6 Klassifikation

6.1 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

$$g : C \mapsto \Omega$$

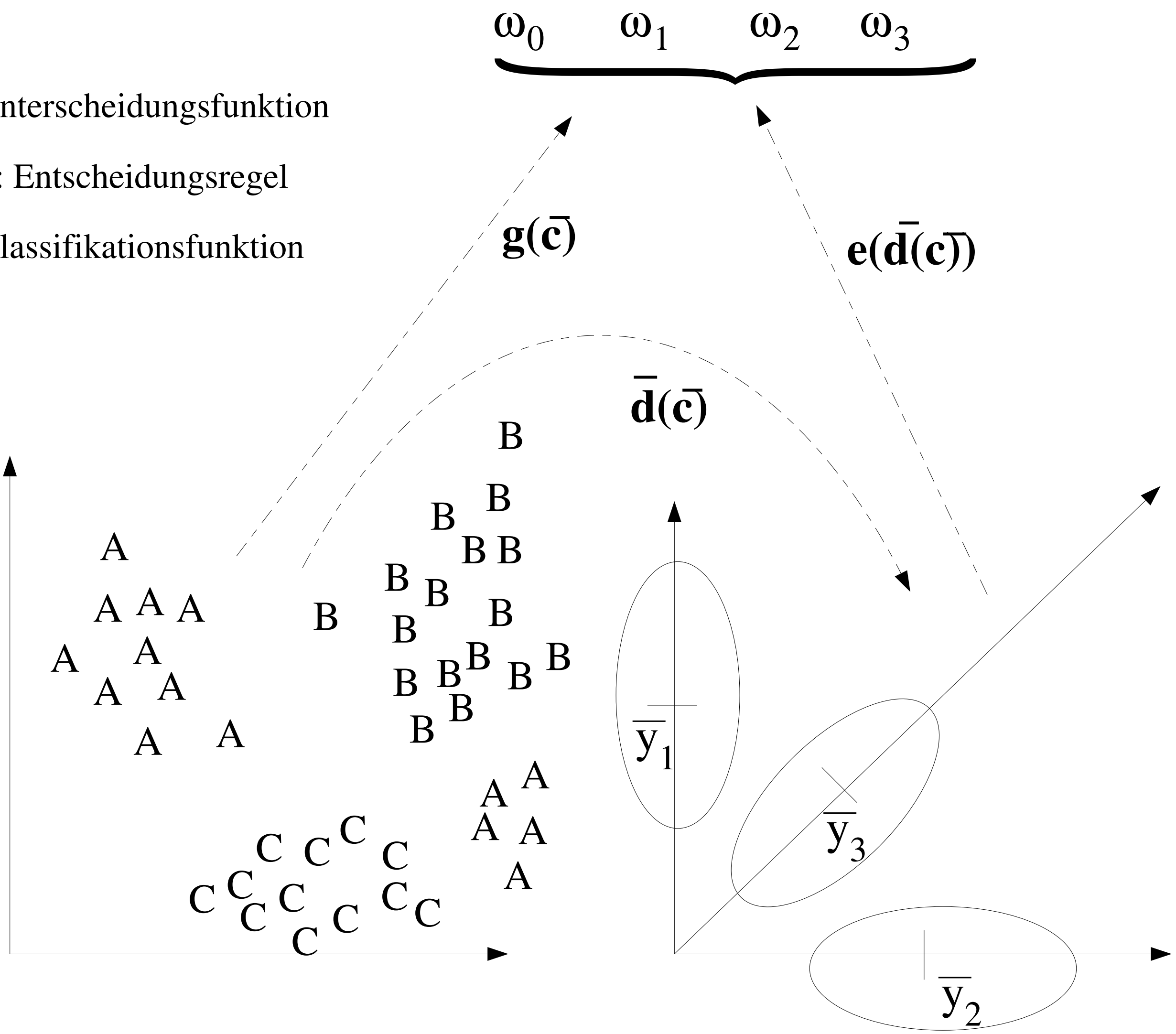
Wegen der einfacheren mathematischen Handhabbarkeit wird diese Funktion meist in zwei Abbildungen zerlegt:

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

$\bar{d}(\bar{c})$: Unterscheidungsfunktion

$e(\bar{d}(\bar{c}))$: Entscheidungsregel

$g(\bar{c})$: Klassifikationsfunktion



hochdimensionaler Merkmalsraum
$C = \mathbb{R}^N$

Unterscheidungsraum
$D = \mathbb{R}^K$ bzw. $\mathbb{R}^{K+1}$

$\mathbb{R}^K$ 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** $\vec{y}_i, i = 1, \ldots, K$ sind also Repräsentanten für ω_i, wobei gilt:

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

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>)

6.2 Statistische Grundlagen

6.2.1 Mustererzeugende Prozesse

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

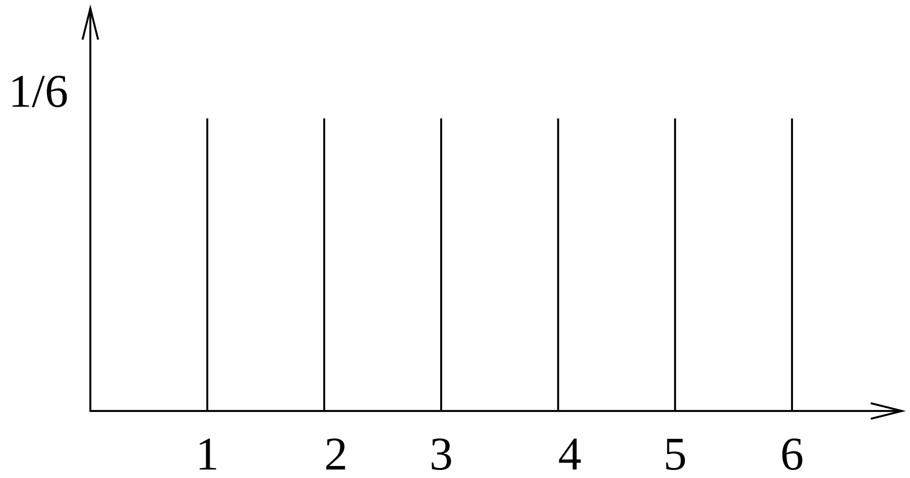
Voraussetzungen für die automatische Klassifikation:

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.

6.2.2 Wahrscheinlichkeiten und Wahrscheinlichkeitsdichten

Einer diskreten Zufallsvariable X wird die Wahrscheinlichkeiten $P(X = x)$ zugeordnet.

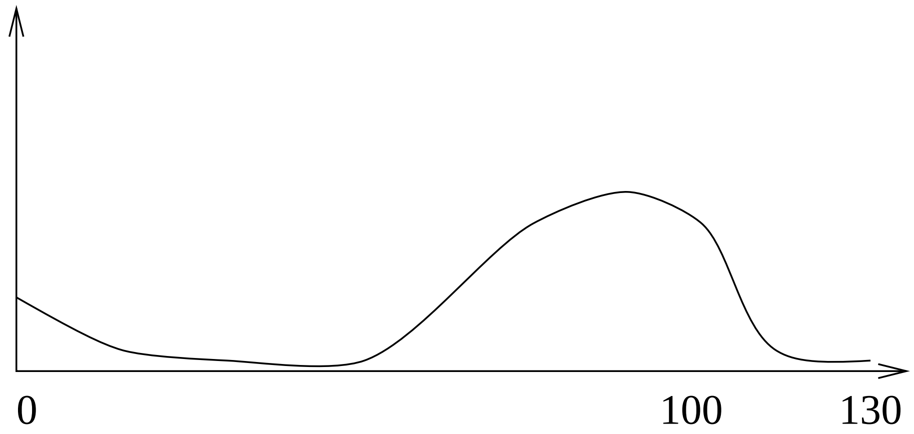
Beispiel: Augenzahl eines Würfels $P(1) = P(2) = \ldots = P(6) = \frac{1}{6}$



