Überblick

- **Grundlagen**
  - Einführung in die automatische Mustererkennung
  - Grundlagen der Wahrscheinlichkeitsrechnung
- **Klassifikation bei bekannter Wahrscheinlichkeitsverteilung**
  - Entscheidungstheorie
  - Bayes-Klassifikator
  - Entscheidungsfunktionen bei gaußverteilten Daten
- **Überwachtes Lernen bei unbekannter Verteilung der Daten**
  - Nichtparametrische Klassifikation
  - Probleme bei hochdimensionalen Daten
  - Lineare Klassifikation, Perzeptron
  - Nicht linear trennbare Systeme
  - Nichtlineare Klassifikatoren
  - Vergleich von Klassifikatoren
- **Unüberwachtes Lernen**
  - Hauptkomponentenanalyse
  - K-Means-Clustering
  - Agglomeratives Clustern

Überblick

- **Wiederholung: Bayes-Klassifikator**
- **Diskriminantenfunktionen**
- **Kontinuierliche Zufallsvariablen**
- **Gaußverteilung**

Bayessche Entscheidungsregel

Nach einer Messung x ist die Fehlerwahrscheinlichkeit

$$p(Fehler \mid x) = \begin{cases} p(\omega_2 \mid x) & bei\ Entscheidung\ für\ \omega_1 \\ p(\omega_1 \mid x) & bei\ Entscheidung\ für\ \omega_2 \end{cases}$$

Die Fehlerwahrscheinlichkeit für gegebenes x wird minimiert durch die Bayessche Entscheidungsregel:

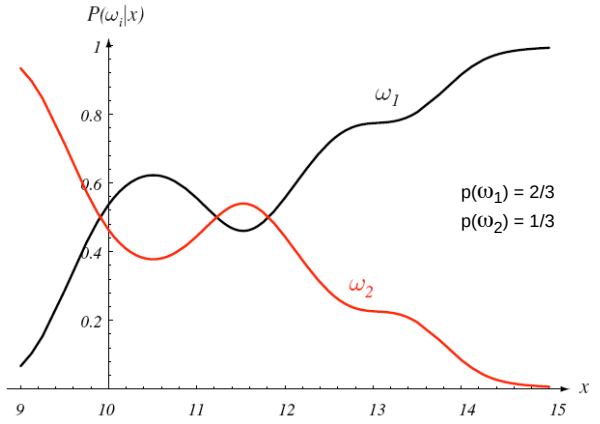
Entscheide für ω_1 wenn $p(\omega_1|x) > p(\omega_2|x)$, sonst für ω_2 bzw.

Entscheide für ω_1 wenn $p(x|\omega_1)\, p(\omega_1) > p(x|\omega_2)\, p(\omega_2)$, sonst für ω_2

Die Bayessche Entscheidungsregel minimiert auch die durchschnittliche Fehlerwahrscheinlichkeit über alle x:

$$p(Fehler) = \frac{1}{|X|}\sum_{x \in X} p(Fehler, x) = \frac{1}{|X|}\sum_{x \in X} p(Fehler \mid x)\, p(x)$$

A-posteriori-Wahrscheinlichkeit (Beispiel)



Verallgemeinerte Bayessche Entscheidungsregel

Bayes-Formel:

$$p(\omega_j \mid \mathbf{x}) = \frac{p(x \mid \omega_j)p(\omega_j)}{p(\mathbf{x})}$$

mit der Evidenz $p(\mathbf{x}) = \sum_{j=1}^{c} p(\mathbf{x} \mid \omega_j)p(\omega_j)$

Die erwarteten Kosten für Aktion α_i (das bedingte Risiko von α_i):

$$R(\alpha_i \mid \mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i \mid \omega_j)p(\omega_j \mid \mathbf{x})$$

Verallgemeinerte Entscheidungsregel (minimiert das erwartete Gesamtrisiko):

„Wähle immer die Aktion α_i, die das bedingte Risiko für die gegebene Beobachtung **x** minimiert."

