Lineare Klassifikatoren

Mustererkennung und Klassifikation, Vorlesung No. 8[1]

M. O. Franz

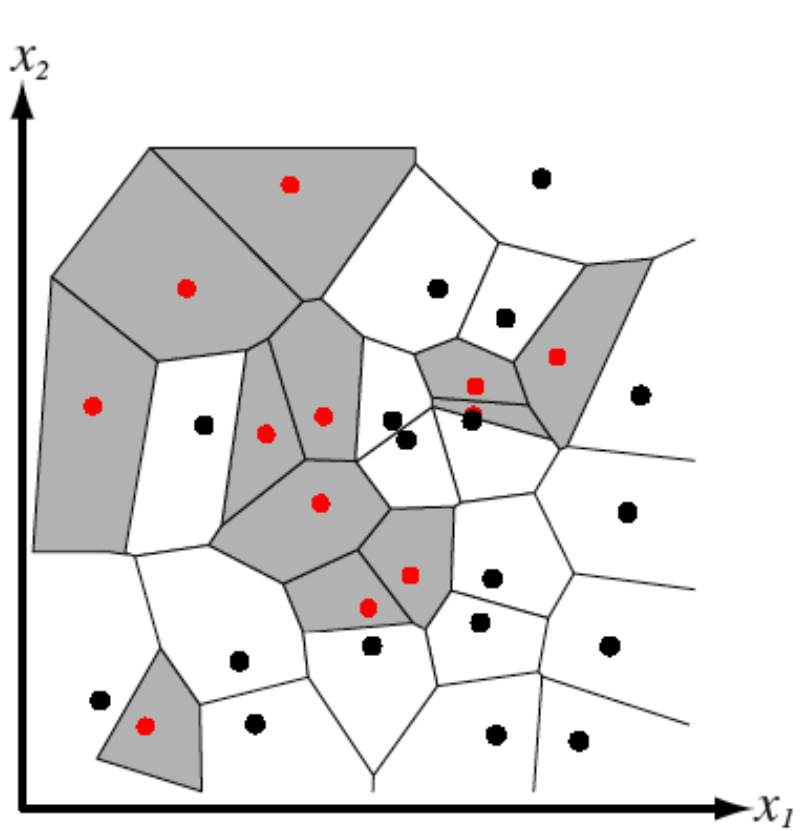
06.12.2007

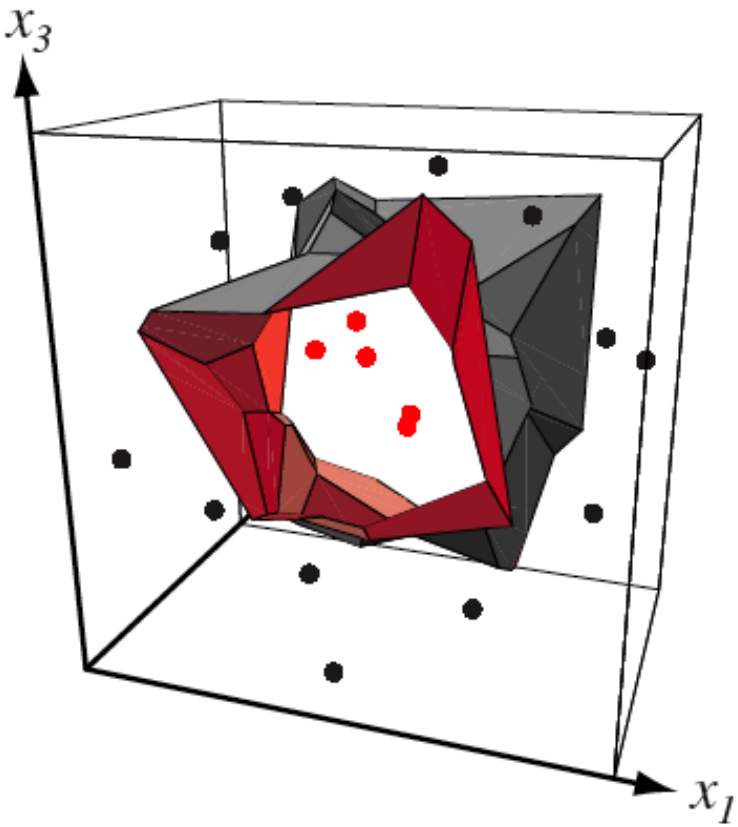
[1] falls nicht anders vermerkt, sind die Abbildungen entnommen aus Duda et al., 2001.

