Appendix zu:

D: Grundlegende Lernverfahren

- **Motivation: Fallbeispiel Sägemühle**
- **Kalkül und Regel von Bayes: Zweiklassenproblem**
- **Generalisierung: Minimale Fehler und Risiken der Klassifikation nach Bayes**
- **Bayes-Kalkül und Diskriminantenfunktionen**
- **Fall der Normalverteilung (NV) als klassenbedingte Wahrscheinlichkeitsdichtefunktion**
- **NV und Diskriminantenfunktionen (linear und quadratisch)**

Bayessche Entscheidungstheorie

(Quelle: R. Duda, P. Hart (1972): Pattern Classification and Scene Analysis. Wiley & Sons)

- Historie
  - Thomas Bayes (1702-1763), brit. Pastor und Mathematiker
  - posthume Veröffentlichung des Kalküls
- Problemstellung: Sägemühle
  - 2-Klassen-Problem: Esche und Birke
  - Problem (Fig. 1.2): Merkmal Helligkeit unzureichend für „sichere“ Entscheidung

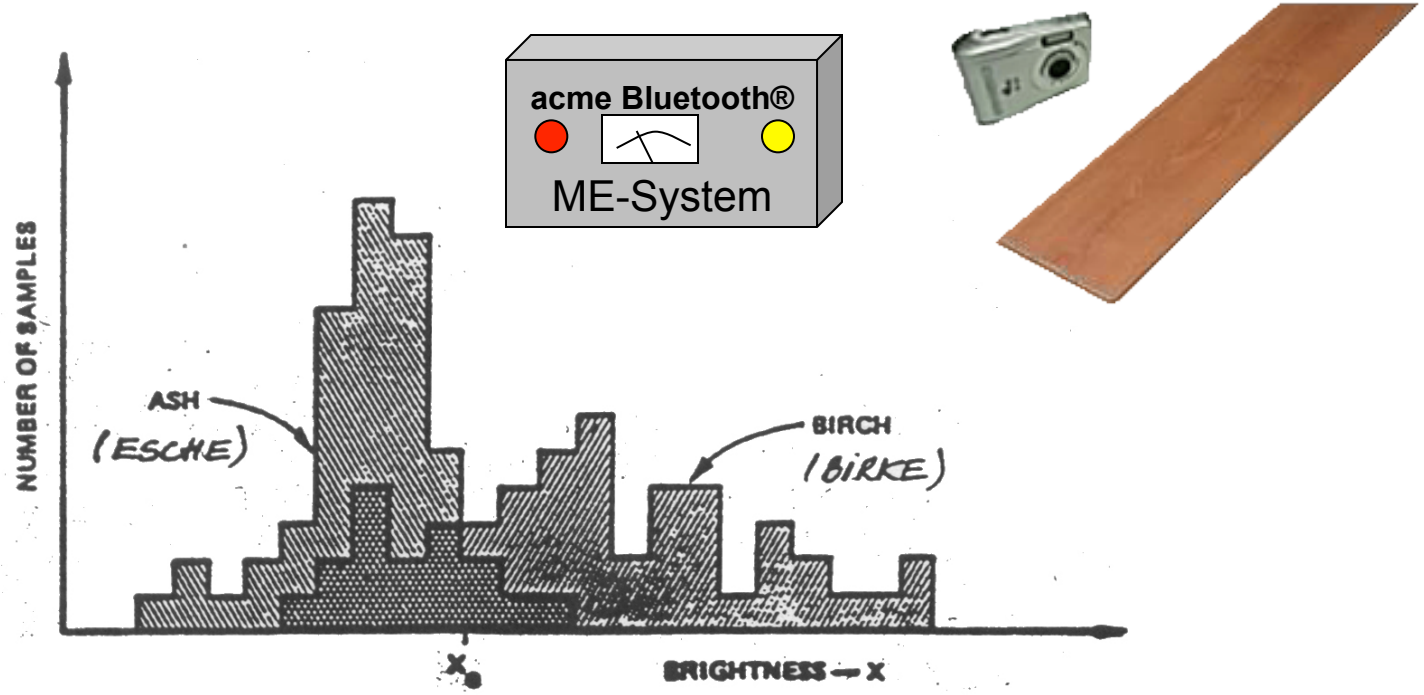


FIGURE 1.2. Histograms for the brightness feature.

(diskrete Häufigkeitsverteilung ⇨ Histogramm der „beobachteten“ Helligkeit)

Bayessche Entscheidungstheorie

- Lösung: Erhöhung der Dimension des Merkmalraumes:
  Helligkeit und Maserung (x_1 und x_2), damit Merkmalvektor $\vec{x} = \left(x_1, x_2\right)^T$

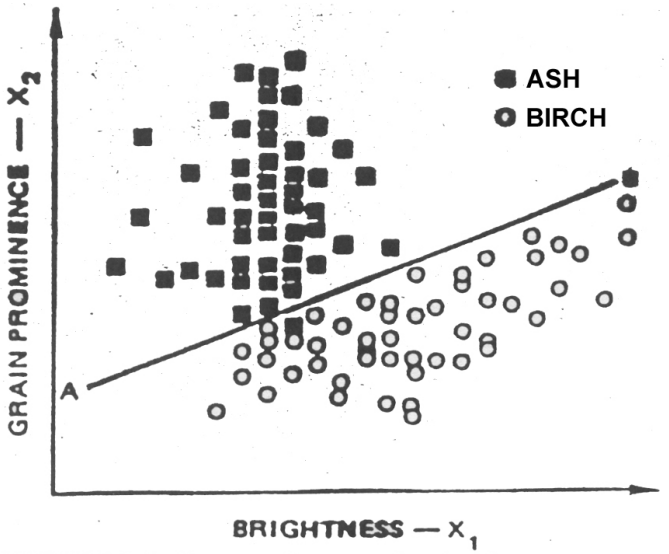


FIGURE 1.3. Scatter diagram for the feature vectors

(Fig. 1.2. und 1.3. aus Duda & Hart, 1972)

(2-d Merkmalraum: Helligkeit vs. Maserung)

jedoch: weiterhin lokaler Fehler beobachtbar!

Lösung: Erhöhung der Dimension
Wahl anderer (d.h. besserer) Merkmale
Verwendung nichtlinearer Trennfunktion

aber: „Kosten“ der Dimensionserhöhung etc., d.h. mathematischer, algorithmischer und technischer Aufwand

Bayessche Entscheidungstheorie

- zentrale wiss.-methodische Fragen
  - Berücksichtigung der statistischen Eigenschaften der Daten (hier: Elemente der Stichprobe als Vektoren im Merkmalraum)?
  - Klassifikation mit Minimierung der Fehlerwahrscheinlichkeit?
  - Bezug zu Diskriminantenfunktionen?
  - Perzeptron als statistischer Klassifikator?
  - theoretische Grenzen?

  → mathematisches Modell von Bayes für Entscheiden/Klassifizieren bei statistischer Unsicherheit mit minimalem Fehler

- Bayes ‘sches Kalkül
  - Annahmen (allgemein)
    1) wahrscheinlichkeitstheoretische Formulierung des Problems
    2) probabilistische Parameterwerte vollständig bekannt (!)

Bayessche Entscheidungstheorie

„*Apriori versus Aposteriori*

Es ist, wie man weiß, IMMANUEL KANT gewesen, der in seinen kritischen Schriften die Grenzen der Vernunft und der Urteilskraft untersuchte[26]; und er schälte aus ihnen jene Voraussetzungen heraus, die nicht aus der Erfahrung heraus stammen können, weil sie selbst für den elementarsten Erfahrungs-Gewinn die Voraussetzung sind. Dies sind die *Apriori* der Vernunft und der Urteilskraft. Ermutigend in KANTS Ergebnis ist nur die Präzision. Wir erfahren nun sehr genau, worin das Dilemma der Vernunft besteht. Das Problem der Vernunft aber wird anerkanntermaßen nicht gelöst, es wird präzisiert. Denn das *Apriori*, was wörtlich: ›vom früheren her‹ bedeutet, läßt sich nicht hinterfragen. Die Kette der Voraussetzungen von den Voraussetzungen kann kein Ende finden. Und damit erweist sich gerade das, worauf sich all unsere Vernunft gründen muß, durch eben diese Vernunft als durchaus unbegründbar.

