6. Allgemeine Problemstellungen

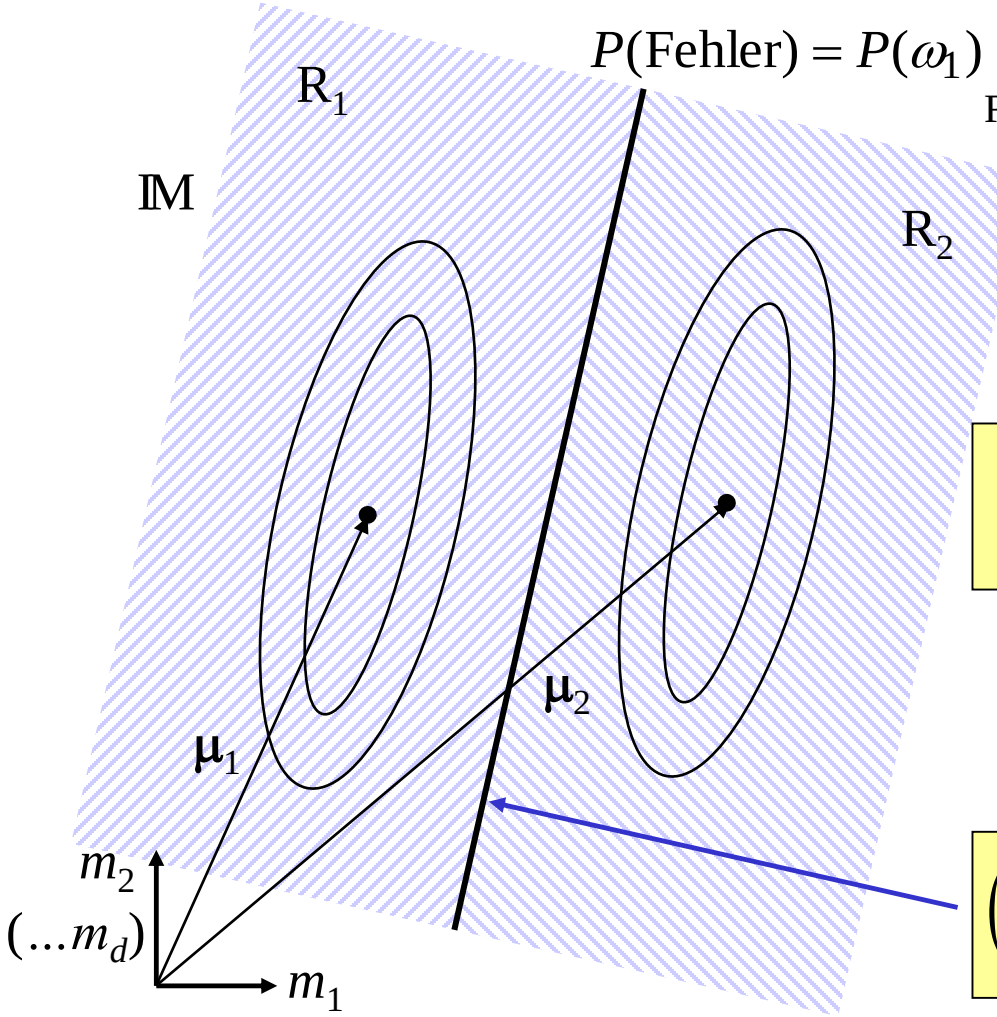
6.1. Dimension des Merkmalsraumes

Einfluss der Merkmalsraumdimension auf die Klassifikationsgüte

Beispiel: Unterscheidung zweier Klassen, $c = 2$.

$$\mathbf{m} \mid \omega_i \sim p(\mathbf{m} \mid \omega_i) = \mathrm{N}(\mathbf{m}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}) \qquad i = 1,2 \qquad P(\omega_1) = P(\omega_2) = \tfrac{1}{2}$$

$$P(\text{Fehler}) = P(\omega_1) \int_{\mathrm{R}_2} p(\mathbf{m} \mid \omega_1)\mathrm{d}\mathbf{m} + P(\omega_2) \int_{\mathrm{R}_1} p(\mathbf{m} \mid \omega_2)\mathrm{d}\mathbf{m}$$



Bayes'sche Optimalentscheidung, Entscheidungsfunktionen:

$$k_i(\mathbf{m}) := -\frac{1}{2}(\mathbf{m} - \boldsymbol{\mu}_i)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{m} - \boldsymbol{\mu}_i)$$

Entscheidungsgrenze:

$$k_1(\mathbf{m}) = k_2(\mathbf{m}) \;\Leftrightarrow$$

$$\left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right)^{\mathrm{T}} \left(\mathbf{m} - \tfrac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right) = 0$$

6.1. Dimension des Merkmalsraumes

Beispiel: Unterscheidung zweier Klassen, $c = 2$; *Fortsetzung*.

$$L(\mathbf{m}) := \frac{P(\omega_2 \mid \mathbf{m})}{P(\omega_1 \mid \mathbf{m})} = \frac{p(\mathbf{m} \mid \omega_2)}{p(\mathbf{m} \mid \omega_1)}$$ Likelihoodverhältnis

$$L(\mathbf{m}) = \frac{\frac{1}{(2\pi)^{d/2} \mid \boldsymbol{\Sigma} \mid^{1/2}} \exp\left[-\frac{1}{2} (\mathbf{m} - \boldsymbol{\mu}_2)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{m} - \boldsymbol{\mu}_2) \right]}{\frac{1}{(2\pi)^{d/2} \mid \boldsymbol{\Sigma} \mid^{1/2}} \exp\left[-\frac{1}{2} (\mathbf{m} - \boldsymbol{\mu}_1)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{m} - \boldsymbol{\mu}_1) \right]} =$$

$$= \exp\left[(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{m} + \frac{1}{2} (\boldsymbol{\mu}_2 + \boldsymbol{\mu}_1)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \right]$$

$\Lambda(\mathbf{m}) := \ln L(\mathbf{m})$ Log-Likelihoodverhältnis

$$\Lambda(\mathbf{m}) = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \left(\mathbf{m} - \frac{1}{2} \boldsymbol{\mu}_2 - \frac{1}{2} \boldsymbol{\mu}_1 \right) = \mathbf{A}\mathbf{m} + \mathbf{b}$$

Entscheidung, durch Vergleich von $\Lambda(\mathbf{m})$ mit der Schwelle 0:

$\hat{\omega} = \omega_1$ wenn $\Lambda(\mathbf{m}) < 0$ $(L(\mathbf{m}) < 1)$ sonst $\hat{\omega} = \omega_2$

Beispiel: Unterscheidung zweier Klassen, $c = 2$; *Fortsetzung*

Λ entsteht durch affine Transformation des normalverteilten Zufallsvektors $\mathbf{m}$. $\Rightarrow \Lambda(\mathbf{m})$ ist normalverteilt.