Übersicht

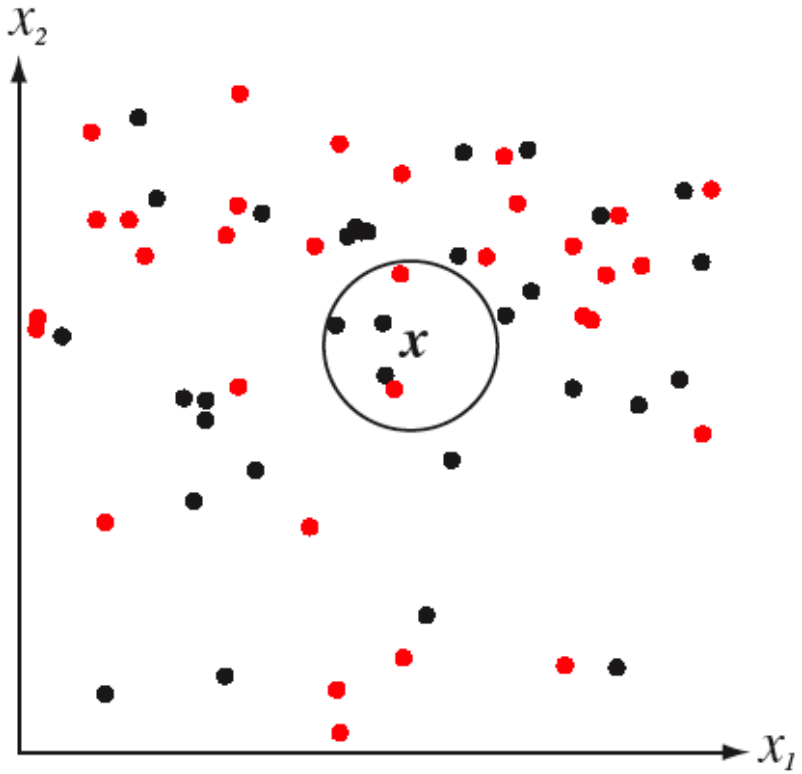
Übersicht

Durch Nächste-Nachbar-Regel erzeugtes Voronoi-Mosaik





k-Nächste-Nachbarn-Regel



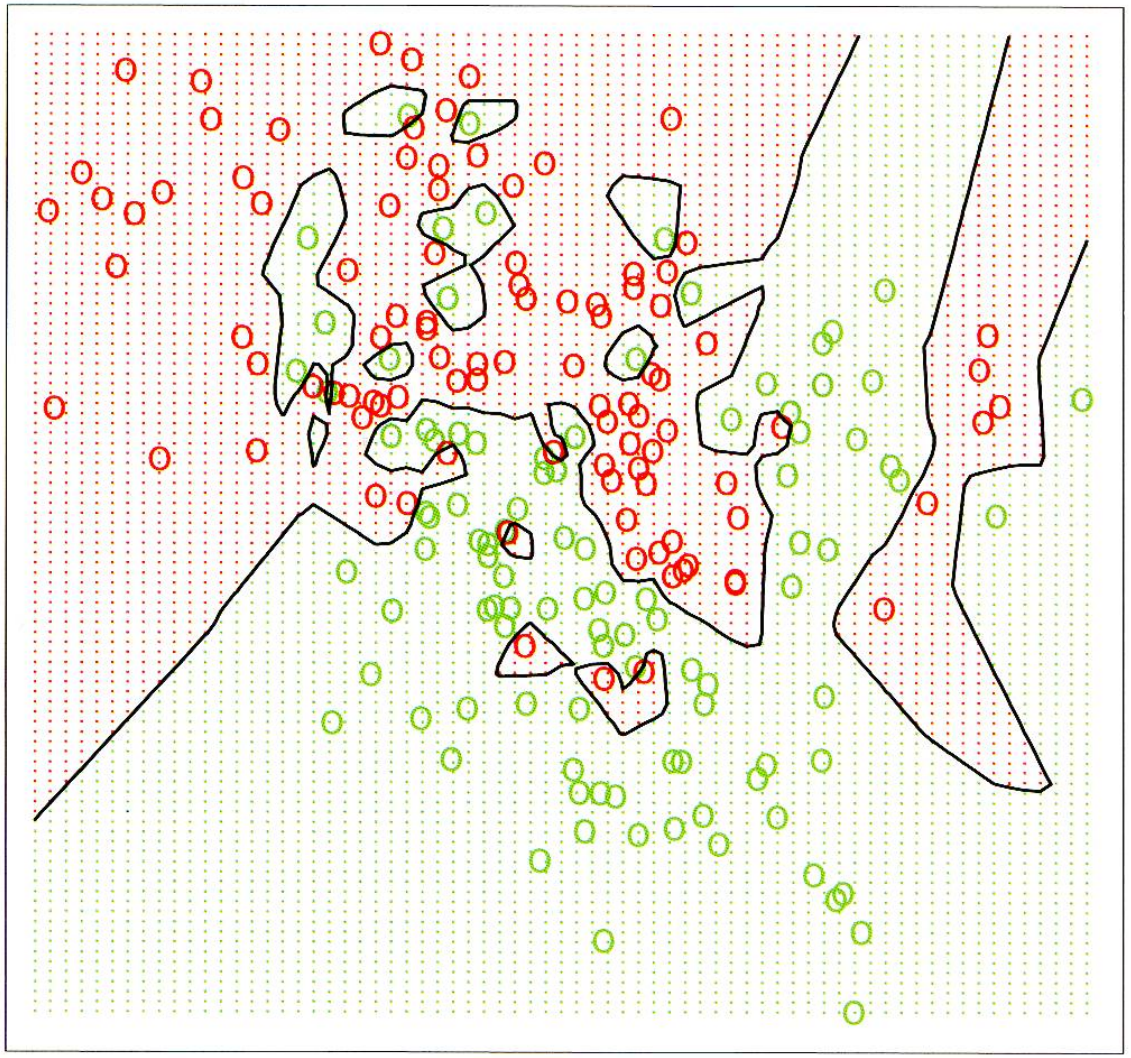
Offensichtliche Erweiterung:

k-Nächste-Nachbarn-Regel

Entscheide immer für die Klasse ω_i der Mehrheit aller Prototypen innerhalb der k-Nächste-Nachbarn-Zelle.

