5 Klassifikation

5.1 Musterklassifikation als mathematische Abbildung

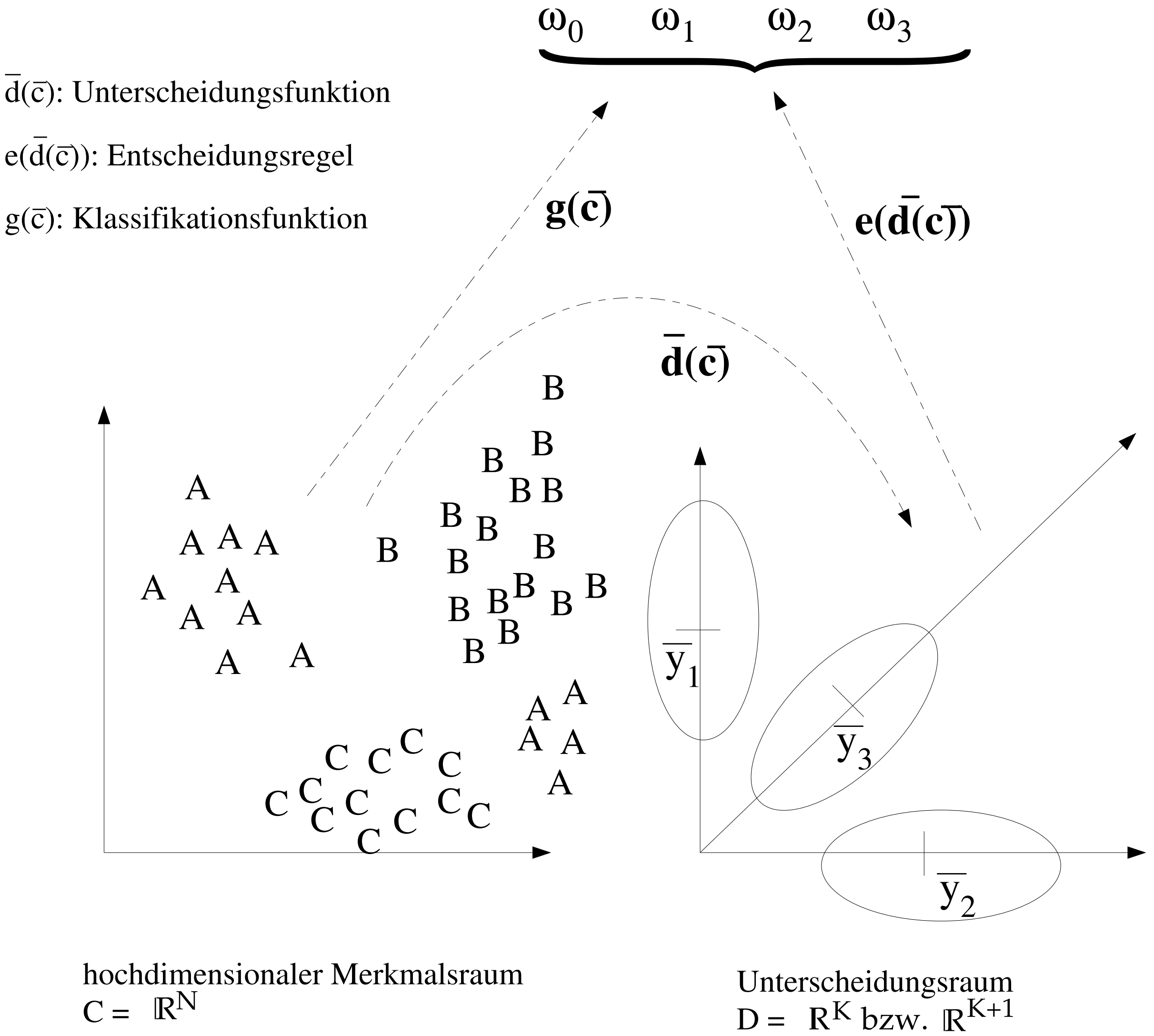
Objekte und Ereignisse werden für die Klassifikation als Merkmalsvektoren $\vec{c}$ beschrieben.

Klassifikation ist die Konstruktion einer Abbildung, die den Merkmalsvektor $\vec{c}$ in ein bestimmtes Symbol aus Ω abbildet, d.h. gesucht wird Klassifikationsfunktion

$$g : C \mapsto \Omega$$

Wegen der einfacheren mathematischen Handhabbarkeit wird diese Funktion meist in zwei Abbildungen zerlegt:

$$g(\vec{c}) = e(\vec{d}(\vec{c}))$$



$\mathbb{R}^K$ bzw. $\mathbb{R}^{K+1}$ ist als Unterscheidungsraum günstiger als $\mathbb{R}^1$: Abbildung der Merkmalsvektoren in einen K- bzw. (K+1)-dimensionalen Raum erfaßt Nachbarschaftsbeziehungen besser

Oft gibt man die Entscheidungsregel $e(\vec{d}(\vec{c}))$ vor:

Klassifiziere in die Klasse ω_i, deren Zielvektor $\vec{y}_i$ minimalen Abstand zur Unterscheidungsfunktion $\vec{d}(\vec{c})$ besitzt.

Die **Zielvektoren** $\vec{y}_i, i = 1, \dots, K$ sind also Repräsentanten für ω_i, wobei gilt:

$$\vec{y}_i = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \begin{array}{l} \\ \\ \\ \leftarrow i - \text{te Komponente enthält } 1 \\ \\ \\ \\ \end{array}$$

Gehört ein Merkmalsvektor $\vec{c}$ zur Klasse ω_i, so wird der zugehörige Zielvektor $\vec{y}_i$ auch als $\vec{y}(\vec{c})$ bezeichnet.

(Achtung: dies ist <u>keine Funktion</u>)

5.2 Statistische Grundlagen

5.2.1 Mustererzeugende Prozesse

Muster sind Wertepaare $(\vec{c}, \omega_i)$, die die Erscheinungsform (repräsentiert durch den Merkmalsvektor $\vec{c}$) mit der Bedeutung des Musters (repräsentiert durch die Klasse ω_i) verbinden.

Der mustererzeugende Prozeß (MEP) wird als stochastischer Prozeß modelliert, der mit der Wahrscheinlichkeit $P(\vec{c}, \omega_i)$ zufällig aber nicht regellos Muster generiert.

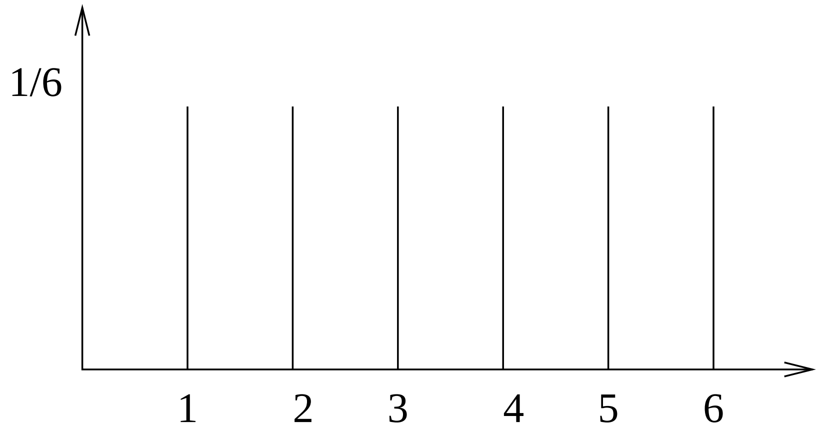
Voraussetzungen für die automatische Klassifikation:

1. Die statistischen Eigenschaften des MEP sind stationär.
2. Statistische Eigenschaften in der Lernphase müssen sich auf die Arbeitsphase übertragen lassen, also wird eine repräsentative Stichprobe benötigt.

5.2.2 Wahrscheinlichkeiten und Wahrscheinlichkeitsdichten

Einer diskreten Zufallsvariable X wird die Wahrscheinlichkeit $P(X)$ zugeordnet.