Erwartung:
$$\mathrm{E}\{\Lambda(\mathbf{m}) \mid \omega_1\} = \mathrm{E}\left\{(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}\mathbf{m} + \frac{1}{2}(\boldsymbol{\mu}_2 + \boldsymbol{\mu}_1)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \mid \omega_1\right\}$$
$$= (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}\mathrm{E}\{\mathbf{m} \mid \omega_1\} + \frac{1}{2}(\boldsymbol{\mu}_2 + \boldsymbol{\mu}_1)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$
$$= -\frac{1}{2}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$$
$$\mathrm{E}\{\Lambda(\mathbf{m}) \mid \omega_2\} = +\frac{1}{2}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$$

Varianz:
$$\mathrm{Var}\{\Lambda(\mathbf{m}) \mid \omega_i\} = \mathrm{Var}\left\{(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}\mathbf{m} + \frac{1}{2}(\boldsymbol{\mu}_2 + \boldsymbol{\mu}_1)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \mid \omega_i\right\}$$
$$= (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}\mathrm{Var}\{\mathbf{m} \mid \omega_i\}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$$
$$= (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$$
$$i = 1,2$$

6.1. Dimension des Merkmalsraumes

Beispiel: Unterscheidung zweier Klassen, $c = 2$.

WDFen des Log-Likelihoodverhältnisses:

$$\Lambda(\mathbf{m}) \mid \omega_1 \sim \mathrm{N}\left(\Lambda; -\tfrac{1}{2}s^2, s^2\right) \qquad \Lambda(\mathbf{m}) \mid \omega_2 \sim \mathrm{N}\left(\Lambda; \tfrac{1}{2}s^2, s^2\right)$$

$$s := \| \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1 \|_m = \sqrt{(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)}$$

Mahalanobis Distanz der Erwartungswerte

$$P(\text{Fehler}) = \tfrac{1}{2} P(\Lambda(\mathbf{m}) \geq 0 \mid \omega_1) + \tfrac{1}{2} P(\Lambda(\mathbf{m}) < 0 \mid \omega_2) = P(\Lambda(\mathbf{m}) \geq 0 \mid \omega_1)$$

$$= \frac{1}{\sqrt{2\pi}s} \int_0^\infty \exp\left(-\tfrac{1}{2} \frac{(\Lambda + \frac{s^2}{2})^2}{s^2} \right) \mathrm{d}\Lambda$$

$$= \frac{1}{\sqrt{2\pi}} \int_{s/2}^\infty \exp\left(\frac{-u^2}{2} \right) \mathrm{d}u$$

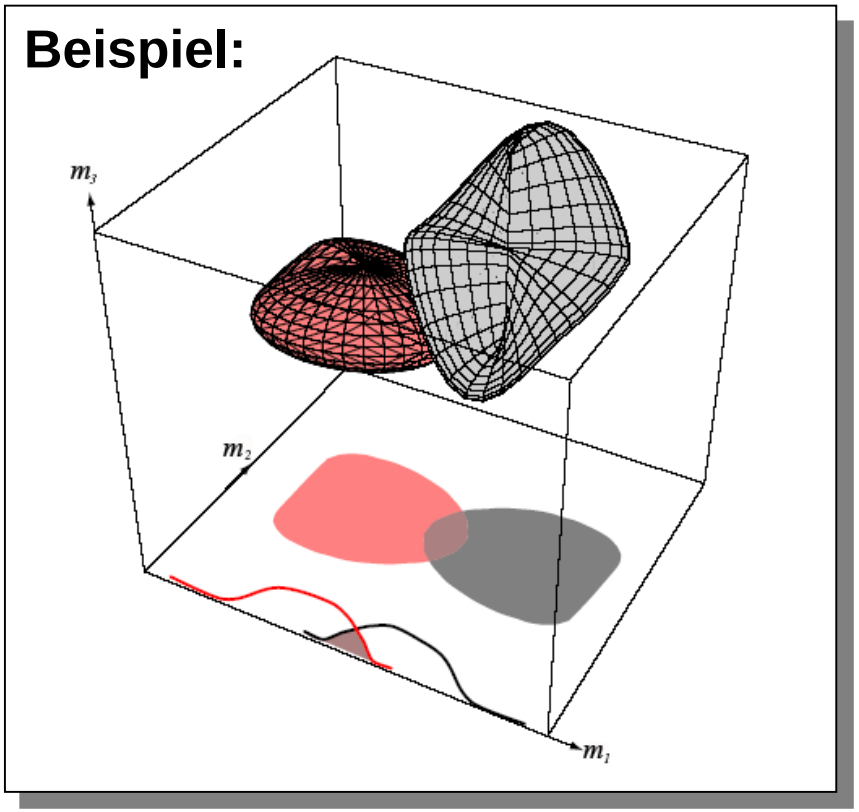
$$P(\text{Fehler}) \xrightarrow[\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_m \to \infty]{} 0 \Rightarrow \text{Jedes weitere Merkmal reduziert den Fehler.}$$

$\Rightarrow$ Je höher die Dimension des Merkmalsraumes desto besser ! (?)

6.1. Dimension des Merkmalsraumes

Ideal: Mit jedem hinzu kommenden Merkmal vergrößert sich der Abstand zwischen den Klassen und ihre „Überlappung“ wird geringer.

