Klassifizierung

Klassifizierung: Beispiel

Automatische Unterscheidung zw. Lachs und Seebarsch (aus Duda et. al.: Pattern classification)

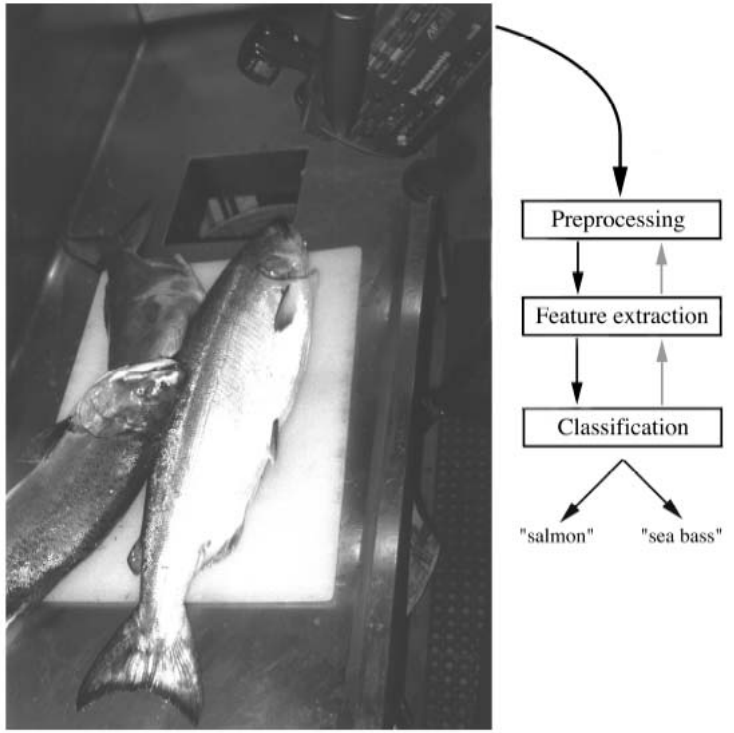


FIGURE 1.1. The objects to be classified are first sensed by a transducer (camera), whose signals are preprocessed. Next the features are extracted and finally the classification is emitted, here either "salmon" or "sea bass." Although the information flow is often chosen to be from the source to the classifier, some systems employ information flow in which earlier levels of processing can be altered based on the tentative or preliminary response in later levels (gray arrows). Yet others combine two or more stages into a unified step, such as simultaneous segmentation and feature extraction. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Unterscheidung durch Schwellwert

Länge:

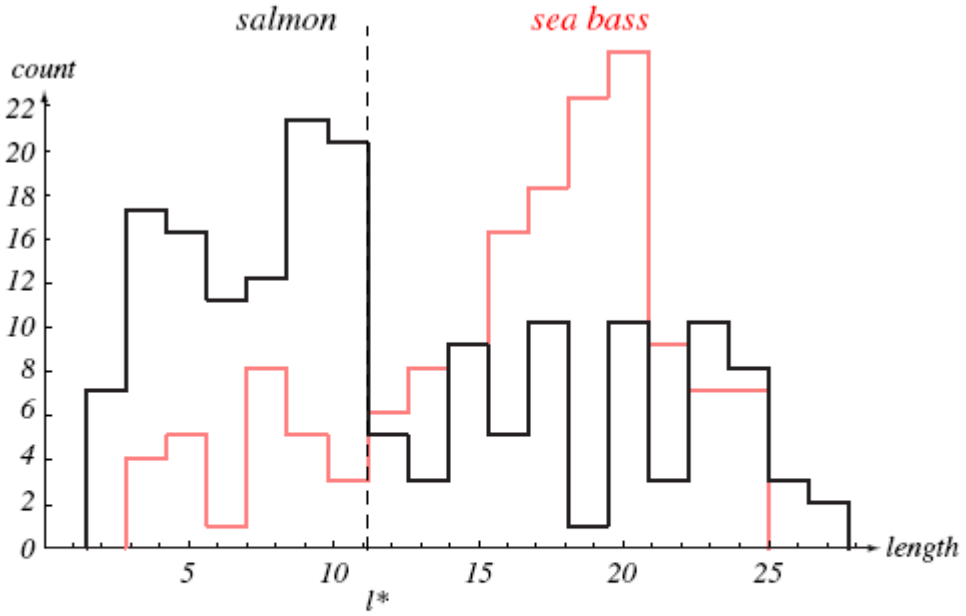


FIGURE 1.2. Histograms for the length feature for the two categories. No single threshold value of the length will serve to unambiguously discriminate between the two categories; using length alone, we will have some errors. The value marked *l** will lead to the smallest number of errors, on average. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Kosten

Wahl der Schwelle:

- Berücksichtigung von Kosten für Falschentscheidung bzw. Gewinn f. richtige Entscheidung
- Wähle Schwelle, die minimale Kosten verursacht