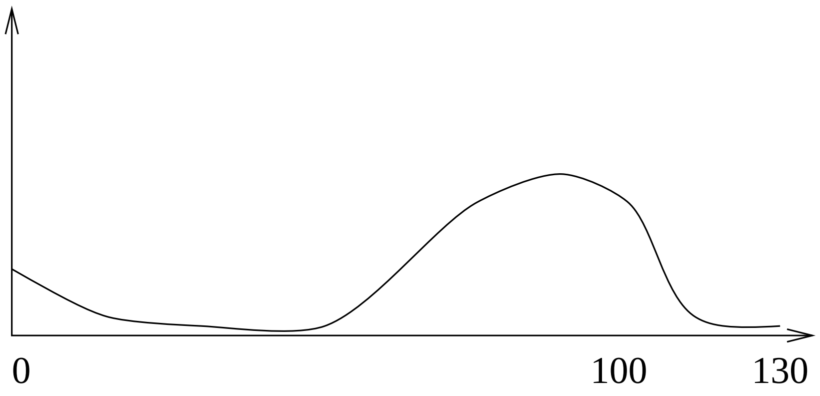
Beispiel: Augenzahl eines Würfels $P(1) = P(2) = \ldots = P(6) = \frac{1}{6}$



Einer kontinuierlichen Zufallsvariable X wird die Wahrscheinlichkeitsdichte $P(X)$ zugeordnet.

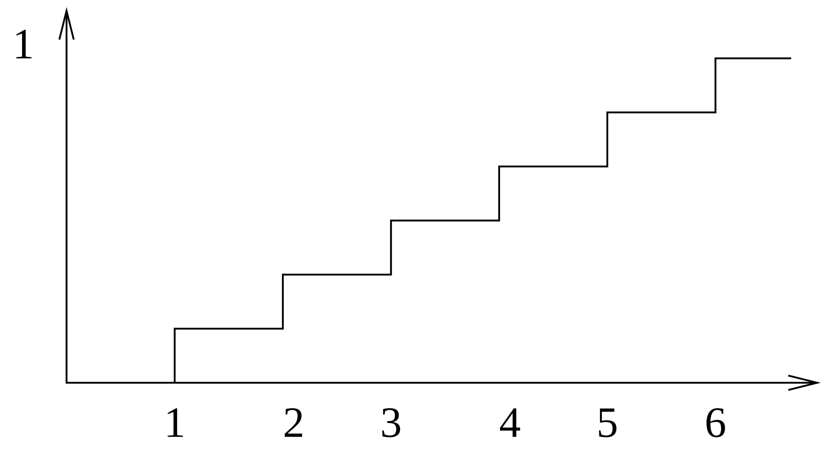
Beispiel: Lebensalter von Menschen

$$P(67 \leq X \leq 68) = \int_{67}^{68} P(X = x)dx$$

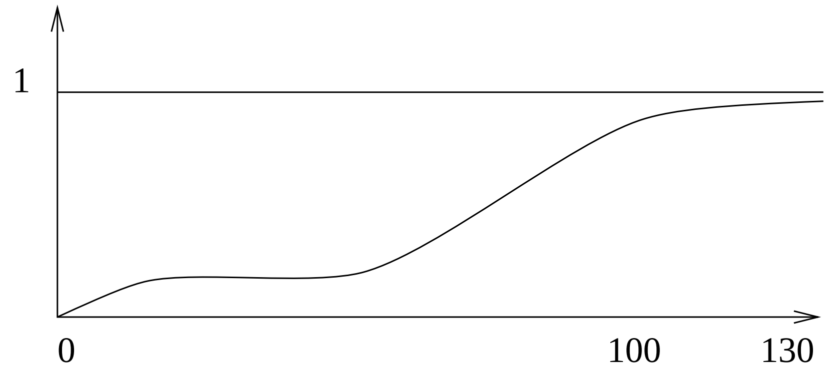


Die Funktion $F(x) = P(X \leq x)$ heißt Verteilungsfunktion.

Beispiel: Augen eines Würfels: $F(2,5) = P(X \leq 2,5) = P(1) + P(2) = \frac{1}{3}$



Beispiel Lebensalter: $F(67,5) = P(x \leq 67,5) = \int\limits_{0}^{67,5} P(X = x)dx$



wie erwähnt, wird der MEP wird durch die Dichte $P(\vec{c}, \omega_i)$ beschrieben.

Daraus lassen sich folgende Dichten und Wahrscheinlichkeiten (WK) ableiten (Randdichten):

- $P(\omega_i) = \int\limits_{\vec{c}} P(\vec{c}, \omega_i) d\vec{c}$ a priori WK der Klasse ω_i
- $P(\vec{c}) = \sum\limits_{\omega_i} P(\vec{c}, \omega_i)$ Dichte der Merkmale (unabhängig von der Bedeutung)

Bedingte Wahrscheinlichkeit und *bedingte* Dichte:

- *Klassenspezifische Dichte, likelihood*: $p(\vec{c} \mid \omega_i)$
- *A posteriori*-Wahrscheinlichkeit oder *Rückschlußwahrscheinlichkeit*: $P(\omega_i \mid \vec{c})$

Nach dem **Gesetz von Bayes** gilt:

$$P(\vec{c}, \omega_i) = P(\vec{c} \mid \omega_i) P(\omega_i) = P(\omega_i \mid \vec{c}) P(\vec{c})$$

5.3 Minimierung des Klassifikationsrisikos

5.3.1 Allgemeiner Ansatz

- für Klassifikationssysteme (KS) ist nur $\vec{c}$, nicht aber ω_i sichtbar
- wir benötigen also eine Klassifikationsfunktion

$$g(\vec{c}) = \widehat{\omega}$$

  die jedem Merkmalsvektor eine Klasse aus $\{\omega_1, \ldots, \omega_K)$ zuordnet oder als nicht gültig zurückweist.
- es kann durchaus vorkommen, daß die ermitteltete Klasse $\widehat{\omega}$ nicht der tatsächlich vorliegenden Klasse ω_i entspricht
- eine solche Fehlklassifikation verursacht Kosten.
  Um diese Kosten näher zu bestimmen, stellt man eine Verlustmatrix $V(\omega_i, \widehat{\omega})$ auf. Hierbei ist $V(\omega_i, \widehat{\omega})$ der Verlust, der entsteht, wenn man sich für $\widehat{\omega}$ entscheidet, obwohl ω_i vorliegt. V ist abhängig von der Anwendung und muß per Hand bestimmt werden.

- Zur Optimierung des KS setzt man eine bereits klassifizierte Stichprobe ein.
- Ziel ist es, den durchschnittlichen Verlust (das *Risiko*) zu minimieren, d.h.
  minimiere $R = E\{V\} = E\{V(\omega_i, \widehat{\omega})\} = E\{V(\omega_i, g(\vec{c}))\}$.
- Wie berechnet man nun den Erwartungwert einer Zufallsvariable? Hierzu zwei Beispiele:

  1. Quadrat der Augenzahlen eines Würfels:
  $$E\{f(X) = X^2\} = 1^2 \cdot P(1) + 2^2 \cdot P(2) + \ldots + 6^2 \cdot P(6) = \sum_{i=1}^{6} i^2 \cdot P(i) = 15.1\bar{6}$$
  2. Lebensalter:
  $$E\{X\} = \int_0^\infty x \cdot P(X = x)\, dx$$

Auf diese Weise wird nun auch das Risiko berechnet:

$$
\begin{aligned}
R = E\{V(\omega, g(\vec{c}))\} &= \int\limits_{\vec{c}} \sum_{\omega_i} V(\omega_i, g(\vec{c})) \cdot P(\vec{c}, \omega_i)\, d\vec{c} \\
\text{nach Bayes} &= \int\limits_{\vec{c}} \sum_{\omega_i} V(\omega_i, g(\vec{c})) \cdot P(\omega_i \mid \vec{c}) \cdot P(\vec{c})\, d\vec{c} \\
&= \int\limits_{\vec{c}} \underbrace{\left(\sum_{\omega_i} V(\omega_i, g(\vec{c})) \cdot P(\omega_i \mid \vec{c}) \right)}_{R_{\vec{c}}(g(\vec{c})) = \text{Risiko für } \vec{c} \text{ oder lokales Risiko}} \cdot P(\vec{c})\, d\vec{c} \\
&= \int_{\vec{c}} R_{\vec{c}}(g(\vec{c})) \cdot P(\vec{c})\, d\vec{c}
\end{aligned}
$$

Das Gesamtrisiko ist natürlich minimal, falls $R_{\vec{c}}(g(\vec{c}))$ minimal für jedes $\vec{c}$ ist.

Da keine Klassifikation oft besser als eine falsche Klassifikation ist, gilt i.a. $\widehat{\omega} = g(\vec{c}) \in \{\omega_0, \omega_1, \ldots, \omega_K\}$.