Was begründet unser Vertrauen auf Wahrscheinlichkeit, die uns eine, wenn auch nur ungefähre, Voraussicht vorspiegelt, über eine Voraussicht, die wir nicht besitzen können? Wie viele weiße Schwäne müssen wir sehen, um in unserem Schluß: alle Schwäne werden weiß sein, dennoch zu irren? Wer macht uns vertrauen, daß Mehrheit Wahrheit bedeuten könnte? »Der Verrückte, der sich für ein Rührei hält«, sagt der verzweifelte BERTRAND RUSSELL, wäre dann »nur deshalb abbzulehnen, weil er sich in der Minderheit befindet.«[27] Hierin liegt schon das Problem der *a-priori*-Wahrscheinlichkeit, sowie das HUME-KANT-POPPERsche Induktionsproblem, auf dem alles Schließen vom Speziellen auf das Allgemeine ruht. Ein Schluß also, mit dessen Berechtigung alle induktive Wissenschaft, das ist die ganze Naturwissenschaft, stehen oder fallen muß. Keine Gewißheit, kein zureichender Grund ist, wie RUDOLF CARNAP und WOLFGANG STEGMÜLLER zeigen[28], selbst in der modernen induktiven Logik für diesen Schluß zu finden. Und KARL POPPER erklärt ihn als einen Widerspruch in sich selbst.[29].“

(aus: R. Riedl (1981): Biologie der Erkenntnis, Verlag P. Parey)

Bayessche Entscheidungstheorie

- Bayessches Kalkül
  - Annahmen (allgemein)
    1) wahrscheinlichkeitstheoretische Formulierung des Problems
    2) probabilistische Parameterwerte vollständig bekannt (!)
  - 2-Klassen-Problem (z.B. Sägemühle)
    - Klassenvektor ⇔ Zustandsvektor $\vec{\omega} = (\omega_1, \omega_2)^T$
    - a priori-Wahrscheinlichkeit eines Zustandes $P(\omega_i)$, mit

$$P(\omega_1), P(\omega_2) \geq 0$$
$$P(\omega_1) + P(\omega_2) = 1$$

      d.h. Wahrscheinlichkeit, dass sich das System im Zustand ω_i befindet
    - Merkmalsvektor $\vec{x}$ als kontinuierliche Zufallsvariable
    - klassenbedingte Wahrscheinlichkeitsdichtefunktion $p\left(\vec{x} \middle| \omega_j\right)$, d.h. die Wahrscheinlichkeit, $\vec{x}$ zu beobachten, wenn ω_j vorliegt (oder: wenn das System im Zustand ω_j ist)
    - spezielle Annahmen
      - $P(\omega_i)$ und $p\left(\vec{x} \middle| \omega_j\right)$ seien bekannt
      - $\vec{x}$ wurde gemessen bzw. beobachtet

Bayessche Entscheidungstheorie

- Bayessche Regel

$$P\left(\omega_j \middle| \vec{x}\right) = \frac{p\left(\vec{x} \middle| \omega_j\right) \cdot P\left(\omega_j\right)}{p\left(\vec{x}\right)} \quad \text{mit } p\left(\vec{x}\right) = \sum_{j=1}^{2} p\left(\vec{x} \middle| \omega_j\right) \cdot P\left(\omega_j\right)$$ [1]

a posteriori-Wahrscheinlichkeit; d.h. Wahrscheinlichkeit, dass Klasse ω_j vorliegt (bzw. Wahrscheinlichkeit, dass sich das System im Zustand ω_j befindet), nachdem (!) $\vec{x}$ beobachtet/gemessen wurde.

$$\Rightarrow \begin{array}{ll} P\left(\omega_1 \middle| \vec{x}\right) > P\left(\omega_2 \middle| \vec{x}\right), & \text{dann } \vec{x} \sim \omega_1 \\ P\left(\omega_1 \middle| \vec{x}\right) < P\left(\omega_2 \middle| \vec{x}\right), & \text{dann } \vec{x} \sim \omega_2 \end{array}$$ [2]

mit minimaler mittlerer Fehlerwahrscheinlichkeit, d.h. über alle möglichen Entscheidungen gemittelte Wahrscheinlichkeit für eine falsche Entscheidung wird minimal!

- Fehler

$$P\left(\text{Fehler} \middle| \vec{x}\right) = \begin{cases} P\left(\omega_1 \middle| \vec{x}\right), & \text{für } \vec{x} \sim \omega_2 \\ P\left(\omega_2 \middle| \vec{x}\right), & \text{für } \vec{x} \sim \omega_1 \end{cases}$$

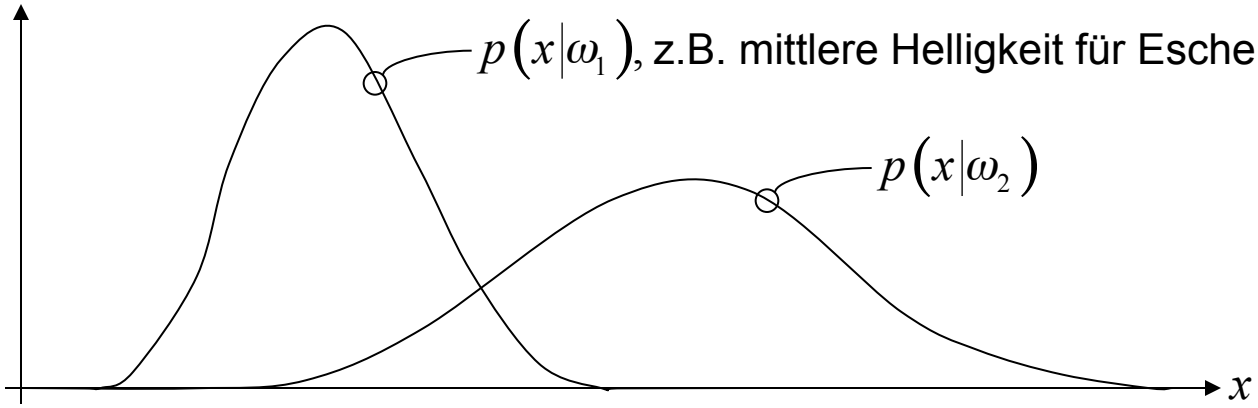
---

[1]: nach Duda & Hart (1972) "... just a scale factor", damit $P\left(\omega_1 \middle| \vec{x}\right) + P\left(\omega_2 \middle| \vec{x}\right) = 1$

[2]: lies $\vec{x} \sim \omega_1 : \vec{x}$ wird der Klasse ω_1 zugeordnet

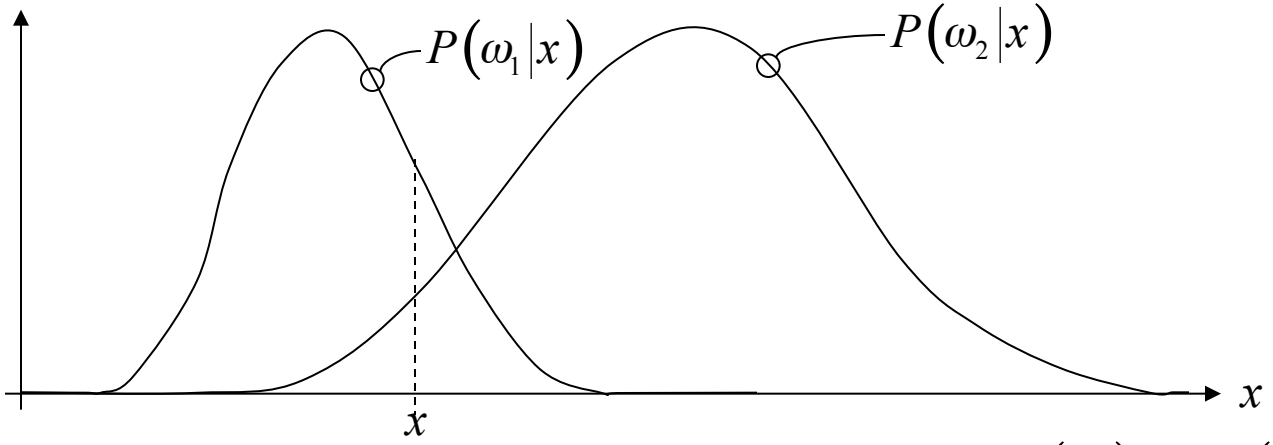
Bayessche Entscheidungstheorie

- Wahrscheinlichkeits-Funktionen (nach: Duda & Hart (1972), S.12)



(hypothetische) klassenbedingte Wahrscheinlichkeits-Dichtefunktion

(mit: x entspricht Helligkeit im Sägemühlen-Bsp.), d.h. „Streuung“ des/der beobachteten Merkmale(s) in der Stichprobe in Abhängigkeit der bekannten Klasse



(hypothetische) *a posteriori*-Wahrscheinlichkeiten für $P(\omega_1) = \frac{2}{3}, P(\omega_2) = \frac{1}{3}$, d.h. $x \sim \omega_1$, da $P(\omega_1|x) > P(\omega_2|x)$ mit Fehler $P(\text{Fehler}|x) = P(\omega_2|x)$