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

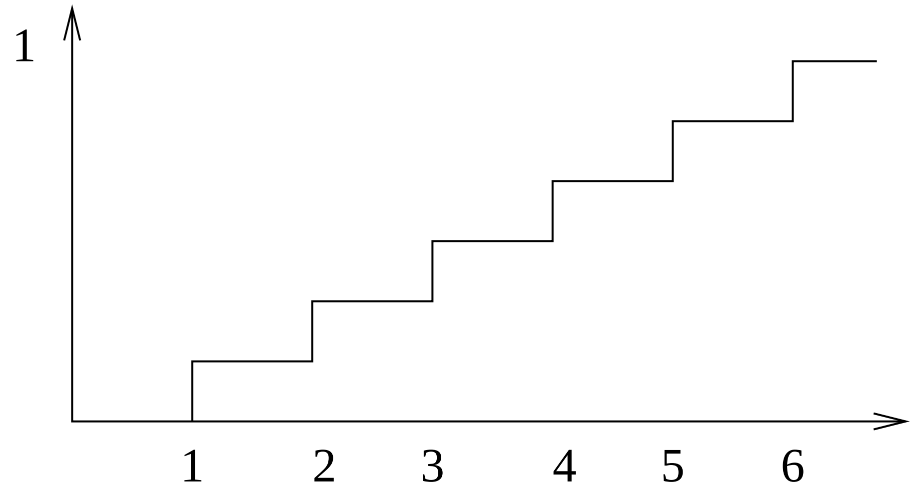
Beispiel: Lebensalter von Menschen

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

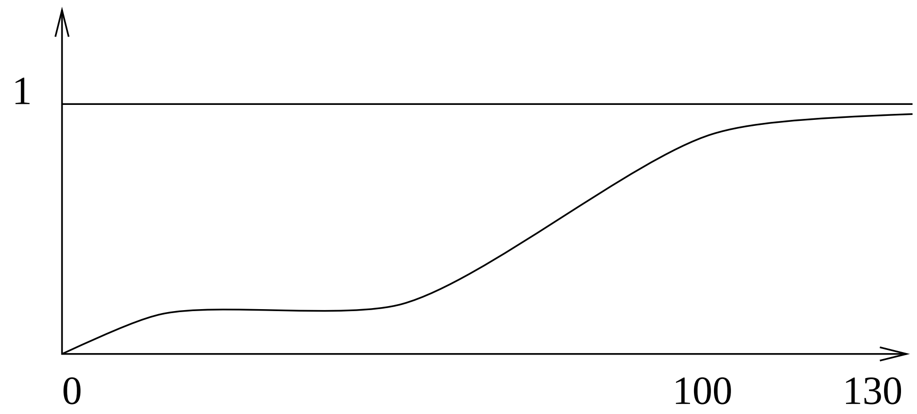


Die Funktion $F(x) = P(X \leq x)$ heißt Verteilungsfunktion.

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 *Rückschluß*wahrscheinlichkeit: $P(\omega_i \mid \vec{c})$

Nach dem **Gesetz von Bayes** gilt:

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

6.3 Minimierung des Klassifikationsrisikos

6.3.1 Allgemeiner Ansatz

- für Klassifikationssysteme (KS) ist nur $\vec{c}$, nicht aber ω_i sichtbar
- wir benötigen also eine Klassifikationsfunktion $g(\vec{c})$ 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 $g(\vec{c})$ nicht der tatsächlich vorliegenden Klasse ω_{soll} entspricht
- eine solche (Fehl-)Klassifikation verursacht Kosten:

  Um diese Kosten näher zu bestimmen, stellt man eine Verlustmatrix $V(\omega_{\text{soll}}, \omega_{\text{ist}})$ auf.

  Hierbei ist $V(\omega_{\text{soll}}, \omega_{\text{ist}})$ der Verlust, der entsteht, wenn man sich für ω_{ist} entscheidet, obwohl ω_{soll} 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 $R = E\{V\} = E\{V(\omega, g(\vec{c}))\}$.
- Wie berechnet man nun den Erwartungwert einer Zufallsvariable? Hierzu zwei Beispiele:

  1. Quadrat der Augenzahlen eines Würfels:

     $$E\{X^2\} = 1^2 \cdot P(1) + 2^2 \cdot P(2) + \ldots + 6^2 \cdot P(6) = \sum_{i=1}^{6} i^2 \cdot P(i) = 15.1\bar{6}$$

  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.

$$g(\vec{c}) \in \{\omega_0, \omega_1, \ldots, \omega_K\}.$$

Es wird nun also das Minimum des lokalen Risikos

$$R_{\vec{c}}(g(\vec{c})) = \sum_{\omega_i} V(\omega_i, g(\vec{c})) \cdot P(\omega_i \mid \vec{c}) \tag{6.1}$$

