Signalentdeckungstheorie, Dichteschätzung

Mustererkennung und Klassifikation, Vorlesung No. 6[1]

M. O. Franz

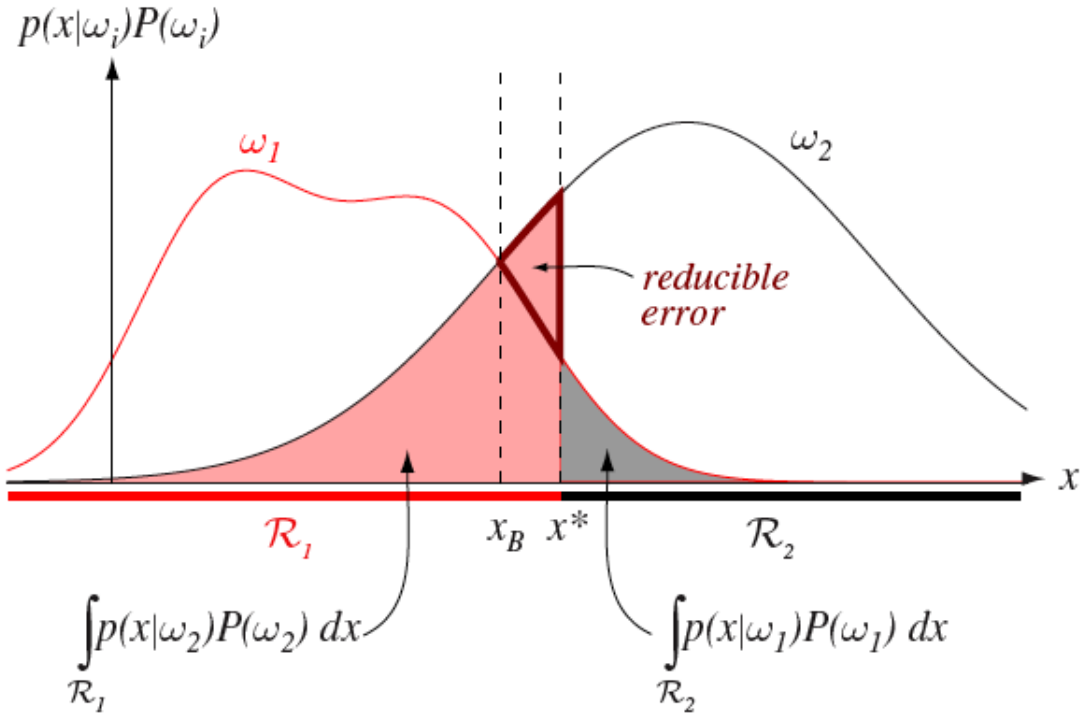
15.11.2007

[1] falls nicht anders vermerkt, sind die Abbildungen entnommen aus Duda et al., 2001.

Übersicht

Übersicht

Klassifikationsfehler



Es gibt zwei Möglichkeiten für Fehler:

- Beobachtung $\mathbf{x}$ fällt in in $\mathcal{R}_2$, der wahre Naturzustand ist ω_1.
- Beobachtung $\mathbf{x}$ fällt in in $\mathcal{R}_1$, der wahre Naturzustand ist ω_2.

Bayesfehler

Fehlerwahrscheinlichkeit:

$$\begin{aligned} p(\text{Fehler}) &= p(\mathbf{x} \in \mathcal{R}_2, \omega_1) + p(\mathbf{x} \in \mathcal{R}_1, \omega_2) \\ &= p(\mathbf{x} \in \mathcal{R}_2 | \omega_1) p(\omega_1) + p(\mathbf{x} \in \mathcal{R}_1 | \omega_2) p(\omega_2) \\ &= \int_{\mathcal{R}_1} p(\mathbf{x}|\omega_1) p(\omega_1)\, d\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}|\omega_2) p(\omega_2)\, d\mathbf{x} \end{aligned}$$

Der minimale Fehler wird erreicht, wenn die Entscheidungsgrenze auf dem Punkt liegt, an dem beide A-posteriori-Wahrscheinlichkeiten gleich groß sind ($\Rightarrow$ Bayessche Entscheidungsregel).

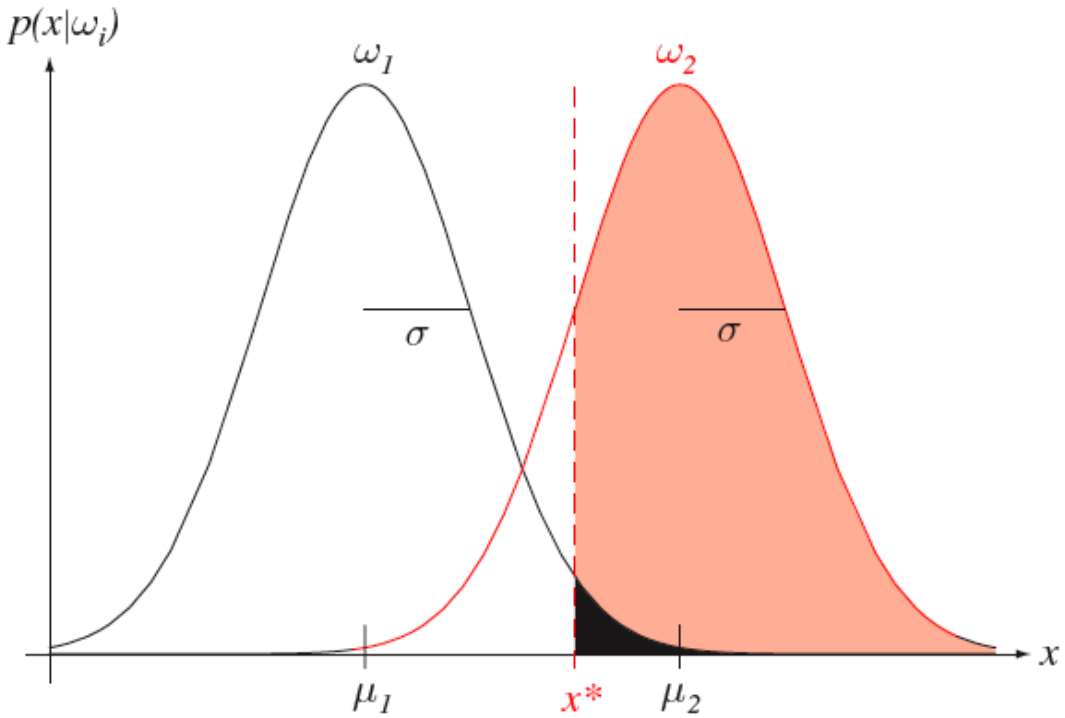
Der dadurch erreichbare minimale Fehler heißt Bayesfehler.

