Parameterschätzung - Bayes‘sche Schätzung

Parameterschätzung

Was haben wir bis jetzt behandelt?

- Klassifizierung
- Repräsentation unseres Wissens (durch Wahrscheinlichkeiten und Verteilungen)

Lernen:

- Gewinnen des Wissens aus Trainingsdaten

Parameterschätzung

Relativ unproblematisch ist normalerweise die Schätzung der a-Priori-Wahrscheinlichkeiten aus Beispieldaten

Schwieriger ist die Bestimmung der Likelihoodfunktionen $p(\mathbf{x} \mid \omega_i)$.

Einfacher wird das Problem, wenn nur Parameter einer parametrisierten Funktion bestimmt werden müssen, z.B. bei Normalverteilung $\boldsymbol{\mu}_i$ und $\mathbf{C}_i$.

Parameterschätzung

Methoden für diese Parameterschätzung:

- Maximum likelihood
- Bayes‘sche Schätzung

Was ist der Unterschied?

- Maximum likelihood such die Parameter, die die Trainingsdaten am besten erklären

- Bayes‘sche Schätzung liefert nicht einen Wert, sondern Wahrscheinlichkeitsverteilung für Parameter.

Parameterschätzung

Maximum likelihood

Wir bestimmen likelihood für die Trainingsdaten $\mathcal{D}$:

$$P(\mathcal{D}\,|\,\boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k\,|\,\boldsymbol{\theta})$$

und suchen den Parametervektor $\hat{\boldsymbol{\theta}}$, der $P(\mathcal{D}\,|\,\boldsymbol{\theta})$ maximiert.

Einfacher ist es oft, die log-likelihood zu maximieren

$$l(\boldsymbol{\theta}) = \ln P(\mathcal{D}\,|\,\boldsymbol{\theta}) = \sum_{k=1}^{n} \ln\, p(\mathbf{x}_k\,|\,\boldsymbol{\theta})$$

Parameterschätzung

maximum
likelihood

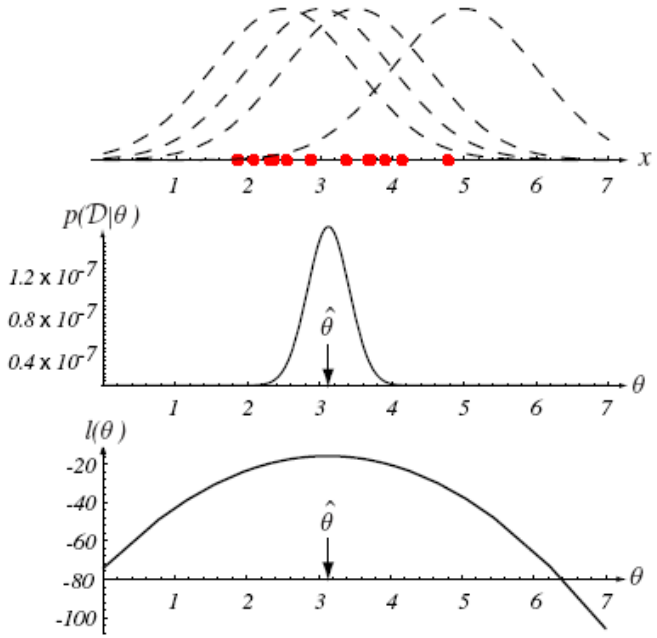


FIGURE 3.1. The top graph shows several training points in one dimension, known or assumed to be drawn from a Gaussian of a particular variance, but unknown mean. Four of the infinite number of candidate source distributions are shown in dashed lines. The middle figure shows the likelihood $p(\mathcal{D}|\theta)$ as a function of the mean. If we had a very large number of training points, this likelihood would be very narrow. The value that maximizes the likelihood is marked $\hat{\theta}$; it also maximizes the logarithm of the likelihood—that is, the log-likelihood $l(\theta)$, shown at the bottom. Note that even though they look similar, the likelihood $p(\mathcal{D}|\theta)$ is shown as a function of θ whereas the conditional density $p(x|\theta)$ is shown as a function of x. Furthermore, as a function of θ, the likelihood $p(\mathcal{D}|\theta)$ is not a probability density function and its area has no significance. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*.

