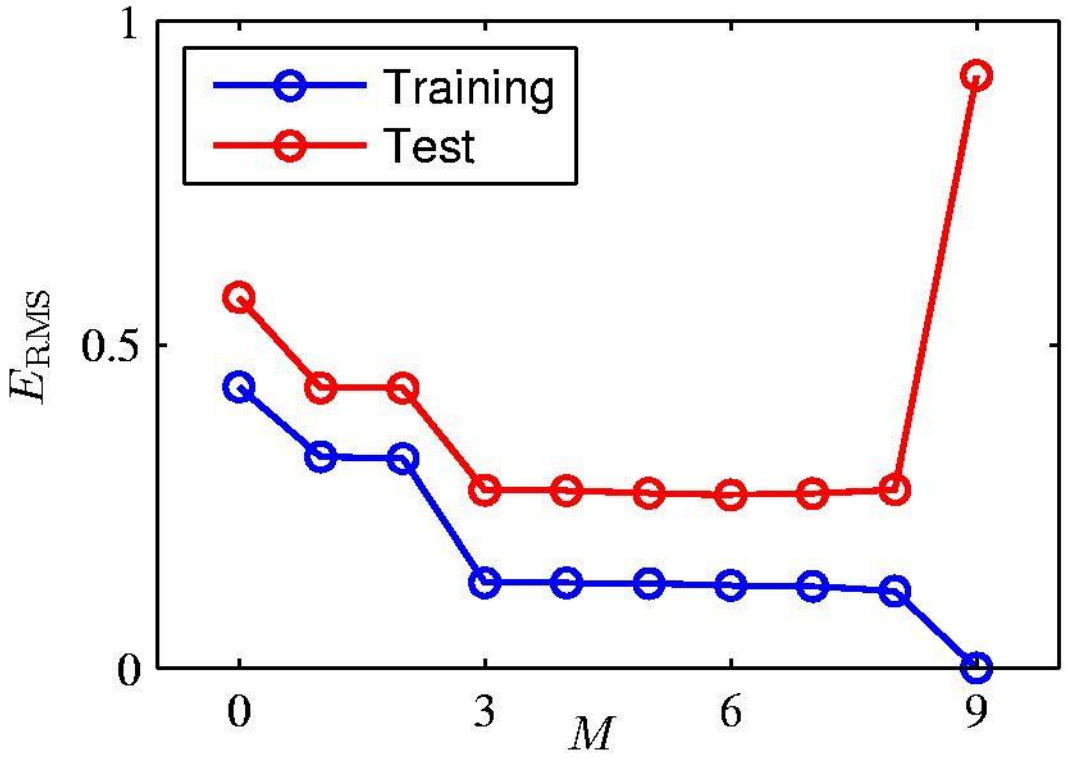
$$y(x, \boldsymbol{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_8 x^8 + w_9 x^9$$

Starke „Oszillation“



Overfitting

Trainingsdaten N = 10, Testdaten N = 100



Root-Mean-Square (RMS) Error:

$$E_{\mathrm{RMS}} = \sqrt{2E(\mathbf{w}^\star)/N}$$

Normalisierung

Koeffizienten

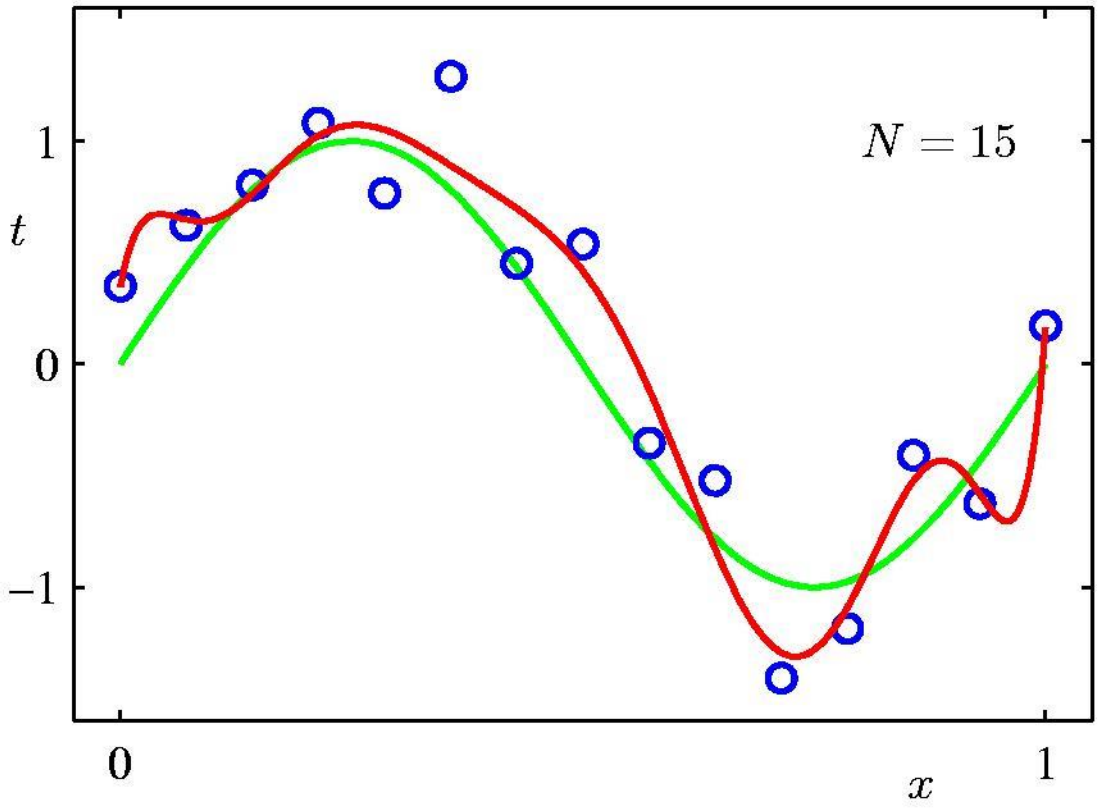
	$M=0$	$M=1$	$M=3$	$M=9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

Overfitting
→ komplexes Modell
→ Betrag der Koeffizienten wird groß

Umso mehr Freiheitsgrade ein Polynom hat, desto stärker wird es durch Störungen (noise) in den Trainingsdaten beeinflusst.

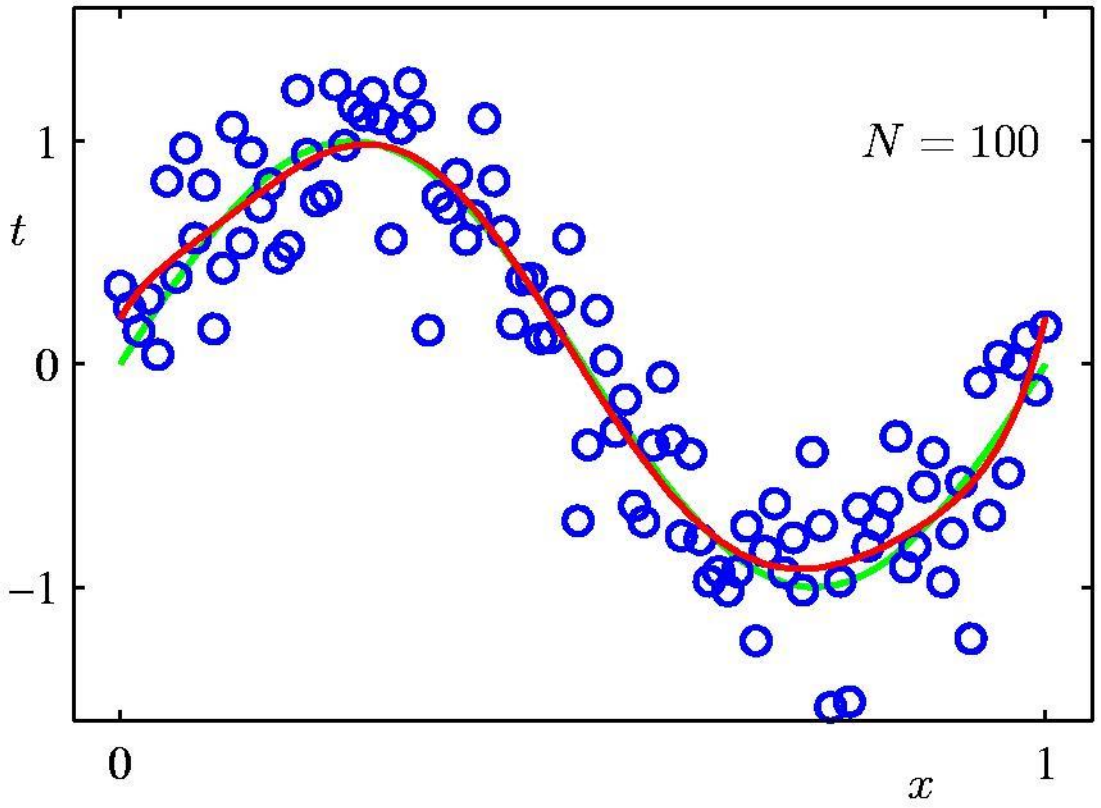
Mehr Trainingsdaten …

Trainingsdaten N = 15



Mehr Trainingsdaten …

Trainingsdaten N = 100



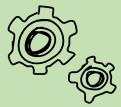
Mehr Trainingsdaten …

… verringern das „Overfitting-Problem“, bei gleich bleibendem Grad des Polynoms.

Umso höher die Komplexität des Modells, desto mehr Trainings-daten sind notwendig, um „Overfitting“ zu verhindern.

Faustregel:

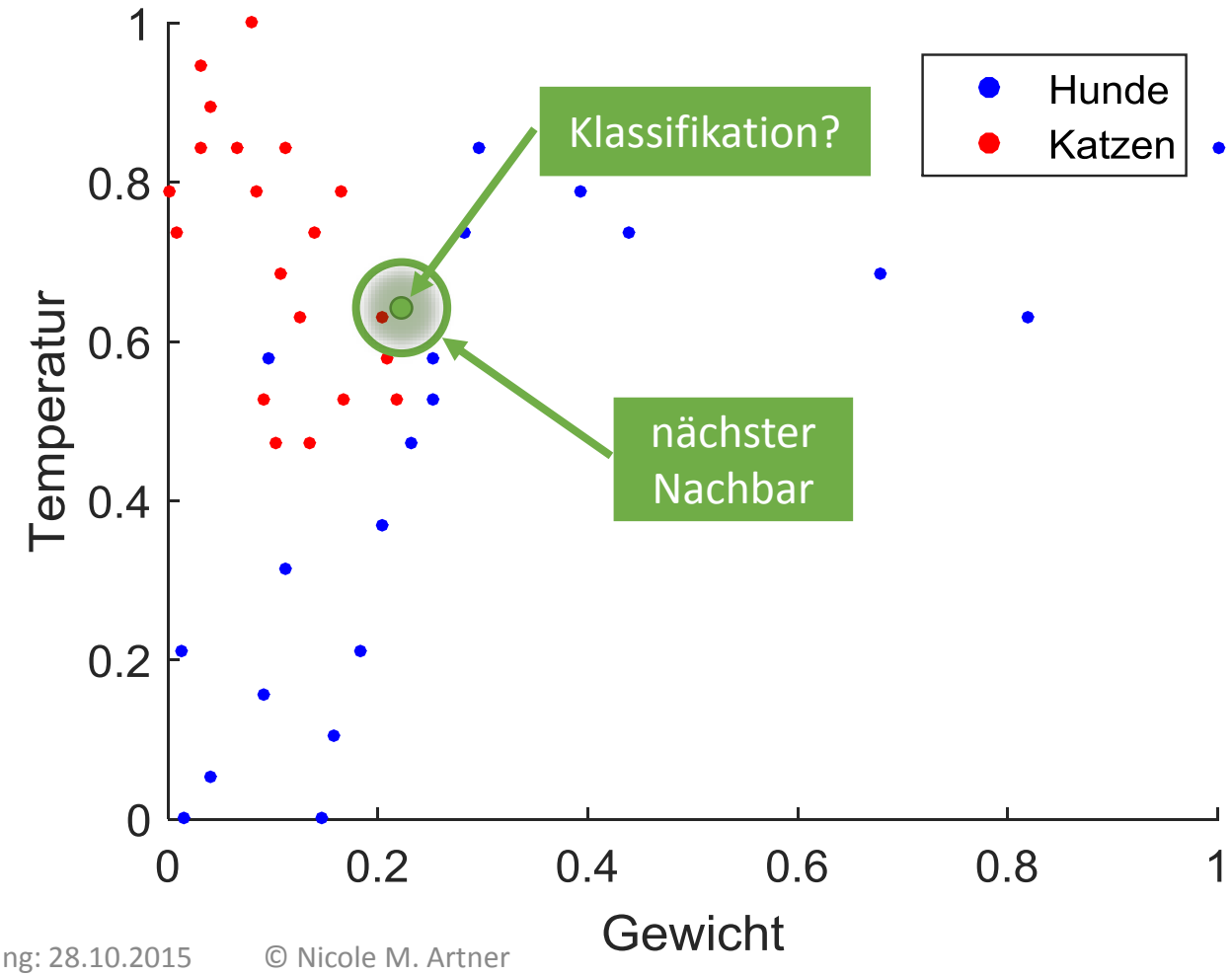
$$N = \alpha(M + 1) \qquad 5 \leq \alpha \leq 10$$



II. Klassifikator: Nearest Neighbor

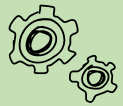
Nearest Neighbor (NN)

… nächstliegender Nachbar (freie Übersetzung)



Wie es funktioniert …

- sei $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ ein Trainingsdatensatz bestehenden aus N Vektoren (Prototypen) mit Klassenlabel
- sei $\boldsymbol{x}'$ ein Vektor aus dem Testdatensatz
- dann findet der NN-Klassifikator den Prototypen $\boldsymbol{x}_i$ der $\boldsymbol{x}'$ am nächsten ist
- $\boldsymbol{x}'$ wird das Klassenlabel von $\boldsymbol{x}_i$ zugewiesen



Wie misst man Nähe?

