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

$$g : C \mapsto \Omega \quad |\Omega| = K$$

Oft ist es sinnvoll, die gwwünschte Klasse ω_i als **Zielvektor** $\vec{y}_i, i = 1, \ldots, K$, zu repräsentieren

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

Gehört ein Merkmalsvektor $\vec{c}$ zur Klasse ω_i, so wird der zugehörige Zielvektor mit $\vec{y}$ bezeichnet.

In der Stichprobe haben wir N Muster(paare) $(\vec{c}_n, \omega_n)$ bzw. $(\vec{c}_n, \vec{y}_n)$

6.2 Statistische Grundlagen

6.2.1 Mustererzeugende Prozesse

Muster sind Wertepaare $(\vec{c}, \omega)$, die

- die Erscheinungsform, repräsentiert durch den Merkmalsvektor $\vec{c} \in \mathcal{R}^M$
- mit der Bedeutung des Musters, repräsentiert durch die Klasse $\omega \in \Omega$

verbinden.

Oft bezeichnen wir auch $\vec{c}$ alleine als das Muster, die Klasse ist unbekannt.

Der mustererzeugende Prozeß (MEP) wird als stochastischer Prozeß modelliert, der mit der Wahrscheinlichkeit $P(\vec{c}, \omega)$ 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 $P(X = 1) = P(X = 2) = \ldots = P(X = 6) = \frac{1}{6}$
- die Klassenzugehörigkeit

Eine kontinuierlichen Zufallsvariable X wird durch Wahrscheinlichkeitsdichte $P(X)$ beschrieben.

Beispiel: Lebensalter von Menschen

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

Die Funktion $F_X(x) = P(X \leq x)$ heißt **Verteilungsfunktion** der ZV X.

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

Beispiel Lebensalter: $F_X(67,5) = P(X \leq 67,5) = \int\limits_0^{67{,}5} P(x)dx$

Der **Erwartungswert** einer Zufallsvariablen X oder einer Funktion $f(X)$ der Zufallsvariablen ist definiert als:

$$E_{P(X)}[f(X)] \;=\; \sum_x P(X=x)f(x) \quad \text{bzw.}$$

$$E_{P(X)}[f(X)] \;=\; \int_x P(x)f(x)\,dx$$

Die **Varianz** gibt die erwartete quadratische Abweichung vom Mittelwert

$$Var[f(X)] = E[(f(X) - E[f(X)])^2] = E[(f(X))^2] - (E[f(X)])^2$$

Die **Kovarianz** eines Vectors $\vec{X}$ von Zufallsvariablen ist

$$Cov[\vec{X}] = E[(\vec{X} - E[\vec{X}])(\vec{X} - E[\vec{X}])^T]$$

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

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

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

Bedingte Wahrscheinlichkeit und *bedingte* Dichte:

- *Klassenspezifische Dichte, likelihood*: $P(\vec{c} \mid \omega)$
- *A posteriori*-Wahrscheinlichkeit oder *Rückschluß*wahrscheinlichkeit: $P(\omega \mid \vec{c})$

Nach dem **Gesetz von Bayes** gilt:

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

6.3 Minimierung des Klassifikationsrisikos

6.3.1 Allgemeiner Ansatz

- für Klassifikationssysteme (KS) ist nur $\vec{c}$, nicht aber ω 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.

- Ziel ist es, den durchschnittlichen Verlust (das *Risiko*) zu minimieren, d.h.

  minimiere $R = E[V] = E[V(\omega, g(\vec{c}))]$.

- Zur Bestimmiung von $g(\vec{c})$ setzen wir eine klassifizierte Stichprobe ein.

Zunächst aber Berechung des Risikos unter der Annahme, wir kennen die Verteilung des MEP.

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

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

Da wir $g(\vec{c})$ für jedes $\vec{c}$ unabhängig wählen dürfen (wir haben keine Einschränkungen bezüglich der Funktion g gemacht), klappt das in diesem Fall.