Es wird nun also das Minimum des lokalen Risikos

$$R_{\vec{c}}(g(\vec{c})) = \sum_{\omega_i} V(\omega_i, g(\vec{c})) \cdot P(\omega_i \mid \vec{c}) \tag{5.1}$$

gesucht.

Setzt man die Unterscheidungsfunktion $\vec{d}(\vec{c}) := \begin{pmatrix} R_{\vec{c}}(\omega_0) \\ \vdots \\ R_{\vec{c}}(\omega_i) \\ \vdots \\ R_{\vec{c}}(\omega_K) \end{pmatrix}$,

so wird das Minimum erreicht, falls man folgende Entscheidungsregel anwendet:

$$g(\vec{c}) = \hat{\omega} = e(\vec{d}(\vec{c})) = \omega_l, \quad \text{falls } l \textit{ minimale} \text{ Komponente von } \vec{d}(\vec{c}) \tag{5.2}$$

dieser Bayes-Klassifikator heißt auch MAP-Klassifikator (maximun a posteriori)

Folgende Werte werden dazu benötigt:

- $V(\omega_i, \widehat{\omega})$: Diese sind für jede Anwendung von Hand zu bestimmen.
- $P(\omega_i \mid \vec{c})$: Diese Wahrscheinlichkeiten werden üblicherweise aus einer repräsentativen und klassifizierten Stichprobe geschätzt:

$$P(\omega_i \mid \vec{c}) = \frac{p(\vec{c} \mid \omega_i) \cdot P(\omega_i)}{P(\vec{c})}$$

5.3.2 Bayes-Klassifikator

Der Bayes-Klassifikator hat eine spezielle, symmetrische Kostenfunktion:

$$V(\omega_i, \widehat{\omega}) = \begin{cases} 0, & \text{falls } \widehat{\omega} = \omega_i \\ V_f, & \text{falls } \widehat{\omega} \neq \omega_i \wedge \widehat{w} \neq \omega_0 \\ V_r, & \text{falls } \widehat{\omega} = \omega_0 \end{cases}$$

Setzt man dies nun in Gleichung (5.1) ein, so erhält man für das lokale Risiko bei einer Fehlentscheidung:

$$\begin{aligned} R_{\vec{c}}(\widehat{\omega} \neq \omega_0) &= \sum_{\omega_i} V(\omega_i, \widehat{\omega}) P(\omega_i \mid \vec{c}) \\ &= 0 \cdot P(\widehat{\omega} \mid \vec{c}) + \sum_{\omega_i \neq \widehat{\omega}} V_f \, P(\omega_i \mid \vec{c}) \\ &= V_f \sum_{\omega_i \neq \widehat{\omega}} P(\omega_i \mid \vec{c}) \\ &= V_f \, (1 - P(\widehat{\omega} \mid \vec{c})) \end{aligned}$$

und bei einer Rückweisung:

$$\begin{aligned} R_{\vec{c}}(\widehat{\omega} = \omega_0) &= \sum_{\omega_i} V(\omega_i, \widehat{\omega}) P(\omega_i \mid \vec{c}) \\ &= V_r \sum_{\omega_i} P(\omega_i \mid \vec{c}) \\ &= V_r \end{aligned}$$

Setzt man $\vec{d}(\vec{c}) = \begin{pmatrix} \frac{V_f - V_r}{V_f} \\ P(\omega_1 \mid \vec{c}) \\ \vdots \\ P(\omega_i \mid \vec{c}) \\ \vdots \\ P(\omega_K \mid \vec{c}) \end{pmatrix}$,

so wird das Risiko minimiert, falls man folgende Entscheidungsregel anwendet:

$$g(\vec{c}) = \hat{\omega} = e(\vec{d}(\vec{c})) = \omega_l, \quad \text{falls } l \textit{ maximale} \text{ Komponente von } \vec{d}(\vec{c}) \tag{5.3}$$

5.3.3 Maximum Likelihood Klassifikator

Beim Bayes-Klassifikator werden seltene Klassen „benachteiligt“. Um dies zu vermeiden verändert man die Kostenfunktion:

$$V(\omega_i, \widehat{\omega}) = \begin{cases} 0, \ falls \ \widehat{\omega} = \omega_i \\ \frac{V_f}{P(\omega_i)}, \ falls \ \widehat{\omega} \neq \omega_i \wedge \widehat{w} \neq \omega_0 \\ V_r, \ falls \ \widehat{\omega} = \omega_0 \end{cases}$$

Setzt man dies nun in Gleichung (5.1) ein, so erhält man für das lokale Risiko bei einer Fehlentscheidung:

$$\begin{aligned}
R_{\vec{c}}(\widehat{\omega} \neq \omega_0) &= \sum_{\omega_i} V(\omega_i, \widehat{\omega}) P(\omega_i \mid \vec{c}) \\
&= 0 \cdot P(\widehat{\omega} \mid \vec{c}) + \sum_{\omega_i \neq \widehat{\omega}} \frac{V_f}{P(\omega_i)} P(\omega_i \mid \vec{c}) \\
&= V_f \sum_{\omega_i \neq \widehat{\omega}} \frac{1}{P(\omega_i)} \cdot \frac{P(\vec{c} \mid \omega_i) P(\omega_i)}{p\,(\vec{c})} \\
&= \frac{V_f}{P(\vec{c})} \left[\left(\sum_{\omega_i} P(\vec{c} \mid \omega_i) \right) - P(\vec{c} \mid \widehat{\omega}) \right]
\end{aligned}$$

und bei einer Rückweisung:

$$\begin{aligned} R_{\vec{c}}(\widehat{\omega} = \omega_0) &= \sum_{\omega_i} V(\omega_i, \widehat{\omega}) P(\omega_i \mid \vec{c}) \\ &= V_r \sum_{\omega_i} P(\omega_i \mid \vec{c}) \\ &= V_r \end{aligned}$$

Setzt man $\vec{d}(\vec{c}) = \begin{pmatrix} \sum_{\omega_i} P(\vec{c} \mid \omega_i) - \frac{V_r P(\vec{c})}{V_f} \\ P(\vec{c} \mid \omega_1) \\ \vdots \\ P(\vec{c} \mid \omega_i) \\ \vdots \\ P(\vec{c} \mid \omega_K) \end{pmatrix}$, so wird das Risiko minimiert, falls man folgende Entscheidungsregel anwendet:

$$g(\vec{c}) = \hat{\omega} = e(\vec{d}(\vec{c})) = \omega_l, \quad \text{falls } l \text{ \textit{maximale} Komponente von } \vec{d}(\vec{c}) \tag{5.4}$$

5.4 Quadratmittelansatz

5.4.1 Optimierungsansatz

- In Abschnitt 5.3 wurde
  - sowohl die Unterscheidungsfunktion $\vec{d}(\vec{c})$,
  - als auch die Entscheidungsregel $e(\vec{d}(\vec{c}))$

  optimiert.
- jetzt wird bei konstanter Entscheidungsregel $e(\vec{c})$ nur die Unterscheidungsfunktion $\vec{d}(\vec{c})$ optimiert.
  - Die feste Entscheidungsregel lautet:
    $\hat{\omega} = e(\vec{d}(\vec{c})) = \omega_l, \quad$ falls l ist maximale Komponente von $\vec{d}(\vec{c})$

- In Abschnitt 5.3 war die Optimierung ausgerichtet auf die Minimierung des Klassifikationsrisikos $E\{V\}$.
- jetzt wird der euklidsche Abstand
  - der Unterscheidungsfunktion $\vec{d}(\vec{c})$ und
  - dem zum Zielvektor $\vec{y}(\vec{c})$
    (erinnere: $\vec{y}(\vec{c})$ ist keine Funktion!)

  minimiert

  d.h. minimiere den mittleren quadratischen Fehler
  $S^2 = E\{(\vec{y}(\vec{c}) - \vec{d}(\vec{c}))^2\}$
- Die Unterscheidungsfunktion $\vec{d}(\vec{c})$ soll also dem Zielvektor $\vec{y}(\vec{c})$ möglichst ähnlich sein.

5.4.2 Lösung über Variationsrechnung

Unter Annahme, die optimale Lösung $\vec{d}(\vec{c})$ sei bekannt, verschlechtert sich das Optimierungskriterium S^2 durch jede Abweichung $\delta\vec{d}(\vec{c})$, das heißt, daß

$$S^2\left(\vec{d}\,(\vec{c}) + \delta\vec{d}\,(\vec{c})\right) \geq S^2\left(\vec{d}\,(\vec{c})\right) \quad \forall\delta \in \mathbb{R} \setminus \{0\} \tag{5.5}$$

(Im weiteren wird $\vec{d}$ anstelle von $\vec{d}\,(\vec{c})$ und $\vec{y}$ anstelle von $\vec{y}\,(\vec{c})$ geschrieben.)

