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 $\vec{\mu}_l$ 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
- 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:

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

- 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

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

- 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:

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

- 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:

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

- Überprüfung der Annahme normalverteilter Klassen:
  - statistische Tests
  - Annahme anhand des Klassifikationsergebnisses auf einer Teststichprobe evaluieren

6.6.4 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:

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

- 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 Unterscheidungsfunktion $\vec{d}(\vec{c})$ den gleichen (maximalen bzw. minimalen) Wert annehmen:

$$\begin{aligned} d_i(\vec{c}) &= d_l(\vec{c}) \Leftrightarrow \\ d_i(\vec{c}) - d_l(\vec{c}) &= 0 \end{aligned}$$

- Einsetzen für Bayesklassifikators mit Normalverteilungen ergibt:

$$d_i(\vec{c}) - d_l(\vec{c}) = b_{0i} + (\vec{c} - \vec{\mu}_i)^T \underline{K}_i^{-1}(\vec{c} - \vec{\mu}_i) - b_{0l} - (\vec{c} - \vec{\mu}_l)^T \underline{K}_l^{-1}(\vec{c} - \vec{\mu}_l) = 0$$

- 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:

$$d_i(\vec{c}) - d_l(\vec{c}) \;=\; b_{1i} - b_{2i}\vec{c} - b_{1l} + b_{2l}\vec{c}$$

  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:
  $$p(\vec{c}) = \sum_{l=1}^{L} a_l \cdot \mathcal{N}_{\vec{c}}(\vec{\mu}_l, \underline{K}_l)$$
  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 (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
  (In um Konvergenz zu beweisen)

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

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

- 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 $R_l, l = 1, \dots, L$
  - $N_l := 0; \quad \hat{\vec{\mu}}_l := \vec{0}, \quad \hat{\underline{M}}_l := O$
- 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.
    $\hat{\vec{\mu}}_l := \hat{\vec{\mu}}_l + \vec{c}_i \qquad \hat{\underline{M}}_l := \hat{\underline{M}}_l + \vec{c}_i\vec{c}_i^T$
  - $N_l := N_l + 1$
- $H^{(t)} := H^{(t)}/N$
- FOR alle Gebiete $R_l, l = 1, \dots, L$
  - $P(R_l) = \frac{N_l}{N}, \quad \vec{\mu}_l := \hat{\vec{\mu}}_l/N_l; \quad \underline{K}_l := \hat{\underline{M}}_l/N_l - \vec{\mu}_l\vec{\mu}_l^T$

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

Weiche Vektorquantisierung

- 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 N):

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

- 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

6.6 Klassifikation bei gegebener Wahrscheinlichkeitsverteilung

- 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)
- 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 $\vec{c}_i$ der Stichprobe
    - berechne $p(\vec{c}_i \mid R_l)$ und $p(R_l \mid \vec{c}_i)$ 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.
        $\hat{\vec{\mu}}_l := \hat{\vec{\mu}}_l + \vec{c}_i \cdot P(R_l \mid \vec{c}_i) \qquad \underline{\hat{M}}_l := \underline{\hat{M}}_l + \vec{c}_i \vec{c}_i^T \cdot P(R_l \mid \vec{c}_i)$
      - $N_l := N_l + P(R_l \mid \vec{c}_i)$
  - $H^{(t)} := H^{(t)}/N$
  - 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 $(H^{(t)} - H^{(t-1)}) / \mid H^{(t)} \mid \leq \varepsilon$

LBG-Algorithmus für Dichten

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

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

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

- 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 $\mathcal{N}(\vec{\mu}_l, \underline{K}_l)$
- bestimme die Gewichtsparameter der obigen Dichten für die Klasse ω_k aus einer (evtl. kleineren) **klassifizierten** Stichprobe wie folgt:

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

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

- $$P(\vec{c} \mid \omega_k) = \sum_{l=1}^{L} a_l^k \mathcal{N}(\vec{\mu}_l, \underline{K}_l)$$
- 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)
-
$$p(\vec{c} \mid \omega_k) = \prod_{i=1}^{K} p(c_\nu \mid \omega_i)$$
- 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 relaive Häufigkeiten tabelliert werden

Parzenschätzung

- **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)

$$p(\vec{c} \mid \omega_k) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{N}\left(\vec{\mu}_l^k, K\right)$$

wobei $K = \sigma I$, 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