Aufgabe 3.1 - verallgemeinerte Bayesregel

Sie installieren Ihr neues Bildverarbeitungssystem in einer Chipfabrik, das 95% aller kaputten Chips entdeckt, allerdings auch bei 1% der funktionsfähigen Chips Alarm schlägt. Insgesamt sind 1% aller Chips Ausschuß. Die Herstellungskosten für einen Chip betragen 5 €, der Verkaufpreis 20 €. Die Folgekosten für einen irrtümlich durchgelassenen kaputten Chip sind sehr hoch, nämlich 1995 €.

Ihr Chef will von Ihnen das erwartete Gesamtrisiko bei Anwendung der optimalen Bayes-Entscheidungsregel wissen, d.h. wieviel Gewinn oder Verlust er in diesem Fall pro Chip machen wird.

Spezialfall: 2 Klassen, 2 Aktionen

Vereinfachte Kostenfunktion: $\lambda_{ij} = \lambda(\alpha_i \mid \omega_j)$

Bedingtes Risiko:

$$R(\alpha_1 \mid \mathbf{x}) = \lambda_{11}p(\omega_1 \mid \mathbf{x}) + \lambda_{12}p(\omega_2 \mid \mathbf{x})$$
$$R(\alpha_2 \mid \mathbf{x}) = \lambda_{21}p(\omega_1 \mid \mathbf{x}) + \lambda_{22}p(\omega_2 \mid \mathbf{x})$$

Entscheidungsregel: Entscheide für α_1, wenn

$$(\lambda_{21} - \lambda_{11})p(\omega_1 \mid \mathbf{x}) > (\lambda_{12} - \lambda_{22})p(\omega_2 \mid \mathbf{x})$$

bzw.

$$(\lambda_{21} - \lambda_{11})p(\mathbf{x} \mid \omega_1)p(\omega_1) > (\lambda_{12} - \lambda_{22})p(\mathbf{x} \mid \omega_2)p(\omega_2)$$

Unter der Annahme, daß $\lambda_{21} > \lambda_{11}$, entscheide für α_1

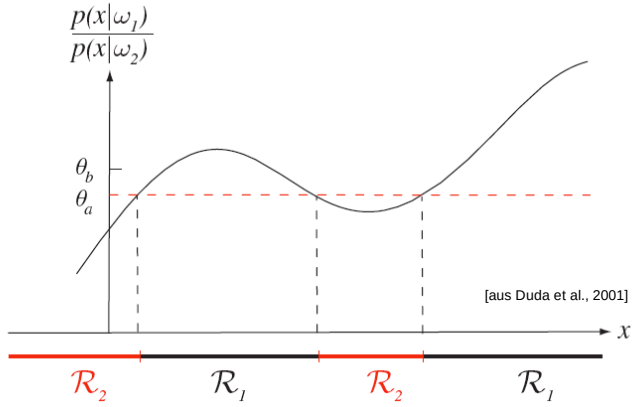
$$\frac{p(\mathbf{x} \mid \omega_1)}{p(\mathbf{x} \mid \omega_2)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{p(\omega_2)}{p(\omega_1)}$$

(likelihood ratio)

Aufgabe 3.2 - verallgemeinerte Bayesregel

Leiten Sie die Entscheidungsregel von Aufgabe 3.2. mit Hilfe der *Likelihood Ratio* ab.

Likelihood ratio (Beispiel)

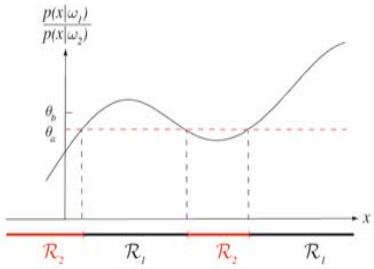


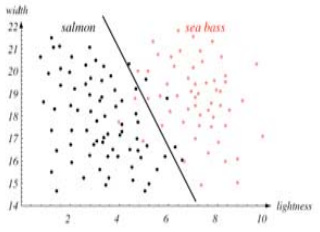
[aus Duda et al., 2001]

Überblick

- **Wiederholung: Bayes-Klassifikator**
- **Diskriminantenfunktionen**
- **Kontinuierliche Zufallsvariablen**
- **Gaußverteilung**

Beispiele für Diskriminantenfunktionen





[aus Duda et al., 2001]

Bayes-Klassifikator:

„Entscheide für ω_1 wenn $p(\omega_1|x) > p(\omega_2|x)$, sonst für ω_2"

Linearer Klassfikator (Trennlinie)

Diskriminantenfunktionen

Diskriminantenfunktionen sind die häufigste Art, einen Klassifikator zu beschreiben.