Parameterschätzung

Die Parameter $\hat{\boldsymbol{\theta}}$, die die maximum likelihood Bedingung erfüllen, sind also:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \sum_{k=1}^{n} \ln p(\mathbf{x}_k \mid \boldsymbol{\theta})$$

Sie müssen folgende Bedingung erfüllen:

$$\nabla l(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \frac{\partial l(\boldsymbol{\theta})}{\partial \theta_1} \\ \vdots \\ \frac{\partial l(\boldsymbol{\theta})}{\partial \theta_p} \end{pmatrix} = \sum_{k=1}^{n} \nabla \ln p(\mathbf{x}_k \mid \boldsymbol{\theta}) = \mathbf{0}$$

Maximum Likelihood

Beispiel: Gaußverteilung $p\left(\mathbf{x}_k \mid \boldsymbol{\theta}\right) = N\left(\boldsymbol{\mu}, \mathbf{C}\right)\left(\mathbf{x}_k\right)$

Eindimensionale log-likelihood für einzelnen Datenpunkt:

$$\ln p\left(x_k \mid \boldsymbol{\theta}\right) = -\frac{1}{2}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}\left(x_k - \mu\right)^2$$

Die Ableitungen nach μ und σ^2:

$$\nabla \ln p\left(x_k \mid \boldsymbol{\theta}\right) = \begin{pmatrix} \frac{1}{\sigma^2}\left(x_k - \mu\right) \\ -\frac{1}{2\sigma^2} + \frac{\left(x_k - \mu\right)^2}{2\sigma^4} \end{pmatrix}$$

Maximum Likelihood

Für die gesamte log-likelihood (alle Daten) bekommen wir:

$$\sum_{k=1}^{n} \frac{1}{\sigma^2}\left(x_k - \mu\right) = 0 \quad \Rightarrow \quad \hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

$$-\sum_{k=1}^{n} \frac{1}{\sigma^2} + \sum_{k=1}^{n} \frac{\left(x_k - \mu\right)^2}{\sigma^4} = 0 \quad \Rightarrow \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}\left(x_k - \mu\right)^2$$

Maximum Likelihood

Die Schätzung für die Varianz benötigt den korrekten Wert von μ, setzten wir stattdessen $\hat{\mu}$ ein, so erhalten wir ein zu kleines Ergebnis. Man kann zeigen, daß die korrekte Schätzung unter Verwendung von $\hat{\mu}$ lautet:

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{k=1}^{n}\left(x_k - \hat{\mu}\right)^2$$

Bayes‘sche Schätzung

Bei der Bayes‘schen Schätzung betrachtet man statt der wahrscheinlichsten Parameterkombination $\hat{\boldsymbol{\theta}}$ eine Plausibilitätsverteilung der Parameter $p(\boldsymbol{\theta} \mid \mathcal{D})$

Ziel: Berechnung von $P(\omega_i \mid \mathbf{x}, \mathcal{D})$

Bayes Formel:

$$P(\omega_i \mid \mathbf{x}, \mathcal{D}) = \frac{p(\mathbf{x} \mid \omega_i, \mathcal{D}) P(\omega_i \mid \mathcal{D})}{\sum_{j=1}^{c} p(\mathbf{x} \mid \omega_j, \mathcal{D}) P(\omega_j \mid \mathcal{D})}$$

Bayes‘sche Schätzung

Von der gesuchten klassenabhängigen Dichte $p(\mathbf{x} \mid \omega_i, \mathcal{D})$ nehmen wir an, dass sie nur von den Daten der jeweiligen Klasse abhängt und betrachten daher jeweils nur eine Klasse, können also ω_i weglassen.

Unser (Zwischen-)Ziel ist also die Bestimmung von $p(\mathbf{x} \mid \mathcal{D})$, das als Approximation für $p(\mathbf{x})$ dienen soll.

Bayes‘sche Schätzung

Wir setzen voraus, daß die wirkliche Verteilung parametrisch geschrieben werden kann, die Form der Funktion $p(\mathbf{x}\,|\,\boldsymbol{\theta})$ ist also bekannt.

Bei der Bayes‘schen Schätzung verwenden wir für die gesuchte Verteilung $p(\mathbf{x}\,|\,\mathcal{D})$ die Formel:

$$p(\mathbf{x}\,|\,\mathcal{D}) = \int p(\mathbf{x},\boldsymbol{\theta}\,|\,\mathcal{D})\,d\boldsymbol{\theta}$$

$$= \int p(\mathbf{x}\,|\,\boldsymbol{\theta})\,p(\boldsymbol{\theta}\,|\,\mathcal{D})\,d\boldsymbol{\theta}$$

Bayes‘sche Schätzung

Diese Verteilung $p(\mathbf{x}\,|\,\mathcal{D}) = \int p(\mathbf{x}\,|\,\boldsymbol{\theta})\, p(\boldsymbol{\theta}\,|\,\mathcal{D})\, d\boldsymbol{\theta}$ hat also nicht unbedingt die parametrische Form $p(\mathbf{x}\,|\,\boldsymbol{\theta})$, von der wir eigentlich wissen, daß sie stimmt.