Unterscheidung durch Schwellwert

Helligkeit:

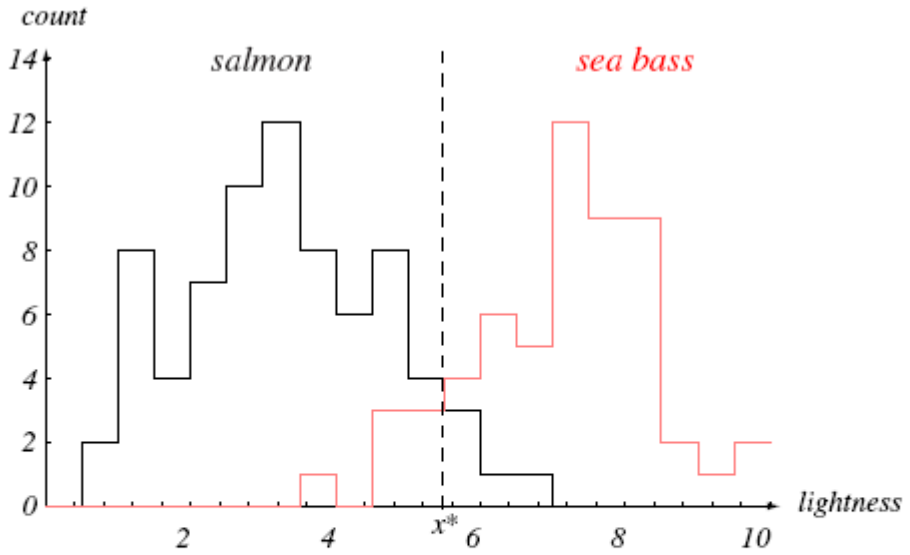


FIGURE 1.3. Histograms for the lightness feature for the two categories. No single threshold value x^* (decision boundary) will serve to unambiguously discriminate between the two categories; using lightness alone, we will have some errors. The value x^* marked will lead to the smallest number of errors, on average. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Mehrdimensionale Feature-Vektoren

Kombination mehrerer Merkmale:

- Finden von Entscheidungsgrenzen in mehrdimensionalen Merkmalsräumen

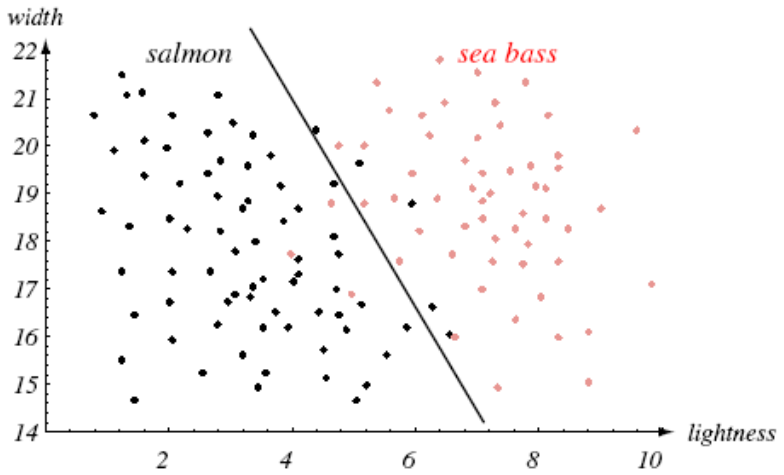


FIGURE 1.4. The two features of lightness and width for sea bass and salmon. The dark line could serve as a decision boundary of our classifier. Overall classification error on the data shown is lower than if we use only one feature as in Fig. 1.3, but there will still be some errors. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Generalisierung

Komplizierte Grenze:

- Perfekt für Trainingsdaten:

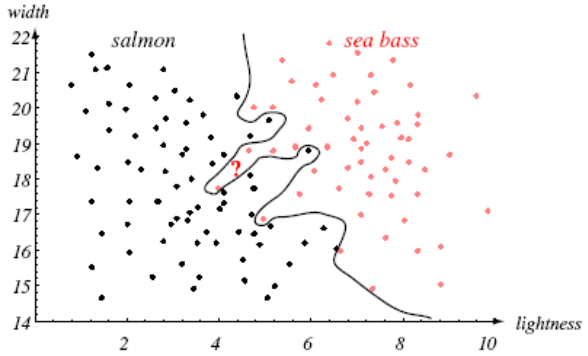


FIGURE 1.5. Overly complex models for the fish will lead to decision boundaries that are complicated. While such a decision may lead to perfect classification of our training samples, it would lead to poor performance on future patterns. The novel test point marked ? is evidently most likely a salmon, whereas the complex decision boundary shown leads it to be classified as a sea bass. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

- Schlecht für neue Daten.