- mit Hilfe von Metriken oder Distanzfunktionen
- Eigenschaften von Metriken für alle Vektoren $\boldsymbol{a}$, $\boldsymbol{b}$ und $\boldsymbol{c}$ gilt:

Positive Definitheit:

$$D(\boldsymbol{a},\boldsymbol{b}) \geq 0 \text{ und } D(\boldsymbol{a},\boldsymbol{b}) = 0 \text{ nur wenn } \boldsymbol{a} = \boldsymbol{b}$$

Symmetrie:

$$D(\boldsymbol{a},\boldsymbol{b}) = D(\boldsymbol{b},\boldsymbol{a})$$

Dreiecksungleichung:

$$D(\boldsymbol{a},\boldsymbol{b}) + D(\boldsymbol{b},\boldsymbol{c}) \geq D(\boldsymbol{a},\boldsymbol{c})$$

Euklidische Distanz

Die euklidische Distanz zwischen zwei Punkten in einem n-dimensionalen euklidischen Raum $\mathbb{R}^n$ ist definiert als:

$$d(x, y) = \|x - y\|_2 = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}\,.$$

Sie wird auch als L_2Norm bezeichnet.

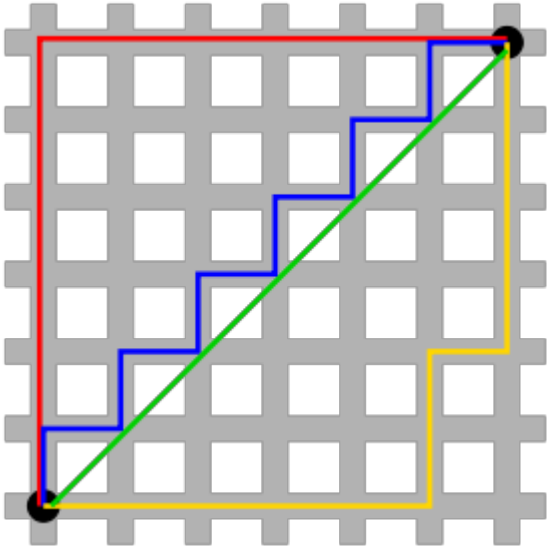
Prüfungsrelevant!

Manhatten Distanz

Die Manhatten Distanz zwischen zwei Punkten im $\mathbb{R}^2$ ist definiert als:

$$d(x, y) = \sum_{i=1}^{2} |x_i - y_i| .$$

Sie wird auch als Mannheimer-, Taxi- oder L_1-Norm bezeichnet.

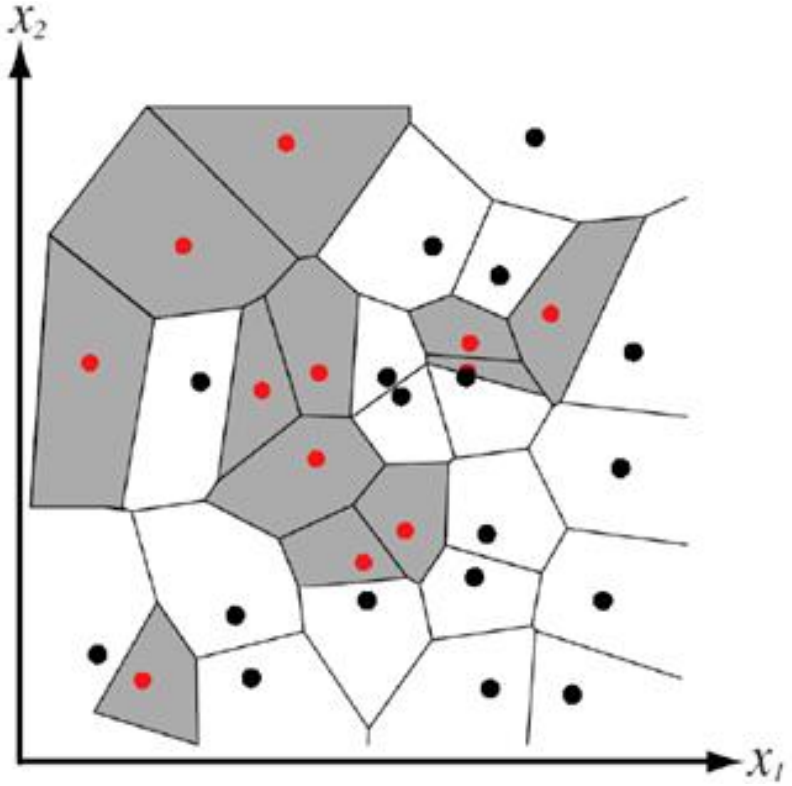


Quelle: wikipedia.org

- roter, blauer und gelber Pfad sind mögliche Pfade (4er-Nachbarschaft) für Manhatten-Distanz
  - alle haben gleiche Länge → 12 „Einheiten“
- grüner Pfad (Vektor) stellt die Euklidische Distanz dar

Voronoi-Diagramm

… Merkmalsraum wird durch Voronoi-Tesselation unterteilt

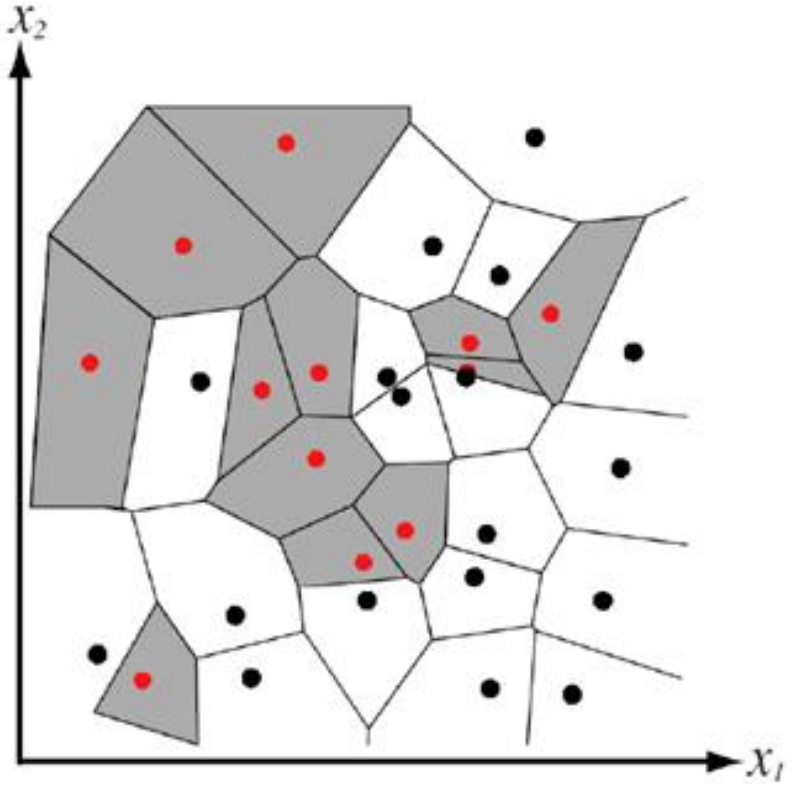


Quelle: Pattern Classification, Duda et al.

Alle Punkte in einer „Zelle“ (Polygon) sind dem zugehörigen Trainingspunkt näher als allen anderen.

Voronoi-Diagramm

… Entscheidungsregionen des NN-Klassifikators

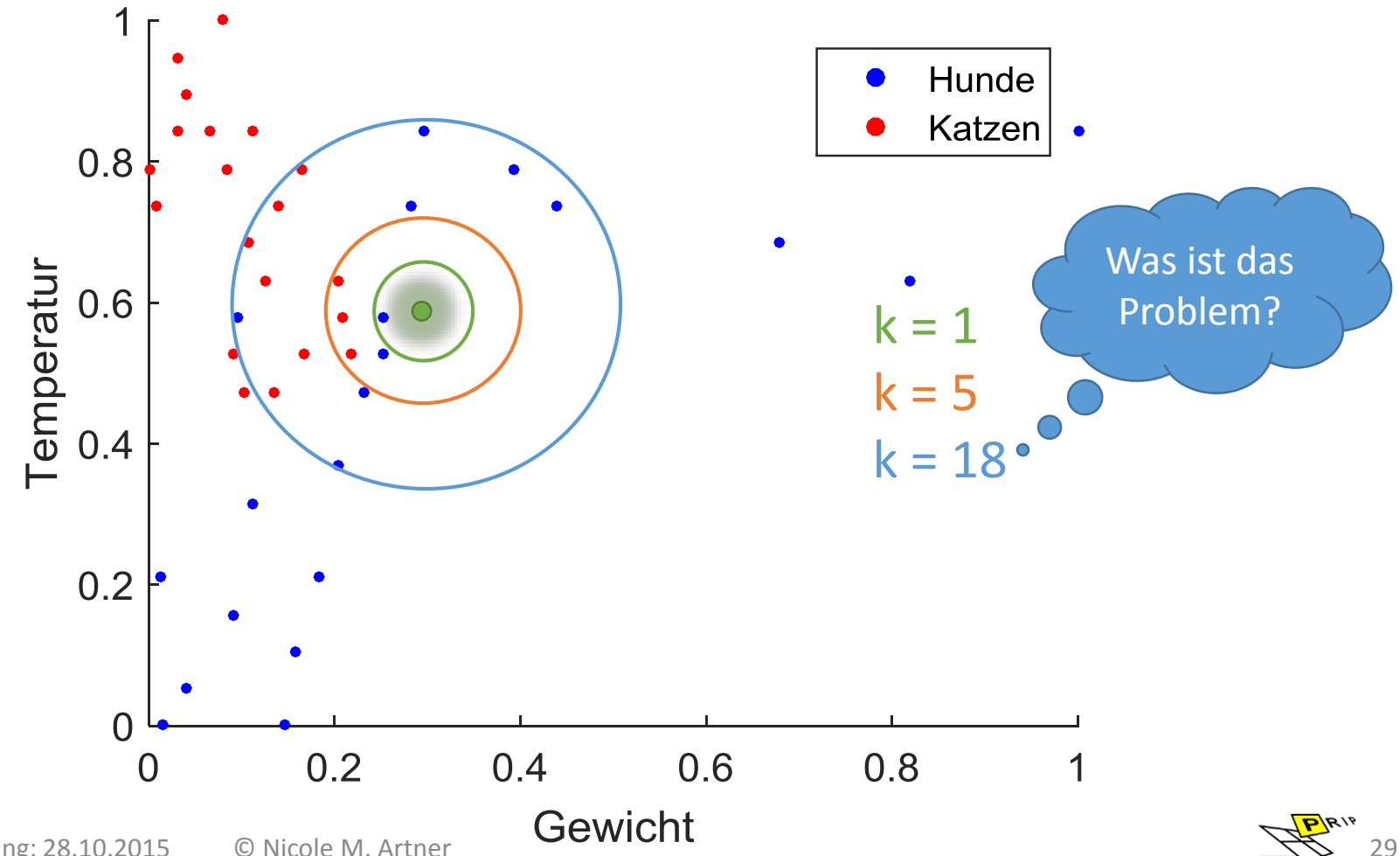


Quelle: Pattern Classification, Duda et al.