Beispiel:



Quelle: R. O. Duda, P. E. Hart, D. G. Stork: Pattern Classification

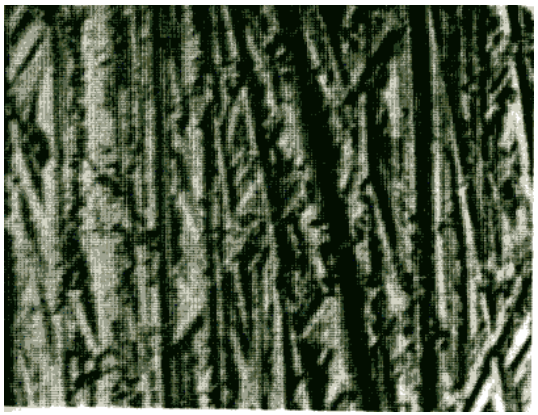
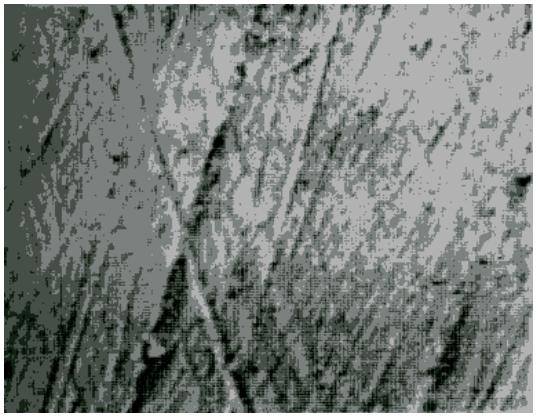
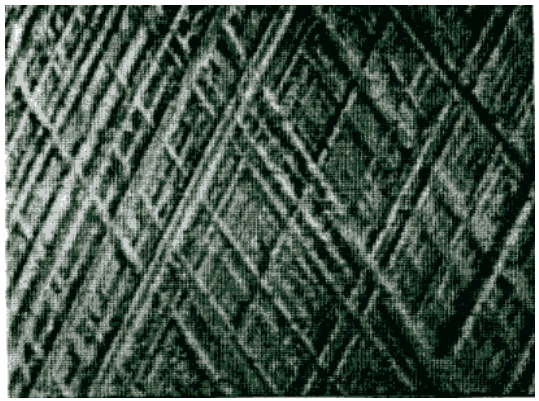
Bemerkungen:

- Aussage stimmt nur im Idealfall, wenn die klassenbedingten WDFen bekannt wären oder eine unendlich große Lernstichprobe vorläge.
- Im Realfall sind die klassenbedingten WDFen unbekannt und die Lernstichprobe nur endlich groß. Die Aussage ist dann falsch.

6.1. Dimension des Merkmalsraumes

Beispiel: Automatische Bewertung von gehonten Zylinderoberflächen

Gegeben: Katalog mit $N = 33$ Texturbildern, die bezüglich ihrer Qualität mit ganzzahligen Noten n von 1 bis 10 bewertet sind (ordinale Klassenstruktur).

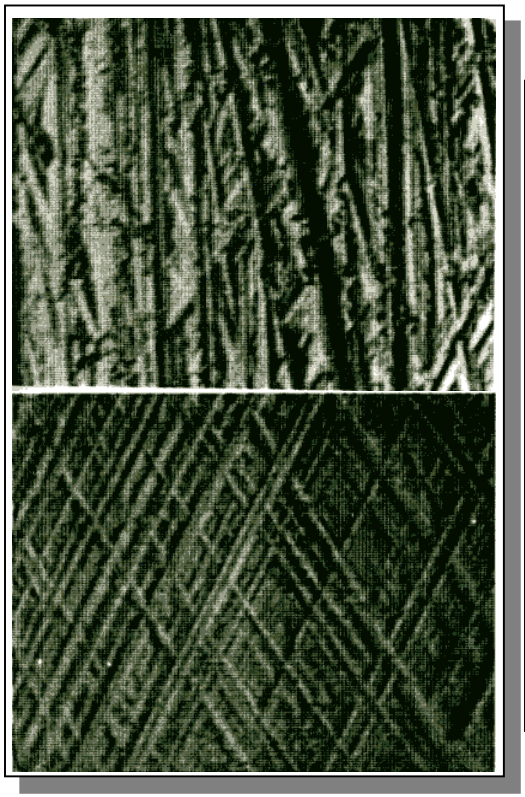
Aufgabe: Automatisierung der Bewertung von Hontexturen auf der Basis dieses Kataloges.

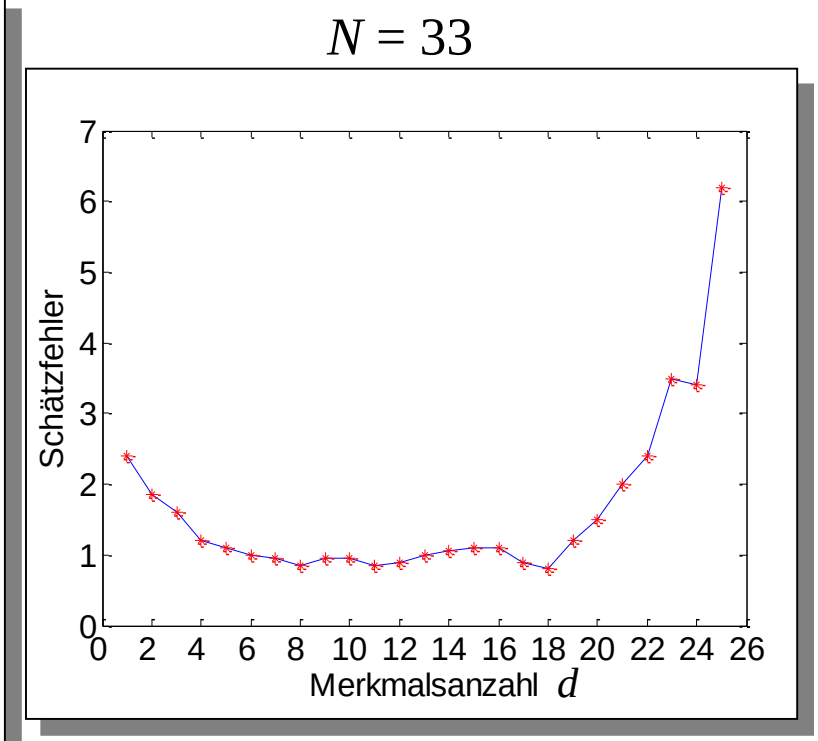
Ansatz:

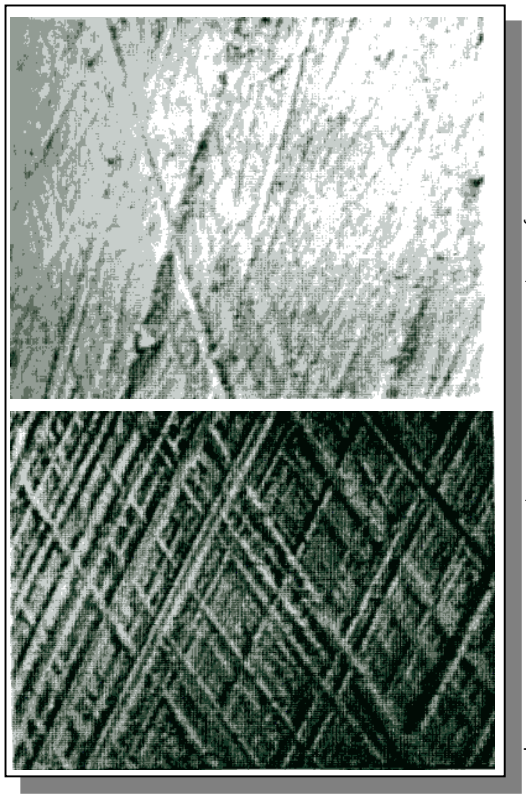
- Heuristisch und modellbasierte Def. von 25 Merkmalen: $\mathcal{M} = \{m_1, ..., m_{25}\}$
- Klassifikation durch Schätzung der Noten mittels linearer Regression $\hat{n} = \mathbf{a}^{\mathrm{T}}\mathbf{m}$ mit kleinstem mittleren quadratischen Fehler.
- Merkmalsauswahl gemäß Unterabschnitt 2.7.5
- Details: Siehe J. Beyerer, Analyse von Riefentexturen, VDI-Verlag 1994

6.1. Dimension des Merkmalsraumes

Beispiel: Automatische Bewertung gehonter Zylinderoberflächen







Beobachtung: Wird die Anzahl der Merkmale erhöht, nimmt der Fehler zunächst ab, nimmt dann aber mit wachsender Dimension d stark zu.

Anschauliche Erklärungen: (A) Jedes Merkmal trägt „Nutzinformation“ und „irrelevante Information“. Ab einer bestimmten Anzahl d kommt nur noch in den berücksichtigten Merkmalen bereits enthaltene Nutzinformation hinzu, aber jeweils deren irrelevante Info sowie deren negativen Einfluss auf die Güte der Schätzung der Parameter des Klassifikators. (B) Mit wachsendem d wird das Verhältnis N/d immer schlechter.

6.1. Dimension des Merkmalsraumes

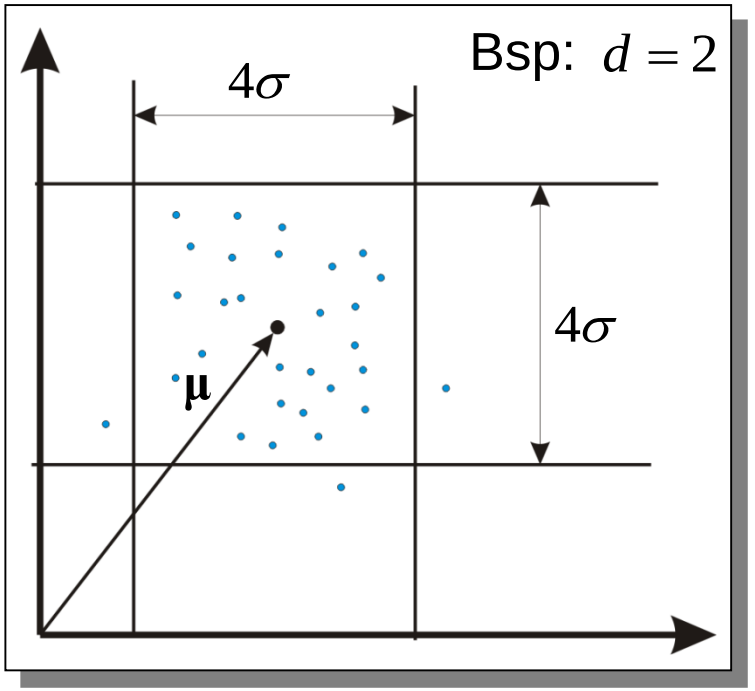
Beispiel: Intervallwahrscheinlichkeiten und Dimension d

Sei $m \sim \mathrm{N}(m;\, \mu,\, \sigma^2) \quad \Rightarrow \quad P(|\, m-\mu \,| < 2\sigma) \approx 0{,}95$

→ Der Großteil der Stichproben liegt innerhalb des Intervalls $[\mu - 2\sigma,\, \mu + 2\sigma]$.

Sei $\mathbf{m} \sim \mathrm{N}(\mathbf{m};\, \boldsymbol{\mu},\, \sigma^2\mathbf{I})$, $d > 1$

$$P(|\, m_1 - \mu_1 \,| < 2\sigma \wedge \ldots \wedge |\, m_d - \mu_d \,| < 2\sigma) \approx (0{,}95)^d$$



Bsp: $d = 2$

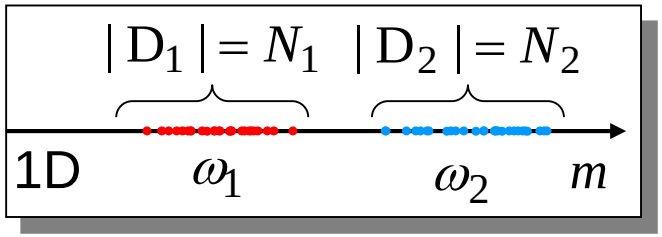
$$P(|\, m_1 - \mu_1 \,| < 2\sigma \wedge |\, m_2 - \mu_2 \,| < 2\sigma) \approx (0{,}95)^2 \approx 0{,}9$$

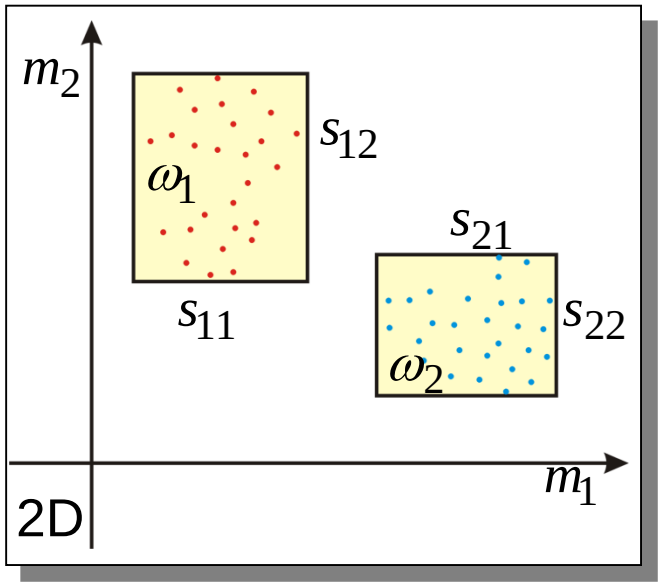
Bsp: $d = 100$

$$P(|\, m_1 - \mu_1 \,| < 2\sigma \wedge \ldots \wedge |\, m_{100} - \mu_{100} \,| < 2\sigma) \approx (0{,}95)^{100} \approx 0{,}0059$$