Bayessche Entscheidungstheorie

- Sei P(Fehler) die mittlere Fehlerwahrscheinlichkeit mit

$$P(\text{Fehler}) = \int_{-\infty}^{\infty} P(\text{Fehler}|x) \cdot p(x)\,dx$$ (für ein Merkmal!)

somit: für $x \sim \omega_2$

$$P(\text{Fehler}) = \int_{-\infty}^{\infty} \underbrace{P(\omega_1|x)}_{\frac{P(\omega_1)\,p(x|\omega_1)}{p(x)}} \cdot p(x)\,dx$$

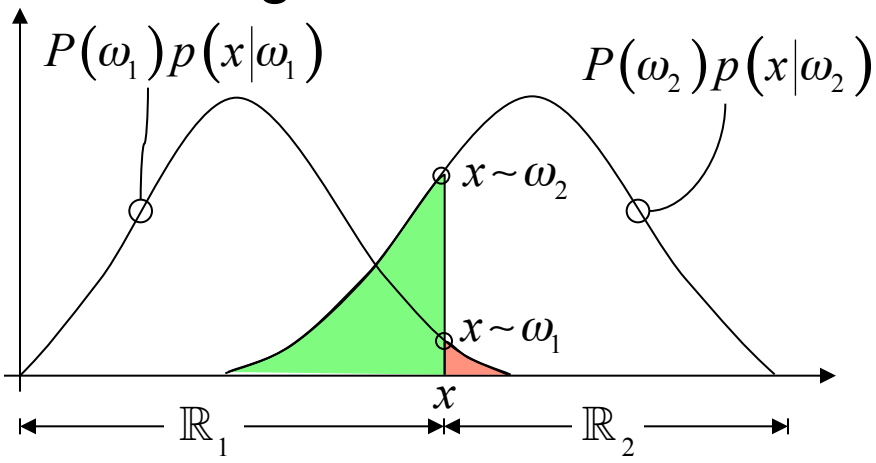
$\hat{=}$ *a posteriori*-Wahrscheinlichkeit

$$\boxed{P(\text{Fehler}) = \int_{-\infty}^{\infty} P(\omega_1) \cdot p(x|\omega_1)\,dx}$$

und für $x \sim \omega_1$

$$\boxed{P(\text{Fehler}) = \int_{-\infty}^{\infty} P(\omega_2) \cdot p(x|\omega_2)\,dx}$$

Bayessche Entscheidungstheorie



d.h. Fehler bei $\begin{cases} x \sim \omega_1: & \int_{\mathbb{R}_1} P(\omega_2)\, p(x|\omega_2)\, dx \\ x \sim \omega_2: & \int_{\mathbb{R}_2} P(\omega_1)\, p(x|\omega_1)\, dx \end{cases}$

da jedoch gemäß Bayesscher Regel gelten muss $x \sim \omega_2$

$$\Rightarrow \underbrace{\int_{\mathbb{R}_1} P(\omega_2)\, p(x|\omega_2)\, dx}_{\text{in Abb. grün dargestellt}} > \underbrace{\int_{\mathbb{R}_2} P(\omega_1)\, p(x|\omega_1)\, dx}_{\text{in Abb. rot dargestellt}}$$

d.h. Bayessche Regel minimiert mittlere Fehlerwahrscheinlichkeit

---

Hinweis: $\mathbb{R}_1$ und $\mathbb{R}_2$ in der obigen Abbildung sind willkürlich festgelegte Entscheidungs-Regionen zur Illustration!

Bayessche Entscheidungstheorie

Grenzfälle:

i) $p(x|\omega_1) = p(x|\omega_2)$: Entscheidung hängt nur von *a priori*-Wahrscheinlichkeiten ab!

ii) $P(\omega_1) = P(\omega_2)$: Entscheidung hängt nur von klassenbedingter Wahrscheinlichkeits-Dichtefunktion ab!

iii) $P(\omega_1|x) = P(\omega_2|x)$: „zufällige“ Entscheidung mit falls gleiche Fehler-Wahrscheinlichkeit!

Bayessche Entscheidungstheorie

- Generalisierung (für den kontinuierlichen Fall)
  - mehr als ein Merkmal: Merkmalsvektor $\vec{x}$
  - mehr als 2 Klassen/Zustände: n-Klassen-Problem
  - nicht nur Klassifikation, sondern auch andere Formen von „actions“, z.B. Zurückweisungen
  - nicht nur Betrachtung der Fehler-Wahrscheinlichkeit, sondern Berücksichtigung von Kosten-(Verlust-)Funktionen ('loss functions')
  - $\Omega = \{\omega_1, ..., \omega_n\}$: endliche Menge von Klassen

    $A = \{\alpha_1, ..., \alpha_n\}$: endliche Menge von „actions“ (zu verstehen als z.B. Klassenzuweisung, Systemabbruch etc.)

    $\lambda(\alpha_i|\omega_j)$: Kosten (Verlust) bei der Entscheidung für α_i, falls „das System im Zustand ω_j“

    $\vec{x}$: vektorielle Zufallsvariable mit d Komponenten, $(x_1, x_2, ..., x_d)^T$

    $p\left(\vec{x}\middle|\omega_j\right)$: klassenbedingte Wahrscheinlichkeits-Dichtefunktion

    $P(\omega_j)$: a priori-Wahrscheinlichkeit

    a posteriori-Wahrscheinlichkeit (bei s Klassen):

$$P\left(\omega_j\middle|\vec{x}\right) = \frac{p\left(\vec{x}\middle|\omega_j\right)\cdot P\left(\omega_j\right)}{p\left(\vec{x}\right)} \quad \text{mit} \quad p\left(\vec{x}\right) = \sum_{j=1}^{s} p\left(\vec{x}\middle|\omega_j\right)\cdot P\left(\omega_j\right)$$

Bayessche Entscheidungstheorie

Annahme: $\vec{x}$ wird beobachtet;

Entscheidung für 'action' α_i wird in Erwägung gezogen, ω_j ist jedoch wahre(r) Klasse (bzw. Zustand),

dann ist $\lambda(\alpha_i|\omega_j)$ 'loss' ↦ Kosten/Verlust

- bedingtes (mittleres) Risiko über alle s Klassen

$$R\left(\alpha_i \middle| \vec{x}\right) = \sum_{j=1}^{s} \lambda\left(\alpha_i \middle| \omega_j\right) \cdot P\left(\omega_j \middle| \vec{x}\right)$$

nach der Beobachtung/Messung von $\vec{x}$ (wegen $P\left(\omega_j \middle| \vec{x}\right)$)

- Gesamtrisiko

für jedes $\vec{x}$ mit Entscheidungs-Regel $\alpha\left(\vec{x}\right): \vec{x} \sim \left\{\alpha_1, \ldots, \alpha_a\right\}$:

$$\boxed{\mathbf{R} = \int_{-\infty}^{\infty} R\left(\alpha\left(\vec{x}\right)|\vec{x}\right) p\left(\vec{x}\right) d\vec{x}} \quad \rightarrow \min,$$

d.h. Wahl von $\alpha(\vec{x})$ so, dass $R(\alpha(\vec{x})|\vec{x})$ minimal wird für jedes $\vec{x}$.

- optimale Bayessche Entscheidungsregel (Bayes-Klassifikator):

„Um das Gesamtrisiko zu minimieren, berechne das bedingte Risiko $R(\alpha_i(\vec{x})|\vec{x})$, $i = 1, ..., a$, und wähle α_i so, daß $R(\alpha_i(\vec{x})|\vec{x})$ ein Minimum ist.“

! minimales Gesamtrisiko ↦ Bayessches Risiko **!**

Bayessche Entscheidungstheorie

Beispiel: 2-Klassen-Problem

α_1: Entscheidung für Klasse ω_1

α_2: Entscheidung für Klasse ω_2

$\lambda_{ij} = \lambda(\alpha_i|\omega_j)$: Kosten für Entscheidung α_i (d.h. für ω_i), obwohl ω_j vorliegt

$$\mathbf{R}(\alpha(\vec{x})|\vec{x}) = \mathbf{L}\,\mathbf{P}(\vec{\omega}|\vec{x}) \quad \Rightarrow \quad R(\alpha_1|\vec{x}) = \lambda_{11}P(\omega_1|\vec{x}) + \lambda_{12}P(\omega_2|\vec{x})$$
$$R(\alpha_2|\vec{x}) = \lambda_{21}P(\omega_1|\vec{x}) + \lambda_{22}P(\omega_2|\vec{x})$$

mit Verlust-/Kosten („loss“-) -Matrix $\mathbf{L} = \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{pmatrix}$ [1]

$\vec{x} \sim \omega_1$ iff $R(\alpha_1|\vec{x}) < R(\alpha_2|\vec{x})$

bzw. iff $(\lambda_{21} - \lambda_{11})\underbrace{P(\omega_1|\vec{x})}_{} > (\lambda_{12} - \lambda_{22})\underbrace{P(\omega_2|\vec{x})}_{}$

bzw. iff $(\lambda_{21} - \lambda_{11})p(\vec{x}|\omega_1)P(\omega_1) > (\lambda_{12} - \lambda_{22})p(\vec{x}|\omega_2)P(\omega_2)$

bzw. iff $\underbrace{\frac{p(\vec{x}|\omega_1)}{p(\vec{x}|\omega_2)}}_{} > \frac{(\lambda_{12} - \lambda_{22})P(\omega_2)}{(\lambda_{21} - \lambda_{11})P(\omega_1)}$ [1] $\overset{!}{\equiv}$ konstanter Schwellwert

'likelihood ratio': Bayessche Regel in Abhängigkeit eines von der Beobachtung $\vec{x}$ unabhängigen Schwellwertes!