Wir betrachten den allgmeinen Fall, dass eine Rückweisung zugelassen ist, d.h.

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

Es wird nun also das Minimum des lokalen Risikos

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

gesucht.

Folgende Werte werden dazu benötigt:

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

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

(mehr dazu später)

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 (beachte: $\omega_{\text{ist}} = g(\vec{c})$):

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

und bei einer Rückweisung:

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

durch Umformung erhalten wir äquivalent: $\vec{d}(\vec{c}) = \begin{pmatrix} \frac{V_f - V_r}{V_f} \\ P(\omega_1 \mid \vec{c}) \\ \vdots \\ P(\omega_i \mid \vec{c}) \\ \vdots \\ P(\omega_K \mid \vec{c}) \end{pmatrix}$,

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

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

6.3.3 Maximum Likelihood Klassifikator

Beim Bayes-Klassifikator werden seltene Klassen „benachteiligt“.

$$P(\omega|\vec{c}) = \frac{P(\vec{c}|\omega)P(\omega)}{P(\vec{c})} \propto P(\vec{c}|\omega)P(\omega)$$

Um dies zu vermeiden verändert man die Kostenfunktion:

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

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

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

und bei einer Rückweisung:

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

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

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

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

6.4 Verteilungsfreie Klassifikatoren

Wir wollen jetzt die Klassifikation als Funktionsapproxmation betrachten.

Gesucht ist eine Funktion $\vec{d}(\vec{c})$, eine sogenannte **Diskriminatorfunktion**, die jeden Merkmalsvektor eindeutig einer Klasse zuordnet:

- im allgemeinen Fall soll der Zielvektor $\vec{y}$ approximiert werden
- im 2-Klassenfall oft stattdessen 0 für ω_1 und 1 für ω_2

6.4.1 Quadratmittelansatz

- In Abschnitt 6.3 war die Optimierung ausgerichtet auf die Minimierung des Klassifikationsrisikos $E\{V\}$, und wir haben die statistischen Eigenschaften des MEP explizit betrachtet.

- jetzt wird der euklidsche Abstand
  - der Diskriminatorfunktion $\vec{d}(\vec{c})$ und
  - dem zum Zielvektor $\vec{y}$

  minimiert
- genauer: minimiere den erwarteten quadratischen Fehler

$$S^2 = E_{P(\vec{c},\vec{y})}[(\vec{y} - \vec{d}(\vec{c}))^2]$$

  für Muster(paare) $(\vec{c}, \vec{y})$

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

$$S^2\left(\vec{d}\,(\vec{c}) + \delta\vec{d}\,(\vec{c})\right) \geq S^2\left(\vec{d}\,(\vec{c})\right) \quad \forall\delta \in \mathbb{R} \setminus \{0\} \tag{6.4}$$

(Im weiteren wird $\vec{d}$ anstelle von $\vec{d}\,(\vec{c})$ 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:

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

Setzt man nun die erhaltenen Werte von $S^2(d)$ und $S^2(d + \delta d)$ in die Ungleichung (6.4) ein, so ergibt sich

$$E\left\{(\vec{y}-\vec{d})^2\right\} - 2E\left\{\delta\vec{d}^T(\vec{y}-\vec{d})\right\} + E\left\{(\delta\vec{d})^2\right\} \;\geq\; E\left\{(\vec{y}-\vec{d})^2\right\} \quad \Leftrightarrow$$

$$\underbrace{E\left\{(\delta\vec{d})^2\right\}}_{>0} - 2E\left\{\delta\vec{d}^T(\vec{y}-\vec{d})\right\} \;\geq\; 0$$

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

(Beachte: wir müssen diesen Term auch Null werden lassen, da δ beliebig klein werden darf)

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

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

wir haben – nach den abstandsmessenden Klassifikatoren – zwei prinzipielle Wege zur Konstruktion eines Klassifikators kennengelernt:

1. Minimieren des Risikos

   hierzu benötigen wir die a posteriori Klassenwahrscheinlichkeit $P(\omega|\vec{c})$

   wir können

   (a) $P(\vec{c}, \omega)$ oder äquivalent $P(\vec{c}|\omega)$ und $P(\omega)$ schätzen

   (dies wird das Schätzen eines generativen Modells genannt, da es uns das sampeln, d.h. ziehen, von Mustern des MEP erlaubt;

   wir erhalten damit auch eine Schätzung der statistischen Eignschaften des MEP)

   (b) $P(\omega|\vec{c})$ direkt schätzen

   (diese werden auch diskriminative Modelle genannt)

2. verteilungsfrei eine Diskriminatorfunktion bestimmen

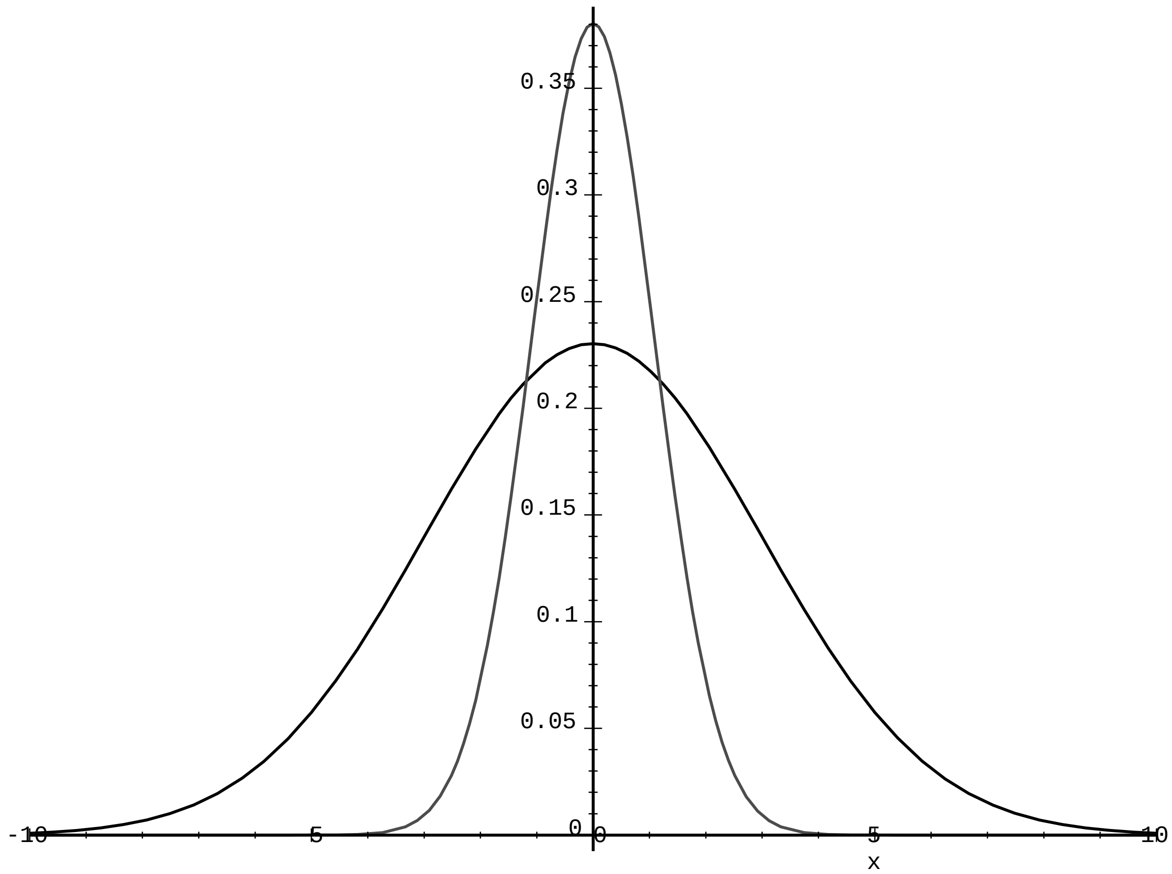
6.6 Schätzung der klassenspezifischen Dichten

6.6.1 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.