Entscheidungsregion für Klasse 1 (rot) und Klasse 2 (schwarz) ergibt sich aus der Vereinigung aller zugehörigen Voronoi-Polygone.

k-NN Klassifikator

… k nächste Nachbarn (freie Übersetzung)



Wie es funktioniert …

Erweiterung des NN-Klassifikators

- k wird vom Benutzer festgelegt
- k ist üblicherweise ungerade

Um $\boldsymbol{x}'$ zu klassifizieren:

- finde die k nächsten Prototypen aus $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$
- weise $\boldsymbol{x}'$ das Klassenlabel zu, dass am häufigsten in der gewählten Nachbarschaft vorkommt

Eigenschaften von k-NN

- es gibt kein Training im eigentlichen Sinne
- Training (Learning): speichern der Trainingsdaten als „Referenz-Menge“
- k-NN ist nicht-parametrisch (non-parametric)
  - Anzahl der Parameter wächst mit der Anzahl der Trainingsdaten
  - die a posteriori Wahrscheinlichkeit wird direkt geschätzt → mehr dazu später beim Bayes Theorem
- Speicher- und Laufzeitaufwand wachsen linear mit der Anzahl der Trainingsdaten $O(N)$

III. Statistische Grundlagen für diskrete Zufallsvariablen

Zufallsvariablen & Wahrscheinlichkeiten

Zufallsvariable x heißt diskret, wenn sie nur endlich viele verschiedene Werte aus der Menge $X = \{v_1, \dots, v_m\}$ annehmen kann.

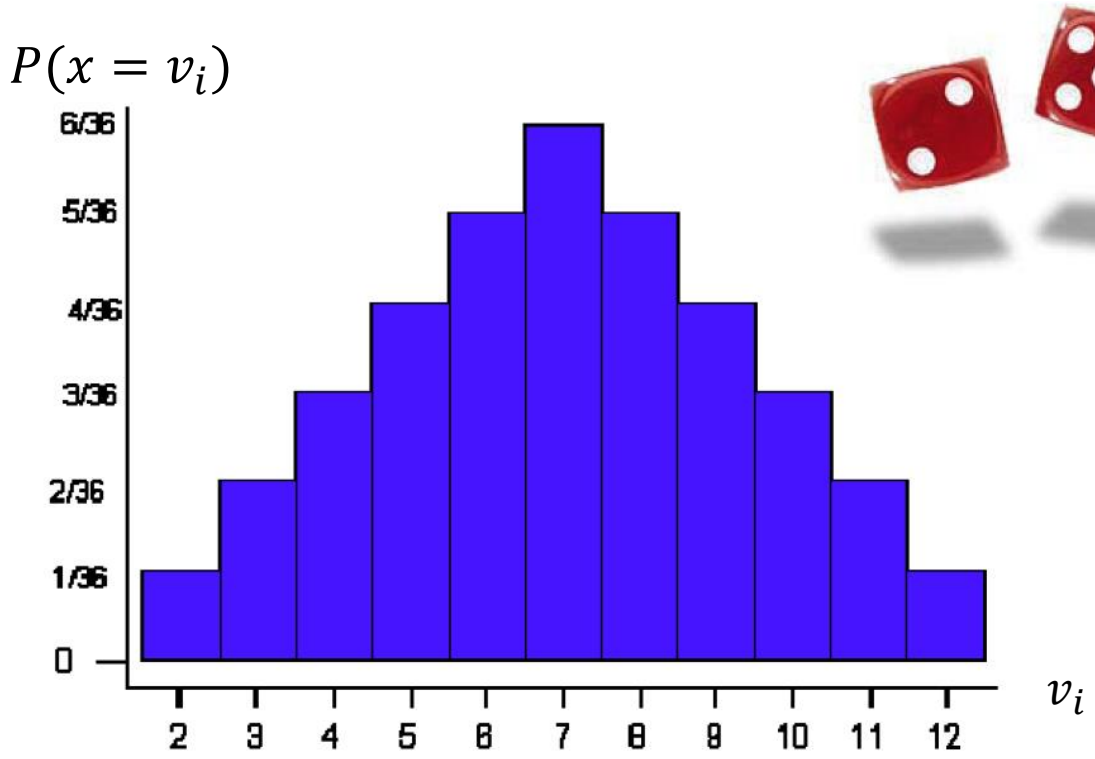
Die Wahrscheinlichkeit für das Eintreten eines Ereignisses $\mathrm{x} = v_i$ ist durch die **Wahrscheinlichkeitsfunktion** $p_i = P(x = v_i), i = 1, \dots, m$ gegeben.

Wahrscheinlichkeiten müssen folgende **Voraussetzungen** erfüllen:

$$P(x) \geq 0 \qquad \sum_{x \in X} P(x) = 1$$

Beispiel: Würfel

Verteilung einer diskreten Zufallsvariable



Kenngrößen von Verteilungen

- werden verwendet um Verteilungen zu beschreiben

- mit Stichproben (Trainingsdaten) kann man Kenngrößen schätzen

- Beispiele für Kenngrößen:
  - Mittelwert (mean)
  - Varianz (variance)
  - Standardabweichung (std, standard deviation)

Erwartung

Die Erwartung $\varepsilon[]$ einer Funktion $h(x)$ einer Zufallsvariable x ist definiert als

$$\varepsilon[h(x)] = \sum_{x \in X} h(x)P(x)$$

- $h(x) = x^i$ ist das Moment i-ter Ordnung der Verteilung

- für $i = 1$ erhält man den Mittelwert μ

Mittelwert

- wird auch oft als Erwartung oder Moment 1. Ordnung bezeichnet

Der **Mittelwert** einer Zufallsvariable x ist definiert als

$$\varepsilon[x] = \mu = \sum_{x \in X} x\, P(x) = \sum_{i=1}^{m} v_i p_i$$

Beispiel: Würfel

Berechnung des Mittelwerts:

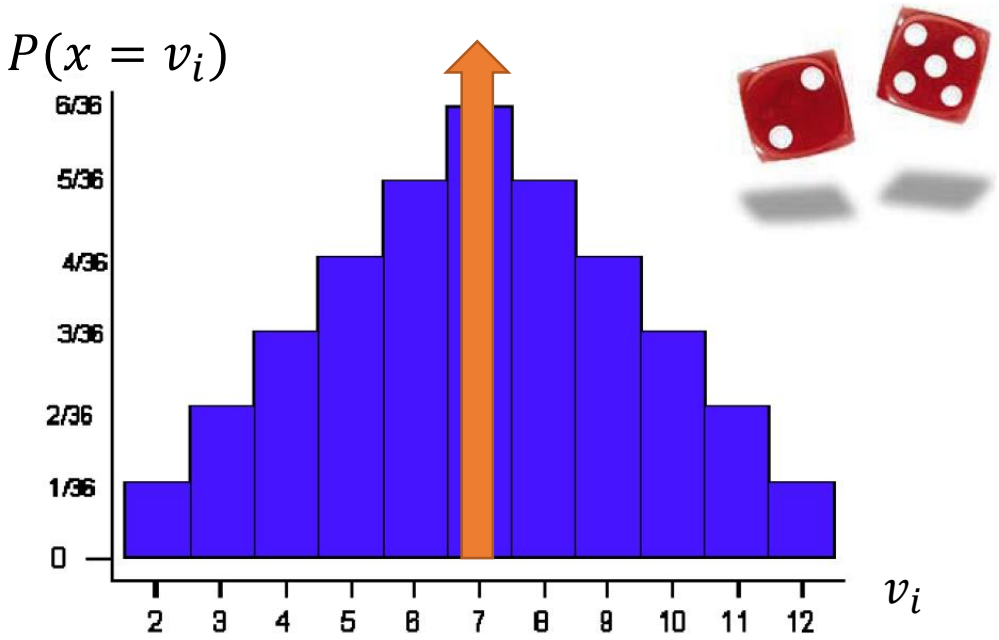
$$\varepsilon(x) = 2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} + 4 \cdot \frac{3}{36} + 5 \cdot \frac{4}{36} + 6 \cdot \frac{5}{36} + 7 \cdot \frac{6}{36}$$
$$+8 \cdot \frac{5}{36} + 9 \cdot \frac{4}{36} + 10 \cdot \frac{3}{36} + 11 \cdot \frac{2}{36} + 12 \cdot \frac{1}{36} = 7$$

Wahrscheinlichkeit des Ereignisses

Ereignis

Interpretation: Mittelwert

- die Wahrscheinlichkeitsfunktion p_i steht für das Gewicht eines Ereignisses $x = v_i$
- der Mittelwert ist somit der **Schwerpunkt**



Varianz

- zentrales Moment 2. Ordnung

Die **Varianz** σ^2 ist definiert als

$$\sigma^2 = \varepsilon[(x-\mu)^2] = \sum_{x \in X} (x-\mu)^2 P(x)$$

- es gilt auch

$$\sigma^2 = \varepsilon[x^2] - (\varepsilon[x])^2$$

Standardabweichung

- beschreibt wie stark die Werte einer Verteilung vom Mittelwert abweichen
- wird aus der Varianz berechnet

$$\sigma = \sqrt{\varepsilon[(x-\mu)^2]}$$

- Faustregel (Normalverteilung):
  - 68% der x-Werte liegen im Intervall $|x-\mu| \leq \sigma$
  - 95% der x-Werte liegen im Intervall $|x-\mu| \leq 2\sigma$
  - 99.7% der x-Werte liegen im Intervall $|x-\mu| \leq 3\sigma$

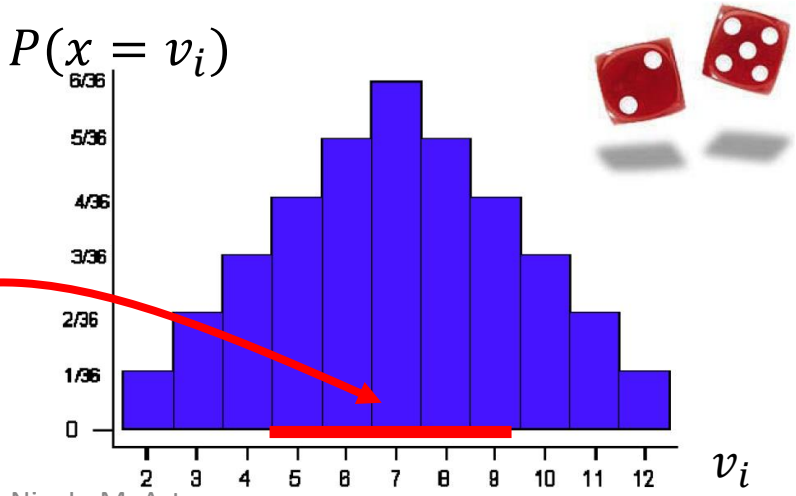
Beispiel: Würfel

Berechnung: Varianz σ^2 und Standardabweichung σ

$$\sigma^2 = (2-7)^2 \cdot \frac{1}{36} + (3-7)^2 \cdot \frac{2}{36} + (4-7)^2 \cdot \frac{3}{36} + (5-7)^2 \cdot \frac{4}{36}$$
$$+(6-7)^2 \cdot \frac{5}{36} + (7-7)^2 \cdot \frac{6}{36} + (8-7)^2 \cdot \frac{5}{36} + (9-7)^2 \cdot \frac{4}{36}$$
$$+(10-7)^2 \cdot \frac{3}{36} + (11-7)^2 \cdot \frac{2}{36} + (12-7)^2 \cdot \frac{1}{36} = 5{,}83$$

Mittelwert μ

$$\sigma = \sqrt{\sigma^2} = 2.42$$



Paare diskreter Zufallsvariablen

- seien x und y zwei diskrete Zufallsvariablen
  - x nimmt Werte aus der Menge $X = \{v_1, \dots, v_m\}$ an
  - y nimmt Werte aus der Menge Y= $\{w_1, \dots, w_m\}$ an

Die **Verbundwahrscheinlichkeit** (joint probability) $p_{ij} = P(x = v_i, y = w_j)$ gibt an, wie wahrscheinlich es ist, dass das Wertepaar (v_i, w_j) auftritt.

- es muss gelten:

$$P(x,y) \geq 0 \qquad \sum_{x \in X} \sum_{y \in Y} P(x,y) = 1$$

Beispiel: Hund oder Katze?