4 mögliche Ausgänge eines Detektionsexperimentes

2 Zustände in der Signaldetektion (Dichotomie): ω_1 - zu detektierendes Signal ist tatsächlich da, ω_2 - Signal nicht vorhanden.

- Hit: Vorhandenes Signal wurde detektiert (auch richtig positiv). Detektionswahrscheinlichkeit (Sensitivität, Richtig-Positiv-Rate): $p(x \in \mathcal{R}_1|\omega_1)$.
- Korrekte Rückweisung: Signal ist nicht vorhanden und wurde auch nicht detektiert (auch richtig negativ). Richtig-Negativ-Rate (Spezifität): $p(x \in \mathcal{R}_2|\omega_2)$.
- Fehlalarm: Detektion trotz nicht vorhandenem Signal (auch falsch positiv, Fehler 1. Art). Fehlalarmrate (Falsch-Positiv-Rate): $p(x \in \mathcal{R}_1|\omega_2)$.
- Miss: Vorhandenes Signal wurde nicht detektiert (auch falsch negativ, Fehler 2. Art). Falsch-Negativ-Rate: $p(x \in \mathcal{R}_2|\omega_1)$.

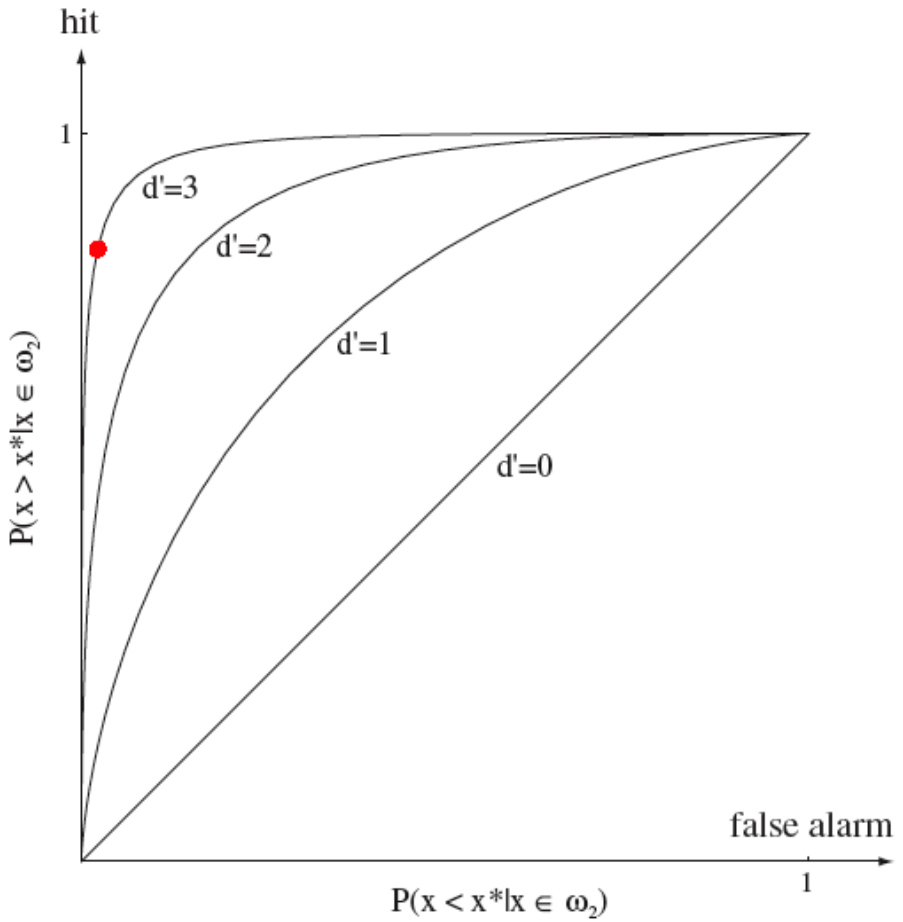
Beispiel: Signaldetektion bei gaußverteiltem Rauschen