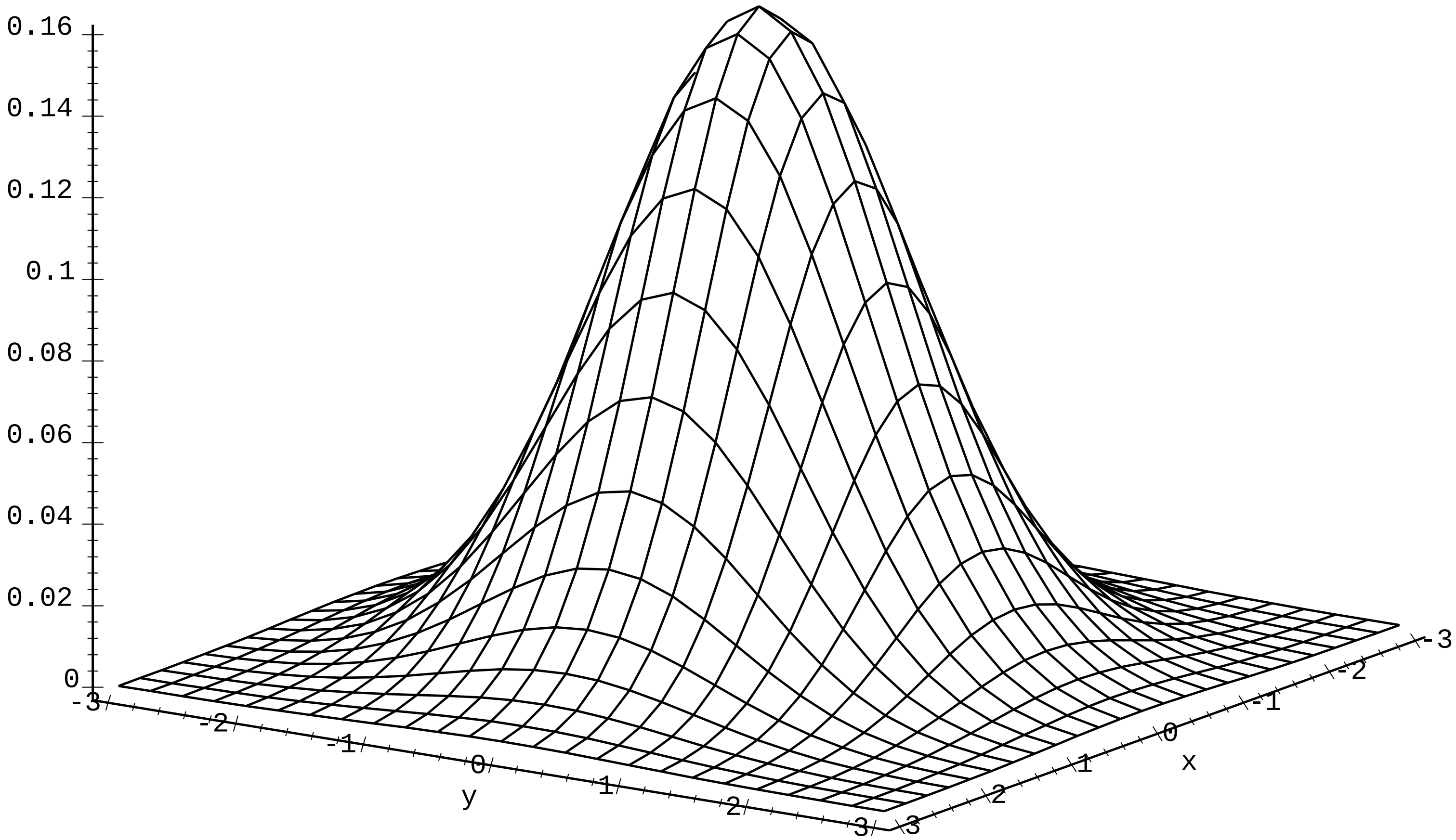
Beispiel zweier Normalverteilungen im $\mathbb{R}^1$