- Zufallsvariablen beschreiben Verteilung der Merkmale
  - Gewicht x
  - Temperatur y

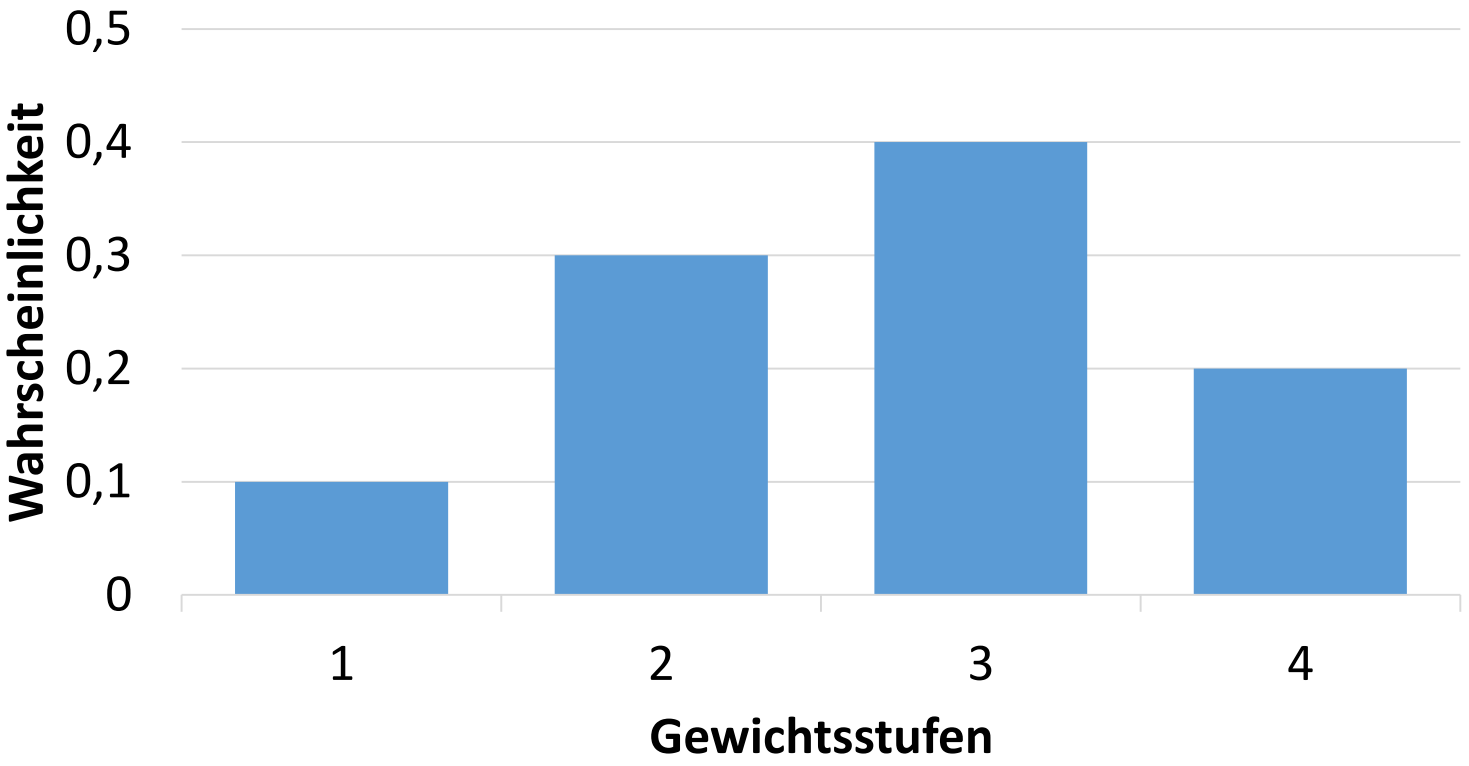
- $n_x = 4$ Gewichtsstufen $X = \{1,2,3,4\}$
- $n_y = 2$ Temperaturstufen $\mathrm{Y} = \{1,2\}$

Beispiel: Hund oder Katze?

- Wahrscheinlichkeitsfunktionen:
  - für Gewicht $p_i = P(x = i)$
  - für Temperatur $p_j = P(y = j)$
- beide Temperaturstufen sind gleich wahrscheinlich:
  - $P(y = 1) = 0{,}5$
  - $P(y = 2) = 0{,}5$

Beispiel: Hund oder Katze?

- die Gewichtsstufen verteilen sich wie folgt:



Beispiel: Hund oder Katze?

- Wahrscheinlichkeitsfunktionen p_i und p_j

	1	2	3	4
p_i	0,1	0,3	0,4	0,2
p_j	0,5	0,5		

- Verbundwahrscheinlichkeiten p_{ij}

		x			
		1	2	3	4
y	1	0,08	0,12	0,15	0,15
	2	0,02	0,18	0,25	0,05

Randverteilung

- Die **Randverteilung** (marginal distribution) $p_{i,.}$ von x, erhält man aus p_{ij} indem *m*an für jedes Ereignis von x über alle möglichen Ereignisse von y summiert:

$$p_i = p_{i,.} = \sum_{j=1}^{n_y} p_{ij}$$

Analog erhält man $p_{.,j}$

y \ x	1	2	3	4	$p_{.,j}$
1	0,08	0,12	0,15	0,15	0,5
2	0,02	0,18	0,25	0,05	0,5
$p_{i,.}$	0,1	0,3	0,4	0,2	1

	1	2	3	4
p_i	0,1	0,3	0,4	0,2
p_j	0,5	0,5		

Unabhängigkeit

Falls die Zufallsvariablen x und y unabhängig voneinander sind, dann gilt:

$$p_{ij} = p_{i,.} p_{.,j} \quad \forall 1 \leq i \leq n_x \text{ und } \forall 1 \leq j \leq n_y$$

- d.h., man könnte die Verbundwahrscheinlichkeiten als Produkt der korrespondierenden Randverteilungen berechnen
  - z.B.: $0{,}05 = 0{,}1 \cdot 0{,}5$

y \ x	1	2	3	4	$p_{.,j}$
1	0,05	0,15	0,2	0,1	0,5
2	0,05	0,15	0,2	0,1	0,5
$p_{i,.}$	0,1	0,3	0,4	0,2	1

Bedingte Wahrscheinlichkeit

Die Ereignisse $x = i$ und $y = j$ kann man kurz als A und B bezeichnen.

Die **bedingte Wahrscheinlichkeit** (conditional probability) $P(A \mid B)$ von A unter B, ist die Wahrscheinlichkeit, dass A eintritt, nachdem B bereits eingetreten ist.

Berechnung der bedingten Wahrscheinlichkeit:

$$P(A|B) = \frac{P(A,B)}{P(B)} = \frac{P(x = i, y = j)}{P(y = j)} = \frac{p_{ij}}{p_{.,j}}$$

Prüfungsrelevant!

Am Beispiel …

y \ x	1	2	3	4	$p_{.,j}$
1	0,08	0,12	0,15	0,15	0,5
2	0,02	0,18	0,25	0,05	0,5
$p_{i,.}$	0,1	0,3	0,4	0,2	1

$$P(A|B) = \frac{P(A,B)}{P(B)} = \frac{P(x=i, y=j)}{P(y=j)} = \frac{p_{ij}}{p_{.,j}}$$

$$P(x=1|y=1) = \frac{P(x=1, y=1)}{P(y=1)} = \frac{0{,}08}{0{,}5} = \mathbf{0,16}$$

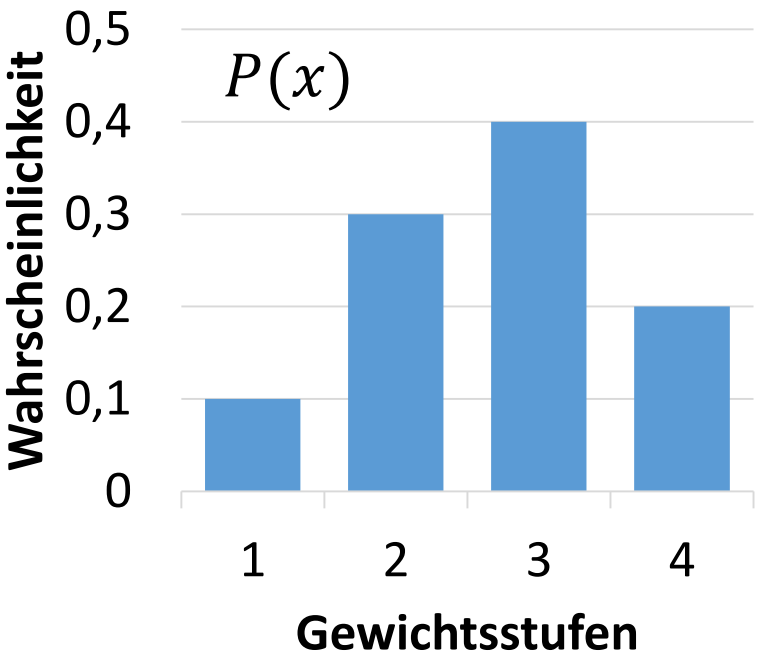
Am Beispiel …

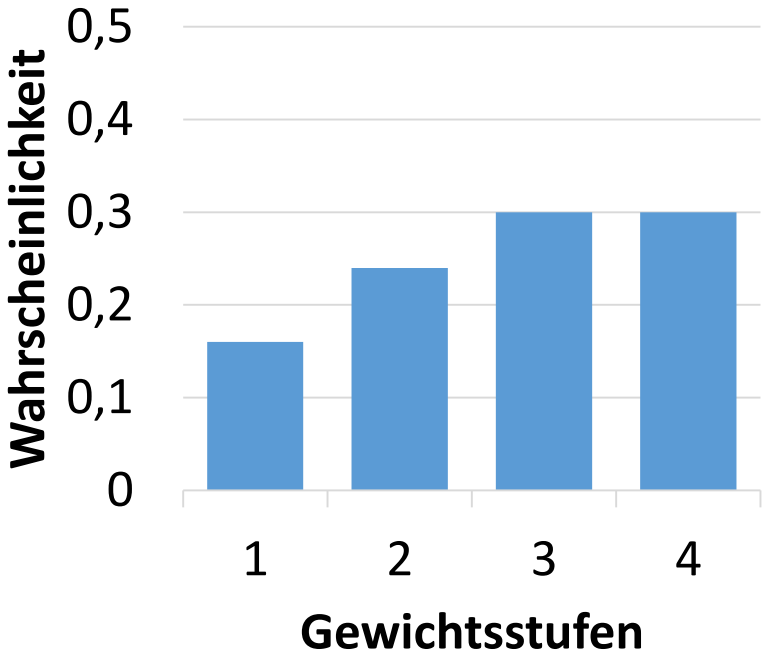
… alle restlichen Werte

y \ x	1	2	3	4	$p_{.,j}$
1	0,08	0,12	0,15	0,15	0,5
2	0,02	0,18	0,25	0,05	0,5
$p_{i,.}$	0,1	0,3	0,4	0,2	1

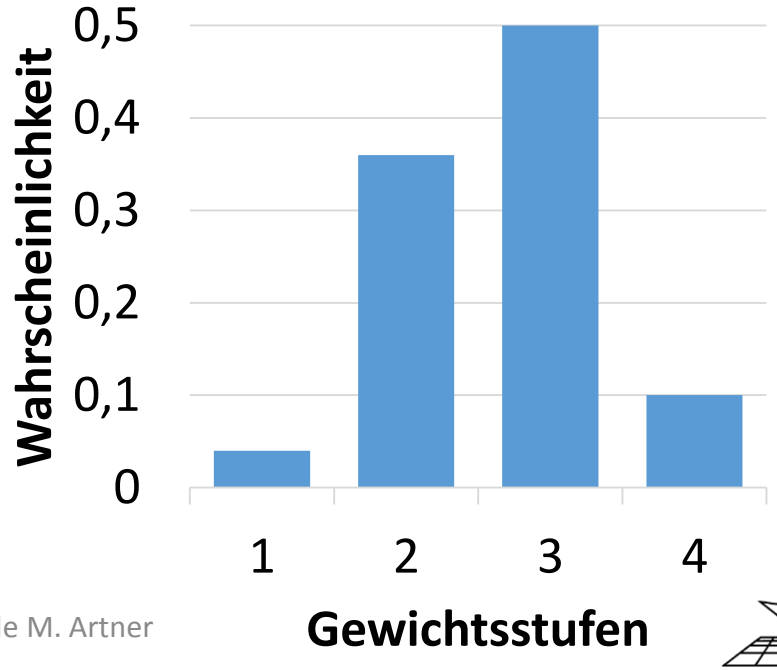
	1	2	3	4	
$P(x = i \mid y = 1)$	0,16	0,24	0,30	0,30	1
$P(x = i \mid y = 2)$	0,04	0,36	0,50	0,10	1