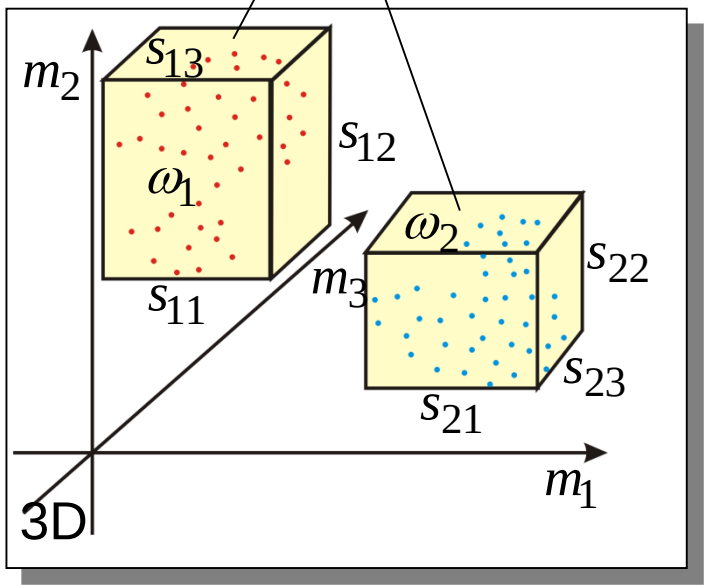
→ Der Großteil der Stichproben liegt außerhalb des Intervalls $[\mu - 2\sigma,\, \mu + 2\sigma]^d$.

Stichproben-Dichte und Dimension des Merkmalsraumes





Umschreibende Quader



Volumina: $V_i = \prod_{j=1}^{d} s_{ij}$ Lokale Stichprobendichten: $\gamma_i := N_i / V_i$

$$\gamma_i = \frac{N_i}{\bar{s}_i^{\,d}} \overset{!}{=} \text{konstant} \Leftrightarrow N_i = N_i(d) \propto \bar{s}_i^{\,d} \quad \text{mit} \quad \bar{s}_i := \sqrt[d]{\prod_{j=1}^{d} s_{ij}}$$

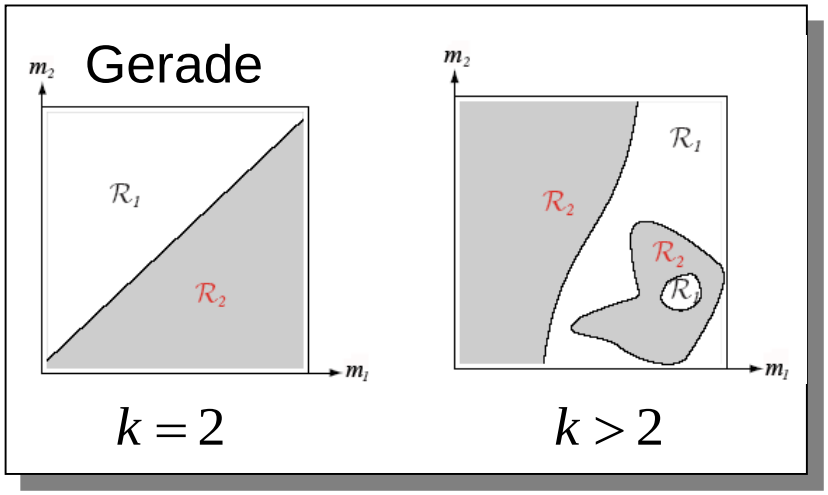
Entscheidungsgebietsgrenzen und Dimension des Merkmalsraumes

- Entscheidungsgebiete im Merkmalsraum sind begrenzt durch Hyperflächen (Dimension $d-1$).
- Die Grenzen der Entscheidungsgebiete sind durch Gleichheiten von Entscheidungsfunktionen bestimmt: $k_l(\mathbf{m}) = k_j(\mathbf{m})$.
- Zur parametrisierten Beschreibung der Grenzen der Entscheidungsgebiete werden i. Allg. mathematische Modelle $f(\mathbf{m};\boldsymbol{\theta}) = 0$ über einem Parameterraum Θ der Dimension $k \geq d$ benötigt.
- Die Parameter $\boldsymbol{\theta}$ müssen aus der Lernstichprobe D geschätzt werden. Der damit einhergehende Schätzfehler beeinflusst die Güte der Entscheidungsgebietsgrenzen.
- Welchen Einfluss hat die Dimension d des Merkmalsraumes auf den Fehler der Schätzung der Parameter $\boldsymbol{\theta}$?

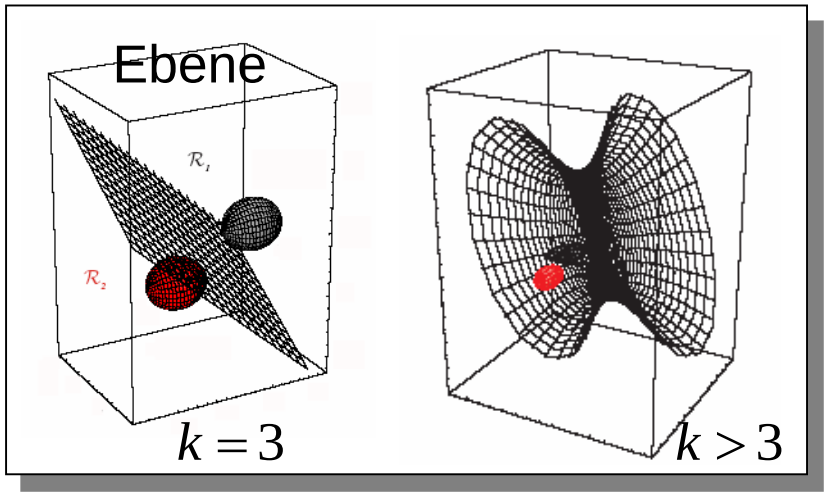
Entscheidungsgebietsgrenzen und Dimension des Merkmalsraumes

$c = 2$

$d = 2 \quad k \geq 2$



$d = 3 \quad k \geq 3$



Quelle: R. O. Duda, P. E. Hart, D. G. Stork: Pattern Classification

$\vdots$

$k \geq d$

6.1. Dimension des Merkmalsraumes

Mehrdimensionale Cramer-Rao-Schranke (CRB) für erwartungstreue Schätzer (klassische Statistik), ohne Beweis:

$\mathrm{D} = \{\mathbf{m}_1, ..., \mathbf{m}_N\}$, Beobachtungen $\mathbf{m}_i$ i.i.d.