  Kriterium : Einfachheit der Grenze

Generalisierung

Einfachere Grenze:

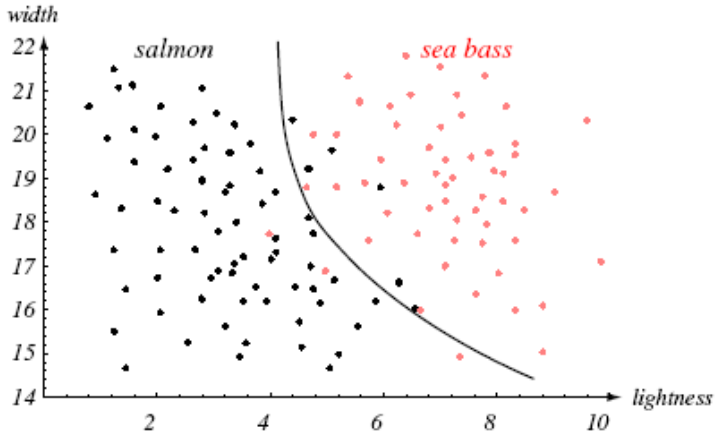


FIGURE 1.6. The decision boundary shown might represent the optimal tradeoff between performance on the training set and simplicity of classifier, thereby giving the highest accuracy on new patterns. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Bayes’sche Entscheidungstheorie

Parameterschätzung

Grundlagen nach „Data Analysis – A Bayesian Tutorial“ von D. S. Sivia, Oxford University Press 1996

Parameterschätzung

Allgemeine Auffassung oft:

- Wahrscheinlichkeitsrechnung = Mathematik
- Statistik = Rezepte aus obskuren Kochbüchern
- Zitat aus Numerical Recipes:

> In other words, we identify the probability of the data given the parameters (which is a mathematically computable number), as the *likelihood* of the parameters given the data. This identification is entirely based on intuition. It has no formal mathematical basis in and of itself; as we already remarked, statistics is *not* a branch of mathematics!

Einführung

- Lösung: Bayes‘sche Theorie

Reverend Thomas Bayes (1702-1761)

Grundlagen

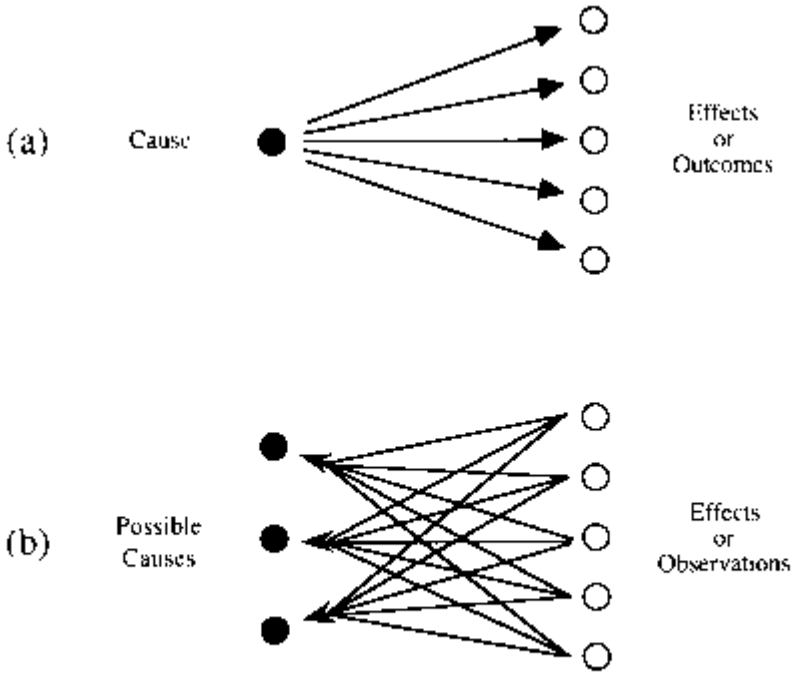


Fig. 1.1 A schematic representation of (a) deductive logic, or pure mathematics, and (b) plausible reasoning, or inductive logic.

[aus Sivia: Data Analysis, Oxford 1996]

Grundlagen

- Logisch konsistenter Rahmen für quantitative Festlegung der „Glaubwürdigkeit“ (Cox 1946 auf der Grundlage von Bayes (1763) und Laplace (1812)
- Reelle Zahl als Maß für das Vertrauen, daß eine Aussage richtig ist
- Zwei einfache Axiome:
  - Festlegung wie glaubhaft A ist, legt fest wie glaubhaft (nicht A) ist.
  - Festlegung, wie glaubhaft Y ist und wie glaubhaft X ist, wenn Y wahr ist legt fest, wie glaubhaft (X und Y) ist.