---

[1] Beachte: Kosten für einen Fehler sind normalerweise höher als für eine richtige Entscheidung, daher $\lambda_{21} > \lambda_{11}$ und $\lambda_{12} > \lambda_{22}$ im allgemeinen Fall

Bayessche Entscheidungstheorie

- Klassifikation mit minimaler Fehlerrate
  - symmetrische (oder 'zero-one') Kostenfunktion

$$\ddot{e}\left(\alpha_i \middle| \omega_j\right) = \begin{cases} 0, & i = j \\ 1, & i \neq j \end{cases} \quad i, j = 1, \ldots, c \text{ (Klassen)}$$

$$\Rightarrow \mathbf{L} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

0: keine Kosten

1: Einheitskosten bei Fehler

  - d.h. Entscheidung für α_i ist gefallen, ω_j liegt tatsächlich vor, dann ist Entscheidung <u>korrekt</u> für $i = j$ (wegen Nullkosten)
  - für das 2-Klassen-Problem gilt nun

$$\lambda_{11} = \lambda_{22} = 0 \quad \text{und} \quad \lambda_{12} = \lambda_{21} = 1$$

und

$$R\left(\alpha_1 \middle| \vec{x}\right) = \lambda_{12} P\left(\omega_2 \middle| \vec{x}\right) = P\left(\omega_2 \middle| \vec{x}\right)$$
$$R\left(\alpha_2 \middle| \vec{x}\right) = \lambda_{21} P\left(\omega_1 \middle| \vec{x}\right) = P\left(\omega_1 \middle| \vec{x}\right)$$

  - da nach Bayes-Regel gilt:

$\vec{x} \sim \omega_1$ iff $P\left(\omega_1 \middle| \vec{x}\right) > P\left(\omega_2 \middle| \vec{x}\right)$ mit Fehler $P\left(\text{Fehler} \middle| \vec{x}\right) = P\left(\omega_2 \middle| \vec{x}\right)$

muss auch gelten: $R\left(\alpha_1 \middle| \vec{x}\right) < R\left(\alpha_2 \middle| \vec{x}\right)$

und somit ist die Forderung nach der <u>Minimierung</u> des Risikos erfüllt!

Bayessche Entscheidungstheorie

- d.h. das mit dieser Kosten-/Verlust-Funktion verbundene Risiko entspricht der mittleren Fehler-Wahrscheinlichkeit, da

$$R(\alpha_i|\vec{x}) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j) \cdot P(\omega_j|\vec{x})$$
$$= \sum_{j \neq i} P(\omega_j|\vec{x})$$
$$= 1 - \underbrace{P(\omega_i|\vec{x})}$$

bedingte Wahrscheinlichkeit, dass α_i korrekt ist!

- ergo: „Um die mittlere Fehler-Wahrscheinlichkeit zu minimieren, muss i so gewählt werden, dass die a posteriori- Wahrscheinlichkeit $P(\omega_i|\vec{x})$ maximiert wird“, d.h. minimale Fehlerrate nur für $\vec{x} \sim \omega_i$ wenn $P(\omega_i|\vec{x}) > P(\omega_j|\vec{x}),\ \forall\ i \neq j$

- Klassifikatoren, Diskriminantenfunktionen und Entscheidungs-(Trenn-)Hyperflächen (siehe: Teil C „Diskriminantenfunktionen...“ zur Einführung)
  - Diskriminantenfunktionen $g_i(\vec{x})$
  - Entscheidungsregel $\vec{x} \sim \omega_i$ wenn $g_i(\vec{x}) > g_j(\vec{x}),\ \forall\ i \neq j,\ i = 1, 2, \ldots, c$ (Klassen)

    d.h. Aufteilung des Merkmalraumes durch Trenn-Hyperflächen in c „Entscheidungs-gebiete“ Ω_1, ..., Ω_c, so dass $\vec{x}$ innerhalb von Ω_i und $g_i(\vec{x}) = g_j(\vec{x})$ ist Gleichung der Trennfläche(n)

Bayessche Entscheidungstheorie

- Beziehung zu Bayes-Klassifikator
  - allgemeiner Fall: $\underbrace{g_i(\vec{x})}_{\text{maximale Diskriminantenfunktion}} = \underbrace{-R(\alpha_i|\vec{x})}_{\text{minimales bedingtes Risiko}}$
  - Fall der minimalen Fehlerrate: $\underbrace{g_i(\vec{x})}_{\text{maximale Diskriminantenfunktion}} = \underbrace{P(\omega_i|\vec{x})}_{\text{maximale a posteriori-Wahrscheinlichkeit}}$

! Multiplikation von $g_i(\vec{x})$ mit einer positiven Konstanten oder Addition einer Konstanten beeinflusst nicht die Entscheidung, d.h. z.B. Ersetzung von $g_i(\vec{x})$ durch $f(g_i(\vec{x}))$, wobei f eine streng monoton ansteigende Funktion sei, hat keinen Einfluss auf Klassifikationsresultat.

⇒ z.B. verschiedene Ansätze zur '*minimum error rate*'-Klassifikation mit verschiedenem **Aufwand** aber identischen Klassifikationsergebnissen:

$$g_i(\vec{x}) = \underbrace{P(\omega_i|\vec{x})}$$

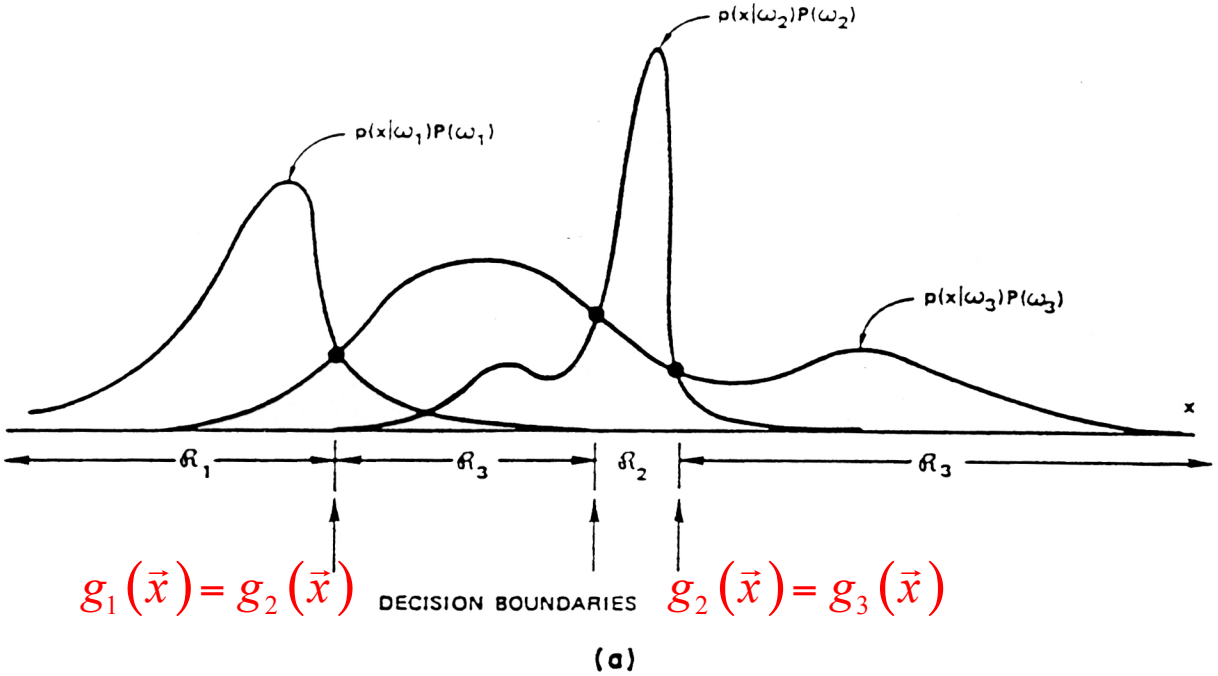
$$g_i(\vec{x}) = \frac{p(\vec{x}|\omega_i)\cdot P(\omega_i)}{\sum_{j=1}^{c} p(\vec{x}|\omega_j)\cdot P(\omega_j)}$$