Vorgehensweise:

- Zu jeder Klasse ω_i wird eine Diskriminantenfunktion $g_i(\boldsymbol{x})$ definiert.
- Der Klassifikator klassifiziert den Inputvektor $\boldsymbol{x}$ als zur Klasse ω_i zugehörig, wenn

$$g_i(\boldsymbol{x}) > g_j(\boldsymbol{x}) \quad \text{für alle } j \neq i$$

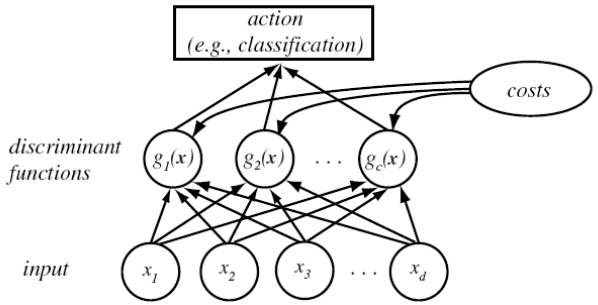
Beispiel Bayes-Klassifikator: „Wähle immer die Aktion α_i, die das bedingte Risiko für die gegebene Beobachtung **x** minimiert.", d.h. entscheide für Klasse ω_i, wenn gilt

$$R(\alpha_i \mid \mathbf{x}) < R(\alpha_j \mid \mathbf{x}) \; bzw. \; p(\omega_i \mid \mathbf{x}) > p(\omega_j \mid \mathbf{x}) \quad für\, alle \quad j \neq i$$

d.h. die Diskriminantenfunktion ist hier

$$g_i(\mathbf{x}) = -R(\alpha_i \mid \mathbf{x}) \quad bzw. \quad g_i(\mathbf{x}) = p(\omega_i \mid \mathbf{x})$$

Allgemeiner Aufbau eines statistischen Klassifikators



[aus Duda et al., 2001]

Eindeutigkeit der Diskriminantenfunktion

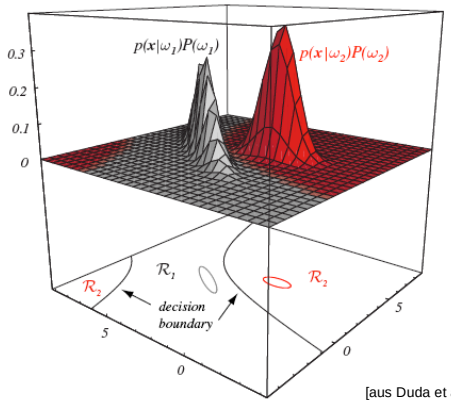
- die Wahl der Diskriminantenfunktion ist nicht eindeutig.
- die Entscheidungsregel bleibt gleich, wenn alle Diskriminantenfunktionen mit der gleichen positiven Konstante multipliziert werden oder wenn zu allen die gleiche Konstante addiert wird.
- Allgemein gilt: Wenn *f(.)* eine monoton steigende Funktion ist, dann bleibt die Klassifikationsregel unverändert, wenn alle Diskriminantenfunktionen $g_i(\mathbf{x})$ durch $f(g_i(\mathbf{x}))$ ersetzt werden.
- z.B. ergeben die folgenden Diskriminantenfunktionen dieselben Entscheidungsregeln:

$$g_j(\mathbf{x}) = p(\omega_j \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_j)\, p(\omega_j)}{p(\mathbf{x})}$$

$$g_j(\mathbf{x}) = p(\mathbf{x} \mid \omega_j)\, p(\omega_j)$$

$$g_j(\mathbf{x}) = \ln p(\mathbf{x} \mid \omega_j) + \ln p(\omega_j)$$

Entscheidungsgrenzen und Entscheidungsregionen

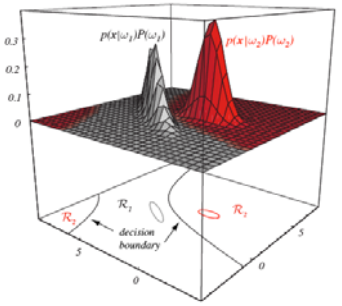


Die Entscheidungsregel unterteilt den Inputraum in Entscheidungsregionen.