Maß für Unterscheidbarkeit (unabhängig vom Klassifikator):

$$d' = \frac{|\mu_2 - \mu_1|}{\sigma}$$

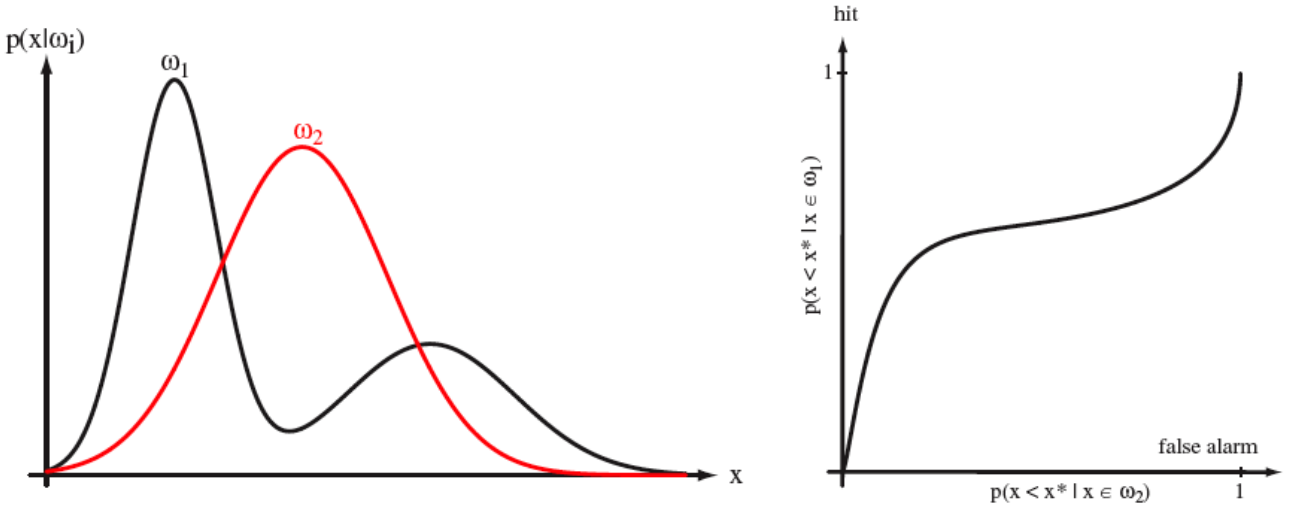
ROC-Kurve



ROC: Receiver Operating Characteristic

Darstellung der Detektions- vs. Fehlalarmrate für alle Schwellwerte

ROC-Kurven bei mehrdimensionalen Problemen



Bei mehrdimensionalen Problemen gibt es i.A. zu jeder Hit-Rate viele mögliche Entscheidungsflächen mit unterschiedlichen Fehlalarmraten.

ROC-Kurven werden ebenfalls über einen Parameter gewonnen, der in beiden Extremen durch den Ursprung und durch (1,1) geht.

Fläche unter ROC-Kurve (AUC - area under curve) ist ein häufiges Maß für die Qualität eines Klassifikators.

Aufgabe 5 (Signalentdeckungstheorie)

In einer Sortiermaschine sollen Kaffeebohnen von irrtümlich beigemischten Steinen getrennt werden. Kaffee kommt in 60 % aller Fälle vor. Die Wahrscheinlichkeiten der Längen für beide Kategorien sind in folgender Tabelle aufgetragen:

Länge l (mm)	5	6	7	8	9	10	11	12
p(l | Kaffee)	0.074	0.221	0.368	0.221	0.074	0.037	0.005	0.000
p(l | Stein)	0.000	0.005	0.037	0.074	0.221	0.368	0.221	0.074

1. Tragen Sie beide A-posteriori-Wahrscheinlichkeiten in einem Histogramm auf.

2. Ab welcher Länge würden Sie dasc Objekt wegwerfen?

3. Für den von Ihnen gewählten Schwellwert: was ist die zugehörige Detektionswahrscheinlichkeit für Steine? Was ist die Fehlalarmrate? Wie wahrscheinlich sind korrekte Identifikationen von Steinen? Wie häufig kommen Steine durch?

4. Tragen Sie die zugehörige ROC-Kurve auf. Was ist der AUC-Wert?
In einer zweiten Charge befinden sich etwas kleinere Kaffebohnen mit gleicher Häufigkeit:

Länge l (mm)	5	6	7	8	9	10	11	12
p(l | Kaffee)	0.221	0.368	0.221	0.100	0.070	0.010	0.005	0.000
p(l | Stein)	0.000	0.005	0.037	0.074	0.221	0.368	0.221	0.074

5. Wie ist hier der AUC-Wert? Welches Klassifikationsproblem ist leichter?

Übersicht

Parametrische Klassifikation

Die bisher vorgestellten Techniken waren parametrisch, d.h. die Form der zugrundeliegenden Verteilungen wurde als bekannt vorausgesetzt:

- Bayes-Klassifikator: A-posteriori-Wahrscheinlichkeit $p(\omega_i|x)$ muß analytisch oder als Histogramm gegeben sein.
- Spezialfall Gaußverteilung: $p(\omega_i|x) = \mathcal{N}(\mu, \sigma^2)$

Es gibt eine Vielzahl weiterer parametrischer Techniken, bei denen einzelne Parameter von Verteilungen geschätzt werden (z.B. μ, σ in $\mathcal{N}(\mu, \sigma^2)$): Maximum-Likelihood- oder Bayessche Parameterschätzung.

Problem: Die meisten parametrisierten Verteilungsmodelle passen nicht auf die tatsächlich vorkommenden Dichten.

Nichtparametrische Klassifikation

Nichtparametrische Klassifikationstechniken können bei beliebigen Verteilungen, da sie (fast) keine Annahmen über die vorkommenden Verteilungen machen.