Im Gegensatz dazu verwendeten wir bei maximum likelihood:

$$p(\mathbf{x}\,|\,\hat{\boldsymbol{\theta}})$$

mit

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} P(\mathcal{D}\,|\,\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta})$$

Bayes‘sche Schätzung

Wie rechnen wir nun

$$p(\mathbf{x} \mid \mathcal{D}) = \int p(\mathbf{x} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathcal{D})\, d\boldsymbol{\theta}$$

aus?

Wir brauchen: $p(\boldsymbol{\theta} \mid \mathcal{D})$

und verwenden dafür wieder die Bayes‘sche Formel:

$$p(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{\int p(\mathcal{D} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}}$$

Bayes‘sche Schätzung

Beispiel: Normalverteilung $p(x \mid \mu) = N(\mu, \sigma^2)$; dabei ist σ bekannt, μ ist unbekannt.

$$p(\mu \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mu)\, p(\mu)}{\int p(\mathcal{D} \mid \mu)\, p(\mu)\, d\mu}$$

Irgendwie müssen wir den Prior festlegen, und tun das z.B. so: $p(\mu) = N(\mu_0, {\sigma_0}^2)$

Wir erhalten: $p(\mu \mid \mathcal{D}) = \alpha \prod_{k=1}^{n} p(x_k \mid \mu) p(\mu)$

Bayes‘sche Schätzung

Als Ergebnis kommt raus:

$$p(\mu \mid \mathcal{D}) = N\left(\mu_n, {\sigma_n}^2\right)$$

mit

$$\mu_n = \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0$$

und

$$\sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}$$

Bayes‘sches Lernen

Als Ergebnis kommt raus: $p(\mu \mid \mathcal{D}) = N(\mu_n, {\sigma_n}^2)$

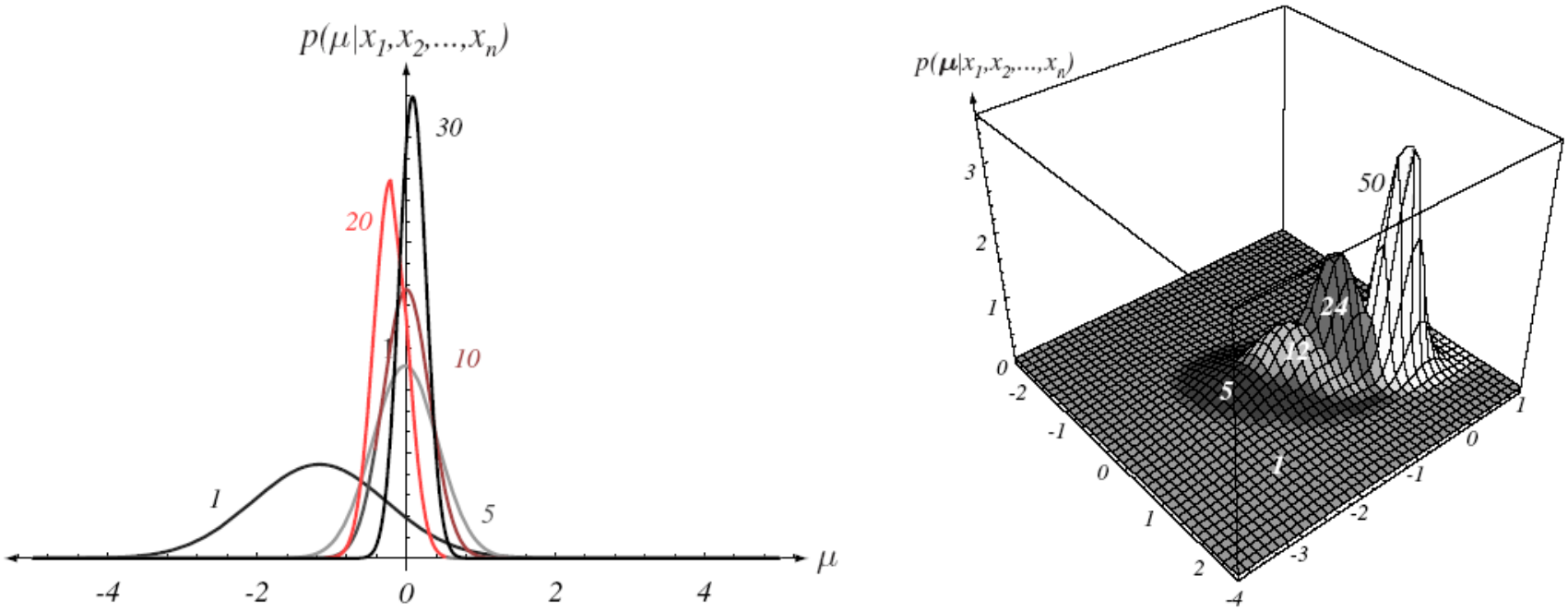


FIGURE 3.2. Bayesian learning of the mean of normal distributions in one and two dimensions. The posterior distribution estimates are labeled by the number of training samples used in the estimation. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Bayes‘sches Lernen

Um klassifizieren zu können müssen wir nur noch $p(x \mid \mathcal{D})$ berechnen:

$$p(x \mid \mathcal{D}) = \int p(x \mid \mu)\, p(\mu \mid \mathcal{D})\, d\mu$$

und erhalten: $p(x \mid \mathcal{D}) = N\left(\mu_n, \sigma^2 + \sigma_n^2\right)$

Bayes‘sches Lernen

Allgemein: Wie konvergiert $p(\mathbf{x} \mid \mathcal{D})$ gegen $p(\mathbf{x})$?

Nehmen wir unabhängige Samples an, so erhalten wir

$$p(\mathcal{D}^n \mid \boldsymbol{\theta}) = p(\mathbf{x}^n \mid \boldsymbol{\theta})\, p(\mathcal{D}^{n-1} \mid \boldsymbol{\theta})$$

Eingesetzt in Bayes Formel:

$$p(\boldsymbol{\theta} \mid \mathcal{D}^n) = \frac{p(\mathbf{x}^n \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathcal{D}^{n-1})}{\int p(\mathbf{x}^n \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathcal{D}^{n-1})\, d\boldsymbol{\theta}}$$

Bayes‘sches Lernen

Wir fangen also mit keinen Daten an und erhalten unseren Prior:

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\theta} \mid \mathcal{D}^0)$$

und bekommen dann nacheinander:

$$p(\boldsymbol{\theta} \mid \mathcal{D}^1) = p(\boldsymbol{\theta} \mid \mathbf{x}_1)$$
$$p(\boldsymbol{\theta} \mid \mathcal{D}^2) = p(\boldsymbol{\theta} \mid \mathbf{x}_1, \mathbf{x}_2)$$
$$p(\boldsymbol{\theta} \mid \mathcal{D}^3) = p(\boldsymbol{\theta} \mid \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$$
$$\vdots$$

Man nennt diese Art der Schätzung rekursive Bayes-Schätzung, ein Beispiel für ein inkrementelles oder On-Line-Lernverfahren

Bayes‘sches Lernen

Beispiel: Uniforme Verteilung

$$p(x|\theta) = U(0,\theta) = \begin{cases} 1/\theta & 0 \le x \le \theta \\ 0 & \text{sonst} \end{cases}$$

Als Prior nehmen wir an: $p(\theta) = U(0,10)$

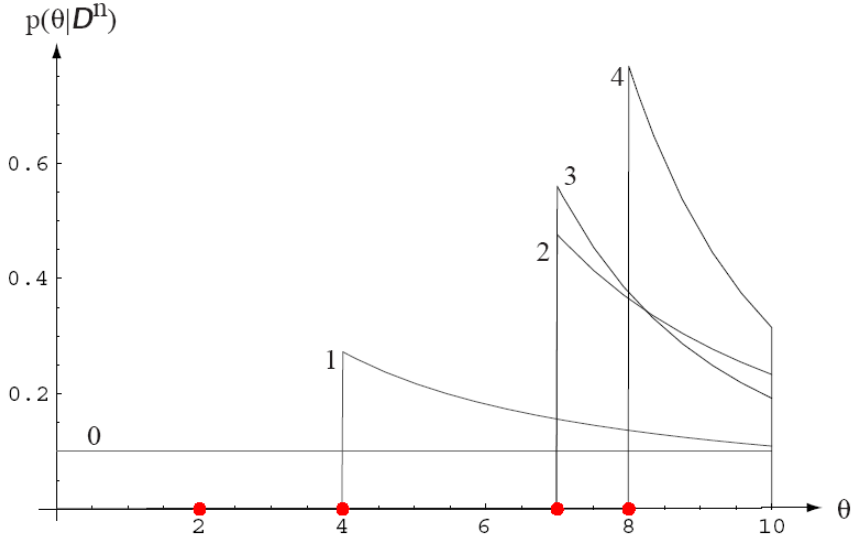
Bei einem Datenpunkt $x_1 = 4$ erhalten wir:

$$p(\theta|\mathcal{D}^1) \propto p(x|\theta)\, p(\theta|\mathcal{D}^0) = \begin{cases} 1/\theta & 4 \le \theta \le 10 \\ 0 & \text{sonst} \end{cases}$$

Bayes‘sches Lernen

Der zweite Datenpunkt $x_2 = 7$ liefert:

$$p\left(\theta \mid \mathcal{D}^2\right) \propto p\left(x \mid \theta\right) p\left(\theta \mid \mathcal{D}^1\right) = \begin{cases} 1/\theta^2 & 7 \le \theta \le 10 \\ 0 & \text{sonst} \end{cases}$$



Bayes‘sches Lernen

Aus der Verteilung für θ können wir $p\left(x \mid \mathcal{D}^2\right)$ berechnen:

$$p\left(x \mid \mathcal{D}^2\right) = \int_0^{10} p\left(x \mid \theta\right) p\left(\theta \mid \mathcal{D}^2\right) d\theta$$

$$\propto \int_7^{10} U\left(0, \theta\right) \frac{1}{\theta^2} d\theta$$

Bayes‘sches Lernen

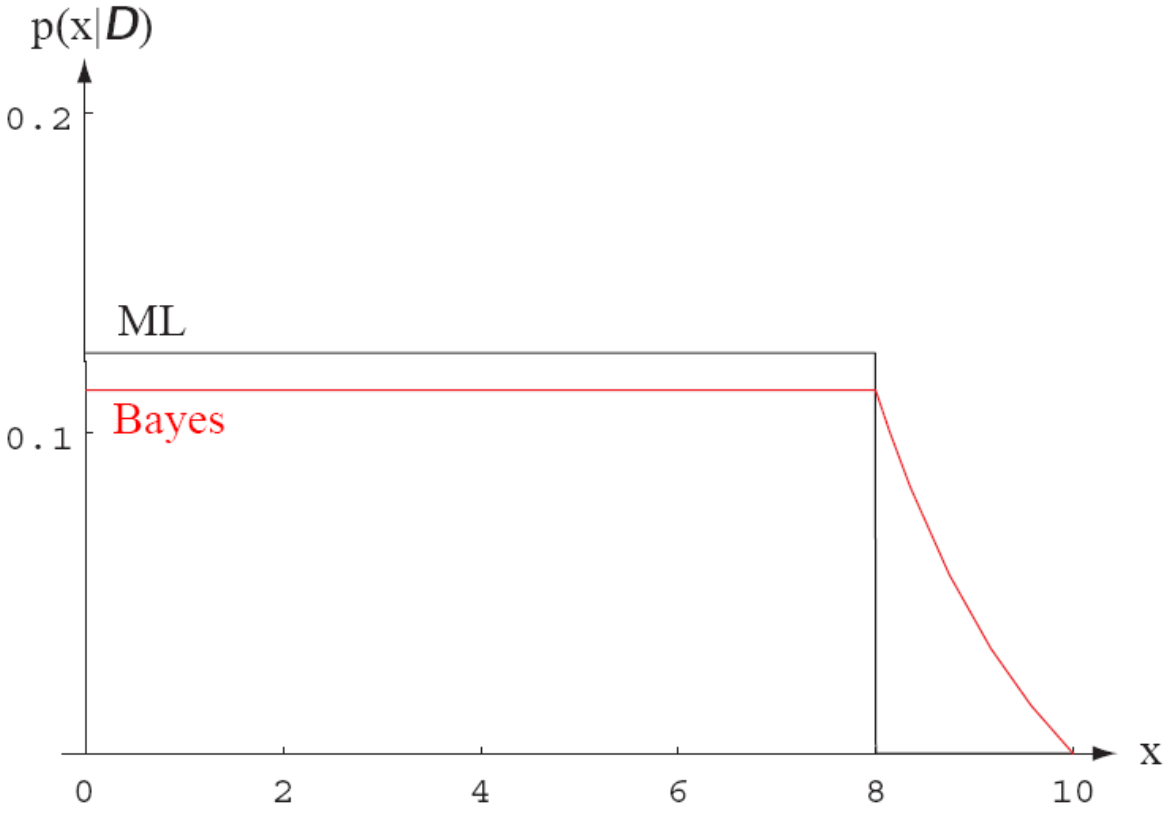
Frage: Was wäre Max Likelihood?