$\boldsymbol{\theta} = (\theta_1, ..., \theta_k)$ aus Lernstichprobe schätzen $\rightarrow \hat{\boldsymbol{\theta}} = (\hat{\theta}_1, ..., \hat{\theta}_k)$

Fishersche Informationsmatrix:

$$\mathbf{J}(\boldsymbol{\theta}) := \mathrm{E}\left\{ \left(\nabla_{\boldsymbol{\theta}} \ln p(\mathbf{m} \mid \boldsymbol{\theta})\right)\left(\nabla_{\boldsymbol{\theta}} \ln p(\mathbf{m} \mid \boldsymbol{\theta})\right)^{\mathrm{T}} \right\}$$

Es gilt: $\mathrm{Cov}\{\hat{\boldsymbol{\theta}}(\mathrm{D})\}$ "$\geq$" $\frac{1}{N}\mathbf{J}^{-1}(\boldsymbol{\theta})$ im Sinne von:

$$\boldsymbol{\alpha}^{\mathrm{T}} \mathrm{Cov}\{\hat{\boldsymbol{\theta}}\}\boldsymbol{\alpha} \geq \frac{1}{N}\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{J}^{-1}(\boldsymbol{\theta})\boldsymbol{\alpha} \quad \forall \boldsymbol{\alpha} \in \mathbb{R}^k$$

$$\mathrm{tr}\{\mathrm{Cov}(\hat{\boldsymbol{\theta}})\} \geq \frac{1}{N}\mathrm{tr}\{\mathbf{J}^{-1}(\boldsymbol{\theta})\}$$

Details siehe z.B.: J. Beyerer: „Verfahren zur quantitativen statistischen Bewertung von Zusatzwissen in der Messtechnik“ VDI-Verlag 1999

6.1. Dimension des Merkmalsraumes

Beispiel zur CRB: $\mathbf{m} \sim \mathrm{N}(\mathbf{m}; \boldsymbol{\mu}, \sigma^2 \mathbf{I})$ $\boldsymbol{\theta} := \boldsymbol{\mu}$ unbekannt, σ^2 bekannt

$$\ln p(\mathbf{m} \mid \boldsymbol{\mu}) = -\frac{1}{2\sigma^2}(\mathbf{m}-\boldsymbol{\mu})^{\mathrm{T}}(\mathbf{m}-\boldsymbol{\mu}) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln \sigma^2 \, |\mathbf{I}|$$

$$\nabla_{\boldsymbol{\mu}} \ln p(\mathbf{m} \mid \boldsymbol{\mu}) = \frac{1}{\sigma^2}(\mathbf{m}-\boldsymbol{\mu})$$

$$\left(\nabla_{\boldsymbol{\mu}} \ln p(\mathbf{m} \mid \boldsymbol{\mu})\right)\left(\nabla_{\boldsymbol{\mu}} \ln p(\mathbf{m} \mid \boldsymbol{\mu})\right)^{\mathrm{T}} = \frac{1}{\sigma^4}(\mathbf{m}-\boldsymbol{\mu})(\mathbf{m}-\boldsymbol{\mu})^{\mathrm{T}}$$

$$\mathrm{E}\left\{\left(\nabla_{\boldsymbol{\mu}} \ln p(\mathbf{m} \mid \boldsymbol{\mu})\right)\left(\nabla_{\boldsymbol{\mu}} \ln p(\mathbf{m} \mid \boldsymbol{\mu})\right)^{\mathrm{T}}\right\} = \frac{1}{\sigma^4}\mathrm{E}\left\{(\mathbf{m}-\boldsymbol{\mu})(\mathbf{m}-\boldsymbol{\mu})^{\mathrm{T}}\right\} = \frac{1}{\sigma^4}\sigma^2\mathbf{I} = \frac{1}{\sigma^2}\mathbf{I}$$

$$\mathbf{J}(\boldsymbol{\mu}) = \sigma^{-2}\mathbf{I}$$

$$\mathbf{J}^{-1}(\boldsymbol{\mu}) = \sigma^2\mathbf{I}$$

$$\mathrm{tr}\left\{\mathbf{J}^{-1}(\boldsymbol{\mu})\right\} = k\sigma^2$$

6.1. Dimension des Merkmalsraumes

Beispiel zur CRB: $\mathbf{m} \sim \mathrm{N}(\mathbf{m}; \boldsymbol{\mu}, \sigma^2 \mathbf{I})$ $\boldsymbol{\theta} := \boldsymbol{\mu}$ unbekannt, σ^2 bekannt

$\mathrm{D} = \{\mathbf{m}_1, ..., \mathbf{m}_N\}$, Beobachtungen $\mathbf{m}_i$ i.i.d.

Summe der quadratischen stochastischen Schätzfehler:

$$\mathrm{tr}\{\mathrm{Cov}(\hat{\boldsymbol{\theta}})\} = \mathrm{tr}\{\mathrm{Cov}(\hat{\boldsymbol{\mu}})\} \geq \frac{k\sigma^2}{N}$$

„Volumen des Spats“ der stochastischen Schätzfehler:

$$\sqrt{|\mathrm{Cov}(\hat{\boldsymbol{\theta}})|} = \sqrt{|\mathrm{Cov}(\hat{\boldsymbol{\mu}})|} \geq \frac{1}{\sqrt{N}} \sqrt{|\mathbf{J}^{-1}(\hat{\boldsymbol{\mu}})|} = \frac{\sigma^k}{\sqrt{N}}$$