Ansätze:

- Schätzung der Wahrscheinlichkeit mit Histogrammen bei Zufallsvariablen mit diskreten Werten
- Dichteschätzung bei kontinuierlichen Zufallsvariablen
- Nächste-Nachbar-Techniken: direkte Schätzung der Entscheidungsfunktion aus den Daten

Schätzung von Wahrscheinlichkeiten (kont. ZV)

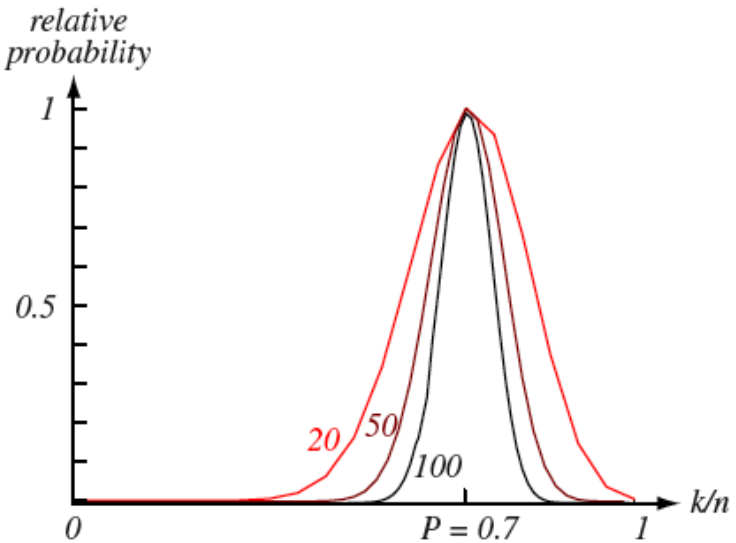
Wahrscheinlichkeit p, daß ein Vektor x in die Region $\mathcal{R}$ fällt:

$$p = \int_{\mathcal{R}} p(x)\, dx$$

Bei n unabhängigen, identisch verteilten (i.i.d.) Versuchen mit k Treffern:

$$p(k|p) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{(Binomialverteilung)}$$

mit Erwartungswert $\mathcal{E}[k] = np$.



$\Rightarrow$ Schätzung von p:
$p = k/n$

Die Verteilung von k/n hat bei großem n eine scharfe Spitze bei p, d.h. k/n ist ein guter Schätzer für p.

Dichteschätzung

Wahrscheinlichkeitsdichte einer kontinuierlichen Variablen ist definiert als Wahrscheinlichkeit pro Volumen. Für hinreichend kleine Regionen mit Volumen V gilt für stetiges $p(x)$

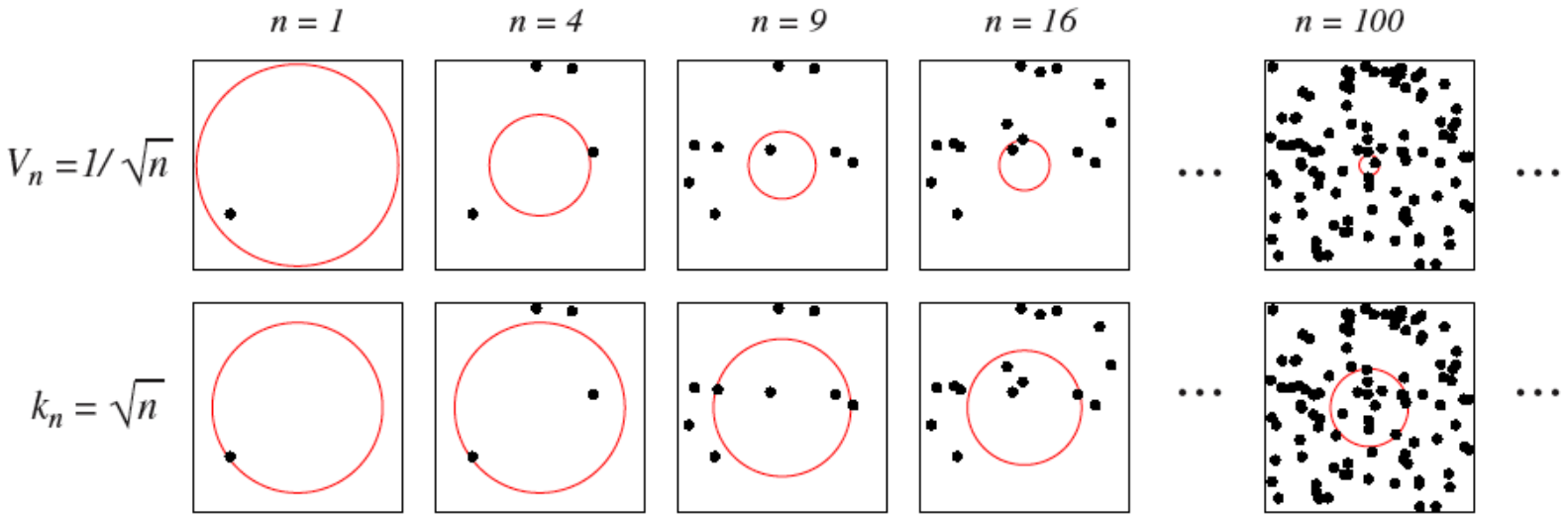
$$p = \int_{\mathcal{R}} p(x)\,dx \approx p(x)V.$$

Mit dem vorherigen Schätzer $p = k/n$ ergibt sich

$$p(x) \approx \frac{k/n}{V}$$

Problem: Wenn man keine räumlich geglättete Schätzung will, muß $V \to 0$ laufen. Bei endlich vielen Daten fällt aber irgendwann kein x_i mehr in $\mathcal{R}$, oder womöglich fällt ein x_i genau auf x. Beide Schätzungen sind nutzlos $\Rightarrow$ Für reale Schätzung ist Glättung und Streuung der k/n unvermeidbar!

Ideale Dichteschätzung



Volumen nimmt in Abhängigkeit von n ab, z.B. mit $V_n = 1/\sqrt{n}$ (oben) oder umschließt einen Teil der Daten, z.B. $k_n = \sqrt{n}$ (unten). In beiden Fällen konvergiert die Schätzung gegen die korrekte Dichte (bei unendlich vielen Daten!).