Mit $S^2(\vec{d}) = E\left\{(\vec{y}-\vec{d})^2\right\} = E\left\{(\vec{y}-\vec{d})^T(\vec{y}-\vec{d})\right\}$ gilt:

$$\begin{aligned}
S^2(\vec{d}+\delta\vec{d}) &= E\left\{(\vec{y}-\vec{d}-\delta\vec{d})^T(\vec{y}-\vec{d}-\delta\vec{d})\right\} \\
&= E\left\{\vec{y}^T\vec{y} - \vec{y}^T\vec{d} - \vec{y}^T\delta\vec{d} - \vec{d}^T\vec{y} + \vec{d}^T\vec{d} + \vec{d}^T\delta\vec{d} - \delta\vec{d}^T\vec{y} + \delta\vec{d}^T\vec{d} + \delta\vec{d}^T\delta\vec{d}\right\} \\
&= \underbrace{E\left\{(\vec{y}-\vec{d})^2\right\}}_{S^2(\vec{d})} - 2E\left\{\delta\vec{d}^T(\vec{y}-\vec{d})\right\} + E\left\{(\delta\vec{d})^2\right\}
\end{aligned}$$

Setzt man nun die erhaltenen Werte von $S^2(d)$ und $S^2(d+\delta d)$ in die Ungleichung (5.5) ein, so ergibt sich

$$
\begin{aligned}
E\left\{(\vec{y}-\vec{d})^2\right\} - 2E\left\{\delta\vec{d}^T(\vec{y}-\vec{d})\right\} + E\left\{(\delta\vec{d})^2\right\} \;\geq\; E\left\{(\vec{y}-\vec{d})^2\right\} \quad &\Leftrightarrow \\
\underbrace{E\left\{(\delta\vec{d})^2\right\}}_{>0} - 2E\left\{\delta\vec{d}^T(\vec{y}-\vec{d})\right\} \;\geq\; 0 &
\end{aligned}
$$

Diese Ungleichung ist auf jeden Fall erfüllt, falls $E\left\{\delta\vec{d}^T(\vec{y}-\vec{d})\right\} = \vec{0}$ ist .

$$
\begin{aligned}
E\left\{\delta\vec{d}^T(\vec{y}-\vec{d})\right\} &= \int_{\vec{c}} \sum_{\vec{y}} \delta\vec{d}^T(\vec{y}-\vec{d}) \cdot P(\vec{c},\vec{y})\, d\vec{c} \\
&= \int_{\vec{c}} \sum_{\vec{y}} \delta\vec{d}^T(\vec{y}-\vec{d}) \cdot P(\vec{y} \mid \vec{c}) \cdot P(\vec{c})\, d\vec{c} \\
&= \int_{\vec{c}} \delta\vec{d}^T \left[\sum_{\vec{y}} (\vec{y}-\vec{d}) P(\vec{y} \mid \vec{c})\right] P(\vec{c})\, d\vec{c} \stackrel{!}{=} \vec{0}
\end{aligned}
$$

Dies ist nur dann für beliebige $\delta\vec{d}$ erfüllt, falls gilt:

$$\begin{aligned}
\sum_{\vec{y}} (\vec{y} - \vec{d}) P(\vec{y} \mid \vec{c}) &= \vec{0} \Leftrightarrow \\
\sum_{\vec{y}} (\vec{y} \cdot P(\vec{y} \mid \vec{c})) - \sum_{\vec{y}} (\vec{d} \cdot P(\vec{y} \mid \vec{c})) &= \vec{0} \Leftrightarrow \\
\sum_{\vec{y}} (\vec{y} \cdot P(\vec{y} \mid \vec{c})) - \vec{d} \underbrace{\sum_{\vec{y}} P(\vec{y} \mid \vec{c})}_{=1} &= \vec{0} \Leftrightarrow \\
\sum_{\vec{y}} (\vec{y} \cdot P(\vec{y} \mid \vec{c})) &= \vec{d} \Rightarrow
\end{aligned}$$

$$
\begin{aligned}
\vec{d}(\vec{c}) &= \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} P(\omega_1 \mid \vec{c}) + \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} P(\omega_2 \mid \vec{c}) + \ldots + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} P(\omega_K \mid \vec{c}) \\
&= \begin{pmatrix} P(\omega_1 \mid \vec{c}) \\ P(\omega_2 \mid \vec{c}) \\ \vdots \\ P(\omega_K \mid \vec{c}) \end{pmatrix}
\end{aligned}
$$

Die Optimierung des Quadratmittelansatzes entspricht also der des Bayes-Klassifikators ohne Rückweisung.

5.5 Zusammenfassung

Ansatz	optimiert	Klassifikationsfunktion
Risikominimierung mit beliebiger Kostenmatrix	U & E	$g(\vec{c}) = \hat{\omega} = e(\vec{d}(\vec{c})) = \omega_l$, falls l minimale Komponente von $\vec{d}(\vec{c}) = (R_{\vec{c}}(\omega_0), \ldots, R_{\vec{c}}(\omega_K))^T$ mit $R_{\vec{c}}(\omega_i) = \sum_{\omega_i} V(\omega_i, g(\vec{c})) \cdot P(\omega_i \mid \vec{c})$
Bayes-Klassifikator einfache Kostenmatrix: $V(\omega_i, \widehat{\omega}) = \begin{cases} 0, & \widehat{\omega} = \omega_i \\ V_f, & \widehat{\omega} \neq \omega_i \wedge \widehat{w} \neq \omega_0 \\ V_r, & \widehat{\omega} = \omega_0 \end{cases}$	U & E	$g(\vec{c}) = \hat{\omega} = e(\vec{d}(\vec{c})) = \omega_l$, falls l maximale Komponente von $\vec{d}(\vec{c}) = \left(\frac{V_f - V_r}{V_f}, P(\omega_1 \mid \vec{c}), \ldots, P(\omega_K \mid \vec{c})\right)^T$
Maximum-Likelihood-Klassifikator umgekehrt proport. Kostenmatrix: $V(\omega_i, \widehat{\omega}) = \begin{cases} 0, & \widehat{\omega} = \omega_i \\ \frac{V_f}{P(\omega_i)}, & \widehat{\omega} \neq \omega_i \wedge \widehat{w} \neq \omega_0 \\ V_r, & \widehat{\omega} = \omega_0 \end{cases}$	U & E	$g(\vec{c}) = \hat{\omega} = e(\vec{d}(\vec{c})) = \omega_l$, falls l maximale Komponente von $\vec{d}(\vec{c}) =$ $\left(\sum_{\omega_i} P(\vec{c} \mid \omega_i) - \frac{V_r P(\vec{c})}{V_f}, P(\vec{c} \mid \omega_1), \ldots, P(\vec{c} \mid \omega_K)\right)^T$
Quadratmittelansatz: quadr. Fehler zwischen Unterscheidungsfunktion und Zielvektor	**nur** U	analog Bayes-Klassifikator ohne Rückweisung