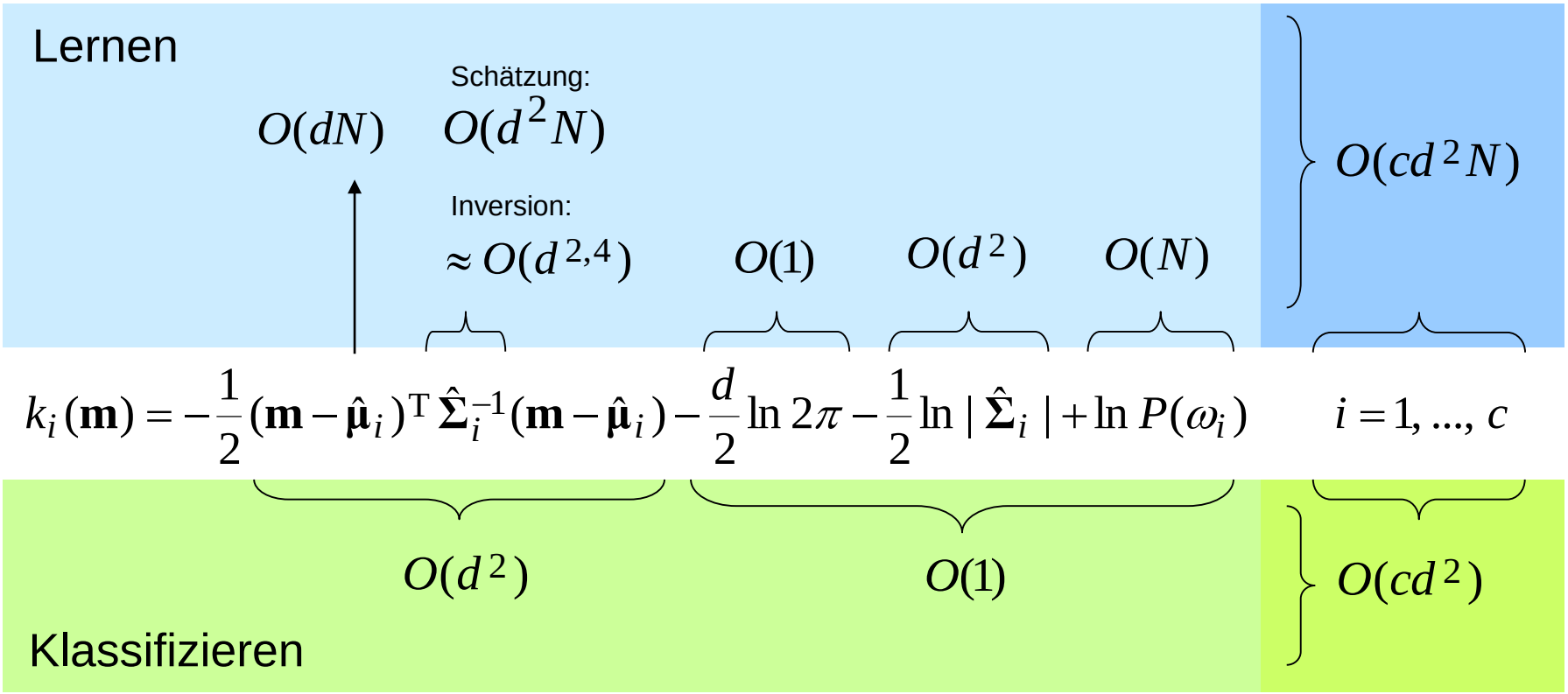
Ergebnis:

$$\mathrm{tr}\{\mathrm{Cov}(\hat{\boldsymbol{\theta}})\} \geq \frac{1}{N} \mathrm{tr}\{\mathbf{J}^{-1}(\boldsymbol{\theta})\} \xrightarrow[d \to \infty \Rightarrow k \to \infty]{} \infty$$

6.1. Dimension des Merkmalsraumes

Komplexität:

Beispiel MAP-Klassifikation (Bayes'sche Klassifikation) bei normalverteilten Merkmalen



6.1. Dimension des Merkmalsraumes

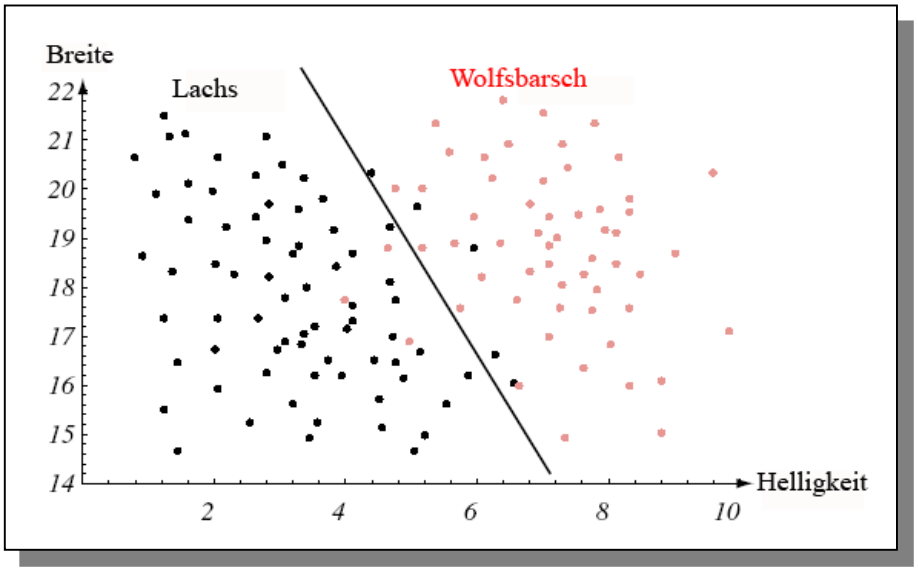
Zusammenfassung:

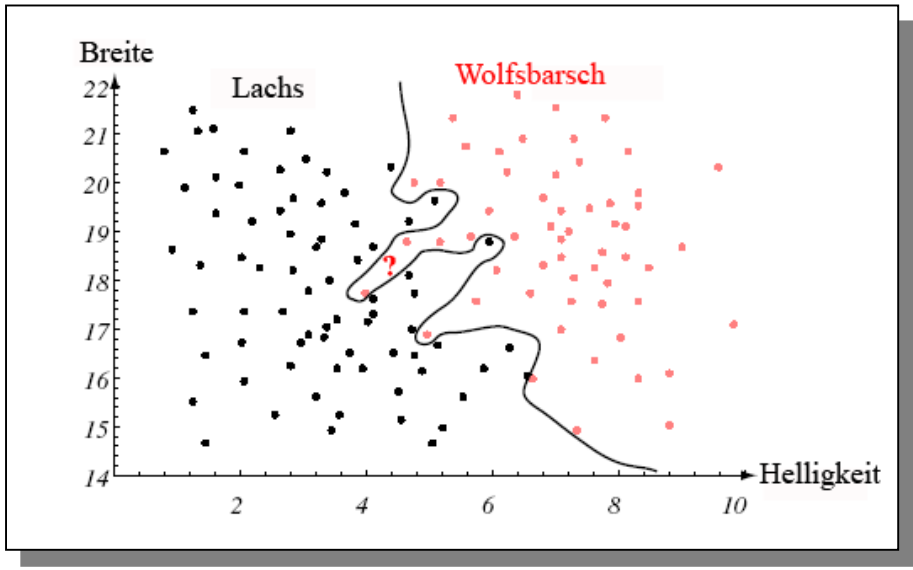
Dimension des Merkmalsraumes: $d \uparrow \Rightarrow$