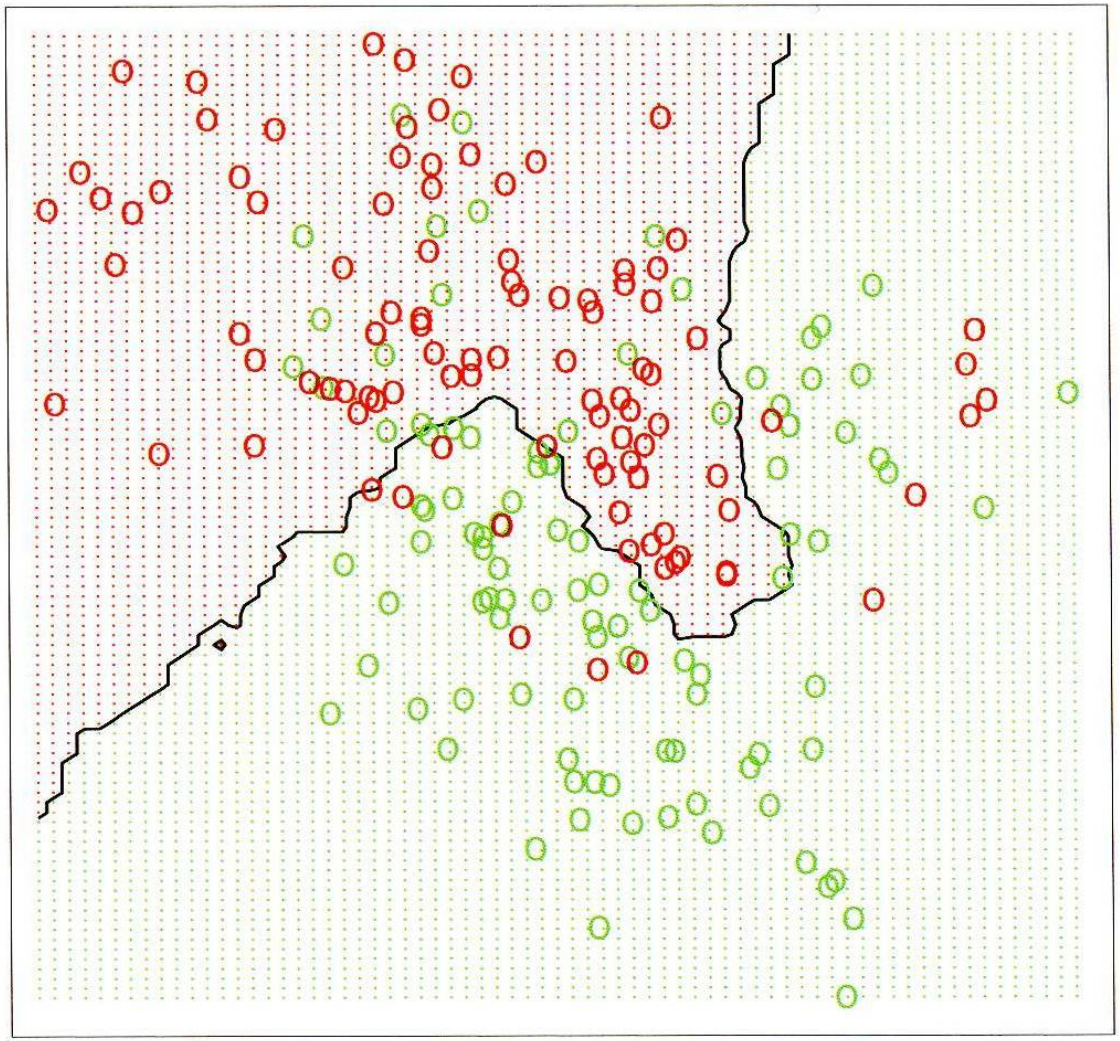
Im Limit unendlich vieler Daten führt die k-Nächste-Nachbarn-Regel auf kleinere Fehlerraten als die Nächste-Nachbar-Regel. Die Schätzungen sind robuster (wegen Mehrheitswahl), aber glätten über eine größere Umgebung.