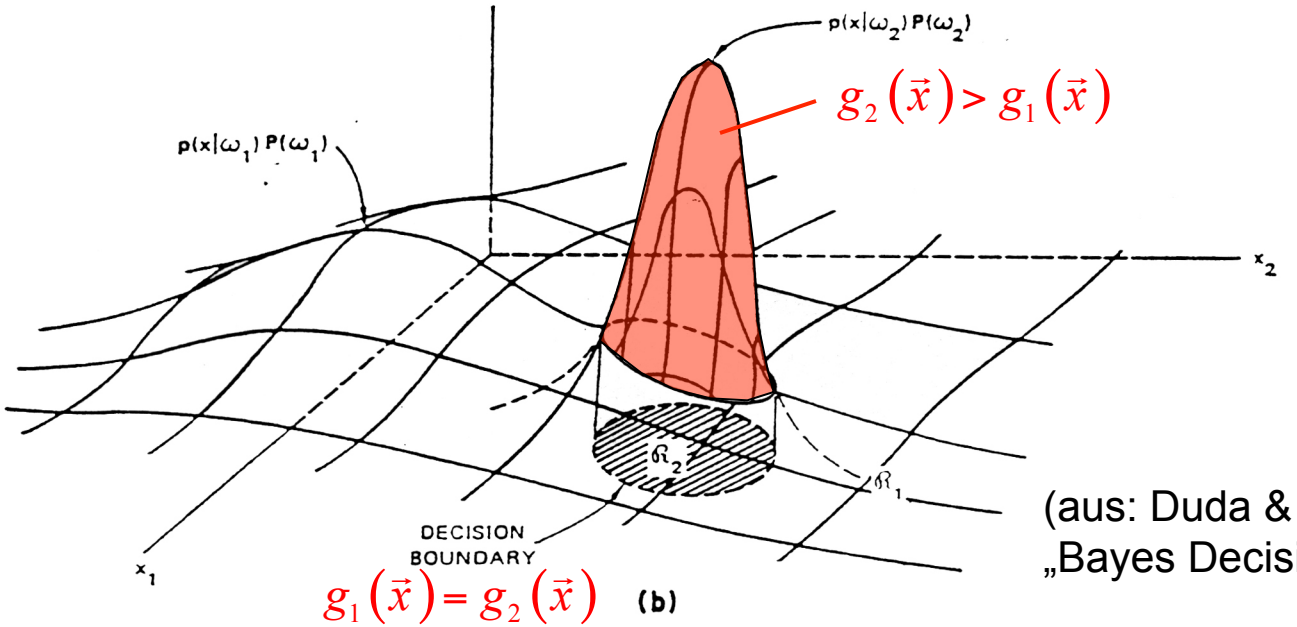
$$g_i(\vec{x}) = p(\vec{x}|\omega_i)\cdot P(\omega_i)$$

$$g_i(\vec{x}) = \log p(\vec{x}|\omega_i) + \log P(\omega_i)$$

Aufwandsreduktion!
(Klassengrenzen verschieben sich nicht!)

Beispiel:





(aus: Duda & Hart, 1973, S. 19, Kap. 2 „Bayes Decision Theory“)

FIGURE 2.4. Examples of decision boundaries and decision regions.

Bayessche Entscheidungstheorie

Beispiel: 2-Klassen-Fall

anstatt: $\vec{x} \sim \omega_1$ wenn $g_1(\vec{x}) > g_2(\vec{x})$

üblicher: $g(\vec{x}) = g_1(\vec{x}) - g_2(\vec{x})$ so dass

$\vec{x} \sim \omega_1$ wenn $g(\vec{x}) > 0$

technische Aufwandsreduktion, da nur Vorzeichenprüfung (anstatt Maximumberechnung)!

mit 'minimum error rate'-Diskriminantenfunktion $g(\vec{x})$*:

$$g(\vec{x}) = P(\omega_1 | \vec{x}) - P(\omega_2 | \vec{x})$$

oder $$g(\vec{x}) = \log \frac{p(\vec{x}|\omega_1)}{p(\vec{x}|\omega_2)} + \log \frac{P(\omega_1)}{P(\omega_2)}$$

(*... „following two are particularly convenient...“, Duda & Hart, 1973)

Bayessche Entscheidungstheorie

- Fehler-Wahrscheinlichkeiten und Fehler-Integrale
  - Entscheidungs-„Gebiete“ $\mathbb{R}_1$ und $\mathbb{R}_2$ im 2-Klassen-Fall

  - 2 Fehlertypen:
    i) $\vec{x}$ fällt in $\mathbb{R}_2$, aber wahrer Zustand ist ω_1
    ii) $\vec{x}$ fällt in $\mathbb{R}_1$, aber wahrer Zustand ist ω_2

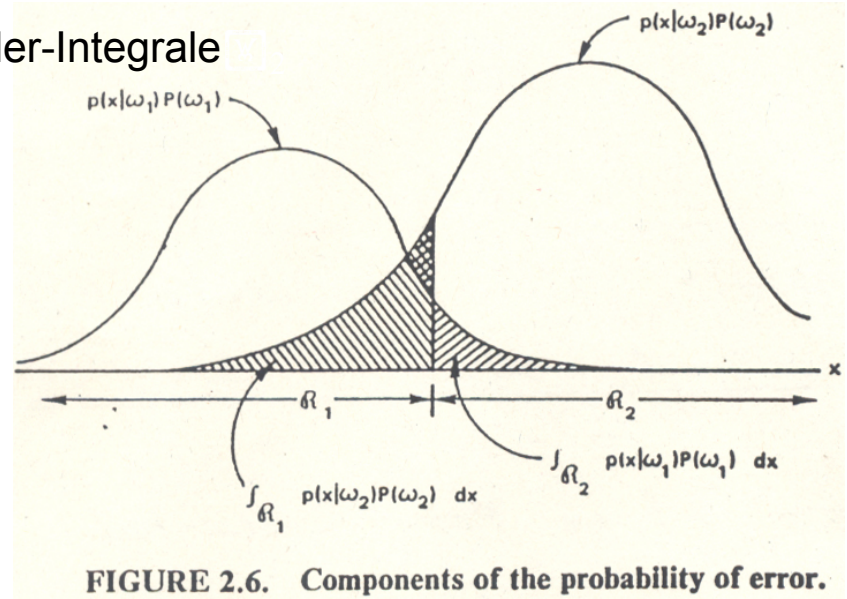


FIGURE 2.6. Components of the probability of error.

damit:
$$P(\text{Fehler}) = P(\vec{x} \in \mathbb{R}_2, \omega_1) + P(\vec{x} \in \mathbb{R}_1, \omega_2)$$
$$= P(\vec{x} \in \mathbb{R}_2 | \omega_1) \cdot P(\omega_1) + P(\vec{x} \in \mathbb{R}_1 | \omega_2) \cdot P(\omega_2)$$

$$\Rightarrow \boxed{P(\text{Fehler}) = \int_{\mathbb{R}_2} p(\vec{x}|\omega_1) \cdot P(\omega_1) d\vec{x} + \int_{\mathbb{R}_1} p(\vec{x}|\omega_2) \cdot P(\omega_2) d\vec{x}}$$

„This result is illustrated in the one-dimensional case in Figure 2.6. The two terms in the sum are merely the areas in the tails of the functions $p(\mathbf{x}|\omega_i)P(\omega_i)$. Because the regions R_1 and R_2 were chosen arbitrarily, the probability of error is not as small as it might be. By moving the decision boundary to the left, it is clear that we can eliminate the dark „triangular“ area and reduce the probability of error. In general, if $p(\mathbf{x}|\omega_1)P(\omega_1) > p(\mathbf{x}|\omega_2)P(\omega_2)$, it is advantageous to have $\mathbf{x}$ be in R_1 so that the smaller quantity will contribute to the integral; this is exactly what the Bayes decision rule achieves.“ (Duda & Hart, 1973, S. 21)

Bayessche Entscheidungstheorie

- Alternative, z.B. im Mehr-Klassen-Fall

$$P(\text{Korrektheit}) = \sum_{i=1}^{c} P\left(\vec{x} \in \mathbb{R}_i, \omega_i\right) \quad \text{(Summation über alle Klassen!)}$$

$$= \sum_{i=1}^{c} P\left(\vec{x} \in \mathbb{R}_i \middle| \omega_i\right) \cdot P\left(\omega_i\right)$$

$$\boxed{P(\text{Korrektheit}) = \sum_{i=1}^{c} \int_{\mathbb{R}_i} p\left(\vec{x} \middle| \omega_i\right) P\left(\omega_i\right) d\vec{x}}$$

! Bayes-Klassifikator maximiert $P(\text{Korrektheit})$, da er $\mathbb{R}_i$ so festlegt, dass Integranden maximal (und damit $P(\text{Fehler})$ minimal) werden.