Grundlagen

Ergebnis:

- Zahlen für „Glaubhaftigkeit“ müssen sich abbilden lassen auf positive Zahlen, die der Summenregel und der Produktregel der Wahrscheinlichkeitsrechnung gehorchen.

Grundlagen

Summenregel:

$$\text{prob}\left(X \mid I\right) + \text{prob}\left(\overline{X} \mid I\right) = 1$$

Produktregel:

$$\text{prob}\left(X, Y \mid I\right) = \text{prob}\left(X \mid Y, I\right) \times \text{prob}\left(Y \mid I\right)$$

Daraus lassen sich alle anderen Regeln unter Verwendung der Boole‘schen Algebra herleiten.

Boole’sche Algebra

Für Aussagen gelten die Regeln der Boole’schen Algebra. Diese können wir mit einfachen Axiomen definieren: Eine Boole’schen Algebra ist eine Menge mit zwei Verknüpfungen (Produkt und Summe) auf dieser Menge so dass für Elemente A und B aus dieser Menge gilt:

Es gibt zwei Elemente 0 and 1, die nicht gleich sind.		**[A1]**
AB = BA	**A+B = B+A**	**[A2]**
A(B+C) = (AB)+(AC)	**A+(BC) = (A+B)(A+C)**	**[A3]**
1A = A	**0+A = A**	**[A4]**
A¬A = 0	**A+¬A = 1**	**[A5]**

In einer Zeile stehen jeweils zwei duale Axiome, die sich durch Vertauschung von Produkt und Summe, sowie von 0 und 1 ergeben. Im folgenden werden die linken Axiome mit a, die rechten mit b gekennzeichnet([A2a] bedeutet also AB = BA).

Aus Harri Valpola: http://users.ics.aalto.fi/harri/thesis/bayesformulas.html

Grundlagen

Aus den Axiomen werden folgende Lemmata abgeleitet:

¬¬A = A		[L1]
AA = A	A+A = A	[L2]
¬1 = 0	¬0 = 1	[L3]
AB = 0 & A+B = 1 => B = ¬A		[L4]
0A = 0	1+A = 1	[L5]
A(A+B) = A	A+AB = A	[L6]
A(BC) = (AB)C	A+(B+C) = (A+B)+C	[L7]
¬A(AB) = 0	¬A+(A+B) = 1	[L8]
¬(AB) = ¬A+¬B	¬(A+B) = ¬A¬B	[L9]
AB = 1 => A = 1	A+B = 0 => A = 0	[L10]

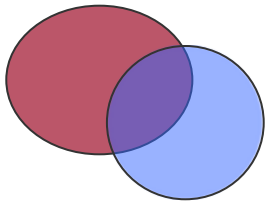
Wenn nur die Elemente 0 und 1 verwendet werden, erhält man die Boole'sche Logik. 0 wird dabei als "falsch", 1 als "wahr" interpretiert. Produkt steht für das "und", die Summe für das "oder".

Aus Harri Valpola: http://users.ics.aalto.fi/harri/thesis/bayesformulas.html

Grundlagen

Beispiel: Herleitung der erweiterten Summenregel:

P(A+B | C) =	[L1]
P(¬¬(A+B)) | C) =	[L7b]
P(¬(¬A¬B) | C) =	[Sum Rule]
1 - P(¬A¬B | C) =	[Product Rule]
1 - P(¬A | C) P(¬B | ¬AC) =	[Sum Rule]
1 - P(¬A | C) [1 - P(B | ¬AC)] =	
1 - P(¬A | C) + P(¬A | C) P(B | ¬AC) =	[Sum Rule]
P(A | C) + P(¬A | C) P(B | ¬AC) =	[Product Rule]
P(A | C) + P(¬AB | C) =	[A2a]
P(A | C) + P(B¬A | C) =	[Product Rule]
P(A | C) + P(B | C) P(¬A | BC) =	[Sum Rule]
P(A | C) + P(B | C) [1 - P(A | BC)] =	
P(A | C) + P(B | C) - P(B | C) P(A | BC) =	[Product Rule]
P(A | C) + P(B | C) - P(BA | C) =	[A2a]
P(A | C) + P(B | C) - P(AB | C)	



Aus Harri Valpola: http://users.ics.aalto.fi/harri/thesis/bayesformulas.html

Grundlagen

Aus der Produktregel folgt u.a. der Satz von Bayes:

$$\text{prob}(X \mid Y) = \frac{\text{prob}(Y \mid X) \times \text{prob}(X)}{\text{prob}(Y)}$$

Herleitung:

$$\begin{aligned} &\text{prob}(X, Y) = \text{prob}(X \mid Y) \times \text{prob}(Y) \\ &= \text{prob}(Y, X) = \text{prob}(Y \mid X) \times \text{prob}(X) \end{aligned}$$

Grundlagen

Wichtig für Datenanalyse und Parameterschätzung:

- X: Hypothese
- Y: Daten

$$\text{prob}(Hypothese \mid Daten) = \frac{\text{prob}(Daten \mid Hypothese) \times \text{prob}(Hypothese)}{\text{prob}(Daten)}$$

Grundlagen

Namen:

- A-priori-Wahrscheinlichkeit für Hypothese

$$\text{prob}(Hypothese)$$

- Likelihood-Funktion

$$\text{prob}(Daten \mid Hypothese)$$

- A-posteriori-Wahrscheinlichkeit

$$\text{prob}(Hypothese \mid Daten)$$

Grundlagen

- A-priori-Wahrscheinlichkeit für Daten („evidence“)

$$\text{prob}(\textit{Daten})$$

Grundlagen

Da $\text{prob}(Daten)$ unabhängig von der Hypothese ist, erhalten wir:

$$\text{prob}(Hypothese \mid Daten)$$
$$\propto \text{prob}(Daten \mid Hypothese) \times \text{prob}(Hypothese)$$

Grundlagen

„Marginalisation“, d.h. Elimination von nicht benötigten Variablen:

$$\text{prob}(X) = \sum_{k=1}^{M} \text{prob}(X, Y_k)$$

das gilt, wenn die Y_k sich zu eins addieren und gegenseitig ausschließen.

Grundlagen

Kontinuierliche Form der Gleichung, z.B. wenn Y der Wert eines kontinuerlichen Parameters ist:

$$\text{prob}(X) = \int_{-\infty}^{\infty} \text{prob}(X, Y) dY$$

In diesem Fall bezeichnet $\text{prob}(X, Y)$ eine Wahrscheinlichkeitsdichtefunktion:

$$\text{prob}(X, Y = y) = \lim_{\delta y \to 0} \frac{\text{prob}(X, y \leq Y < y + \delta y)}{\delta y}$$

Grundlagen

Klassische Wahrscheinlichkeit vs.
Plausibilität im Bayes‘schen Sinn:

- Klassische Wahrscheinlichkeit als relative Häufigkeit bei unendlich vielen Experimenten
- Wie aber anwenden auf Schätzung z.B. der Masse von Saturn aus wenigen Beobachtungen (Laplace)? Ensemble von Welträumen?

Grundlagen

Laplace berechnete die A-posteriori-Wahrscheinlichkeitsdichte für die Masse unter der Voraussetzung der verfügbaren Daten. Er schloß daraus, daß

'...it is a bet of 11,000 to 1 that the error of this result is not 1/100th of its value'.

und behielt bis heute recht (0.63% daneben).

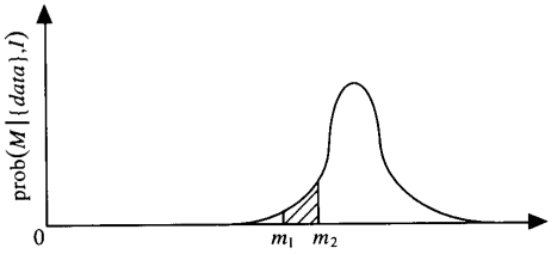


Fig. 1.2 A schematic illustration of the result of Laplace's probability analysis of the mass of Saturn.

Grundlagen

Wahrscheinlichkeit in der Bayes‘schen Auffassung

repräsentiert den Wissensstand über eine physikalische Wirklichkeit, nicht diese selbst.

Beispiel:

- Man zieht verdeckt einen von 5 roten und 7 grünen Bällen: Wahrscheinlichk. f. rot: 5/12
- Beim zweiten Zug (ohne Zurücklegen) hängt Wahrscheinlichkeit vom ersten Zug ab.

Grundlagen

Wie ist Wahrscheinlichkeit beim ersten Zug (dessen Ergebnis wir noch nicht kennen), wenn das Ergebnis des zweiten schon bekannt ist?

Immer noch 5/12?

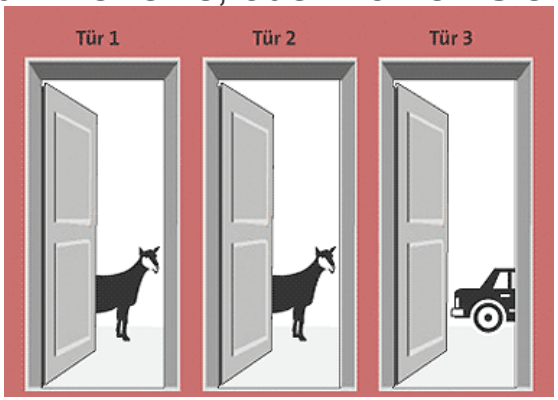
Nein, offensichtlich können wir mehr sagen (deutlich, wenn ursprünglich nur eine rote und eine grüne Kugel da waren).

Die bedingten Wahrscheinlichkeiten stellen also logische, keine kausalen Zusammenhänge dar.