Beispiel: 1-Nächster-Nachbar-Klassifikator



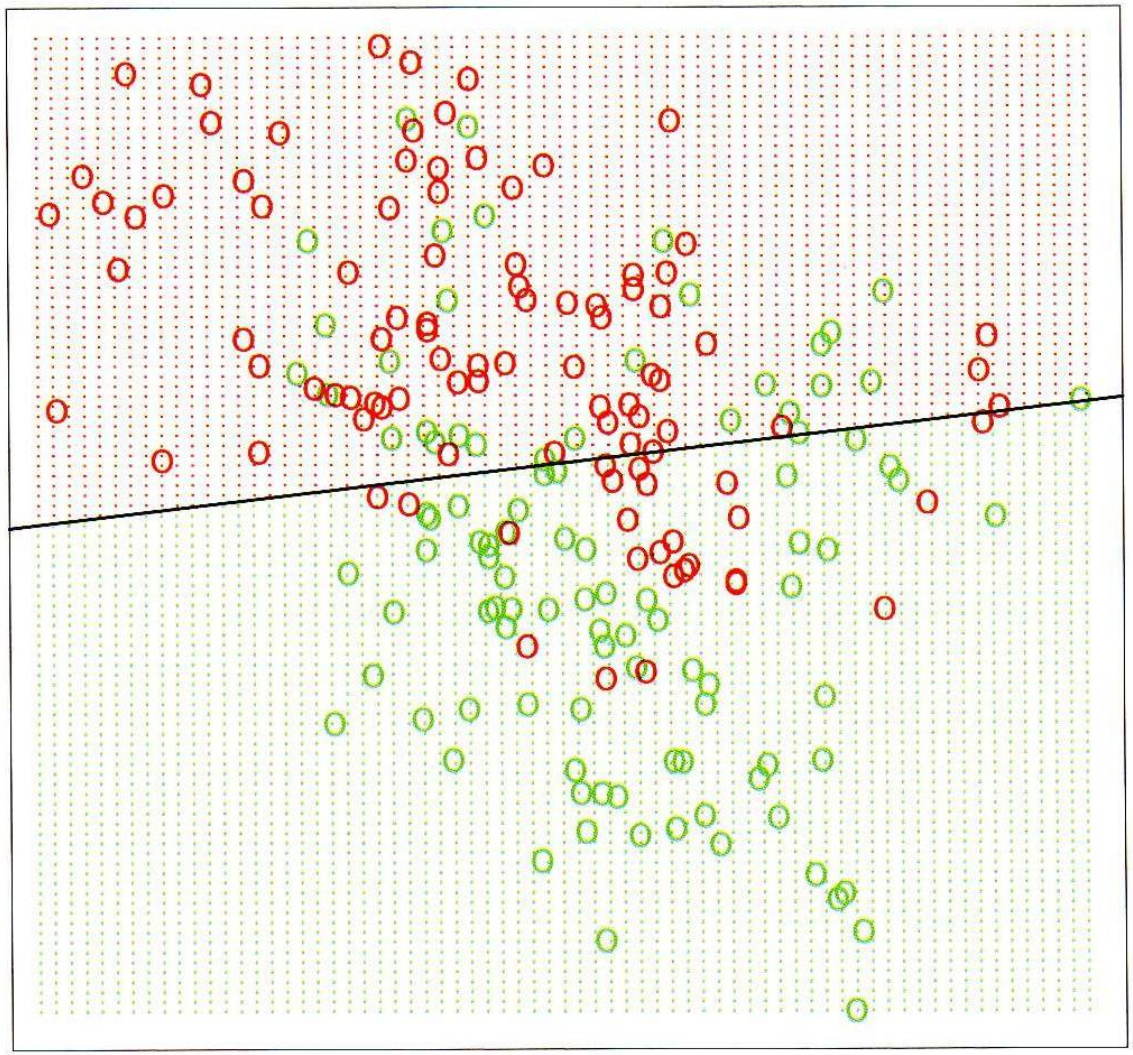
[Hastie et al., 2001]

Beispiel: 15-Nächste-Nachbarn-Klassifikator



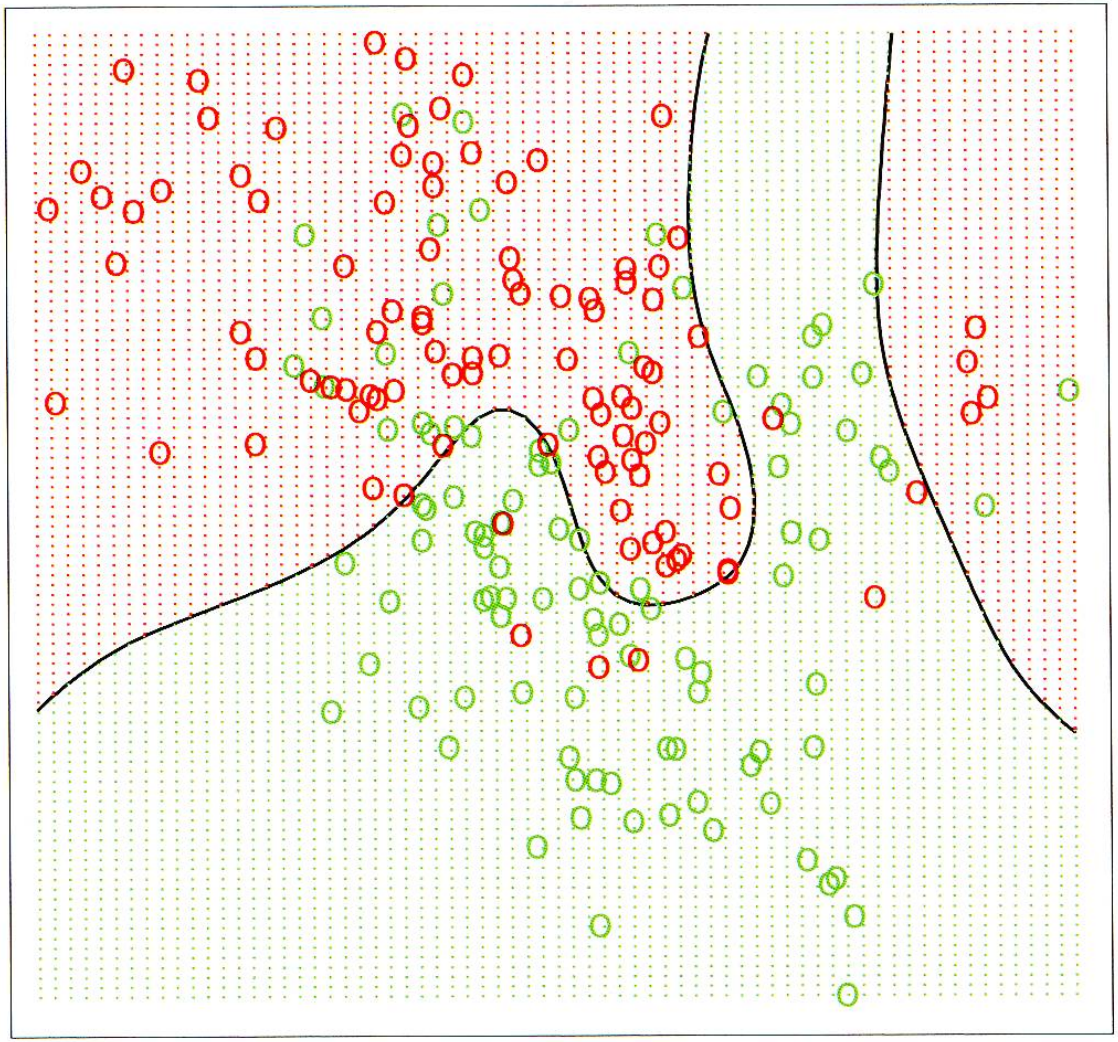
[Hastie et al., 2001]