„In the multiclass case, there are more ways to be wrong than to be right, and it is simpler to compute the probability of being correct.“ (Duda & Hart, 1973, S. 21)

Bayessche Entscheidungstheorie

- Normal-Verteilung (NV) als klassenbedingte Wahrscheinlichkeits-Dichtefunktion
  - Motivation:
    - Bedeutung von $p(\vec{x}|\omega_i)$ für Bayes-Klassifikator
    - analytische „Handhabbarkeit“
    - $\vec{x}$ einer Klasse ω_i streuen um Prototypen $\vec{\mu}_i$ (NV als „Modell“)
  - univariate normalverteilte Dichtefunktion

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

mit μ: Mittelwert,
σ: Standardabweichung,
σ^2: Varianz

wobei $\mu = \int_{-\infty}^{\infty} x \cdot p(x) dx = E(x)$

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot p(x) dx = E\left((x-\mu)^2 \right)$$

Notation[1]: $p(x) \sim N(\mu, \sigma^2)$

---

[1] Zur Erinnerung: $\int_{\mu-2\sigma}^{\mu+2\sigma} p(x) dx \approx 0,95$

Bayessche Entscheidungstheorie

- multivariate normalverteilte Dichtefunktion in d Dimensionen
  - $p\left(\vec{x}\right) \sim N\left(\vec{\mu}, \boldsymbol{\Sigma}\right)$
  - $\vec{x} = \left(x_1, x_2, \ldots, x_d\right)^T$
  - $\vec{\mu} = \left(\mu_1, \mu_2, \ldots, \mu_d\right)^T$
  - $\boldsymbol{\Sigma}$: $d \times d$ – Kovarianzmatrix ($\boldsymbol{\Sigma}^{-1}$: Inverse, $|\boldsymbol{\Sigma}|$: Determinante)

$$p\left(\vec{x}\right) = \frac{1}{\left(2\pi\right)^{d/2} \cdot \left|\boldsymbol{\Sigma}\right|^{1/2}} \cdot \exp\left(-\frac{1}{2}\left(\vec{x} - \vec{\mu}\right)^T \boldsymbol{\Sigma}^{-1}\left(\vec{x} - \vec{\mu}\right)\right)$$

mit $\vec{\mu} = E\left(\vec{x}\right);\ \mu_i = E\left(x_i\right)$

und $\boldsymbol{\Sigma} = E\left(\left(\vec{x} - \vec{\mu}\right)\left(\vec{x} - \vec{\mu}\right)^T\right)$[1]

$$= \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \cdots & \sigma_{nn} \end{pmatrix},$$

σ_{ii} : Varianz von x_i

$\sigma_{ij} = \underbrace{E\left(\left(x_i - \mu_i\right)\left(x_j - \mu_j\right)\right)}_{\text{Kovarianz von } x_j \text{ und } x_j}$

! $\sigma_{ij} = 0$: x_i und x_j sind statistisch unabhängig

! beliebige Linearkombination von N: $\vec{y} = \mathbf{A}^T\vec{x} \quad \Rightarrow \quad p\left(\vec{y}\right) \sim N\left(\mathbf{A}^T\vec{\mu}, \mathbf{A}^T\boldsymbol{\Sigma}\mathbf{A}\right)$

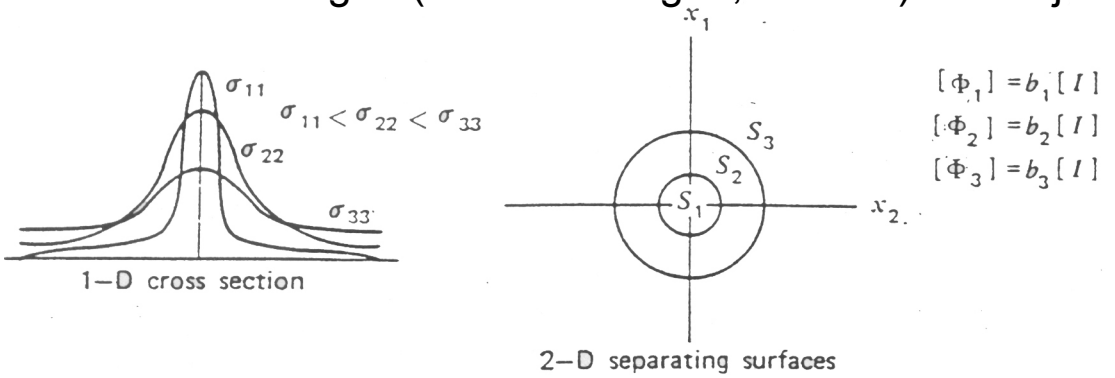
---

[1] $\boldsymbol{\Sigma}$ ist immer symmetrisch und positiv-semidefinit, d.h. $\vec{x}^T\boldsymbol{\Sigma}\vec{x} \geq 0, \forall\, \vec{x}$ (s. Lin. Alg.)

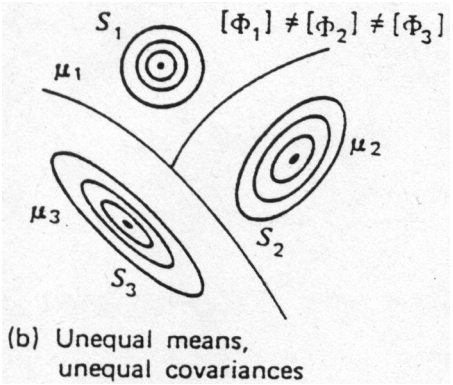
Bayessche Entscheidungstheorie

! Bedeutung der Kovarianzmatrix:

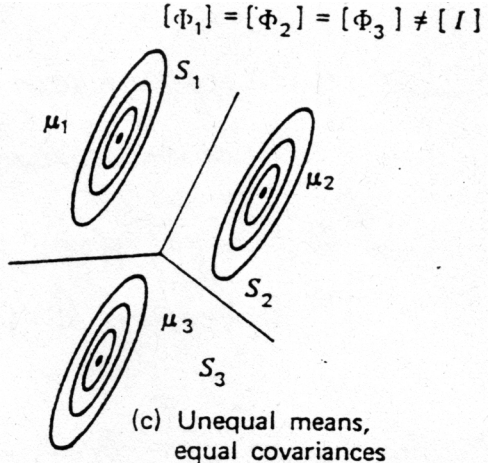
erlaubt Berechnung der Streuung ('dispersion') von Daten (Vektoren) im Merkmalraum bzw. von Datenhäufungen (auch: Ballungen, 'cluster') in z.B. jede beliebige Richtung



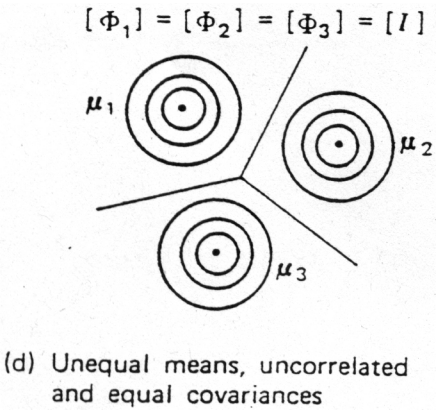
(a) Equal means, uncorrelated covariances