Übersicht

Dichteschätzung mit einem Kern

Beispiel: d-dimensionaler Hyperwürfel mit Kantenlänge h um x:

$$V = h^d$$

Schreibweise mit einem Kern (oder Fensterfunktion):

$$\varphi(u) = \begin{cases} 1 & |u_j| < 1/2 \\ 0 & \text{sonst} \end{cases}$$

d.h. $\varphi((x - x_i)/h)$ ist 1, wenn x_i in den Hyperwürfel fällt, sonst 0.

Anzahl der x_i im Hyperwürfel:

$$k = \sum_{i=1}^{n} \varphi\left(\frac{x - x_i}{h}\right)$$

Dichteschätzung mit $p(x) \approx (k/n)/V$:

$$p(x) \approx \frac{1}{nV} \sum_{i=1}^{n} \varphi\left(\frac{x - x_i}{h}\right)$$

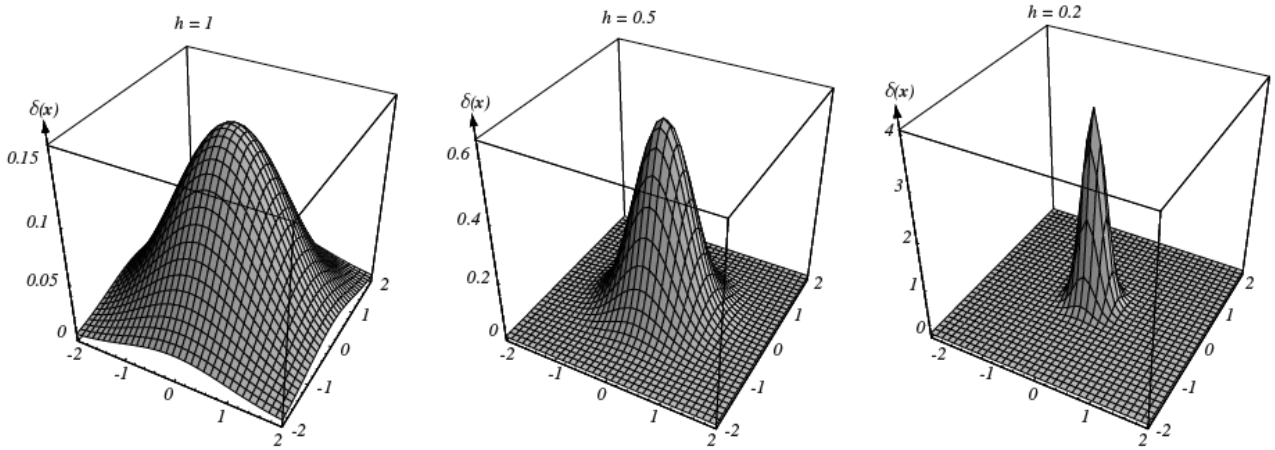
Kerne zur Dichteschätzung

Die Würfelfunktion ist ein Beispiel für einen Kern zur Dichteschätzung. Andere Kerne sind ebenfalls erlaubt (z.B. Gaußfunktionen, Exponentialfunktionen usw.).

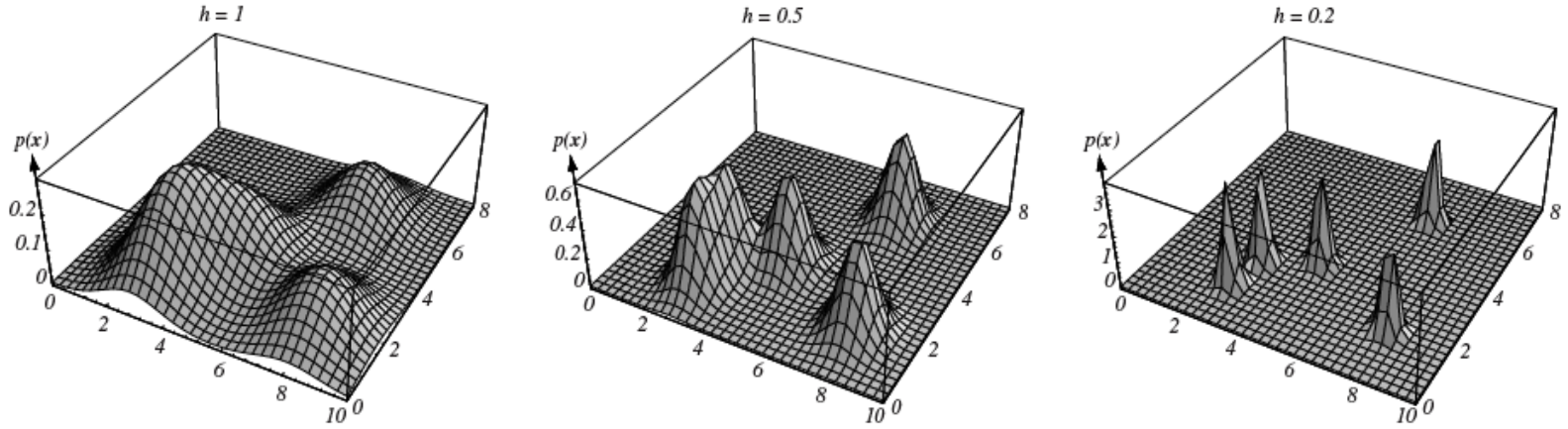
Voraussetzung: Die Schätzung muß eine legitime Dichtefunktion ergeben. Das läßt sich garantieren, wenn $\varphi(x)$ ebenfalls eine Dichtefunktion ist, d.h.