$P(x = i|y = 1)$





$P(x = i|y = 2)$



zur Übung …

… Berechnung der bedingten Wahrscheinlichkeiten von y, d.h. $P(B|A)$

	1	2
$P(y = j \mid x = 1)$		
$P(y = j \mid x = 2)$		
$P(y = j \mid x = 3)$		
$P(y = j \mid x = 4)$		

Berechnung Verbundwahrscheinlichkeit

Die **Verbundwahrscheinlichkeit** kann aus den bedingten Wahrscheinlichkeiten und den Randverteilungen berechnet werden:

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

- wenn die Zufallsvariablen x und y unabhängig sind, dann gilt:

$$P(A,B) = P(A|B) \cdot P(B) = \mathrm{P(A)} \cdot P(B)$$

- und somit gilt: $P(A|B) = \mathrm{P(A)}$

Begriffe

Randverteilung (marginal distribution):

$$P(A) = \sum_{B} P(A, B)$$

— Summenregel (sum rule)

Verbundwahrscheinlichkeit (joint probability):

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

— Produktregel (product rule)

IV. Bayes-Theorem für diskrete Merkmale

Herleitung …

… basierend auf der Produktregel und der Symmetrieeigenschaft $P(A,B) = P(B,A)$, kann man direkt auf folgenden Zusammenhang zwischen den bedingten Wahrscheinlichkeiten schließen:

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

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

Bayes-Theorem

Bayes-Theorem

Das Bayes-Theorem erlaubt es, die bedingte Wahrscheinlichkeit $P(B|A)$ als Funktion der Randverteilungen $P(A)$, $P(B)$ und der bedingten Wahrscheinlichkeit $P(A|B)$ auszudrücken.

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

$P(B)$ → a priori Wahrscheinlichkeit (prior) von B

$P(B|A)$ → a posteriori Wahrscheinlichkeit (posterior) von B unter A

Bayes-Theorem für ME

$$P(w = j|x = i) = \frac{P(x = i|w = j)P(w = j)}{P(x = i)}$$

x … Merkmal

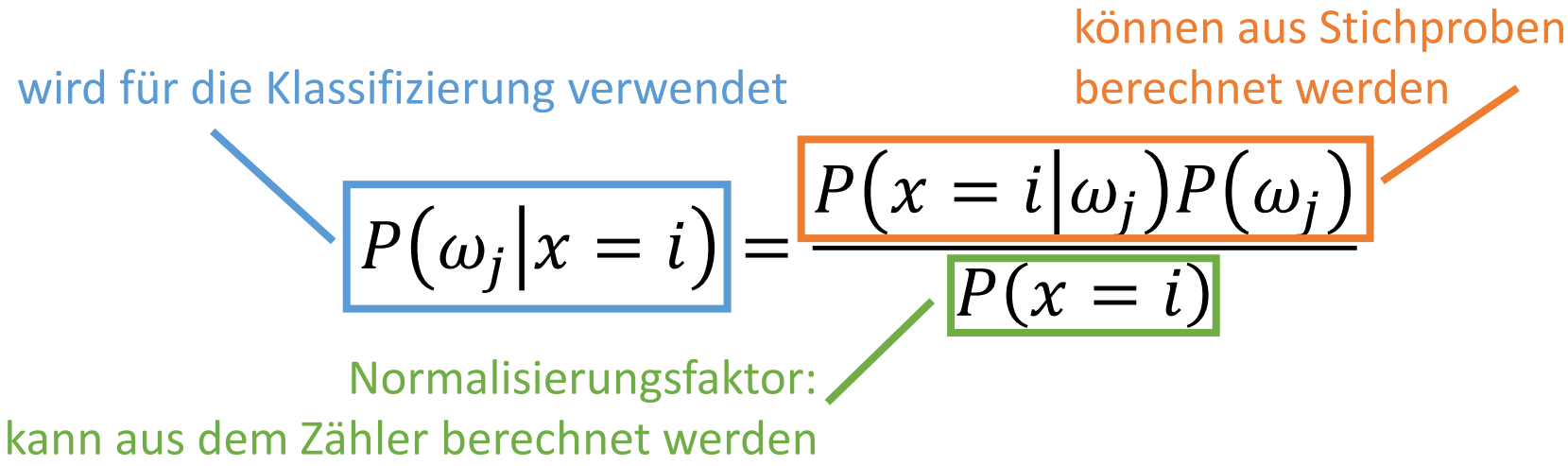
w … Klassenzugehörigkeit von Mustern

Das **Bayes-Theorem** gibt die Wahrscheinlichkeit an, dass das Muster zur Klasse j gehört, basierend auf der Merkmalsausprägung $\mathrm{x} = i$.

Bayes-Theorem für ME

In der Literatur wird oft w_j geschrieben, um anzuzeigen, dass die Zufallsvariable w den Wert j annimmt.

Achtung: Nicht mit der Aussage verwechseln, dass w_j die $j$te Komponente in einem Zufallsvektor darstellt.



Bayes-Entscheidungsregel

… wird auch als „Bayes decision rule“ bezeichnet

$$k = \arg max_j P(\omega_j | x = i)$$

Gegeben die Beobachtung (Merkmalsausprägung) $x = i$, entscheide für die Klasse k, welche die **größte a posteriori Wahrscheinlichkeit** aufweist.

Voraussetzungen

c … Anzahl der Klassen

Summe der a posteriori Wahrscheinlichkeiten über alle Klassen muss 1 sein:

$$\sum_{j=1}^{c} P(\omega_j | x = i) = 1$$

Berechnung des Normalisierungsfaktors (Nenner des Bayes-Theorems):

$$P(x = i) = \sum_{j=1}^{c} P(x = i | \omega_j) P(\omega_j)$$

Beispiel: Hund oder Katze?

Ein Tierarzt behandelt zwei Arten von Kleintieren:

- 60 % Hunde und
- 40% Katzen

Der Arzt will die Patientendaten mit fehlendem Klassenlabel automatisch klassifizieren.

Erste Schritte …

w_1 … ist die Klasse der Hunde
w_2 … ist die Klasse der Katzen

Die a priori Wahrscheinlichkeiten sind bekannt:
$P(\omega_1) = 0{,}6$ und $P(\omega_2) = 0{,}4$.

Klassifizierung anhand der a priori Wahrscheinlichkeiten:
alle Tiere sind Hunde → falsch in 40% der Fälle

Um Fehler zu verkleinern: Berücksichtigung des Gewichts der Patienten mit Hilfe des Bayes-Theorems.

Merkmal: Gewicht

- 4 Gewichtsstufen werden definiert $x = \{1,2,3,4\}$

- für jede Klasse (Hunde und Katzen) wird Verteilung des Merkmals gemessen → $P(x = i|\omega_j)$
  - mit Hilfe der Patienten von denen die Klasse bekannt ist (Besitzer haben dem Arzt ein Bild geschenkt)
  - es wird ermittelt wie viele Patienten der Klasse 1 in die Gewichtsstufe 1 fallen → $P(x = 1|\omega_1)$, usw.

Zwischenergebnisse …

- Merkmalsverteilung in beiden Klassen:

	1	2	3	4
$P(x = i\|\omega_1)$	0,50	0,30	0,15	0,05
$P(x = i\|\omega_2)$	0,30	0,25	0,25	0,20

$$P(\omega_j|x = i) = \frac{P(x = i|\omega_j)P(\omega_j)}{P(x = i)}$$

Erledigt

$$P(x = i) = \sum_{j=1}^{c} P(x = i|\omega_j)P(\omega_j)$$

Berechnung: Normalisierung

$$P(x = i) = \sum_{j=1}^{c} P(x = i|\omega_j) P(\omega_j)$$

a priori

$P(\omega_1) = 0{,}6$

$P(\omega_2) = 0{,}4$

	1	2	3	4
$P(x = i\|\omega_1)$	0,50	0,30	0,15	0,05
$P(x = i\|\omega_2)$	0,30	0,25	0,25	0,20

$$P(x = 1) = 0{,}5 \cdot 0{,}6 + 0{,}3 \cdot 0{,}4 = \mathbf{0,42}$$

	1	2	3	4
$P(x = i)$	0,42	0,28	0,19	0,11

Berechnung: a posteriori

$$P(\omega_j|x=i) = \frac{P(x=i|\omega_j)P(\omega_j)}{P(x=i)}$$

	1	2	3	4
$P(x=i\|\omega_1)$	0,50	0,30	0,15	0,05
$P(x=i\|\omega_2)$	0,30	0,25	0,25	0,20

$P(\omega_1) = 0{,}6$ $P(\omega_2) = 0{,}4$

	1	2	3	4
$P(x=i)$	0,42	0,28	0,19	0,11

$$P(\omega=1|x=1) = \frac{P(x=1|\omega=1)P(\omega=1)}{P(x=1)} = \frac{0{,}5 \cdot 0{,}6}{0{,}42} = \mathbf{0,71}$$

Ergebnisse …

… a posteriori Wahrscheinlichkeiten für jede Klassen-/Merkmalskombination:

	1	2
$P(\omega = i \mid x = 1)$	0,71	0,29
$P(\omega = i \mid x = 2)$	0,64	0,36
$P(\omega = i \mid x = 3)$	0,47	0,53
$P(\omega = i \mid x = 4)$	0,27	0,73

Klassifikation anhand der a posteriori Wahrscheinlich-keiten:

- alle Patienten der Gewichtsstufe 1 → Klasse 1
- alle Patienten der Gewichtsstufe 2 → Klasse 1
- etc.

Einfluss der Merkmalsverteilung

Was passiert bei einer anderen Merkmalsverteilung?

	1	2	3	4
$P(x = i\|\omega_1)$	0,70	0,15	0,15	0,00
$P(x = i\|\omega_2)$	0,00	0,00	0,30	0,70

	1	2
$P(\omega = i\|x = 1)$	1,00	0,00
$P(\omega = i\|x = 2)$	1,00	0,00
$P(\omega = i\|x = 3)$	0,43	0,57
$P(\omega = i\|x = 4)$	0,00	1,00

Die *a* posteriori Wahrscheinlichkeiten sind generell höher.

Fehlerwahrscheinlichkeiten

Alle Patienten der Gewichtsstufe 1 werden der Klasse 1 zugeordnet, da ihre a posteriori Wahrscheinlichkeit am höchsten ist → 0,71

Dabei werden aber 29% der Patienten falsch klassifiziert!