5.6 Klassifikation durch Schätzung der Wahrscheinlichkeitsverteilung

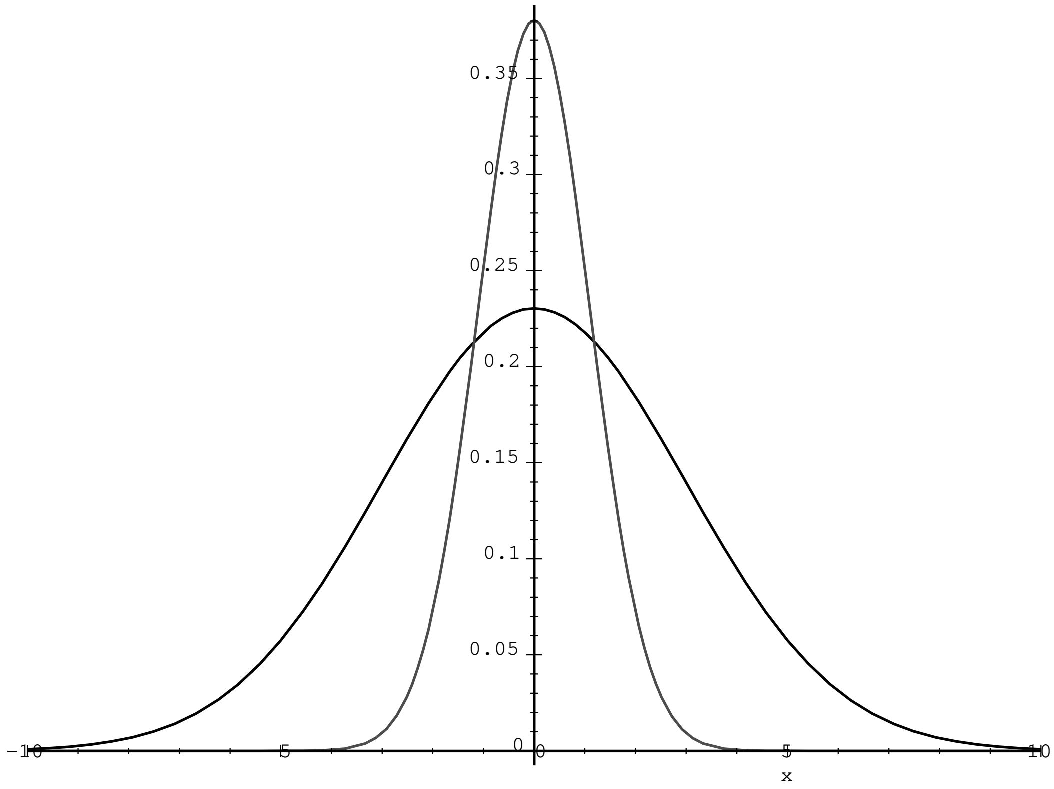
5.6.1 Normalverteilung

Um die Verteilung einer Zufallsvariable X zu modellieren, geht man oft davon aus, daß sie *normalverteilt* ist, d.h.

$$P(X = x) = \mathcal{N}_x(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

- $\mu = E\{X\}$ Erwartungswert von X
  bestimmt das Zentrum der Normalverteilung
- $\sigma^2 = E\left\{(x-\mu)^2\right\}$ Varianz
  bestimmt Breite und Höhe der Kurve
- Ist die Näherung durch eine einfache Normalverteilung nicht geeignet, so lassen sich in den allermeisten Fällen durch einfache Übelagerung mehrerer

Normalverteilungen gute Ergebnisse erzielen.

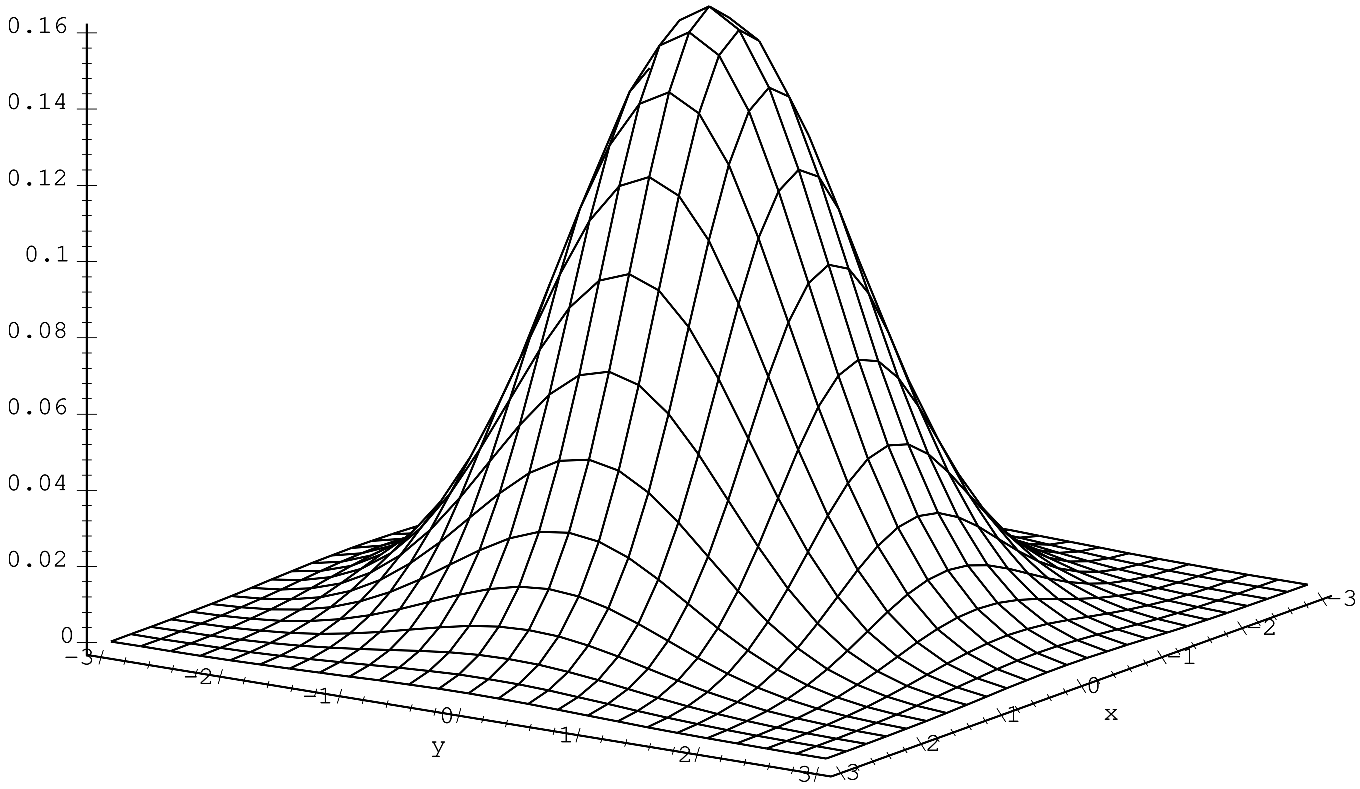


Beispiel zweier Normalverteilungen im $\mathbb{R}^1$

Ist die Zufallsvariable ein N-dimensionaler Vektor $\vec{c}$, so gilt

$$P(\vec{c}) = \mathcal{N}_{\vec{c}}(\vec{\mu}, \underline{K}) = \frac{1}{\sqrt{(2\pi)^N \text{det}\underline{K}}} e^{-\frac{1}{2}(\vec{c}-\vec{\mu})^T \underline{K}^{-1}(\vec{c}-\vec{\mu})}$$

- $\vec{\mu} = E\{\vec{c}\}$ Erwartungswert von $\vec{c}$
- $\underline{K} = E\left\{(\vec{c} - \vec{\mu})(\vec{c} - \vec{\mu})^T\right\}$ die Kovarianzmatrix.
- Zentrum der Normalverteilung ist — wie im $\mathbb{R}^1$ — durch den Erwartungswert $\vec{\mu}$ gegeben
- im $\mathbb{R}^2$ haben Normalverteilungen eine Glockenform
  horizontale Schnitte durch diese Glocke sind entweder kreis- oder ellipsenförmig



Beispiel einer zweidimensionalen Normalverteilung

5.6.2 Parameterschätzung

schätze Mittelwert $\vec{\mu}$ und Kovarianzmatrix $\underline{K}$ als ML-Schätzwert aus einer Stichprobe der Größe I:

$$\widehat{\vec{\mu}} = \frac{1}{I} \sum_{i=1}^{I} \vec{c}_i$$

$$\widehat{\underline{K}} = \frac{1}{I} \sum_{i=1}^{I} (\vec{c}_i - \widehat{\vec{\mu}})(\vec{c}_i - \widehat{\vec{\mu}})^T$$

Rekursive Berechnung

Es kommt häufig vor, daß die Menge der Trainingsmuster erweitert werden soll:

- neues klassifiziertes Trainingsmaterial vorhanden
- druchführen von entscheidungsüberwachtem Lernen

Da es für umfangreiche Stichproben sehr aufwendig ist, bei jeder Erweiterung die Parameter $\widehat{\vec{\mu}}$ und $\widehat{\underline{K}}$ vollständig neu zu berechnen, ist eine rekursive Formel für die

Schätzwerte von großem Vorteil.