$$\varphi(x) \geq 0 \quad \text{und} \quad \int \varphi(u)\, du = 1$$

Verhalten von $1/V\varphi(x/h)$ für verschiedene Breiten h (mit $V = h^d$):

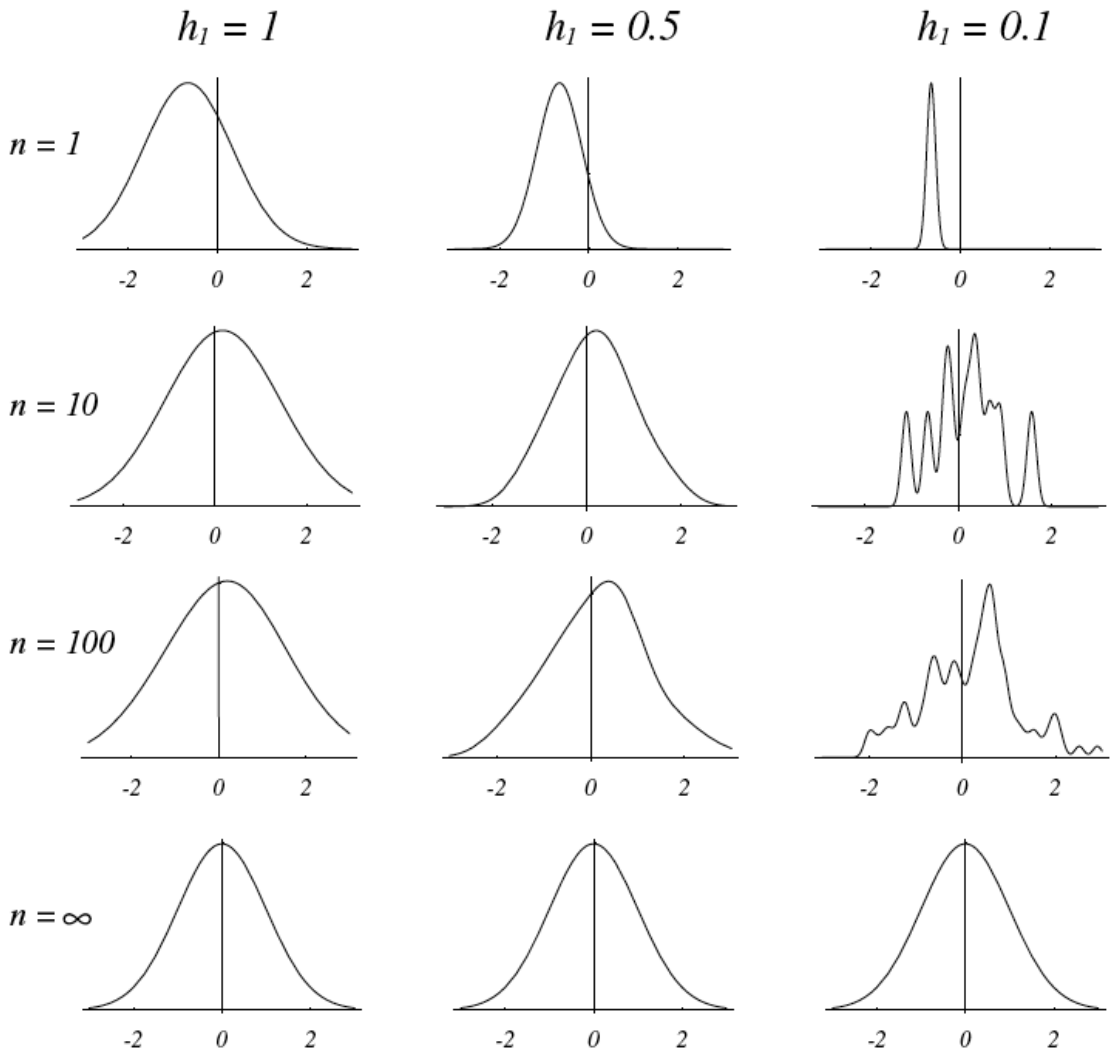


Effekt der Kernbreite auf Dichteschätzung

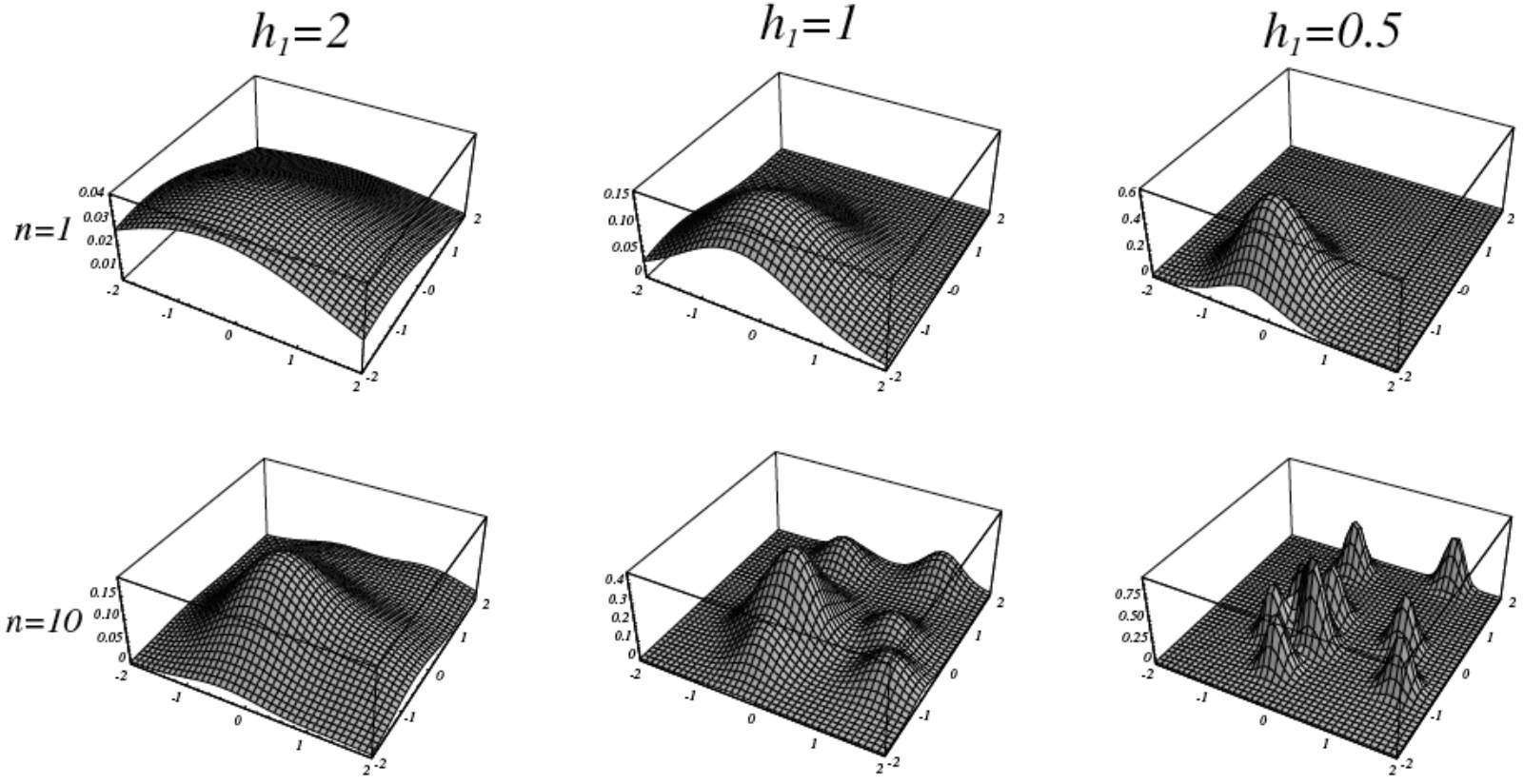


- Je breiter, desto glatter und gröber die Schätzung (Interpolationseffekt)