Beispiel: Linearer Klassifikator



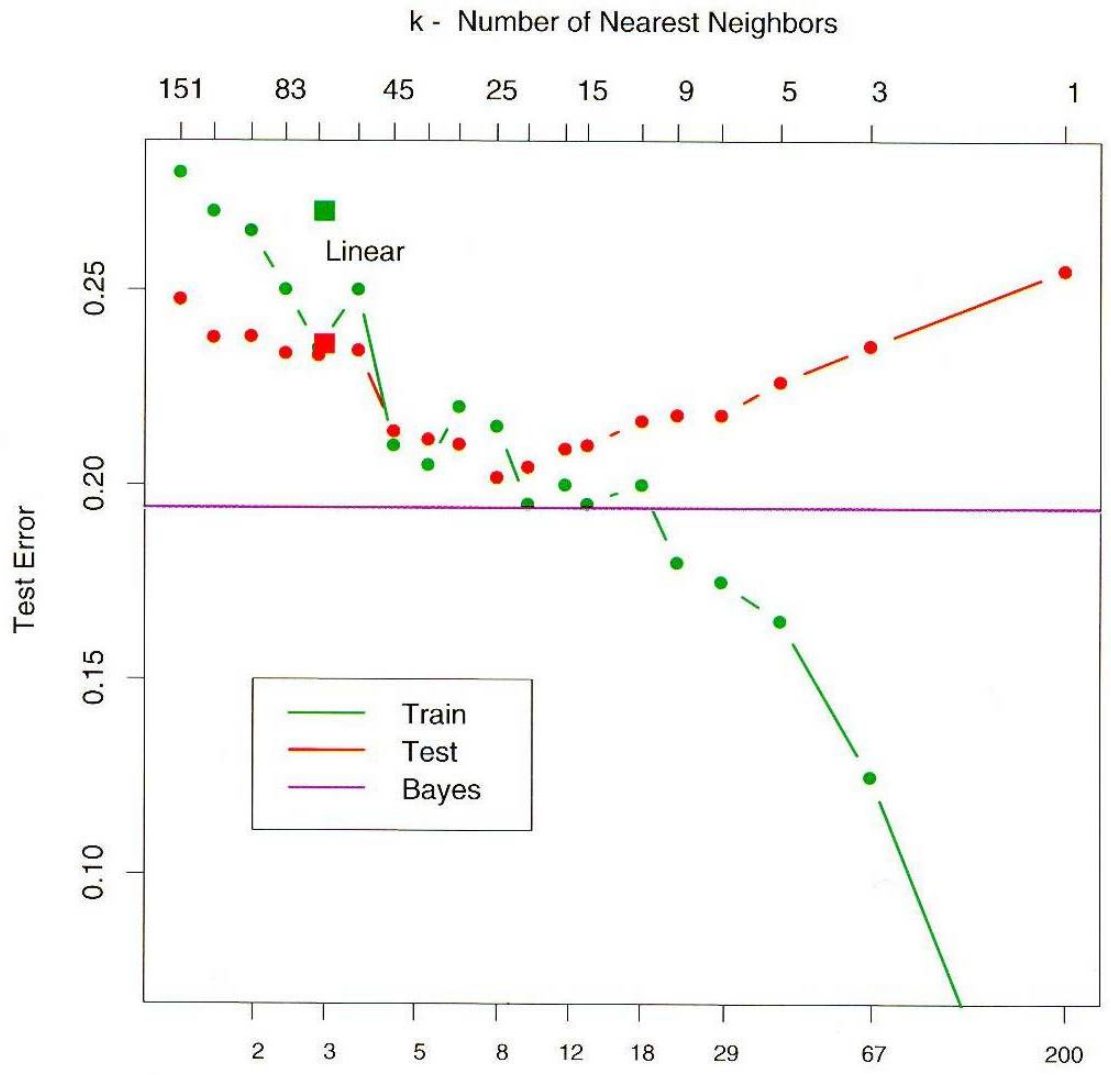
[Hastie et al., 2001]

Beispiel: Bayes-Klassifikator



[Hastie et al., 2001]

Beispiel: Fehlerraten



[Hastie et al., 2001]

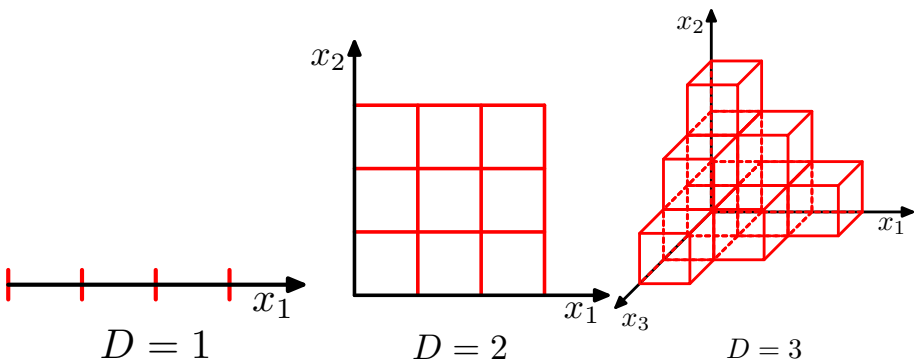
Vergleich Nächste-Nachbarn / lineare Klassifikatoren