Formel für die rekursive Berechnung des Erwartungswertes:

$$\begin{aligned}
\widehat{\vec{\mu}}_I &= \frac{1}{I}\sum_{i=1}^{I}\vec{c}_i \\
&= \frac{1}{I}\left(\sum_{i=1}^{I-1}\vec{c}_i\right) + \frac{1}{I}\vec{c}_I \\
&= \frac{1(I-1)}{I(I-1)}\left(\sum_{i=1}^{I-1}\vec{c}_i\right) + \frac{1}{I}\vec{c}_I \\
&= \frac{I-1}{I}\cdot\left(\frac{1}{I-1}\sum_{i=1}^{I-1}\vec{c}_i\right) + \frac{1}{I}\vec{c}_I \\
&= (1-\frac{1}{I})\cdot\widehat{\vec{\mu}}_{I-1} + \frac{1}{I}\vec{c}_I
\end{aligned}$$

Ähnlich läßt sich die Formel für die Kovarianzmatrix herleiten:

$$\underline{\widehat{K}}_I = (1-\frac{1}{I})[\underline{\widehat{K}}_{I-1} + \frac{1}{I}(\vec{c}_I - \widehat{\vec{\mu}}_{I-1})(\vec{c}_I - \widehat{\vec{\mu}}_{I-1})^T]$$

Sogar für die Inverse der Kovarianzmatrix gibt es eine inverse Formel:

$$\underline{\widehat{K}}_I^{-1} = \frac{I}{I-1}[\underline{\widehat{K}}_{I-1}^{-1} - \frac{1}{I}\frac{\underline{\widehat{K}}_{I-1}^{-1}(\vec{c}_I - \widehat{\vec{\mu}}_{I-1})(\vec{c}_I - \widehat{\vec{\mu}}_{I-1})^T \underline{\widehat{K}}_{I-1}^{-1}}{1 + \frac{1}{I}(\vec{c}_I - \widehat{\vec{\mu}}_{I-1})^T \underline{\widehat{K}}_{I-1}^{-1}(\vec{c}_I - \widehat{\vec{\mu}}_{I-1})}]$$

5.6.3 Bayes-Normalverteilungsklassifikator

Annahme: $P(\vec{c} \mid \omega_i) = \mathcal{N}_{\vec{c}}(\vec{\mu}_i, \underline{K_i})$

- schätze klassenspezifischen Mittelwerte $\vec{\mu}_l$ und Kovarianzmatrizen $\underline{K}_l$ aus klassifizierter Stichprobe
  für Schätzung von $\vec{\mu}_l$ und $\underline{K}_l$ verwende nur die Merkmalsvektoren $\vec{c}_i$, die aus der Klasse ω_l stammen
- schätze a priori Wahrscheinlichkeiten als relative Häufigkeit jeder Klasse in der Stichprobe
- für die Unterscheidungsfunktion des Bayes-Klassifikators benötigt man jedoch die Wahrscheinlichkeiten $P(\omega_i \mid \vec{c})$
  gemäß dem Gesetz von Bayes und Randdichten gilt jedoch:

$$P(\omega_i \mid \vec{c}) = \frac{P(\vec{c} \mid \omega_i)P(\omega_i)}{P(\vec{c})} = \frac{P(\vec{c} \mid \omega_i)P(\omega_i)}{\sum_{l=1}^{K} P(\vec{c}, \omega_l)} = \frac{P(\vec{c} \mid \omega_i)P(\omega_i)}{\sum_{l=1}^{K} P(\vec{c} \mid \omega_l)P(\omega_l)}$$

- verwendet man den Bayes-Klassifikator ohne Rückweisung, so reduziert sich der Rechenaufwand deutlich:

  $P(\vec{c})$ kann als Konstante vernachlässigt werden, damit ergibt sich:

  Die i-te Komponente $d_i(\vec{c})$ der Unterscheidungsfunktion $\vec{d}(\vec{c})$ hat dann die Form

$$d_i(\vec{c}) = p(\vec{c} \mid \omega_i) \cdot P(\omega_i) = \frac{P(\omega_i)}{\sqrt{(2\pi)^N \operatorname{det}(\underline{K}_i)}} e^{-\frac{1}{2}(\vec{c}-\vec{\mu}_i)^T \underline{K}_i^{-1}(\vec{c}-\vec{\mu}_i)}$$

- die Rangordnung der Ergebnisse der Unterscheidungsfunktion ändert sich nicht (bzw. dreht sich um), wenn
  - man sie mit positiven (negativen) Konstanten multipliziert
  - oder Konstanten addiert
  - oder Logiarithmiert

  damit:

$$d'_i(\vec{c}) \; = \; \underbrace{-2\ln(P(\omega_i)) + \ln(\operatorname{det}(\underline{K}_i))}_{\text{Konstante } b_{0i}} + \underbrace{(\vec{c}-\vec{\mu}_i)^T \underline{K}_i^{-1}(\vec{c}-\vec{\mu}_i)}_{\text{je Klass. zu berechnen}}$$

- Bayesklassifikator für normalverteilte Klassen ist also ein *quadratischer* Klassifikator (Aufwand $O(N^2)$ mit $N = \dim(\vec{c})$)
- mit der Unterscheidungsfunktion

$$\vec{d}(\vec{c}) = \begin{pmatrix} b_{01} + (\vec{c} - \vec{\mu}_1)^T \underline{K}_1^{-1} (\vec{c} - \vec{\mu}_1) \\ \vdots \\ b_{0i} + (\vec{c} - \vec{\mu}_i)^T \underline{K}_i^{-1} (\vec{c} - \vec{\mu}_i) \\ \vdots \\ b_{0K} + (\vec{c} - \vec{\mu}_K)^T \underline{K}_K^{-1} (\vec{c} - \vec{\mu}_K) \end{pmatrix}$$

  wird das Risiko mit folgender Entscheidungsregel minimiert:

$$g(\vec{c}) = \hat{\omega} = e(\vec{d}(\vec{c})) = \omega_l, \quad \text{falls } l \textit{ minimale} \text{ Komponente von } \vec{d}(\vec{c})$$

- Berechtigung der Annahme normalverteilter Klassen:
  - statistische Tests
  - Annahme anhand des Klassifikationsergebnisses auf einer Teststichprobe evaluieren
    (diese liefert aber in keinem Fall eine Aussage über die Korrektheit der Normalverteilugns-Annahme)

5.6.4 Vereinfachung durch identische Kovarianzmatrizen

Annahme: die Kovarianzmatrizen aller Klassen sind ω_i identisch

- dann vereinfacht sich $d_i'(\vec{c})$ weiter, da $\ln(\text{det}(\underline{K}))$ nun konstant:

$$\begin{aligned} d_i'(\vec{c}) &= -2\ln(P(\omega_i)) + (\vec{c} - \vec{\mu}_i)^T \underline{K}^{-1} (\vec{c} - \vec{\mu}_i) \\ &= -2\ln(P(\omega_i)) + \vec{c}^T \underline{K}^{-1} \vec{c} - 2\vec{\mu}_i^T \underline{K}^{-1} \vec{c} + \vec{\mu}_i^T \underline{K}^{-1} \vec{\mu}_i \end{aligned}$$

- Der Term $\vec{c}^T \underline{K}^{-1} \vec{c}$ kann ebenfalls vernachlässigt werden:

$$d_i'(\vec{c}) = \underbrace{-2\ln(P(\omega_i)) + \vec{\mu}_i^T \underline{K}^{-1} \vec{\mu}_i}_{b_{1i}} - \underbrace{2\vec{\mu}_i^T \underline{K}^{-1}}_{b_{2i}} \vec{c}$$

- mit der Unterscheidungsfunktion

$$\vec{d}(\vec{c}) = \begin{pmatrix} b_{11} - b_{21}\vec{c} \\ \vdots \\ b_{1i} - b_{2i}\vec{c} \\ \vdots \\ b_{1K} - b_{2K}\vec{c} \end{pmatrix}$$

  so wird das Risiko minimiert, falls man folgende Entscheidungsregel anwendet:

$$g(\vec{c}) = \hat{\omega} = e(\vec{d}(\vec{c})) = \omega_l, \quad \text{falls } l \textit{ minimale} \text{ Komponente von } \vec{d}(\vec{c})$$

- Bayesklassifikator für normalverteilte Klassen mit identischer Kovarianzmatrix ist also ein *linearer* Klassifikator (Aufwand $O(N)$ mit $N = \text{dim}(\vec{c})$)
  $\Rightarrow$ oft deutliche Rechenzeitersparnis, da Merkmalsvektoren mit 10 bis 100 Dimensionen nicht selten

5.6.5 Klassengrenzen

- ein Merkmalsvektor $\vec{c}$ liegt auf einer Klassengrenze, falls mehrere Komponenten der Unterscheidungsfunktion $\vec{d}(\vec{c})$ den gleichen (maximalen bzw. minimalen) Wert annehmen:

$$\begin{aligned} d_\kappa(\vec{c}) &= d_\lambda(\vec{c}) \Leftrightarrow \\ d_\kappa(\vec{c}) - d_\lambda(\vec{c}) &= 0 \end{aligned}$$