	1	2	**Bedingte Fehlerwahrscheinlichkeit**
$P(\omega = i \mid x = 1)$	0,71	0,29	$P(error \mid x = 1) = 0{,}29$
$P(\omega = i \mid x = 2)$	0,64	0,36	$P(error \mid x = 2) = 0{,}36$
$P(\omega = i \mid x = 3)$	0,47	0,53	$P(error \mid x = 3) = 0{,}47$
$P(\omega = i \mid x = 4)$	0,27	0,73	$P(error \mid x = 4) = 0{,}27$

Fehlerrate (error rate)

$$P(error) = \sum_{i=1}^{4} P(error|x = i)P(x = i)$$

Fehlerwahrscheinlichkeit
$P(error\|x = 1) = 0{,}29$
$P(error\|x = 2) = 0{,}36$
$P(error\|x = 3) = 0{,}47$
$P(error\|x = 4) = 0{,}27$

	1	2	3	4
$P(x = i)$	0,42	0,28	0,19	0,11

Wahrscheinlichkeit des Eintretens einer Gewichtsstufe

Ergebnisse …

	1	2
$P(\omega = i \mid x = 1)$	0,71	0,29
$P(\omega = i \mid x = 2)$	0,64	0,36
$P(\omega = i \mid x = 3)$	0,47	0,53
$P(\omega = i \mid x = 4)$	0,27	0,73

$$\left.\begin{array}{l} 0{,}29 \cdot 0{,}42 = 0{,}12 \\ + \\ 0{,}36 \cdot 0{,}28 = 0{,}10 \\ + \\ 0{,}47 \cdot 0{,}19 = 0{,}09 \\ + \\ 0{,}27 \cdot 0{,}11 = 0{,}03 \end{array}\right\} P(error) = 0{,}34$$

Ergebnisse …

	1	2
$P(\omega = i \mid x = 1)$	1,00	0,00
$P(\omega = i \mid x = 2)$	1,00	0,00
$P(\omega = i \mid x = 3)$	0,43	0,57
$P(\omega = i \mid x = 4)$	0,00	1,00

$$\left.\begin{array}{l} 0{,}00 \cdot 0{,}42 = 0{,}00 \\ + \\ 0{,}00 \cdot 0{,}09 = 0{,}00 \\ + \\ 0{,}43 \cdot 0{,}21 = 0{,}09 \\ + \\ 0{,}00 \cdot 0{,}28 = 0{,}00 \end{array}\right\} P(error) = 0{,}09$$

Anmerkungen

- für die Klassifizierung, ist Nenner nicht notwendig, weil $P(x = i)$ für alle j gleich ist.

$$k = \arg max_j P(\omega_j | x = i) = \arg max_j \frac{P(x = i|\omega_j)P(\omega_j)}{P(x = i)}$$

- Sinn der Normalisierung:

$$\sum_{j=1}^{c} P(\omega_j | x = i) = 1$$

Anmerkungen

- im Falle identischer a priori Wahrscheinlichkeiten:

$$P(\omega_i) = P(\omega_j), \forall 1 \leq i, j \leq c$$

$$k = \arg max_j P(\omega_j | x = i) = \arg max_j \frac{P(x = i|\omega_j)P(\omega_j)}{P(x = i)}$$

$$k = \arg max_j P(\omega_j | x = i) = \arg max_j P(x = i|\omega_j)$$

V. Statistische Grundlagen für stetige Zufallsvariablen

Zufallsvariablen

- im ersten Teil der VO haben wir uns mit diskreten Zufallsvariablen beschäftigt
  - Beispiel: „Hund oder Katze“
  - 4 Gewichtsstufen und 2 Temperaturstufen

- Temperatur und Gewicht sind aber eigentlich stetige Merkmale …

Prüfungsrelevant!

Stetige Zufallsvariablen

Zufallsvariable $x \in \mathbb{R}$ heißt stetig, weil sie unendlich viele verschiedene Werte annehmen kann.

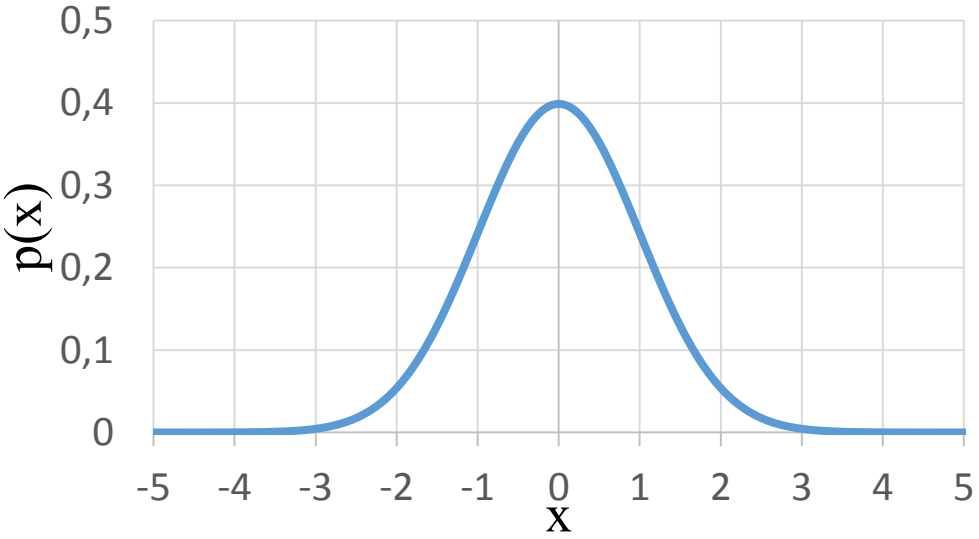
Wichtig: Die Wahrscheinlichkeit, dass x einen bestimmten Wert annimmt, dass das Ereignis $x = a$ eintritt, ist Null:

$$P(x = a) = \int_a^a p(x)dx = 0 \ (\forall a \epsilon \mathbb{R})$$

Dichtefunktion

Für **stetige Zufallsvariablen** wird eine **Dichtefunktion (DF)** $p(x)$ statt einer Wahrscheinlichkeitsfunktion $p_i = P(x = v_i)$ verwendet.

… Dichtefunktion wird auch als „probability density function“ oder kurz „pdf“ bezeichnet, z.B.: Normalverteilung/Gauß-Verteilung

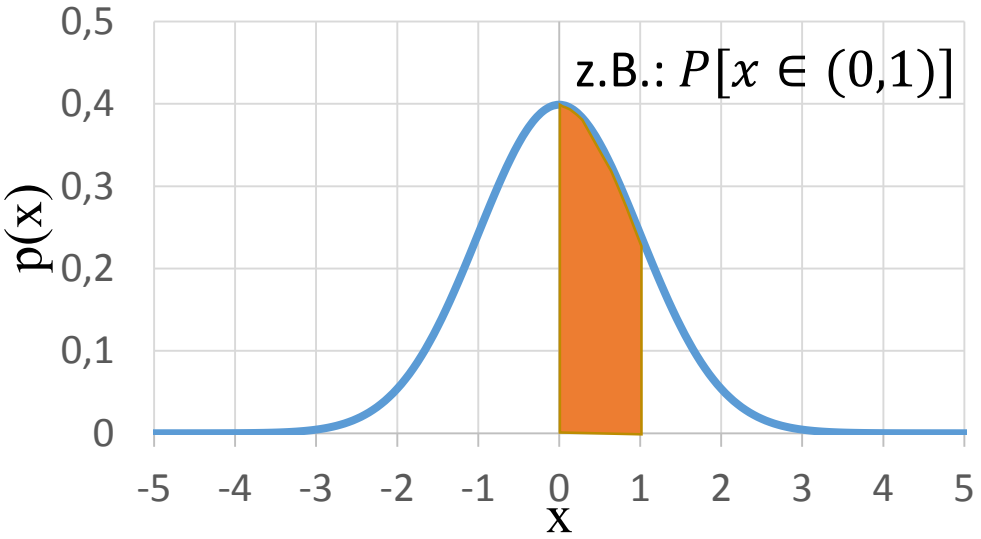


$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Wahrscheinlichkeit

Die Wahrscheinlichkeit P das x in das Intervall $[a, b]$ fällt, kann durch Integration ermittelt werden:

$$P[x \in (a,b)] = \int_a^b p(x)dx$$



Folgende Voraussetzungen muss eine DF erfüllen:

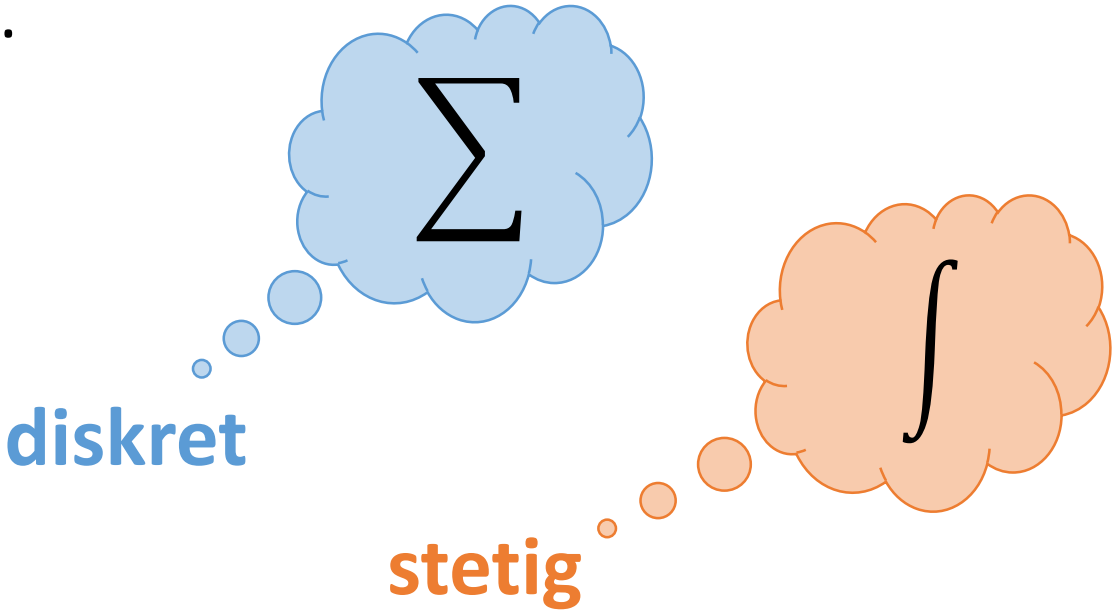
$$p(x) \geq 0$$

$$\int_{-\infty}^{\infty} p(x)dx = 1$$

Kenngrößen von Verteilungen

- Kenngrößen für stetige Verteilungen sind in den meisten Fällen sehr ähnlich zu jenen von diskreten Verteilungen

- meist reicht es aus die Summen durch Integrale zu ersetzten …



Erwartung

Die Erwartung $\varepsilon[]$ einer Funktion $h(x)$ einer stetigen Zufallsvariable x ist definiert als

$$\mathcal{E}[h(x)] = \int_{-\infty}^{\infty} h(x)p(x)dx$$

- zum Vergleich: Erwartung für diskrete Zufallsvariable x

$$\varepsilon[h(x)] = \sum_{x \in X} h(x)P(x)$$

Mittelwert

Der **Mittelwert** einer stetigen Zufallsvariable x ist definiert als

$$\mathcal{E}[x] = \mu = \int_{-\infty}^{\infty} x\, p(x)\, dx$$

- zum Vergleich: Mittelwert für diskrete Zufallsvariable x

$$\varepsilon[x] = \mu = \sum_{x \in X} x\, P(x) = \sum_{i=1}^{m} v_i p_i$$

Varianz

Die **Varianz** σ^2 ist definiert als

$$\sigma^2 = \mathcal{E}[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 p(x) dx$$

- es gilt natürlich auch: $\sigma^2 = \mathcal{E}[x^2] - (\mathcal{E}[x])^2$
- zum Vergleich: Varianz für diskrete Zufallsvariable x

$$\sigma^2 = \varepsilon[(x-\mu)^2] = \sum_{x \in X} (x-\mu)^2 P(x)$$

VI. Bayes-Theorem für stetige Merkmale

Diskret Stetig

- diskret:

$$P(\omega_j|x=i) = \frac{P(x=i|\omega_j)P(\omega_j)}{P(x=i)}$$

- für eine stetige Zufallsvariable/Merkmal x mit zugehöriger Dichtefunktion (pdf) $p(x)$:

Wahrscheinlichkeiten → P
Dichtefunktionen → p

$$P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)}$$

Dichtefunktion einer Klasse

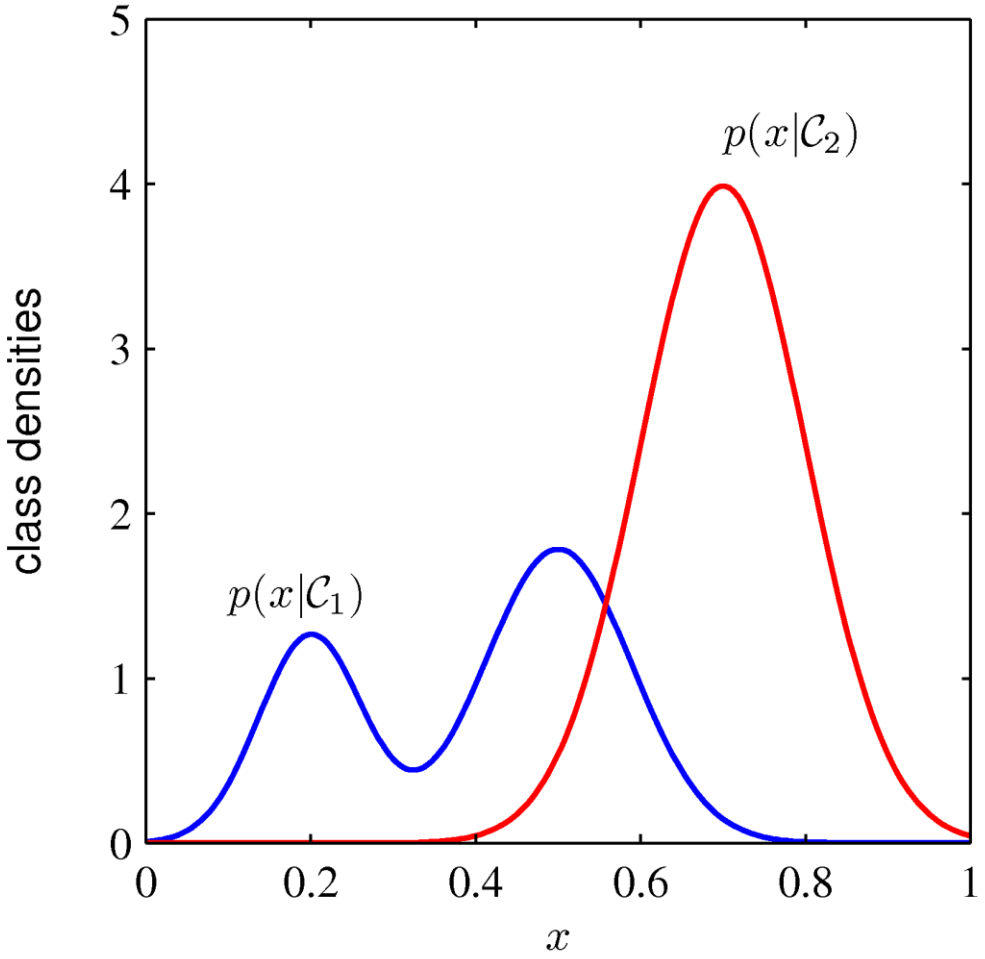
$p(x|\omega_j)$ … wird als „class conditional“ pdf bezeichnet