- Lineare Entscheidungsgrenzen sind glatt und relativ stabil anzufitten, aber sie basieren auf der (anscheinend) einschränkenden Annahme der linearen Trennbarkeit $\Rightarrow$ hoher Bias, geringe Varianz.
- NN-Klassifikatoren machen beinahe keine einschränkenden Annahmen über die Verteilung, aber die Entscheidungsgrenzen hängen sehr stark vom jeweiligen Trainingsdatensatz ab $\Rightarrow$ geringer Bias, hohe Varianz.
- Lineare Klassifikatoren eignen sich besonders für breite, identisch oder isotrop gaußverteilte Klassen, NN für Mischungen vieler schmaler Verteilungen.

Übersicht

Fluch der Dimensionalität

Fluch der Dimensionalität: Die Komplexität von Funktionen mit mehreren Variablen kann exponentiell mit der Dimension wachsen, d.h. die Schätzung einer solchen Funktion erfordert ebenfalls exponentiell viele Daten.



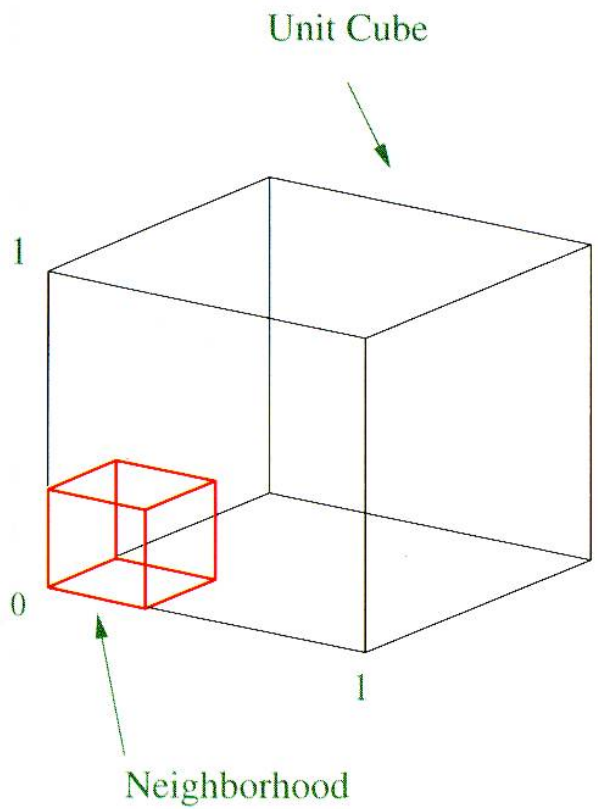
[Bishop, 2006]

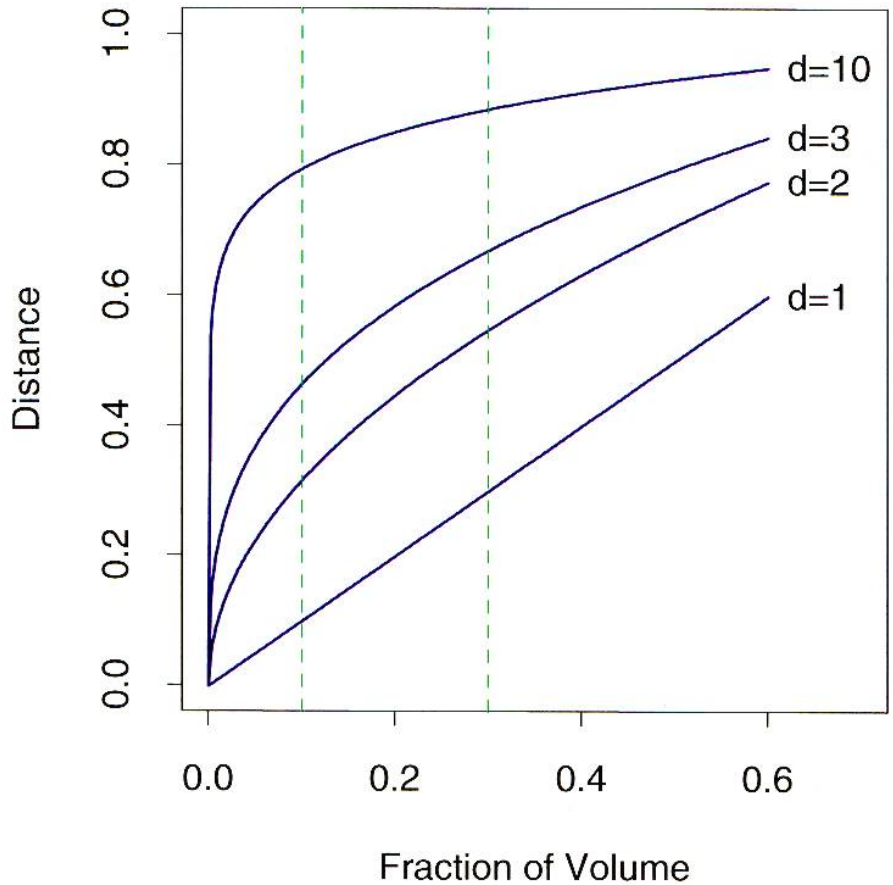
Anzahl w der Würfel Volumen der Seitenlänge L steigt mit der Dimension d bei gleicher Kantenlänge l wie

$$w = \frac{L}{l}^d \quad (\text{hier } w = 3^d)$$

Die Anzahl Datenpunkte für NN mit gleicher Auflösung skaliert ebenso!