- Je schmaler, desto höher die Auflösung, aber auch der Rauscheffekt der einzelnen Beispiele
- Für unendlich viele Beispiele, vernünftige Kerne und langsam schrumpfende Breite ($< 1/n$) läßt sich zeigen, daß der Schätzer gegen die wahre Dichte konvergiert (s. Duda et al., 2001).

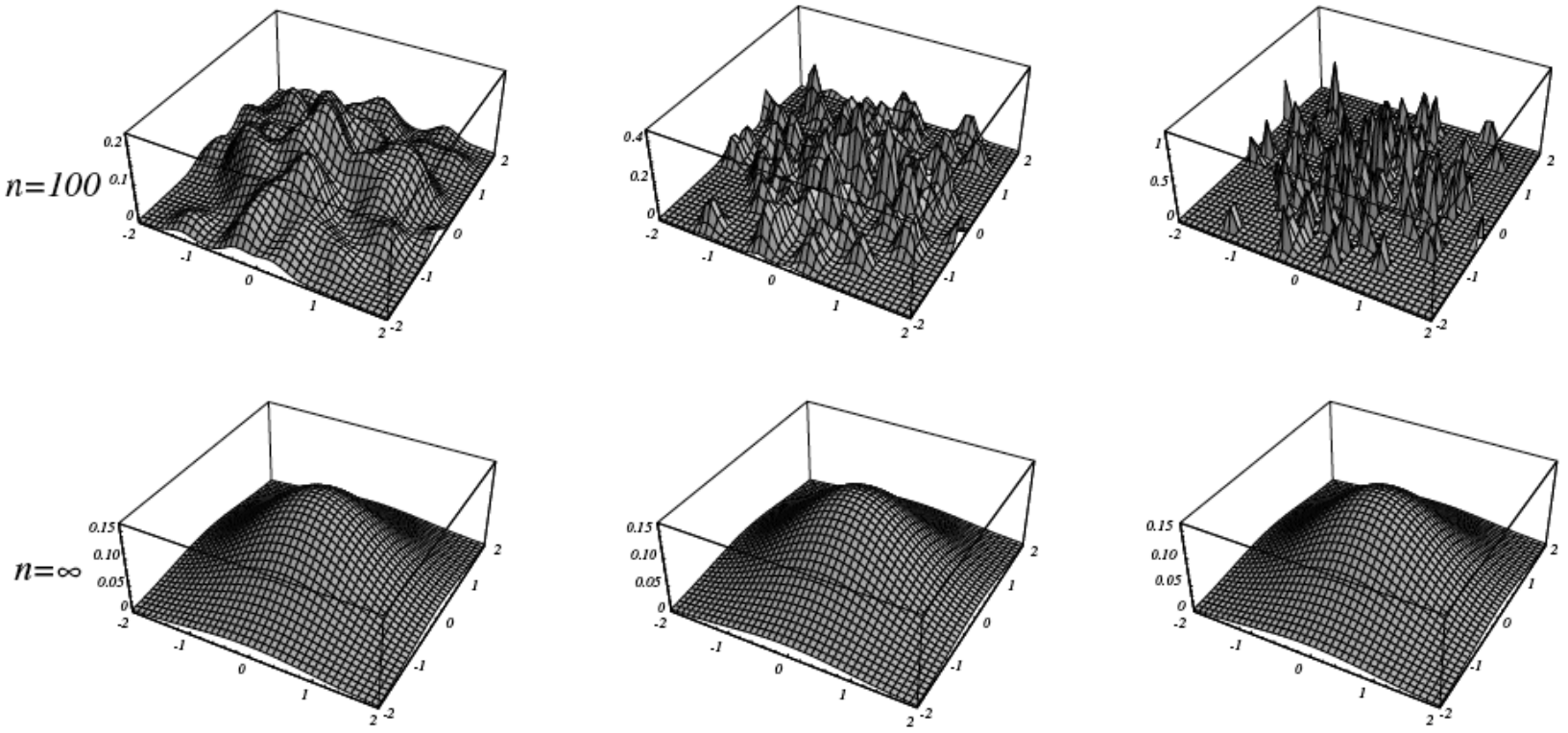
Beispiel: Dichteschätzung einer eindimensionalen Normalverteilung mit Gaußkern