(b) Unequal means, unequal covariances



(c) Unequal means, equal covariances



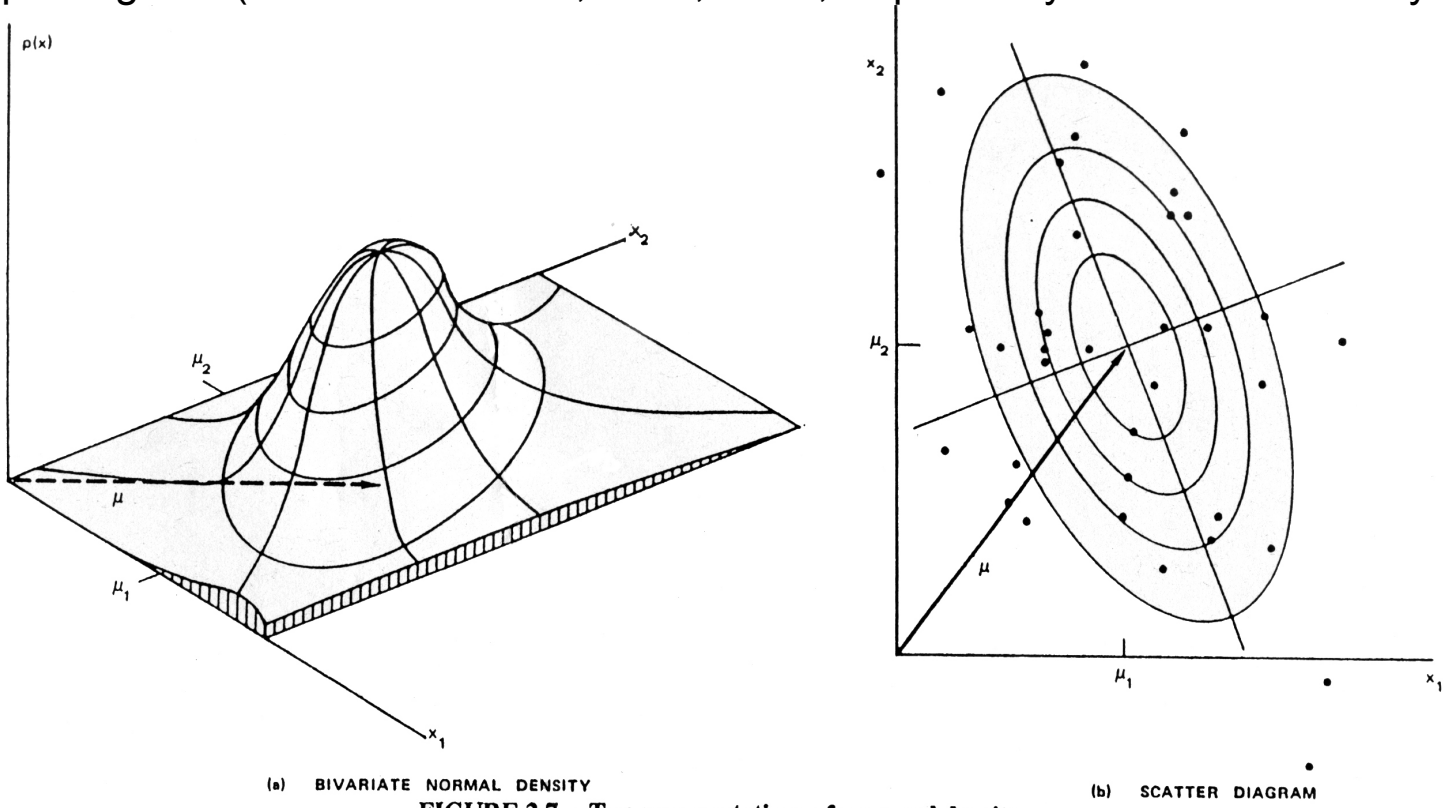
(d) Unequal means, uncorrelated and equal covariances

FIGURE 4.4. Some typical separating surfaces for $N(\mu_k, [\Phi_k])$, (N = 2, k = 3).

aus: AND 72, S. 117, Kap. 4.2 "Parametric Classification"

Bayessche Entscheidungstheorie

Beispiel Fig. 2.7 (aus: Duda & Hart, 1973, S. 25, Kap. 2 "Bayes Decision Theory“)



(a) BIVARIATE NORMAL DENSITY (b) SCATTER DIAGRAM

FIGURE 2.7. Two representations of a normal density.

- Interpretation
  - Zentrum des 'cluster' bei $\vec{\mu} = \left(\mu_1, \mu_2\right)^T$
  - „Form“ des 'cluster' gegeben durch $\boldsymbol{\Sigma}$, d.h. Orte der Punkte gleicher Dichtefunktionswerte sind Hyperellipsoide, für die $\left(\vec{x} - \vec{\mu}\right)^T \boldsymbol{\Sigma}^{-1} \left(\vec{x} - \vec{\mu}\right) = \text{const},$

    wobei 1) deren Hauptachsen definiert sind durch die Eigenvektoren von $\boldsymbol{\Sigma}$

    2) Länge der Hauptachsen definiert durch Eigenwerte von $\boldsymbol{\Sigma}$

Bayessche Entscheidungstheorie

- Mahalanobis-Distanz zwischen $\vec{x}$ und $\vec{\mu}$

$$r^2 = \left(\vec{x} - \vec{\mu}\right)^T \boldsymbol{\Sigma}^{-1} \left(\vec{x} - \vec{\mu}\right)$$

d.h. r^2 ist konstant für Punkte auf Isodichte-Linien!

- Streuung ('scatter') von $\vec{x}$ um $\vec{\mu}$ aus Volumina der Hyperellipsoiden V (für best. r)

$$\boxed{V = V_d \left|\boldsymbol{\Sigma}\right|^{1/2} r^d}$$ als 'scatter'-Maß

mit $V_d = \begin{cases} \dfrac{\pi^{d/2}}{\left(\frac{d}{2}\right)!}, & \text{für } d \text{ gerade} \\ \dfrac{2^d \pi^{d-1/2} \left(\frac{d-1}{2}\right)!}{d!}, & \text{für } d \text{ ungerade} \end{cases}$

(V_d: Volumen einer d-dimensionalen Einheitskugel!)

Anmerkung: Für $d = 2$ gilt: $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$ und $V = \pi \left(\sigma_{11}\sigma_{22} - \sigma_{12}\sigma_{21}\right) r^2$,

d.h. <u>für gegebenes d und beliebiges r</u> hängt Streuung ('scatter') <u>direkt</u> von $|\boldsymbol{\Sigma}|^{1/2}$ ab!!

Bayessche Entscheidungstheorie

- Diskriminantenfunktionen für die normalverteilte Dichtefunktion
  - $g_i(\vec{x}) = \log p(\vec{x}|\omega_i) + \log P(\omega_i)$ für 'minimum-error-rate'-Klassifikation (s.o.)
  - mit $p(\vec{x}|\omega_i) \sim N(\vec{\mu}_i, \boldsymbol{\Sigma}_i)$:

$$g_i(\vec{x}) = -\frac{1}{2}(\vec{x} - \vec{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\vec{x} - \vec{\mu}_i) - \frac{d}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{\Sigma}_i| + \log P(\omega_i)$$

  - Fall 1:
    - Merkmale statistisch unabhängig, $\sigma_{ij} = 0$
    - jedes Merkmal hat gleiche Varianz σ^2

$\Rightarrow \boldsymbol{\Sigma}_i = \sigma^2 I$, d.h. gleichgroße Hyperkugel-"cluster" (pro Klasse) um $\vec{\mu}_i$

- $|\boldsymbol{\Sigma}_i| = \sigma^{2d}$ und $\boldsymbol{\Sigma}_i^{-1} = \left(\frac{1}{\sigma^2}\right) I$

unabhängig von i: additive Konstanten für $g_i(\vec{x})$ werden vernachlässigt!

$$\Rightarrow \boxed{g_i(\vec{x}) = -\frac{\|\vec{x} - \vec{\mu}_i\|^2}{2\sigma^2} + \log P(\omega_i)}$$

mit euklidischer Norm $\|\cdot\|$, d.h. $\|\vec{x} - \vec{\mu}_i\|^2 = (\vec{x} - \vec{\mu}_i)^T (\vec{x} - \vec{\mu}_i)$

- falls $P(\omega_i)$ gleich für alle Klassen:

berechne $\|\vec{x} - \vec{\mu}_i\|$ ($\triangleq$ euklidischer Abstand) von $\vec{x}$ zu allen $\vec{\mu}_i$ bei c Klassen,