- Einsetzen für Bayesklassifikators mit Normalverteilungen ergibt:

$$d_\kappa(\vec{c}) - d_\lambda(\vec{c}) \; = \; b_{0\kappa} + (\vec{c} - \vec{\mu}_\kappa)^T \underline{K}_\kappa^{-1} (\vec{c} - \vec{\mu}_\kappa) - b_{0\lambda} - (\vec{c} - \vec{\mu}_\lambda)^T \underline{K}_\lambda^{-1} (\vec{c} - \vec{\mu}_\lambda) = 0$$

- läßt sich durch geeignete Transformationen zu einem einzigen quadratischen Ausdruck umformen

  $\Rightarrow$ alle Klassengrenzen sind Ellipsen (Ellipsoide), Parabel (Paraboloide) oder Hyperbeln (Hyperboloide)

- Bayesklassifikator für Normalverteilungen mit identischen Kovarianzmatrizen:

$$d_\kappa(\vec{c}) - d_\lambda(\vec{c}) \;=\; b_{1\kappa} - b_{2\kappa}\vec{c} - b_{1\lambda} + b_{2\lambda}\vec{c}$$

  diese Gleichung ist linear
  $\Rightarrow$ Klassengrenzen sind durch Geraden, Flächen oder Hyperflächen gegeben

5.6.6 Mischverteilungen

- Verteilung wird durch Linearkombination von verschiedenen (Normal)verteilungen beschrieben:

  $$P(\vec{c} \mid \omega_i) = \sum_{l=1}^{L} a_l \cdot \mathcal{N}_{\vec{c}}(\vec{\mu}_l, \underline{K}_l)$$

  wobei $\vec{\theta} = (a_1, \vec{\mu}_1, \underline{K}_1, \ldots, a_L, \vec{\mu}_L, \underline{K}_L)$
  die Mischverteilung (z.B. einer Klasse ω) beschreibt
- Schätzen im Prinzip wie k-means, wobei nun jedes Gebiet R_l bzw. jede l-te Normalverteilung (nicht **Klasse**)
  - nicht mehr alleine durch $\vec{\mu}_l$, sondern
  - durch Parameter $\vec{\mu}_l$ und $\underline{K}_l$ einer hochdimensionalen Normalverteilung $\mathcal{N}_{\vec{c}}(\vec{\mu}_l, \underline{K}_l)$ bestimmt ist,
  - $P(\vec{c} \mid R_l) = \mathcal{N}_{\vec{c}}(\vec{\mu}_l, \underline{K}_l)$
  - $P(R_l) = a_l$

Harte Vektorquantisierung

- ordne jeden Merkmalsvektor $\vec{c}_i$ einem Gebiet **hart** zu
- Ziel: durchschnittlichen Likelihood-Wert H maximieren bezüglich der Parameter $\vec{\theta}$ (In um Konvergenz zu beweisen)

$$
\begin{aligned}
H \;=\; & \frac{1}{I} \ln P(\{\vec{c}_1, \ldots, \vec{c}_I\}|\vec{\theta}) = \frac{1}{I} \ln \prod_{i=1}^{I} P(\vec{c}_i|\vec{\theta}) = \frac{1}{I} \sum_{i=1}^{I} \ln P(\vec{c}_i|\vec{\theta}) = \\
& \frac{1}{I} \sum_{i=1}^{I} \ln \max_{R_l} P(\vec{c}_i, R_l|\vec{\theta}) = \frac{1}{I} \sum_{l=1}^{L} \sum_{\vec{c}_i \in R_l} \ln(P(R_l)P(\vec{c}_i \mid R_l))
\end{aligned}
$$

$$\text{mit } \vec{c}_i \in R_l, \;\; \text{falls } p\,(\vec{c}_i \mid R_l) = \max_{k=1,\ldots,L} p\,(\vec{c}_i \mid R_k)$$

- Verfahren zur Optimierung von L Gebieten wähle aufgrund von Vorwissen oder zufällig initiale Parameter $P(R_l), \vec{\mu}_l, \underline{K}_l$ (z.B. $P(R_l) = 1/L$, $\vec{\mu}_l = l$-ter Vektor der Stichprobe, $\underline{K}_l = \mathcal{I}$

$H^0 := -\infty$ (Wert der Likelihood-Funktion in der Iteration 0)
$t := 0$ (Iterationszähler)

- $t := t + 1, \quad H^{(t)} := 0$
- FOR alle Gebiete $R_l, l = 1, \ldots, L$
  - $I_l := 0; \quad \hat{\vec{\mu}}_l := \vec{0}, \quad \underline{\hat{M}}_l := O$
- FOR alle Vektoren $\vec{c}_i$ der Stichprobe
  - bestimme R_l mit maximaler Dichte $P(\vec{c}_i \mid R_l) = \mathcal{N}_{\vec{c}}(\vec{\mu}_l, \underline{K}_l)$
  - $H^{(t)} := H^{(t)} + \ln\left(P\left(R_l\right) p\left(\vec{c} \mid R_l\right)\right)$
  - berechne neue Schätzwerte für den Mittelwert und die Momentenmatrix, d.h.
    $\hat{\vec{\mu}}_l := \hat{\vec{\mu}}_l + \vec{c}_i \qquad \underline{\hat{M}}_l := \underline{\hat{M}}_l + \vec{c}_i\vec{c}_i^T$
  - $I_l := I_l + 1$
- $H^{(t)} := H^{(t)}/I$
- FOR alle Gebiete $R_l, l = 1, \ldots, L$
  - $P(R_l) = \frac{I_l}{I}, \quad \vec{\mu}_l := \hat{\vec{\mu}}_l/I_l; \quad \underline{K}_l := \underline{\hat{M}}_l/I_l - \vec{\mu}_l\vec{\mu}_l^T$

UNTIL $(H^{(t)} - H^{(t-1)})/ \mid H^(t) \mid \leq \varepsilon$

Weiche Vektorquantisierung

- ordne jeden Merkmalsvektor $\vec{c}_i$ mit dem Gewicht seiner a-posteriori-Wahrscheinlichkeit $P\left(R_l \mid \vec{c}_i\right)$ **allen** Gebieten R_l **weich** zu (beachte: $\sum_{l=1}^{L} P(R_l \mid \vec{c}_i) = 1$)
- Ziel: Maximierung (der logarithmierten) Wahrscheinlichkeit H der Produktion der Stichprobe in Abhängigkeit der Mischverteilung (normiert bzgl. der Stichprobengröße I):

$$
\begin{aligned}
H \;=\; & \frac{1}{I} \ln P(\{\vec{c}_1, \ldots, \vec{c}_I\} | \vec{\theta}) = \frac{1}{I} \ln \prod_{i+1}^{I} P(\vec{c}_i | \vec{\theta}) = \frac{1}{I} \sum_{i=1}^{I} \ln P(\vec{c}_i | \vec{\theta}) = \\
& \frac{1}{I} \sum_{i=1}^{I} \ln \sum_{l=1}^{L} P\left(R_l\right) p\left(\vec{c}_i \mid R_l\right)
\end{aligned}
$$

- zur Berechnung von $P(R_l \mid \vec{c}_i)$ folgende Umformungen (Bayes-Gesetzes):

$$
\begin{aligned}
P\left(R_l \mid \vec{c}_i\right) &= \frac{P(R_l)p\left(\vec{c}_i \mid R_l\right)}{p\left(\vec{c}_i\right)} \\
p\left(\vec{c}_i\right) &= \sum_{k=1}^{L} p\left(\vec{c}_i, R_k\right) \\
&= \sum_{k=1}^{L} P(R_k)p\left(\vec{c}_i \mid R_k\right) \\
\Rightarrow P\left(R_l \mid \vec{c}_i\right) &= \frac{P(R_l)p\left(\vec{c}_i \mid R_l\right)}{\sum_{k=1}^{L} P(R_k)p\left(\vec{c}_i \mid R_k\right)}
\end{aligned}
$$

- Verfahren zur Optimierung von L Gebieten

wähle aufgrund von Vorwissen oder zufällig initiale Parameter $P(R_l), \vec{\mu}_l, \underline{K}_l$
(z.B. $P(R_l) = 1/L$, $\vec{\mu}_l = l$-ter Vektor der Stichprobe, $\underline{K}_l = \mathcal{I}$