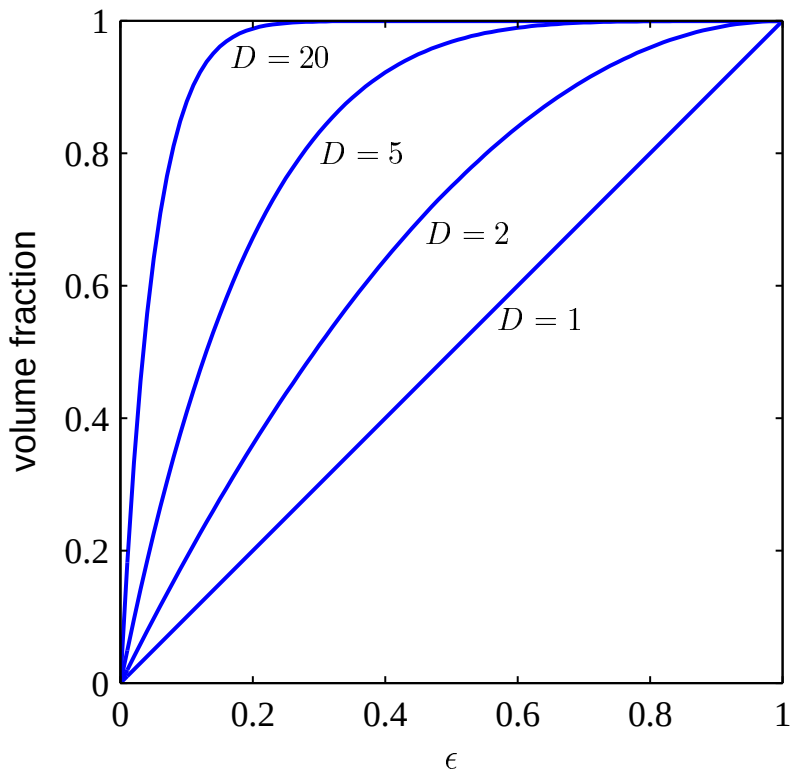
Beispiel: Fluch der Dimensionalität (2)





[Hastie et al., 2001]

Beispiel: Fluch der Dimensionalität (3)



Welcher Volumenanteil einer d-dimensionalen Kugel mit Radius 1 liegt in der Schale von $1 - \epsilon$ und 1?

Volumen der Kugel:

$$V(r) = Kr^d$$

Anteil:

$$\frac{V(1) - V(1-\epsilon)}{V(1)} = 1 - (1-\epsilon)^d$$

[Bishop, 2006]

$\Rightarrow$ In hochdimensionalen Räumen ist das Volumen einer Kugel in einer dünnen Schale an der Oberfläche konzentriert!

Klassifikation in hochdimensionalen Räumen

- Durch den Fluch der Dimensionalität funktionieren nichtparametrische Methoden wie NN oder Kerndichteschätzer nur bei geringer Dimensionalität.
- Klassifikation in hochdimensionalen Räumen erfordert Vorwissen.
- Reale Daten lassen sich dennoch oft klassifizieren, obwohl sie hochdimensional sind, weil sie
  - oft nur einen niedrigdimensionalen Unterraum besetzen,
  - lokale Glattheitseigenschaften aufweisen (d.h. kleine Änderungen am Input führen auch zu kleinen Änderungen des Ergebnisses).
- Lineare Klassifikatoren nutzen Vorwissen über solche lokale Glattheitseigenschaften der Klassenverteilungen.