$\vec{x} \sim \omega_i$ wenn $\|\vec{x} - \vec{\mu}_i\|$ = Minimum

→ „minimaler Abstand“-Klassifikator

Bayessche Entscheidungstheorie

• jedoch: $g_i(\vec{x}) = -\frac{(\vec{x}-\vec{\mu}_i)^T(\vec{x}-\vec{\mu}_i)}{2\sigma^2} + \log P(\omega_i)$

$$= -\frac{1}{2\sigma^2}\left(\underbrace{\vec{x}^T\vec{x}}_{\text{gleich für alle } i \text{ (damit additive Konstante)!}} - 2\vec{\mu}_i^T\vec{x} + \vec{\mu}_i^T\vec{\mu}_i\right) + \log P(\omega_i)$$

somit ist $g_i(\vec{x})$ linearisiert (s.a. Teil C: „Diskriminantenfunktionen...“):

$$\boxed{g_i(\vec{x}) = \vec{w}_i^T\vec{x} + w_{i_0}}$$ „lineare Maschine“

mit $\vec{w}_i = \frac{1}{\sigma^2}\vec{\mu}_i$

und $w_{i_0} = -\frac{1}{2\sigma^2}\vec{\mu}_i^T\vec{\mu}_i + \log P(\omega_i)$

$\Rightarrow$ Trennfläche $g_i(\vec{x}) = g_j(\vec{x})$ (allgemein)

$\vec{w}^T(\vec{x}-\vec{x}_0) = 0$ (Fall 1)

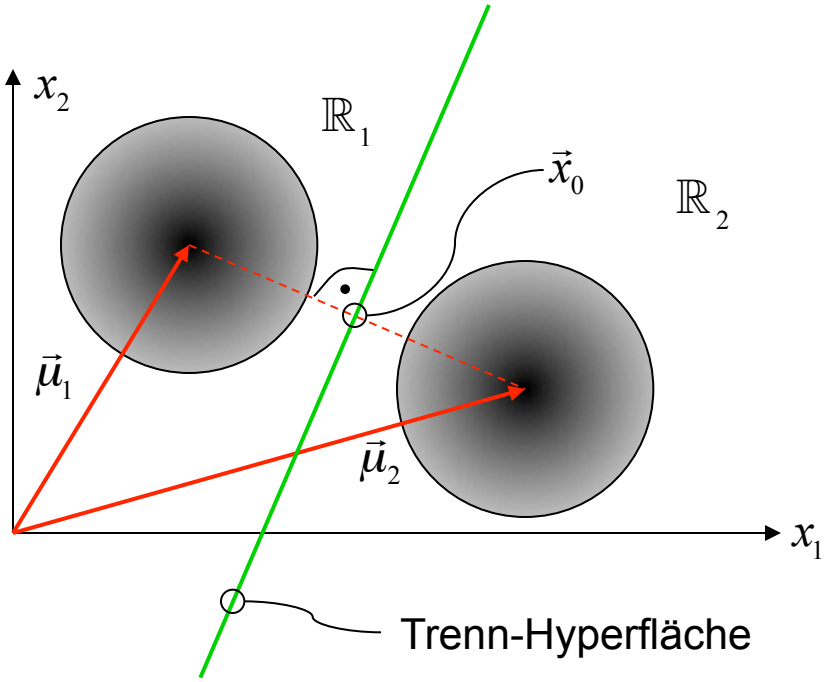
mit $\vec{w} = \vec{\mu}_i - \vec{\mu}_j$

und $\vec{x}_0 = \frac{1}{2}(\vec{\mu}_i + \vec{\mu}_j) - \frac{\sigma^2}{\|\vec{\mu}_i - \vec{\mu}_j\|^2}\log\frac{P(\omega_i)}{P(\omega_j)}\cdot(\vec{\mu}_i - \vec{\mu}_j)$

Bayessche Entscheidungstheorie

d.h. Hyperebene durch $\vec{x}_0$ und orthogonal zu $\vec{w}$;

da $\vec{w} = \vec{\mu}_i - \vec{\mu}_j$, ist Hyperebene orthogonal zu Verbindungslinie zwischen $\vec{\mu}_i$ und $\vec{\mu}_j$ (für $P(\omega_i) = P(\omega_j)$ halbiert $\vec{x}_0$ diese Linie)



Bayessche Entscheidungstheorie

- Fall 2:
  - Kovarianz-Matrizen für alle Klassen gleich
  - d.h. $\mathbf{\Sigma}_i = \mathbf{\Sigma}$: gleiche Hyperellipsoid-"cluster" (pro Klasse um $\vec{\mu}_i$)
  - $|\mathbf{\Sigma}_i|$ und $(d/2)\cdot\log 2\pi$ ebenso unabhängig von i

$$\Rightarrow \boxed{g_i\left(\vec{x}\right) = \underbrace{-\frac{1}{2}\left(\vec{x}-\vec{\mu}_i\right)^T \mathbf{\Sigma}^{-1}\left(\vec{x}-\vec{\mu}_i\right)}_{r^2} + \log P\left(\omega_i\right)}$$

  - falls $P(\omega_i)$ gleich für alle Klassen:

    berechne Mahalanobis-Distanzen r_i^2 von $\vec{x}$ zu allen $\vec{\mu}_i$

    $\vec{x} \sim \omega_i$ wenn $r_i^2 =$ Minimum

  - jedoch: $\vec{x}^T\mathbf{\Sigma}^{-1}\vec{x}$ unabhängig von i und kann vernachlässigt werden,

    somit $g_i\left(\vec{x}\right) = \vec{w}_i^T\vec{x} + w_{i_0}$

    mit $\vec{w}_i = \mathbf{\Sigma}^{-1}\vec{\mu}_i$

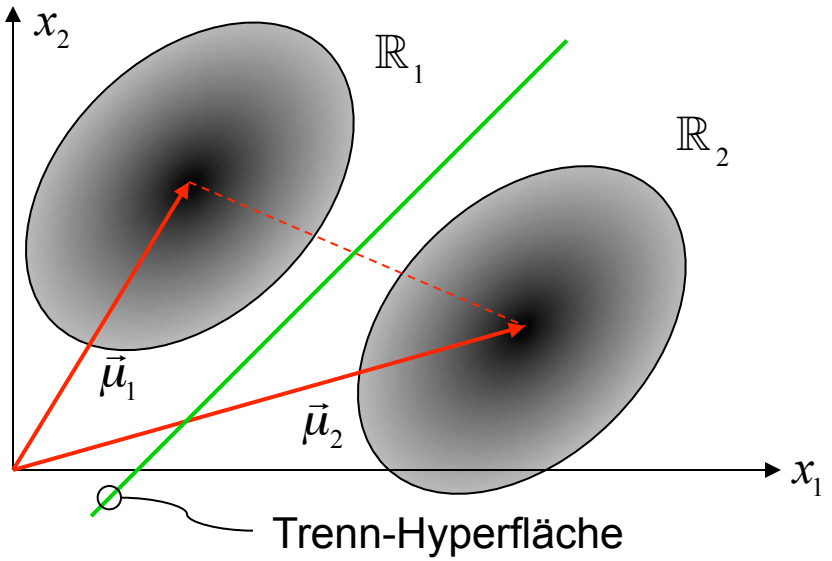
    und $w_{i_0} = -\frac{1}{2}\vec{\mu}_i^T\mathbf{\Sigma}^{-1}\vec{\mu}_i + \log P\left(\omega_i\right)$

    $\Rightarrow g_i\left(\vec{x}\right)$ linearisiert!

Bayessche Entscheidungstheorie

- Trennfläche $\vec{w}^T\left(\vec{x}-\vec{x}_0\right)=0$ mit $\vec{w}=\Sigma^{-1}\left(\vec{\mu}_i-\vec{\mu}_j\right)$ (!)

d.h. Hyperebene zwischen ω_1 und ω_2 i.a. <u>nicht</u> orthogonal, wie im Fall 1!



Bayessche Entscheidungstheorie

- Fall 3:
  - allgemeiner multivariater normalverteilter Fall, d.h. klassenspezifische Kovarianz-Matrizen
  - nur $(d/2)\cdot\log 2\pi$ kann vernachlässigt werden, d.h. <u>quadratische</u> Diskriminanten-Funktion

$$\boxed{g_i(\vec{x}) = \vec{x}^T \mathbf{W}_i \vec{x} + \vec{w}_i^T \vec{x} + w_{i_0}}$$

mit $\mathbf{W}_i = -\frac{1}{2}\mathbf{\Sigma}_i^{-1}$

$$\vec{w}_i = \mathbf{\Sigma}^{-1}\vec{\mu}_i$$

$$w_{i_0} = -\frac{1}{2}\vec{\mu}_i^T\mathbf{\Sigma}^{-1}\vec{\mu}_i - \frac{1}{2}\log|\mathbf{\Sigma}_i| + \log P(\omega_i)$$

$\Rightarrow$ 'hyperquadrics' als Trennflächen:

Paare von Hyperkugeln
oder Hyperellipsoiden
oder Hyperparaboloiden
oder Hyperboloiden und Hyperebenen

für $d = 2$ (also ω_1 und ω_2)

! Grenzen des Perzeptrons bei normalverteilten Daten im Merkmalraum