$H^0 := -\infty$ (Wert der Likelihood-Funktion in der Iteration 0)
$t := 0$ (Iterationszähler)

- $t := t + 1, \quad H^{(t)} := 0$
- FOR alle Gebiete $R_l, l = 1, \ldots, L$
  - $I_l := 0; \quad \hat{\vec{\mu}}_l := \vec{0}, \quad \underline{\hat{M}}_l := O$
- FOR alle Vektoren $\vec{c}_i$ der Stichprobe
  - berechne $P(R_l \mid \vec{c}_i)$ mittels $P(\vec{c} \mid R_l) = \mathcal{N}_{\vec{c}}(\vec{\mu}_l, \underline{K}_l)$ für alle Gebiete R_l
  - $H^{(t)} := H^{(t)} + \ln(\sum_{l=1}^{L} P\left(R_l\right) p\left(\vec{c} \mid R_l\right))$
  - FOR alle Gebiete $R_l, l = 1, \ldots, L$
    - berechne neue Schätzwerte für den Mittelwert und die Momentenmatrix, d.h.
      $\hat{\vec{\mu}}_l := \hat{\vec{\mu}}_l + \vec{c}_i \cdot P(R_l \mid \vec{c}_i) \qquad \underline{\hat{M}}_l := \underline{\hat{M}}_l + \vec{c}_i\vec{c}_i^{\,T} \cdot P(R_l \mid \vec{c}_i)$
    - $I_l := I_l + P(R_l \mid \vec{c}_i)$
- $H^{(t)} := H^{(t)}/I$
- FOR alle Gebiete $R_l, l = 1, \ldots, L$
  - $P(R_l) = \frac{I_l}{I}, \quad \vec{\mu}_l := \hat{\vec{\mu}}_l/I_l; \quad \underline{K}_l := \underline{\hat{M}}_l/I_l - \vec{\mu}_l\vec{\mu}_l^{\,T}$

UNTIL $(H^{(t)} - H^{(t-1)})/ \mid H^{(t)} \mid \leq \varepsilon$

LBG-Algorithmus für Dichten

Anzahl der Gebiete ist $L = 1$, wobei das Gebiet R_1 durch
$P(R_1) = 1.0, \vec{\mu}_1 = \frac{1}{I}\sum_{i=1}^{I}\vec{c}_i$, $\underline{K}_1 = \frac{1}{I}\sum_{i=1}^{I}(\vec{c}_i - \vec{\mu}_1)(\vec{c}_i - \vec{\mu}_1)^T$ repräsentiert ist

- FOR alle Gebiete R_l (oder für die M Gebiete mit kleinstem Likelihood-Wert)
  - zerlege R_l in 2 Normalverteilungen $\mathcal{N}_{\vec{c}}(\vec{\mu}_l + \vec{\delta}, \underline{K}_l)$ und $\mathcal{N}_{\vec{c}}(\vec{\mu}_l - \vec{\delta}, \underline{K}_l)$ mit den jeweiligen a priori Wahrscheinlichkeiten $P(R_l)/2$
- $L := L * 2 \quad (L := L + M)$
- wende den Algorithmus zur harten oder weichen Vektorquantisierung von L Gebieten an

UNTIL Likelihood-Wert $H^{(t)} \geq \varepsilon$ ODER $L \geq$ maximale Klassenanzahl

- $\vec{\delta}$ ist zufälliger betragsmäßig kleiner Vektor
- $\vec{\delta}$ betragsmäßig kleiner Vektor in Richtung der Hauptachse des Verteilungsgebietes, d.h. $\vec{\delta}$ zeigt in Richtung des Eigenvektors zum größten Eigenwert von $\underline{K}_l$

Klassenabhängige Dichten

- zerlege **klassifizierte** Stichprobe in K Stichproben $S_k,\ k = 1, \ldots, K$, die jeweils nur Vektoren aus der Klasse ω_k enthalten
- wende gesondert auf jede dieser Stichproben den LBG-Algorithmus für Dichten an. Man erhält L^k Dichten mit den Parametern $\mathcal{N}\left(\vec{\mu}_l^k, \underline{K}_l^k\right)$
- bestimme aus der klassifizierten Stichprobe die Gewichte wie folgt:

$$a_l^k = \frac{1}{I^k} \sum_{\vec{c}_i \in \omega_k \wedge \vec{c}_i \in R_l^k} 1 \qquad \text{(harte VQ)}$$

$$a_l^k = \frac{1}{I^k} \sum_{\vec{c}_i \in \omega_k} P\left(R_l^k \mid \vec{c}_i\right) \qquad \text{(weiche VQ)}$$

- die Komponenten der Unterscheidungsfunktion ergeben sich nun zu:

$$P(\vec{c} \mid \omega_k) = \sum_{l=1}^{L^k} a_l^k \cdot \mathcal{N}\left(\vec{\mu}_l^k, \underline{K}_l^k\right)$$

Klassenunabhängige Dichten

- wende auf einer **unklassifizierten** Stichprobe den LBG-Algorithmus für Dichten an (weiche oder harte VQ) $\Rightarrow$ man erhält L Normalverteilungen $\mathcal{N}\left(\vec{\mu}_l, \underline{K}_l\right)$
- bestimme die Gewichtsparameter der obigen Dichten für die Klasse ω_k aus einer (evtl. kleineren) **klassifizierten** Stichprobe wie folgt:

$$a_l^k = \frac{1}{I^k} \sum_{\vec{c}_i \in \omega_k \wedge \vec{c}_i \in R_l} 1 \qquad \text{(harte VQ)}$$

$$a_l^k = \frac{1}{I^k} \sum_{\vec{c}_i \in \omega_k} P\left(R_l \mid \vec{c}_i\right) \qquad \text{(weiche VQ)}$$

- $$P(\vec{c} \mid \omega_k) = \sum_{l=1}^{L} a_l^k \mathcal{N}(\vec{\mu}_l, \underline{K}_l)$$
- dieses Vorgehen ist u.U. günstiger, das sich mehrere Klassen eine Normalverteilung “teilen” können

5.6.7 Andere Verteilungen

Statistische Unabhängigkeit

- **Annahme** statistischer Unabhängigkeit der Merkmale
  (die allerdings in der Regel nicht, höchstens approximativ gegeben ist)
- $$P(\vec{c} \mid \omega_k) = \prod_{\nu=1}^{N} P(c_\nu \mid \omega_k)$$
- die $P(c_\nu \mid \omega_k)$ können
  - mit eindimensionalen parametrischen Dichte geschätzt werden
    (größere Auswahl als im höherdimensionalen Fall)
  - nach Diskretisierung der Werte c_ν:
    durch Histogramm als relative Häufigkeiten tabelliert werden

Parzenschätzung $\sim$ radiale Basisfunktionen

- **Motivation:** für höhere Dimensionen ist Diskretisierung des Merkmalsraums bei relativ wenigen Stichprobenelementen schwierig
- $\Rightarrow$ "verschmiere" jeden Beitrag der Stichprobe $\Rightarrow$ um seine Position
- Approximation der Dichte durch Überlagerung von Normalverteilungen, (oder auch andere Fensterfunktionen, z.B. Rechteck)

$$P(\vec{c} \mid \omega_k) = \frac{1}{I} \sum_{i=1}^{I} \mathcal{N}\left(\vec{c}_i, K\right)$$

wobei $K = \sigma I$, und σ die Breite des Fensters (des "Verschmierens") bestimmt

5.7 Alternative Klassifikationsergebnisse

bis jetzt: genau **eine** Klasse oder Rückweisung

Erweiterung — weniger endgültige Entscheidung des Klassifikators,
um nachfolgenden Prozesss die Entscheidung zu überlassen
(wobei dieser weitere Evidenzen eingehen lassen kann)
$\Rightarrow$ *principle of least commitment*

Idee
- ordne die Klassen ω_i gemäß abfallendem $P(\omega_i \mid \vec{c})$
- wähle **Konfidenzschwelle** $\theta \in \left[\frac{1}{K}; 1\right]$
- Ergebniss der Klassifikation sind die ersten Klassen, sodass bei minimaler Anzahl an Klassen gilt:

$$\sum_i P(\omega_i \mid \vec{c}) \geq \theta$$

Bemerkungen durch die Wahl von θ kann die Anzahl an Alternativen gewählt werden:

- $\theta = \frac{1}{K}$: nur beste Klasse, wie bisher
- $\theta = 1$: alle Klaasen