Das Ziegenproblem

Die Ausgangssituation des Ziegenproblems lautet folgendermaßen: Sie sind Kandidat einer Fernsehshow und dürfen eine von drei verschlossenen Türen auswählen. Hinter einer der Türen wartet der Hauptgewinn, ein prachtvolles Auto, hinter den anderen beiden steht jeweils eine meckernde Ziege.

Frohgemut zeigen Sie auf eine der Türen, sagen wir Nummer eins. Doch der Showmaster, der weiß, hinter welcher Tür sich das Auto befindet, lässt sie nicht sofort öffnen, sondern sagt geheimnisvoll: »Ich zeige Ihnen mal was!« Er lässt eine andere Tür öffnen, sagen wir Nummer drei – und hinter dieser steht eine Ziege und glotzt erstaunt ins Publikum. Nun fragt der Showmaster lauernd: »Bleiben Sie bei Tür Nummer eins, oder wählen Sie doch lieber Nummer zwei?« Was sollten Sie tun?



aus: http://www.zeit.de/2004/48/N-Ziegenproblem, wo Sie auch die Lösung finden.

Parameterschätzung

Beispiel: Ist die Münze in Ordnung?

- Wappen oder Zahl, Wahrscheinlichkeit Z für Zahl sollte ½ sein.
- 4 von 11 waren Wappen, ist die Münze in Ordnung?
- Ansatz: Berechne $\text{prob}(Z|\{Daten\})$, also eine Wahrscheinlichkeitsdichtefunktion. Nach Satz von Bayes gilt:

$$\boxed{\text{prob}\left(Z \mid Daten\right) \propto \text{prob}\left(Daten \mid Z\right) \times \text{prob}\left(Z\right)}$$

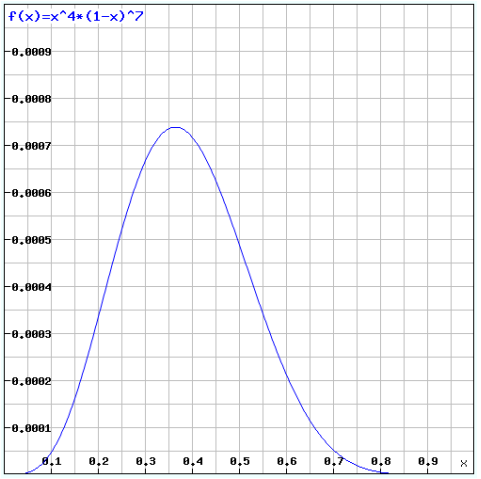
Parameterschätzung

Wir setzen an (A-priori-Wahrscheinlichkeit):

$$\text{prob}(Z) = \begin{cases} 1 & 0 \le Z \le 1 \\ 0 & \text{sonst} \end{cases}$$

Und Wahrscheinl. für R von N mal Zahl (Likelihood-Funktion):

$$\text{prob}(Daten \mid Z) \propto Z^R (1-Z)^{N-R}$$



Was ist wahrscheinlicher bei fairer Münze?

0100100101101101001000000100000101101111001100
1
1
1
0000001101000001010

Diese Ziffernfolge ist die ASCII-Codierung vom Anfang des Johannesevangeliums:
“Im Anfang war das Wort”

oder

000000000000000000000000000000000000000000000
000000000000000000000000000000000000000000
000000000000000000000000000000000000000000
000000000000000000000000000000000000000000
000000000000000000

Bayes‘sche Entscheidungstheorie

Was brauchen wir, um uns zwischen zwei Hypothesen entscheiden zu können, wenn wir keine Daten haben?

- A-priori-Wahrscheinlichkeit (Prior)
- Wenn nicht mehr bekannt, lautet die Entscheidungsregel:

$$\begin{array}{l} \omega_1, \text{ wenn } P(\omega_1) > P(\omega_2), \\ \omega_2 \text{ sonst.} \end{array}$$

Bayes‘sche Entscheidungstheorie

Wenn x bekannt ist, brauchen wir zusätzlich die Likelihood-Funktionen $p(x \mid \omega_1)$ und $p(x \mid \omega_2)$

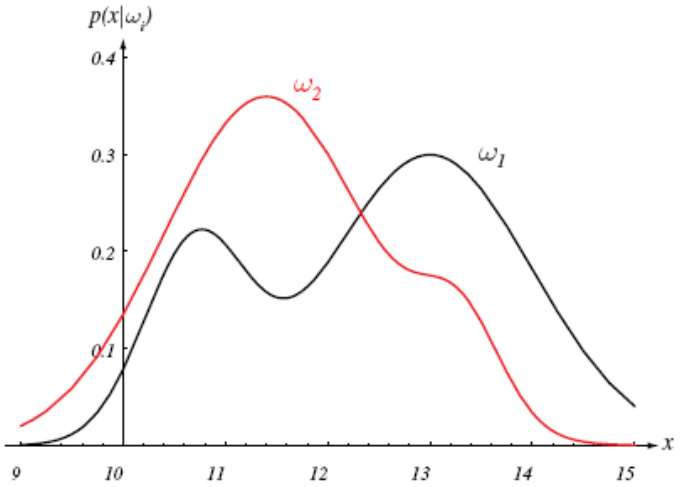


FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i. If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two types of fish. Density functions are normalized, and thus the area under each curve is 1.0. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Bayes‘sche Entscheidungstheorie