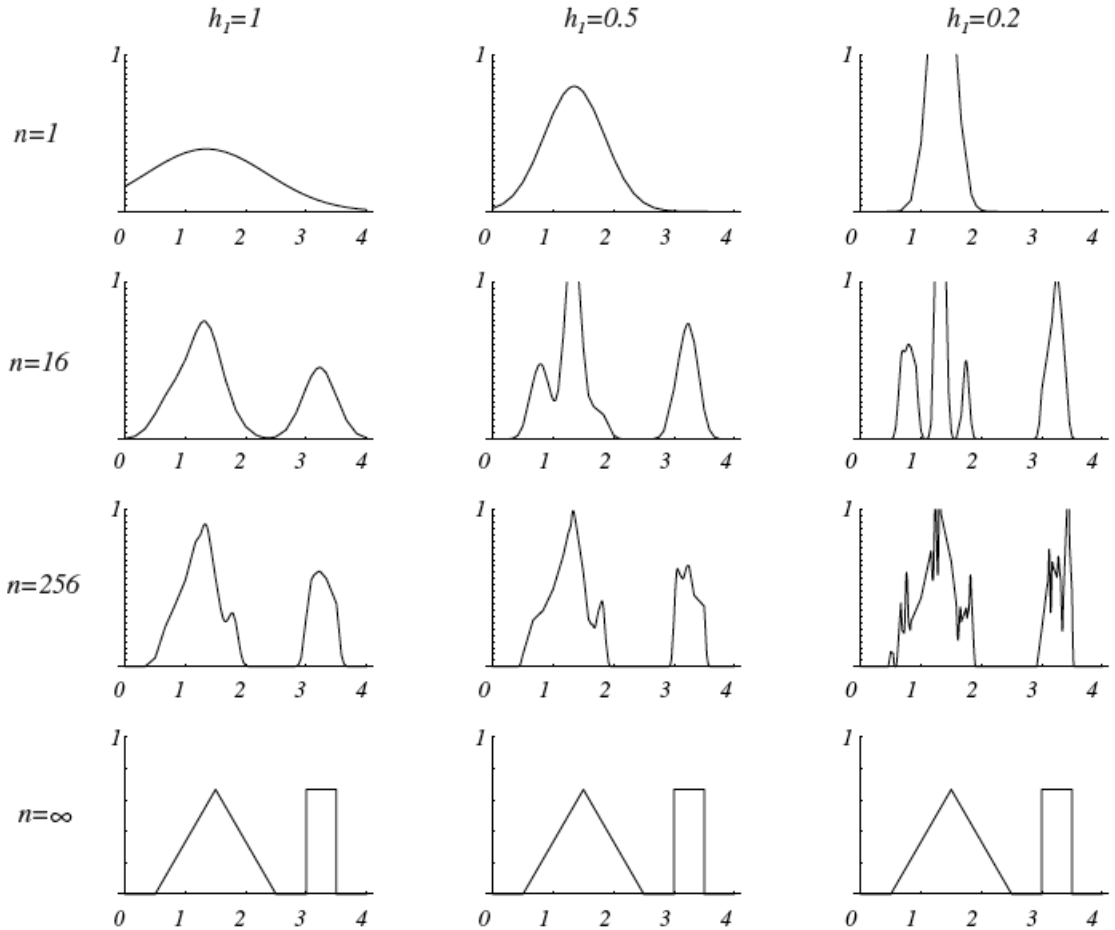
Beispiel: Dichteschätzung einer zweidimensionalen Normalverteilung mit Gaußkern (1)



Beispiel: Dichteschätzung einer zweidimensionalen Normalverteilung mit Gaußkern (2)



Beispiel: Dichteschätzung einer bimodalen Verteilung mit Gaußkern



Klassifikation mit Kerndichteschätzern

1. Schätzung der A-priori-Wahrscheinlichkeiten $p(\omega_i)$ der Kategorien über ihre relativen Häufigkeiten
2. Wahl eines geeigneten Kerns und einer geeigneten Breite (Vorsicht: mit einer zu kleinen Breite lassen sich die Trainingsdaten perfekt modellieren, aber bei der Klassifikation von bisher unbekannten Beispielen ist eine schlechte Leistung zu erwarten!)
3. Schätzung der Likelihood $p(x|\omega_i)$ für jede Kategorie mit Kerndichteschätzer.
4. Klassifikation mit Bayes-Klassifikator

Vorteil: Allgemeingültigkeit - funktioniert im Prinzip für beliebige Verteilung.

Nachteil: braucht riesige Trainingssets, funktioniert nur bei niedrigdimensionalen Daten (”Fluch der Dimensionalität”)

Beispiel: Klassifikation mit Kerndichteschätzern