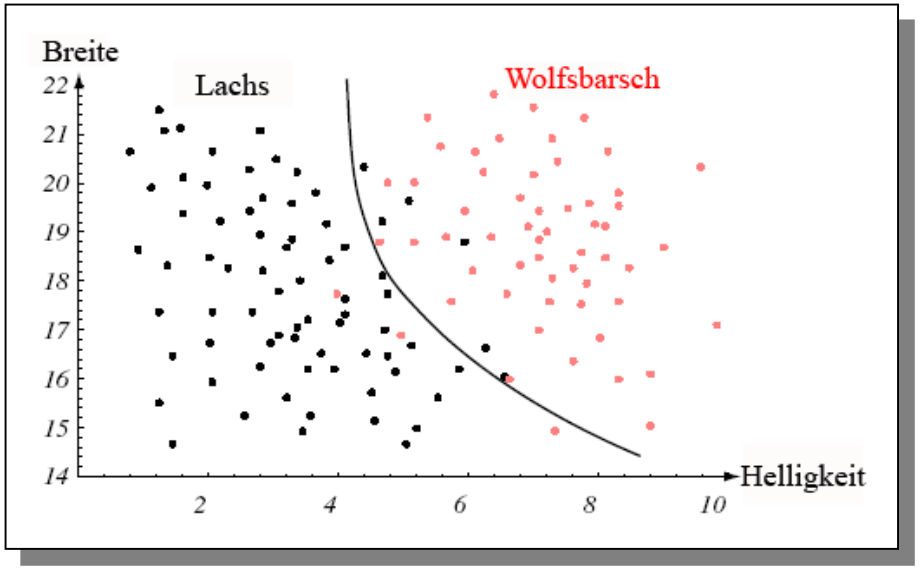
- Wahrscheinlichkeit für Intervalle „konstanter Länge“: $P \downarrow$
- Parametrische Dimension der Gebietsgrenzen im Merkmalsraum: $k \uparrow$
- Stochastischer Gesamtschätzfehler der Parameterschätzung: $\text{Fehler} \uparrow$
- Dichte der Stichproben im Merkmalsraum: $\gamma \downarrow$
- Komplexität $\uparrow$

6.2. Überanpassung (Overfitting)

Beispiel:







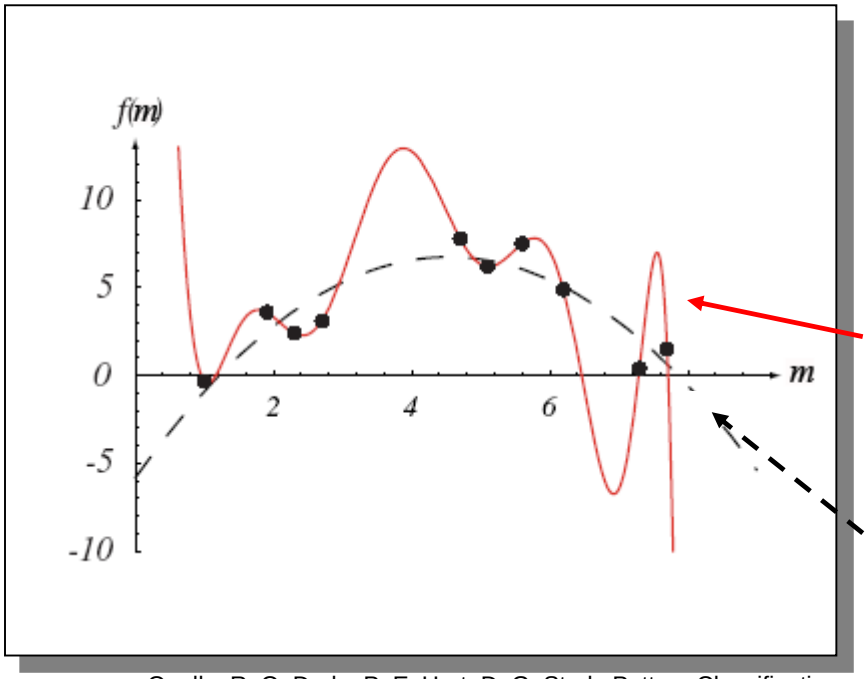
Quelle: R. O. Duda, P. E. Hart, D. G. Stork: Pattern Classification

6.2. Überanpassung (Overfitting)

Beispiel: Approximation von Daten durch eine Funktion

Gegeben: Datensatz mit 10 Punkten $\{(m_j, f(m_j),\ j = 1,\ldots, 10\}$

erzeugt mittels: $f(m) = a_2 m^2 + a_1 m + a_0 + e, \quad e \sim \mathrm{N}(e; 0, \sigma^2)$



Quelle: R. O. Duda, P. E. Hart, D. G. Stork: Pattern Classification

Hypothesenfunktion: $h(m) = \sum_{i=0}^{g-1} a_i m^i$

soll $f(m)$ möglichst gut approximieren.

Modell mit Grad $g = 10$ approximiert perfekt die gegebenen Daten.

Das Modell des Grades $g = 3$ ist für „neue“ Daten besser.

Occam‘s Razor (Ockhams Rasiermesser)*:

Man bevorzugt die einfachste, mit den Daten konsistente Hypothese.

* William von Ockham, englischer Philosoph im 14. Jahrhundert

6.2. Overfitting

Faustregeln:

- Je kleiner die Stichprobe, desto einfacher sollte auch das Modell gewählt werden.
- Je größer die Anzahl der Parameter des Klassifikators, desto größer muss die Anzahl der Lernstichproben sein.