A-posteriori-Wahrscheinlichkeit (Posterior)

$$P\left(\omega_j \mid x\right) = \frac{p\left(x \mid \omega_j\right) \times P\left(\omega_j\right)}{p\left(x\right)}$$

Dabei ist Evidence (Wahrscheinlichkeit für Daten)
Skalierungsfaktor

$$p\left(x\right) = \sum_j p\left(x \mid \omega_j\right) \times P\left(\omega_j\right)$$

Bayes‘sche Entscheidungstheorie

A-posteriori Wahrscheinlichkeiten:

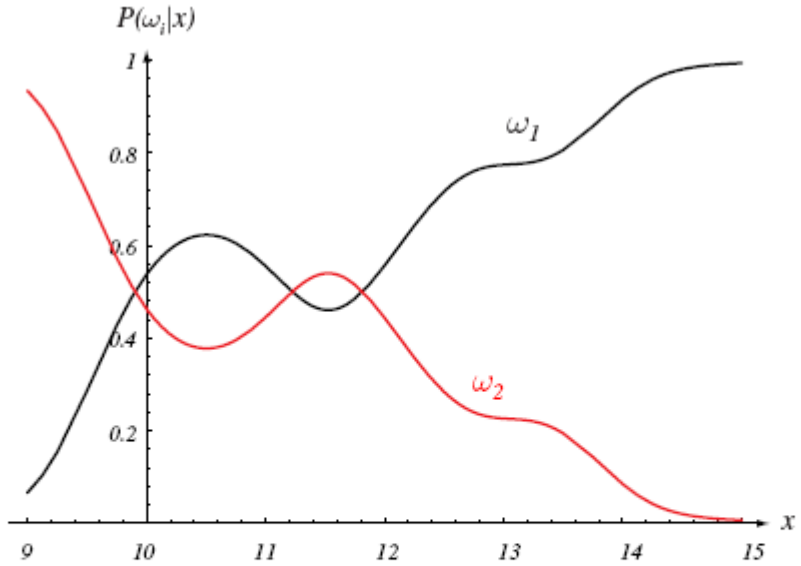


FIGURE 2.2. Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly 0.08, and that it is in ω_1 is 0.92. At every x, the posteriors sum to 1.0. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Bayes‘sche Entscheidungstheorie

Eine mögliche Entscheidungsregel:

minimiere Wahrscheinlichkeit für Fehler:

$$P(error) = \int_{-\infty}^{\infty} p(error, x)\,dx = \int_{-\infty}^{\infty} P(error \mid x)\, p(x)\,dx$$

$$P(error \mid x) = \begin{cases} P(\omega_1 \mid x) & \text{wenn Entscheidung für } \omega_2 \\ P(\omega_2 \mid x) & \text{wenn Entscheidung für } \omega_1 \end{cases}$$

Als Entscheidungsregel erhalten wir somit:

$$\boxed{\begin{array}{l} \omega_1, \text{ wenn } P(\omega_1 \mid x) > P(\omega_2 \mid x), \\ \omega_2 \text{ sonst.} \end{array}}$$

Bayes‘sche Entscheidungstheorie

Für diese Regel ist die Evidence unerheblich, die Division durch $p(x)$ bewirkt nur, daß

$$P(\omega_1 \mid x) + P(\omega_2 \mid x) = 1$$

Bayes‘sche Entscheidungstheorie

Erweiterung durch generalisierte Loss Funktion:

$$\lambda\left(\alpha_i \mid \omega_j\right)$$

sei der Verlust, der mit der Entscheidung α_i verbunden ist, wenn der wirkliche Zustand ω_j ist.

Der Erwartungswert des Verlustes der mit der Entscheidung α_i verbunden ist, wird Risiko (Risk) genannt und ist

$$R\left(\alpha_i \mid \mathbf{x}\right) = \sum_j \lambda\left(\alpha_i \mid \omega_j\right) P\left(\omega_j \mid \mathbf{x}\right)$$

Bayes‘sche Entscheidungstheorie

Entscheidungsregel – Die Entscheidungsfunktion

$$\alpha(\mathbf{x})$$

gibt an, für welchen Zustand aus $\alpha_1 \ldots \alpha_n$ wir uns bei Vorliegen von $\mathbf{x}$ entscheiden.