Ist die Zufallsvariable ein M-dimensionaler Vektor $\vec{X}$, so gilt

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

- $\vec{\mu} = E\{\vec{X}\}$ Erwartungswert von $\vec{X}$
- $\underline{K} = E\left\{\left(\vec{X}-\vec{\mu}\right)\left(\vec{X}-\vec{\mu}\right)^T\right\}$ 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
- $\left(\vec{X}-\vec{\mu}\right)^T \underline{K}^{-1}\left(\vec{X}-\vec{\mu}\right)$ heißt Mahalanobis-Distanz



Beispiel einer zweidimensionalen Normalverteilung

6.6.2 Parameterschätzung

schätze Mittelwert $\vec{\mu}$ und Kovarianzmatrix $\underline{K}$ als Maximum-Likelihood-Schätzwert (ML-Schätzer) aus einer Stichprobe der Größe N (für **eine** Klasse):

Definition: der ML-Schätzer $\hat{\theta}_{ML}$ für den Parameter θ bezüglich einer Stichprobe $\vec{c}_1, \ldots, \vec{c}_N$ ist definiert als:

$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}}\, P(\vec{c}_1, \ldots, \vec{c}_N | \theta)$$

Satz: die ML-Schätzer für eine Normalverteilung sind:

$$\widehat{\vec{\mu}}_{ML} = \frac{1}{N} \sum_{n=1}^{N} \vec{c}_n$$

$$\widehat{\underline{K}}_{ML} = \frac{1}{N} \sum_{n=1}^{N} (\vec{c}_n - \widehat{\vec{\mu}})(\vec{c}_n - \widehat{\vec{\mu}})^T$$

Definition: ein Schätzer $\hat{\theta}$ eines Parameters θ heißt erwartungstreu, gdw:

$$E[\hat{\theta}] = \theta$$

Satz: $\widehat{\vec{\mu}}_{ML}$ ist erwartungstreu.

Satz: $\frac{1}{N-1}\sum_{n=1}^{N}(\vec{c}_n - \widehat{\vec{\mu}})(\vec{c}_n - \widehat{\vec{\mu}})^T$ ist ein erwartungstreuer Schätzer der Kovarianzmatrix einer Normalverteilung

Rekursive Berechnung

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

- neues klassifiziertes Trainingsmaterial vorhanden
- durchführen 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:

$$
\begin{aligned}
\widehat{\vec{\mu}}_N &= \frac{1}{N} \sum_{n=1}^{N} \vec{c}_n \\
&= \frac{1}{N} \left(\sum_{n=1}^{N-1} \vec{c}_n \right) + \frac{1}{N} \vec{c}_N \\
&= \frac{1(N-1)}{N(N-1)} \left(\sum_{n=1}^{N-1} \vec{c}_n \right) + \frac{1}{N} \vec{c}_N \\
&= \frac{N-1}{N} \cdot \left(\frac{1}{N-1} \sum_{n=1}^{N-1} \vec{c}_n \right) + \frac{1}{N} \vec{c}_N \\
&= (1 - \frac{1}{N}) \cdot \widehat{\vec{\mu}}_{N-1} + \frac{1}{N} \vec{c}_N
\end{aligned}
$$

Ähnlich läßt sich die Formel für die Kovarianzmatrix herleiten:

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

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

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

6.6.3 Bayes-Normalverteilungsklassifikator

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

- schätze klassenspezifischen Mittelwerte $\vec{\mu}_i$ und Kovarianzmatrizen $\underline{K}_i$ aus klassifizierter Stichprobe

  für Schätzung von $\vec{\mu}_i$ und $\underline{K}_i$ verwende nur die Merkmalsvektoren $\vec{c}_n$, die aus der Klasse ω_i stammen
- schätze a priori Wahrscheinlichkeiten als relative Häufigkeit jeder Klasse in der Stichprobe
- für die Diskriminatorfunktion 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 Diskriminatorfunktion $\vec{d}(\vec{c})$ hat dann die Form

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

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

  damit:

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

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

$$\vec{d'}(\vec{c}) = \begin{pmatrix} b_{01} + (\vec{c} - \vec{\mu}_1)^T \underline{K}_1^{-1} (\vec{c} - \vec{\mu}_1) \\ \vdots \\ b_{0i} + (\vec{c} - \vec{\mu}_i)^T \underline{K}_i^{-1} (\vec{c} - \vec{\mu}_i) \\ \vdots \\ b_{0K} + (\vec{c} - \vec{\mu}_K)^T \underline{K}_K^{-1} (\vec{c} - \vec{\mu}_K) \end{pmatrix}$$

  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)

6.6.4 Vereinfachung durch identische Kovarianzmatrizen

Annahme: die Kovarianzmatrizen aller Klassen sind ω_i identisch

(diese gemeinsame Kovarinazmatrix wird aus der gesamten Stichprobe geschätzt)

- 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 Diskriminatorfunktion

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

  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 $N = \dim(\vec{c})$)

  $\Rightarrow$ oft deutliche Rechenzeitersparnis, da Merkmalsvektoren mit 10 bis 100 Dimensionen nicht selten

6.6.5 Klassengrenzen

- ein Merkmalsvektor $\vec{c}$ liegt auf einer Klassengrenze, falls mehrere Komponenten der Diskriminatorfunktion $\vec{d}(\vec{c})$ 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:

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

  diese Gleichung ist linear

  $\Rightarrow$ Klassengrenzen sind durch Geraden, Flächen oder Hyperflächen gegeben

6.6.6 Mischverteilungen

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

  $$P(\vec{c} \mid \omega_i) = \sum_{l=1}^{L} a_l \cdot \mathcal{N}_{\vec{c}}(\vec{\mu}_l, \underline{K}_l) \qquad \text{mit } \sum_{l=1}^{L} a_l = 1 \quad \text{und } a_l \geq 0$$

  wobei $\vec{\theta} = (a_1, \vec{\mu}_1, \underline{K}_1, \ldots, a_L, \vec{\mu}_L, \underline{K}_L)$
  die Mischverteilung (hier der Klasse ω_i) 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 (k-means, classifying EM)

- ordne innerhalb jeder Iteration jeden Merkmalsvektor $\vec{c}_n$ einem Gebiet **hart** zu
- Verfahren zur Optimierung von L Gebieten

- wähle aufgrund von Vorwissen oder zufällig initiale Parameter $a_l = P(R_l), \vec{\mu}_l, \underline{K}_l$; (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)
- REPEAT
  - $t := t + 1, \quad H^{(t)} := 0$
  - FOR alle Gebiete $R_l, l = 1, \ldots, L$
    - $N_l := 0; \quad \hat{\vec{\mu}}_l := \vec{0}, \quad \underline{\hat{M}}_l := O$
  - FOR alle Vektoren $\vec{c}_n$ der Stichprobe
    - bestimme R_l mit maximalem $P(R_l \mid \vec{c}_n)$
    - $H^{(t)} := H^{(t)} + \ln\left(P\left(R_l\right) P\left(\vec{c} \mid R_l\right)\right)$
    - berechne neue Schätzwerte für den Mittelwert und die Momentenmatrix, d.h.
      $\hat{\vec{\mu}}_l := \hat{\vec{\mu}}_l + \vec{c}_n \qquad \underline{\hat{M}}_l := \underline{\hat{M}}_l + \vec{c}_n \vec{c}_n^T$
    - $N_l := N_l + 1$
  - $H^{(t)} := H^{(t)}/N$
  - FOR alle Gebiete $R_l, l = 1, \ldots, L$
    - $P(R_l) = \frac{N_l}{N}, \quad \vec{\mu}_l := \hat{\vec{\mu}}_l / N_l; \quad \underline{K}_l := \underline{\hat{M}}_l / N_l - \vec{\mu}_l \vec{\mu}_l^T$
- UNTIL $(H^{(t)} - H^{(t-1)}) / \mid H^(t) \mid \leq \varepsilon$