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:

$$g(\vec{c}) = e(\vec{d}(\vec{c})) = \omega_l, \quad \text{falls } l \textit{ minimale} \text{ Komponente von } \vec{d}(\vec{c}) \tag{6.2}$$

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

Folgende Werte werden dazu benötigt:

- $V(\omega_{\text{soll}}, \omega_{\text{ist}})$: 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:

$$\hat{P}(\omega_i \mid \vec{c}) \propto \hat{P}(\vec{c} \mid \omega_i) \cdot \hat{P}(\omega_i)$$

6.3.2 Bayes-Klassifikator

Der Bayes-Klassifikator hat eine spezielle, symmetrische Kostenfunktion:

$$V(\omega_{\text{soll}}, \omega_{\text{ist}}) = \begin{cases} 0, & \text{falls } \omega_{\text{ist}} = \omega_{\text{soll}} \\ V_f, & \text{falls } \omega_{\text{ist}} \neq \omega_{\text{soll}} \wedge \omega_{\text{ist}} \neq \omega_0 \\ V_r, & \text{falls } \omega_{\text{ist}} = \omega_0 \end{cases}$$

Setzt man dies nun in Gleichung (6.1) ein, so erhält man für das lokale Risiko bei einer Fehlentscheidung:

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

und bei einer Rückweisung:

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

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:

$$g(\vec{c}) = e(\vec{d}(\vec{c})) = \omega_l, \quad \text{falls } l \text{ \textit{maximale} Komponente von } \vec{d}(\vec{c}) \tag{6.3}$$

6.3.3 Maximum Likelihood Klassifikator

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

$$V(\omega_i, \omega_{\text{ist}}) = \begin{cases} 0, \ falls \ \omega_{\text{ist}} = \omega_i \\ \frac{V_f}{P(\omega_i)}, \ falls \ \omega_{\text{ist}} \neq \omega_i \wedge \omega_{\text{ist}} \neq \omega_0 \\ V_r, \ falls \ \omega_{\text{ist}} = \omega_0 \end{cases}$$

Setzt man dies nun in Gleichung (6.1) ein, so erhält man für das lokale Risiko bei einer Fehlentscheidung:

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

und bei einer Rückweisung:

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

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:

$$g(\vec{c}) = e(\vec{d}(\vec{c})) = \omega_l, \quad \text{falls } l \text{ \textit{maximale} Komponente von } \vec{d}(\vec{c}) \tag{6.4}$$

6.4 Quadratmittelansatz

6.4.1 Optimierungsansatz

- In Abschnitt 6.3 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:
    $g(\vec{c}) = e(\vec{d}(\vec{c})) = \omega_l, \quad$ falls l ist maximale Komponente von $\vec{d}(\vec{c})$

- In Abschnitt 6.3 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.

6.4.2 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

$$S^2\left(\vec{d}\,(\vec{c}) + \delta\vec{d}\,(\vec{c})\right) \geq S^2\left(\vec{d}\,(\vec{c})\right) \quad \forall\delta \in \mathbb{R} \setminus \{0\} \tag{6.5}$$

(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 (6.5) 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.

6.5 Zusammenfassung

Ansatz	optimiert	Klassifikationsfunktion
Risikominimierung mit beliebiger Kostenmatrix	U & E	$g(\vec{c}) = \hat{\omega} = e(\vec{d}(\vec{c})) = \omega_l$, falls l min. Komp. 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: $V(\omega_i, \omega_{\text{ist}}) = \begin{cases} 0, & \omega_{\text{ist}} = \omega_i \\ V_f, & \omega_{\text{ist}} \neq \omega_i \wedge \omega_{\text{ist}} \neq \omega_0 \\ V_r, & \omega_{\text{ist}} = \omega_0 \end{cases}$	U & E	$g(\vec{c}) = \hat{\omega} = e(\vec{d}(\vec{c})) = \omega_l$, falls l max. Komp. 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: $V(\omega_i, \omega_{\text{ist}}) = \begin{cases} 0, & \omega_{\text{ist}} = \omega_i \\ \frac{V_f}{P(\omega_i)}, & \omega_{\text{ist}} \neq \omega_i \wedge \omega_{\text{ist}} \neq \omega_0 \\ V_r, & \omega_{\text{ist}} = \omega_0 \end{cases}$	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