Gesucht ist die Funktion, die das Gesamtrisiko

$$R = \int R(\alpha(\mathbf{x}) \mid \mathbf{x})\, p(\mathbf{x}) d\mathbf{x}$$

minimiert.

Bayes‘sche Entscheidungstheorie

Die Lösung lautet:

$$\alpha(\mathbf{x}) = \arg\min\left(R(\alpha \mid \mathbf{x})\right) = \arg\min\left(\sum_j \lambda(\alpha \mid \omega_j) P(\omega_j \mid \mathbf{x})\right)$$

Das resultierende Gesamtrisiko R^* wird Bayes risk genannt und ist das kleinste Gesamtrisiko, das erreicht werden kann (optimale Entscheidungsregel)

Bayes‘sche Entscheidungstheorie

Spezialfall: Zwei Zustände ω_1 und ω_2 .

Wir kürzen ab: $\lambda_{ij} = \lambda\left(\alpha_i \mid \omega_j\right)$

Damit erhalten wir:

$$R\left(\alpha_1 \mid \mathbf{x}\right) = \lambda_{11} P\left(\omega_1 \mid \mathbf{x}\right) + \lambda_{12} P\left(\omega_2 \mid \mathbf{x}\right)$$
$$R\left(\alpha_2 \mid \mathbf{x}\right) = \lambda_{21} P\left(\omega_1 \mid \mathbf{x}\right) + \lambda_{22} P\left(\omega_2 \mid \mathbf{x}\right)$$

Die Entscheidungsregel lautet damit:

$$\alpha\left(\mathbf{x}\right) = \begin{cases} \alpha_1, & \text{wenn } \left(\lambda_{21} - \lambda_{11}\right) P\left(\omega_1 \mid \mathbf{x}\right) > \left(\lambda_{12} - \lambda_{22}\right) P\left(\omega_2 \mid \mathbf{x}\right) \\ \alpha_2 & \text{sonst} \end{cases}$$

Die Ungleichung kann durch die Bayes Regel (und Multiplikation mit Evidence) umgeformt werden:

$$(\lambda_{21}-\lambda_{11})P(\omega_1\,|\,\mathbf{x}) > (\lambda_{12}-\lambda_{22})P(\omega_2\,|\,\mathbf{x})$$

$$(\lambda_{21}-\lambda_{11})\,p(\mathbf{x}\,|\,\omega_1)P(\omega_1) > (\lambda_{12}-\lambda_{22})\,p(\mathbf{x}\,|\,\omega_2)P(\omega_2)$$

Wenn $\lambda_{21} > \lambda_{11}$ können wir stattdessen auch schreiben:

$$\alpha(\mathbf{x}) = \alpha_1, \text{wenn } \frac{p(\mathbf{x}\,|\,\omega_1)}{p(\mathbf{x}\,|\,\omega_2)} > \frac{(\lambda_{12}-\lambda_{22})P(\omega_2)}{(\lambda_{21}-\lambda_{11})P(\omega_1)}$$

Bayes‘sche Entscheidungstheorie

Das Verhältnis $$\frac{p(\mathbf{x} \mid \omega_1)}{p(\mathbf{x} \mid \omega_2)}$$

wird likelihood ratio genannt und hängt nur von $\mathbf{x}$ ab.

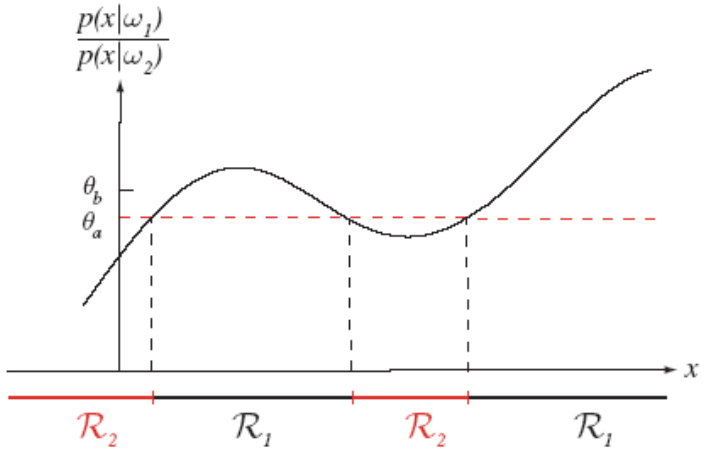


FIGURE 2.3. The likelihood ratio $p(x|\omega_1)/p(x|\omega_2)$ for the distributions shown in Fig. 2.1. If we employ a zero-one or classification loss, our decision boundaries are determined by the threshold θ_a. If our loss function penalizes miscategorizing ω_2 as ω_1 patterns more than the converse, we get the larger threshold θ_b, and hence $\mathcal{R}_1$ becomes smaller. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.