- beschreibt die stetige Verteilung eines Merkmals x für eine gegebene Klasse w_j
- besitzt alle Eigenschaften einer „normalen“ Dichtefunktion

$$p(x|w_j) \geq 0$$

$$\int_{-\infty}^{\infty} p(x|w_j)dx = 1$$

Beispiel: class conditional pdfs



Likelihood

Falls man $p(x|\omega_j)$ als Funktion der Klasse j für ein festes x betrachtet, spricht man von der „**Likelihood**“ von j bezüglich x.

Likelihood ≠ Wahrscheinlichkeit

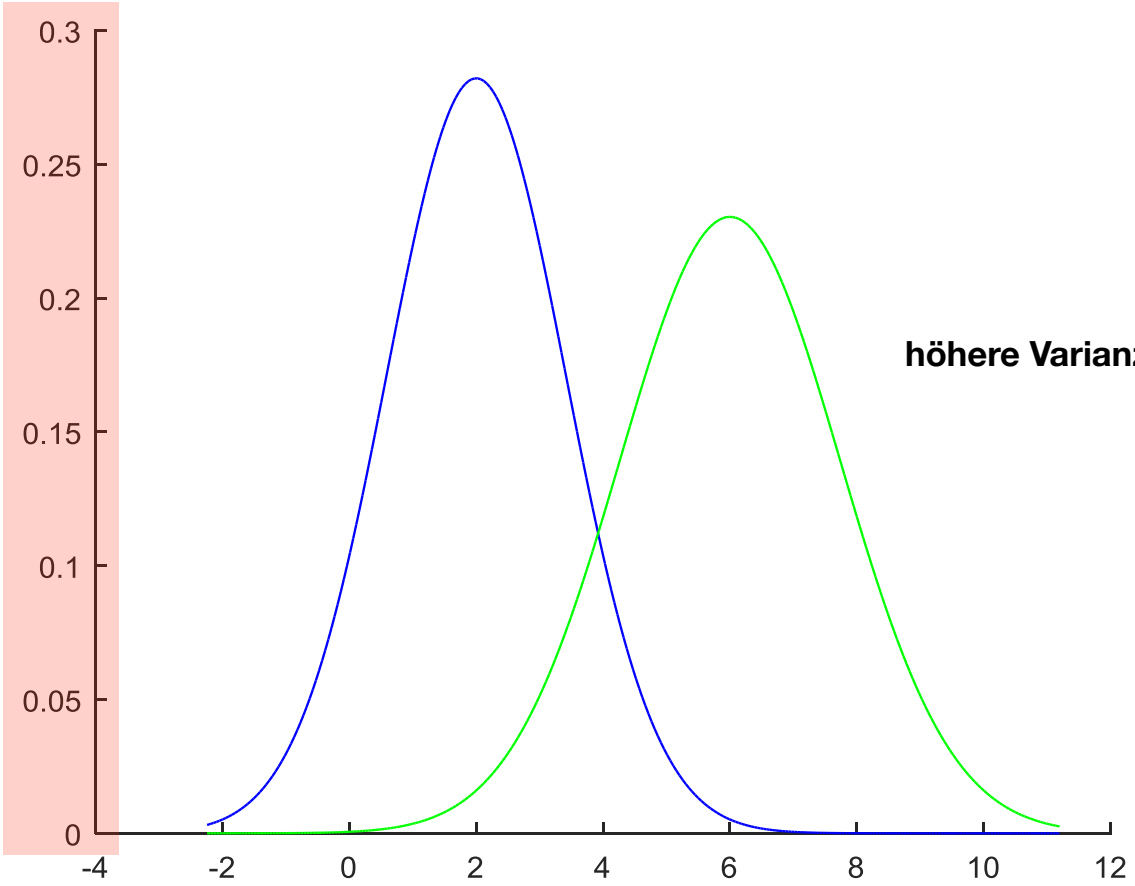
Bayes-Theorem

$$P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)}$$

$$posterior = \frac{likelihood \times prior}{evidence}$$

Beispiel: Likelihood

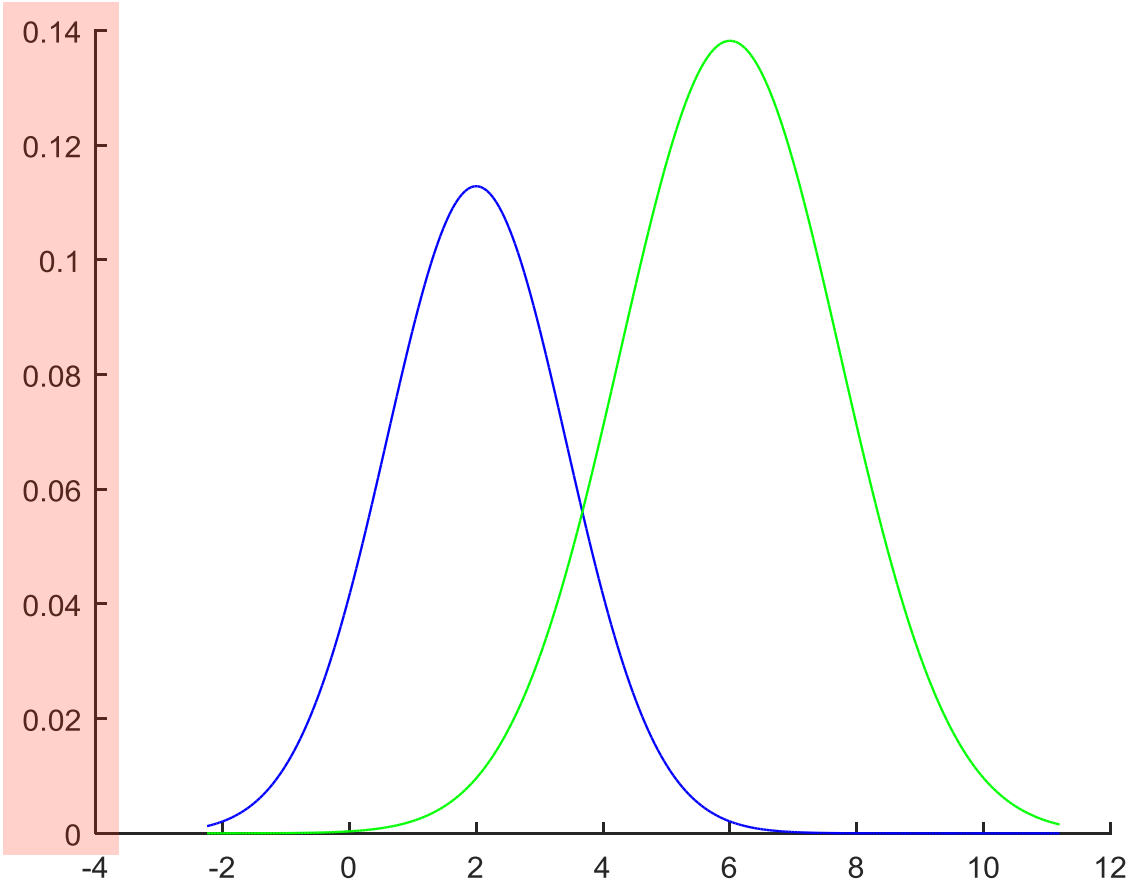
blau: $\mu = 2, \sigma = \sqrt{2}$ grün: $\mu = 6, \sigma = \sqrt{3}$