Antwort: $\hat{\theta} = 7$

und damit: $p\left(x \mid \hat{\theta}\right) = \left(x \mid 7\right) = U\left(0,7\right) = \begin{cases} 1/7 & 0 \le x \le 7 \\ 0 & \text{sonst} \end{cases}$

Bayes‘sches Lernen

Vergleich Maximum Likelihood und Bayes nach der Berücksichtigung aller 4 Datenpunkte:



Parameterschätzung

Grundsätzliches:

- $p(\mathbf{x}|\boldsymbol{\theta})$ wird identifizierbar genannt, wenn $\boldsymbol{\theta}$ eindeutig aus Verteilung bestimmt werden kann.
- Wenn mehrere $\boldsymbol{\theta}$ in Frage kommen, erhält man asymptotisch keine δ-Funktion, sondern mehrere Peaks.
- $p(\mathbf{x})$ ist aber (nach Integration über $\boldsymbol{\theta}$) trotzdem eindeutig.

Parameterschätzung

Grundsätzliches:

Fehlerquellen bei der Klassifizierung

- Bayes Fehler (minimal möglicher Fehler)
- Modell-Fehler: Fehler durch falsche parametrische Form von $p(\mathbf{x})$
- Schätzfehler: Falsches Modell durch zu wenige Trainingsdaten
  - Fehler =0 für unendlich viele Trainingsdaten
  - ML und Bayes asymptotisch äquivalent

Parameterschätzung

Grundsätzliches:

Wahl der A-priori-Verteilung

- Nichtinformativer Prior $p(\boldsymbol{\theta})$
- Kriterium: Invarianz

Parameterschätzung

Dimension:

Was bringt die Hinzunahme weiterer Merkmale?

Beispiel: Zwei Klassen mit gleicher Kovarianz

- Fehlerrate ist: $$P(e) = \int_{r/2}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du$$

mit $$r^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{C}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

Parameterschätzung

Für Eigenwerte von $\mathbf{C}$, die nicht 0 sind, kann also der Fehler verbessert werden, wenn entsprechendes Merkmal dazugenommen werden. Im Grenzfall kann der Fehler sogar beliebig reduziert werden.

$$P(e) = \int_{r/2}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du$$

Mahalanobis Distanz $\quad r^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{C}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$

Parameterschätzung

Beispiel:

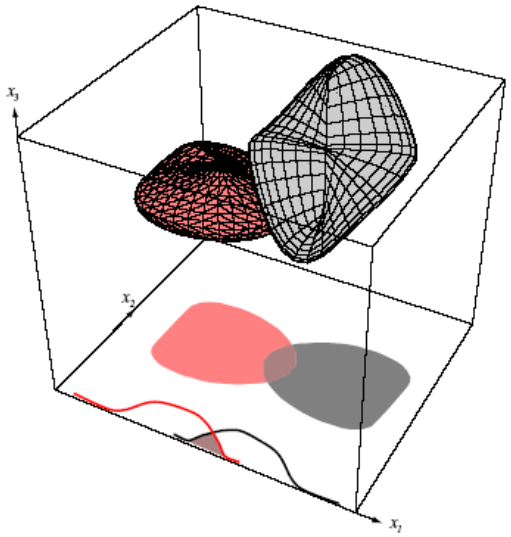


FIGURE 3.3. Two three-dimensional distributions have nonoverlapping densities, and thus in three dimensions the Bayes error vanishes. When projected to a subspace—here, the two-dimensional $x_1 - x_2$ subspace or a one-dimensional x_1 subspace—there can be greater overlap of the projected distributions, and hence greater Bayes error. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

In der Praxis leider manchmal Vergrößerung des Fehlers durch falsches Modell oder zu wenige Trainingsdaten