Weiche Vektorquantisierung, EM-Algorithmus

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

$$H = \frac{1}{N} \ln P(\{\vec{c}_1, \ldots, \vec{c}_N\} | \vec{\theta}) = \frac{1}{N} \ln \prod_{n=1}^{N} P(\vec{c}_n | \vec{\theta}) = \frac{1}{N} \sum_{n=1}^{N} \ln P(\vec{c}_n | \vec{\theta}) =$$

$$\frac{1}{N} \sum_{n=1}^{N} \ln \sum_{l=1}^{L} P\left(R_l\right) p\left(\vec{c}_n \mid R_l\right)$$

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

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

- 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 $R_l, l = 1, \ldots, L$
    - $N_l := 0; \quad \hat{\vec{\mu}}_l := \vec{0}, \quad \underline{\hat{M}}_l := O$
  - FOR alle Vektoren $\vec{c}_n$ der Stichprobe
    - berechne $P(R_l \mid \vec{c}_n)$ mittels $P(\vec{c} \mid R_l) = \mathcal{N}_{\vec{c}}(\vec{\mu}_l, \underline{K}_l)$ und $P(R_l)$ 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.
        $\hat{\vec{\mu}}_l := \hat{\vec{\mu}}_l + \vec{c}_n \cdot P(R_l \mid \vec{c}_n) \qquad \underline{\hat{M}}_l := \underline{\hat{M}}_l + \vec{c}_n \vec{c}_n^T \cdot P(R_l \mid \vec{c}_n)$
      - $N_l := N_l + P(R_l \mid \vec{c}_n)$
  - $H^{(t)} := H^{(t)} / N$
  - FOR alle Gebiete $R_l, l = 1, \ldots, L$
    - $P(R_l) = \frac{N_l}{N}, \quad \vec{\mu}_l := \hat{\vec{\mu}}_l / N_l; \quad \underline{K}_l := \underline{\hat{M}}_l / N_l - \vec{\mu}_l \vec{\mu}_l^T$
- UNTIL $(H^{(t)} - H^{(t-1)}) / \mid H^{(t)} \mid \leq \varepsilon$

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 die Vector Quantization 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:

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

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

- die Komponenten der Diskriminatorfunktion 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

- schätze auf einer **unklassifizierten** Stichprobe (weiche oder harte VQ)
  $\Rightarrow L$ 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:

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

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

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

6.6.7 Andere Verteilungen

Statistische Unabhängigkeit

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

Parzenschätzung $\sim$ radiale Basisfunktionen

- **Motivation:** für höhere Dimensionen ist Diskretisierung des Merkmalsraums schwierig (curse of dimensinality): “fast alle” Zellen bleiben leer
- $\Rightarrow$ “verschmiere” jeden Beitrag der Stichprobe um seine Position
- Approximation der Dichte durch Überlagerung von Normalverteilungen, (oder auch andere Fensterfunktionen, z.B. Rechteck)

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

wobei $K = \sigma^2 \mathcal{I}$, und σ die Breite des Fensters (des “Verschmierens”) bestimmt

6.7 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