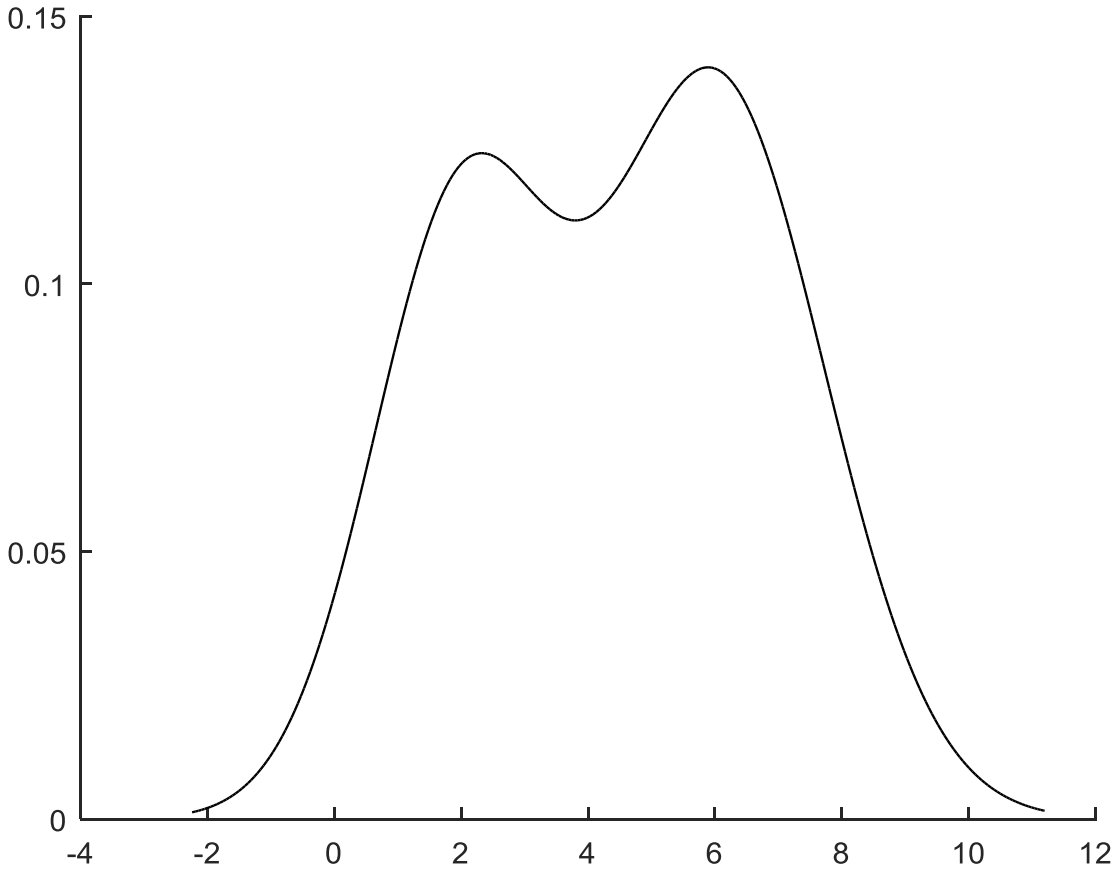
höhere Varianz → niedrigere Kurve

Beispiel: Likelihood x Prior

gewichtete Likelihoods: $P(blau) = 0{,}4 \quad P(grün) = 0{,}6$

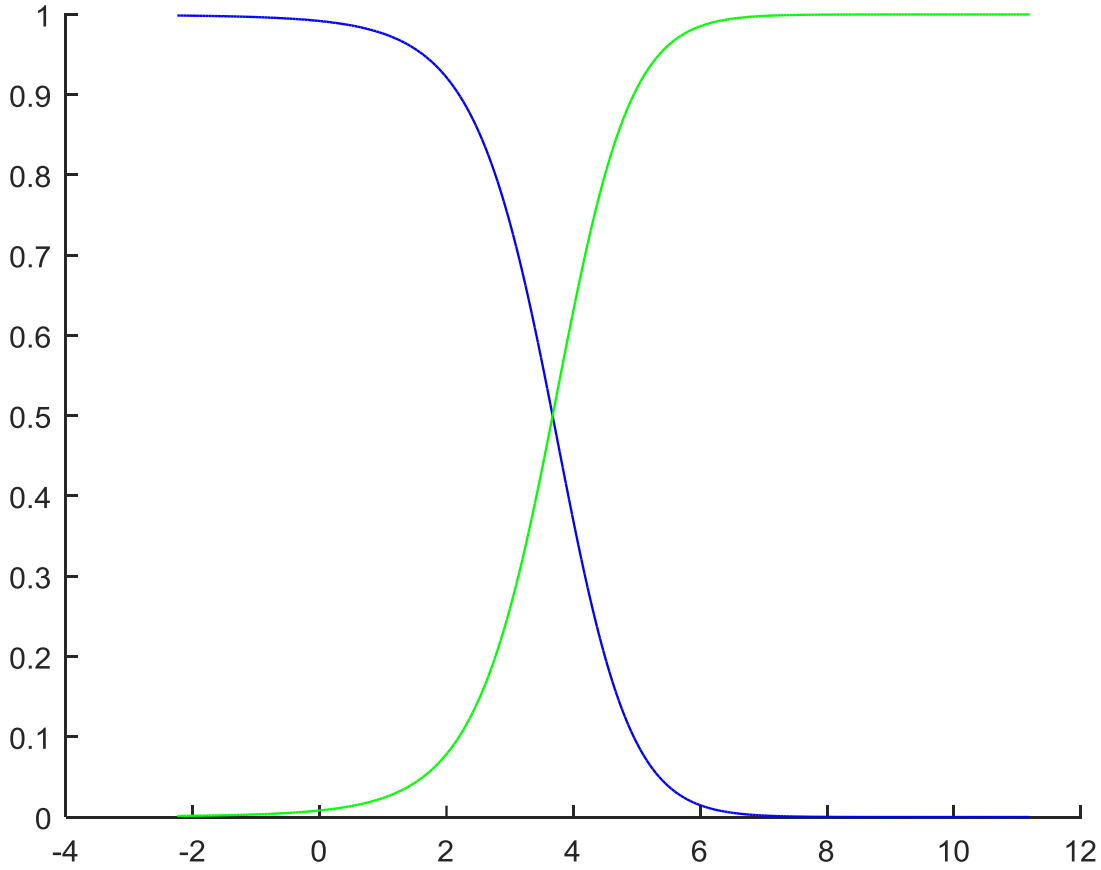


Beispiel: Evidence



Beispiel: Posterior

... Wahrscheinlichkeiten summieren sich auf 1 für jedes Ereignis

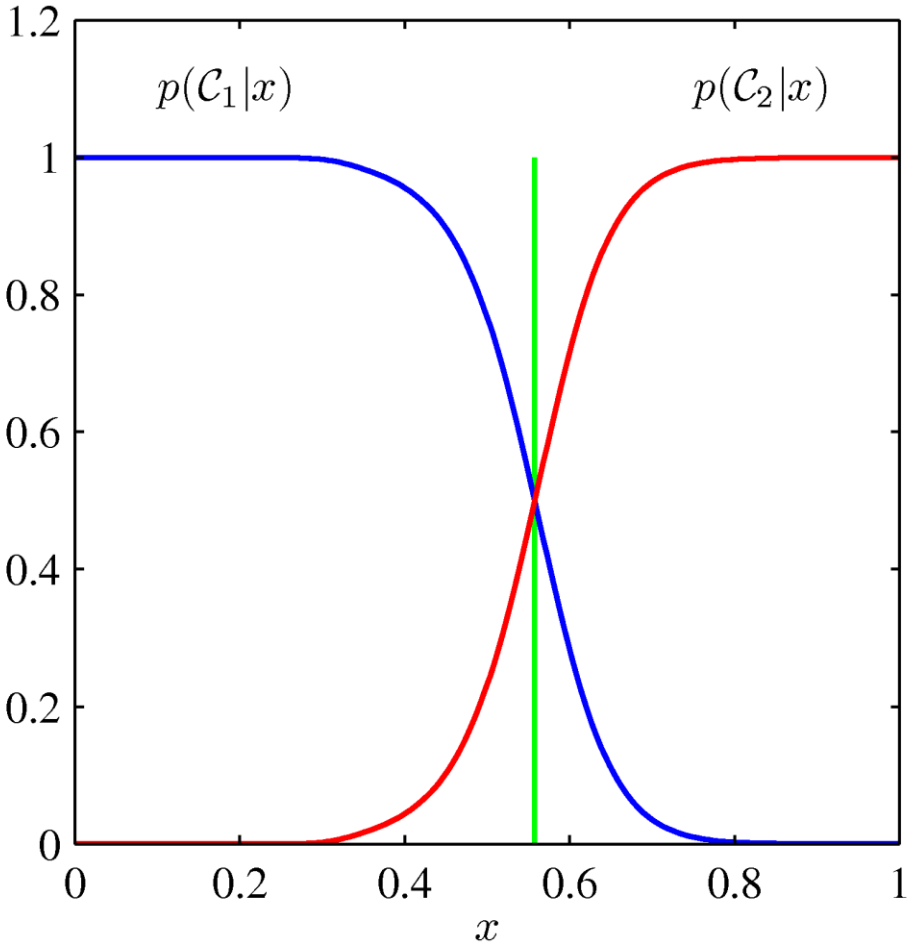


Evidence

- entspricht dem Normalisierungsfaktor (siehe Bayes-Theorem für diskrete Merkmale)
- $p(x)$ ist für alle Klassen identisch → keinen Einfluss auf das Verhältnis der posteriors
- für Bestimmung der Klasse mit größter posteriori Wahrscheinlichkeit ist $likelihood \times prior$ ausreichend
- bei identischen priors wird mit likelihoods klassifiziert

Bayes-Entscheidungsregel

Posterior: falls $P(w_1|x) > P(w_2|x) \rightarrow w_1$

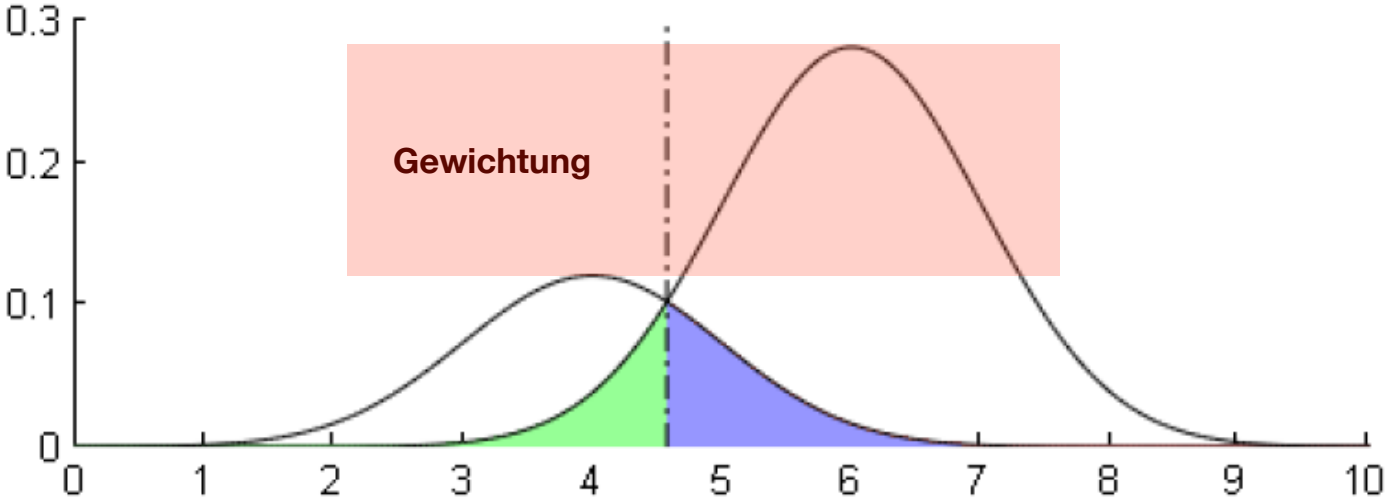


Bayes-Entscheidungsregel

gewichtete Likelihoods:

$$\text{falls } p(x|\omega_1)P(\omega_1) > p(x|\omega_2)P(\omega_2) \rightarrow w_1$$

für $P(\omega_1) = 0{,}3$ und $P(\omega_2) = 0{,}7$

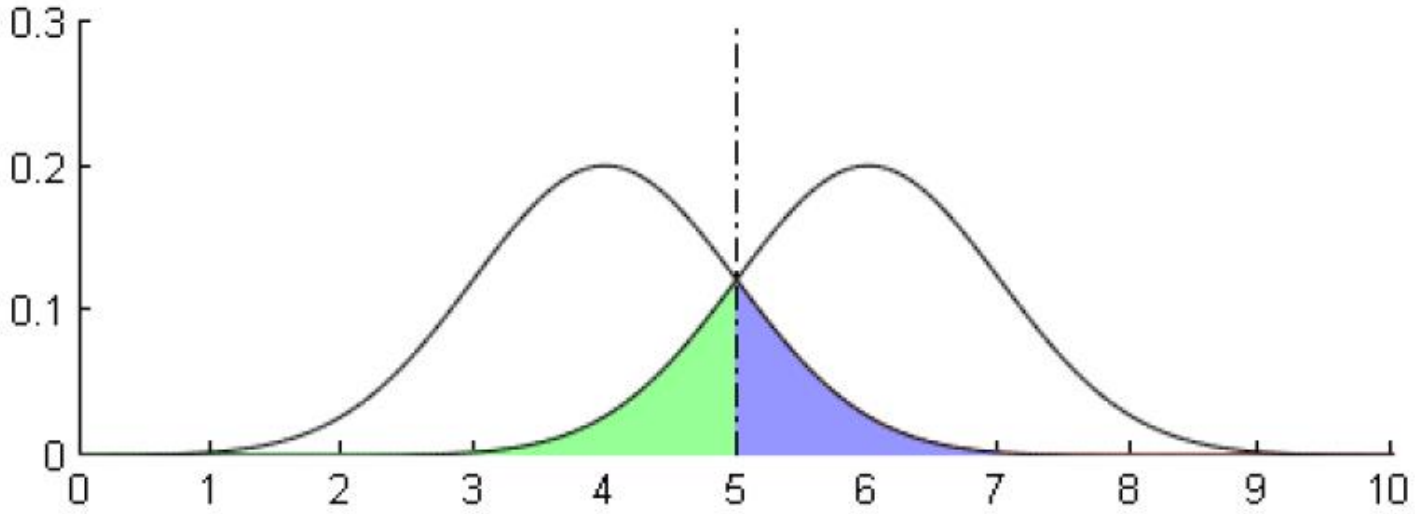


Bayes-Entscheidungsregel

Likelihoods:

$$\text{falls } p(x|\omega_1) > p(x|\omega_2) \rightarrow w_1$$

für $P(\omega_1) = P(\omega_2) = 0{,}5$



Likelihood-Verhältnis

… im Fall von $c = 2$ (zwei Klassen), kann man die Bayes-Entscheidungsregel auch so formulieren:

Entscheide für w_1, falls

$$P(w_1|x) > P(w_2|x)$$

$$p(x|\omega_1)P(\omega_1) > p(x|\omega_2)P(\omega_2)$$

Likelihood Verhältnis (Ratio) — $\frac{p(x|\omega_1)}{p(x|\omega_2)} > \frac{P(\omega_2)}{P(\omega_1)}$ — Schwellwert (Threshold)

Fehlerwahrscheinlichkeit

… die (bedingte) **Wahrscheinlichkeit der Fehlklassifikation** (conditional error) $P(error|x)$ ist

- $P(w_1|x)$, falls man für w_2 entscheidet und
- $P(w_2|x)$, falls man für w_1 entscheidet.

Die **Fehlerrate** (oder der mittlere Fehler) $P(error)$ berechnet sich als

$$P(error) = \int_{-\infty}^{\infty} P(error|x)\, p(x)\, dx$$

Nächste VO
Do, 12.11.2015