Die Entscheidungsregionen sind durch Entscheidungsgrenzen voneinander getrennt.

[aus Duda et al., 2001]

Dichotomien



[aus Duda et al., 2001]

Spezialfall: Entscheidungsprobleme mit 2 Kategorien werden Dichotomien genannt, die entsprechenden Klassifikatoren Dichotomisierer.

Statt 2 Diskriminantenfunktionen $g_1(\mathbf{x})$ und $g_2(\mathbf{x})$ wird stattdessen eine einzige

$$g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x})$$

definiert, mit der Entscheidungsregel „entscheide für ω_1, wenn $g(\mathbf{x}) > 0$, sonst für ω_2 “.

Beispiel: Diskriminantenfunktion für minimale Fehlerrate:

$$g(\mathbf{x}) = p(\omega_1 \mid \mathbf{x}) - p(\omega_2 \mid \mathbf{x})$$

$$g(\mathbf{x}) = \ln \frac{p(\mathbf{x} \mid \omega_1)}{p(\mathbf{x} \mid \omega_2)} + \ln \frac{p(\omega_1)}{p(\omega_2)}$$

Überblick

- **Wiederholung: Bayes-Klassifikator**
- **Diskriminantenfunktionen**
- **Kontinuierliche Zufallsvariablen**
- **Gaußverteilung**

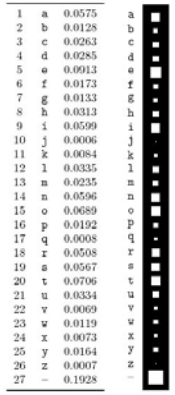
Kontinuierliche Zufallsvariablen (1)

- Bisher: Zufallsvariablen, bei denen ein diskreter Wert a_i aus einem Alphabet mit einer Wahrscheinlichkeit p_i angenommen wird.
- Jetzt: Erweiterung auf Zufallsvariablen mit kontinuierlichen Werten
- Die Wahrscheinlichkeit für einen bestimmten Wert a_i ist bei kontinuierlichen Zufallsvariablen 0!
- Es macht mehr Sinn, davon zu sprechen, daß die ZV einen bestimmten Wert im Intervall $[a,b]$ annimmt.
- Bisher: $p(x \in [a,b]) = \sum_{x_i \in [a,b]} p(x)$

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

(kontinuierliche Wahrscheinlichkeitssdichte)

i	a_i	p_i	
1	a	0.0575	a
2	b	0.0128	b
3	c	0.0263	c
4	d	0.0285	d
5	e	0.0913	e
6	f	0.0173	f
7	g	0.0133	g
8	h	0.0313	h
9	i	0.0599	i
10	j	0.0006	j
11	k	0.0084	k
12	l	0.0335	l
13	m	0.0235	m
14	n	0.0596	n
15	o	0.0689	o
16	p	0.0192	p
17	q	0.0008	q
18	r	0.0508	r
19	s	0.0567	s
20	t	0.0706	t
21	u	0.0334	u
22	v	0.0069	v
23	w	0.0119	w
24	x	0.0073	x
25	y	0.0164	y
26	z	0.0007	z
27	–	0.1928	–



Kontinuierliche Zufallsvariablen (2)

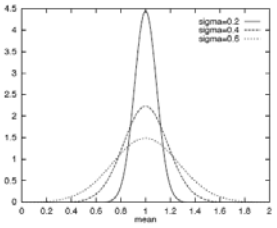
Die bisher benutzten Formeln für diskrete ZV können direkt übernommen werden, indem die Summierung durch ein Integral ersetzt wird:

- Positivität und Normierung: $p(x) \geq 0 \quad und \quad \int_{-\infty}^{\infty} p(x)\, dx = 1$
- Erwartungswert: $E[f(x)] = \int_{-\infty}^{\infty} f(x) p(x)\, dx$
- Mittelwert: $\mu = E[x] = \int_{-\infty}^{\infty} x p(x)\, dx$
- Varianz: $\sigma^2 = E[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)\, dx$
- Bayes-Formel: $p(x \mid y) = \frac{p(y \mid x) p(x)}{\int_{-\infty}^{\infty} p(y \mid x) p(x)\, dx}$

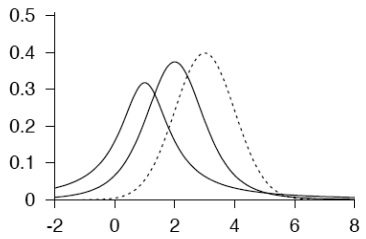
Häufig benutzte Verteilungen (1)

Da im Normalfall die genaue Verteilung der Daten nicht bekannt sind, benutzt man bestimmte Verteilungsfunktionen als Modell der Daten, um damit Vorwissen oder Vorannahmen zu beschreiben.

Beispiel: Verteilungen für ZV, deren Werte unbeschränkt sind:



Gaußverteilung

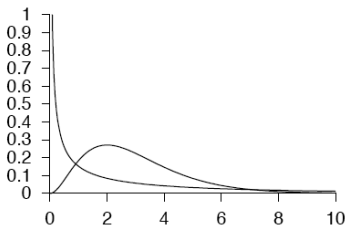


Student-t-Verteilung
(fällt langsamer ab)

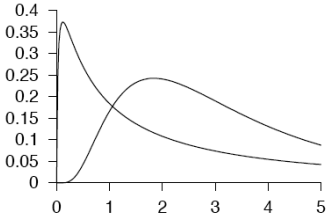
[aus McKay, 2003]

Häufig benutzte Verteilungen (2)

Verteilungen für ZV mit positiven Werten:



Gammaverteilung (x >= 0)



Log-Normalverteilung (x > 0)

Auswahl nach: - Art der Daten (positiv, diskret, ...)
- unimodal, bimodal
- einfache Berechenbarkeit

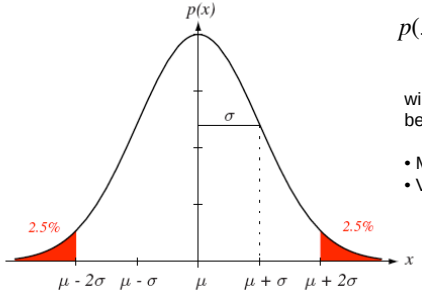
[aus McKay, 2003]

Überblick

- **Wiederholung: Bayes-Klassifikator**
- **Diskriminantenfunktionen**
- **Kontinuierliche Zufallsvariablen**
- **Gaußverteilung**

Gaußverteilung (1)

Zentraler Grenzwertsatz der Wahrscheinlichkeitsrechnung: die Summe einer großen Zahl von (beinahe) beliebig verteilten Zufallsvariablen nähert sich der Gaußverteilung.



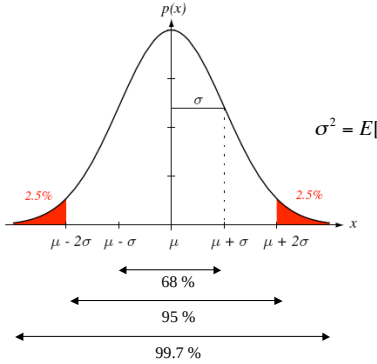
$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

wird durch 2 Parameter beschrieben:

- Mittelwert μ
- Varianz σ^2

Schreibweise: $N(\mu, \sigma^2)$

Gaußverteilung (2)



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

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

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

- Aufgrund des zentralen Grenzwertsatzes ist die Gaußverteilung oft ein gutes Modell.
- Viele Integrale lassen sich geschlossen lösen => häufig in probabilistischen Modellen.