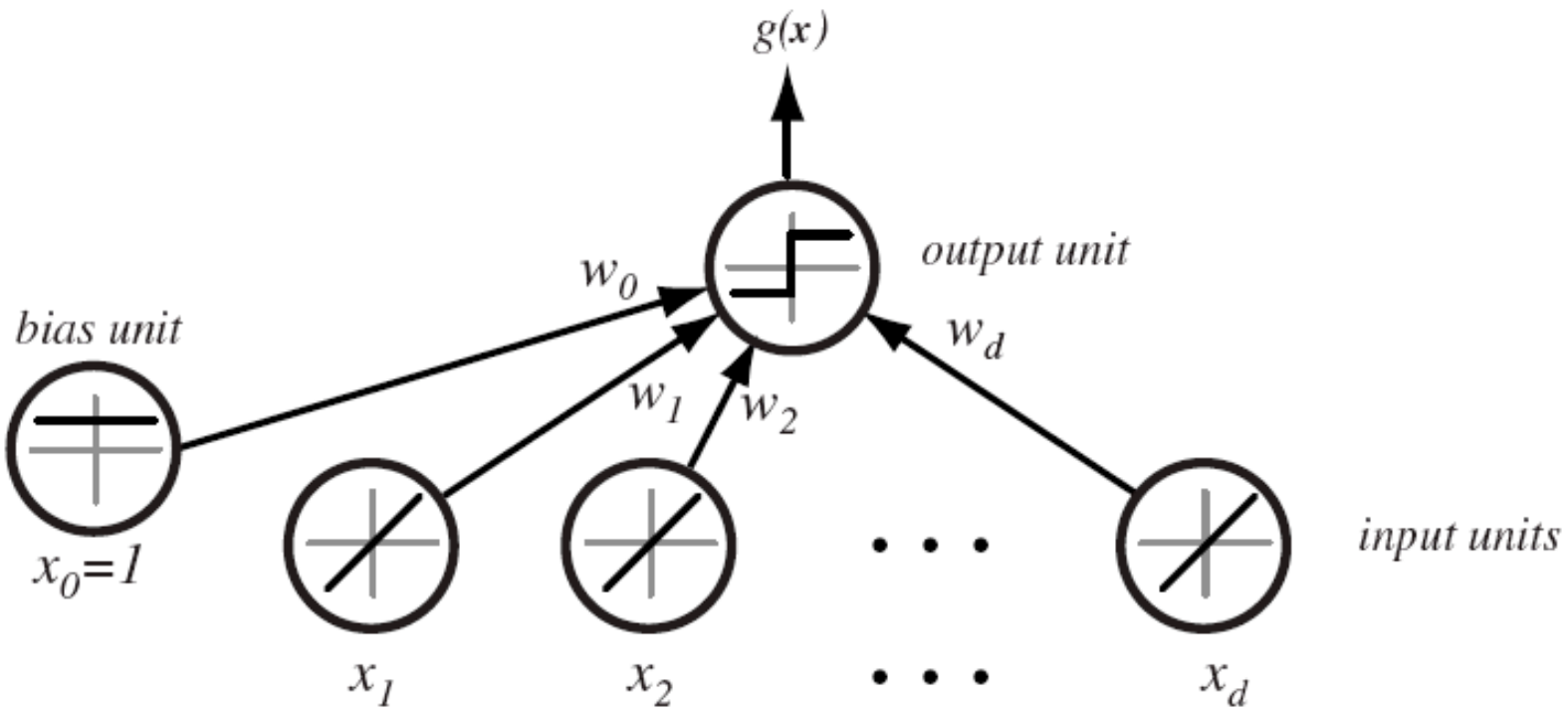
Übersicht

Lineare Diskriminantenfunktionen

Lineare Diskriminantenfunktion:

$$g(x) = w^\top x + w_0$$

mit dem Gewichtsvektor w und Bias oder Schwellwert w_0.



Lineare Entscheidungsgrenzen

Entscheidungsregel: Entscheide für Klasse ω_1 für $g(x) > 0$, für ω_2 für $g(x) < 0$, keine Entscheidung für $g(x) = 0$.

Die Gleichung $g(x) = 0$ definiert eine Entscheidungsfläche, die Punkte von ω_1 und ω_2 trennt. Da $g(x)$ linear ist, ist die Entscheidungsfläche eine Hyperebene.

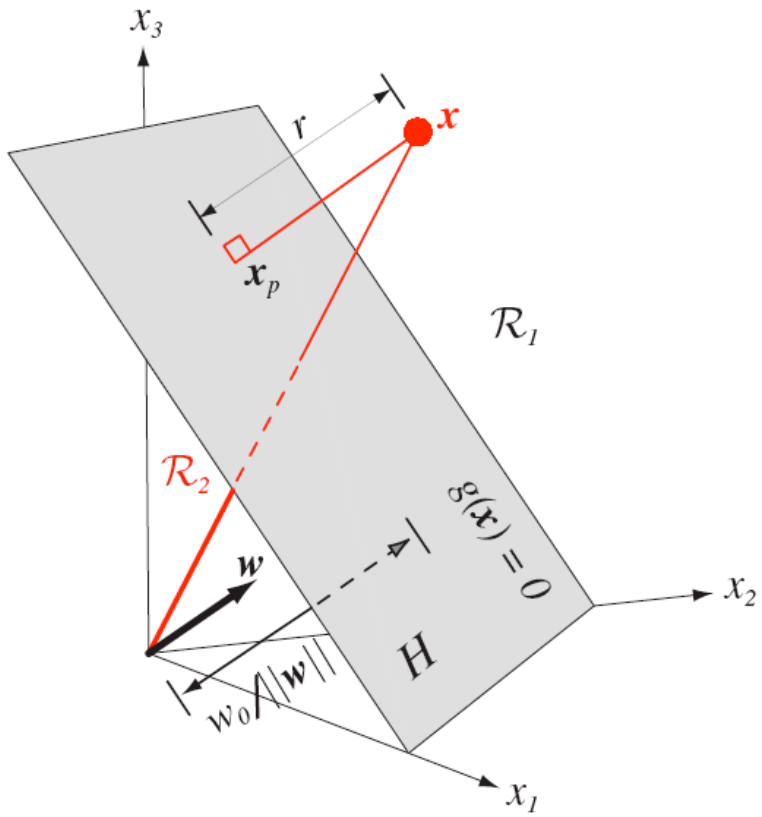
Für 2 Punkte x_1 und x_2 auf der Hyperebene H gilt

$$w^\top x_1 + w_0 = w^\top x_2 + w_0 \quad \text{bzw.} \quad w^\top (x_1 - x_2) = 0.$$

d.h. der Gewichtsvektor w ist der Normalenvektor der Hyperebene H.

H teilt den Merkmalsraum in 2 Halbräume: $\mathcal{R}_1$ für ω_1 bzw. $g(x) > 0$ (d.h. w zeigt Richtung $\mathcal{R}_1$, positive Seite), $\mathcal{R}_2$ für ω_2 bzw. $g(x) < 0$ (negative Seite).

Abstand zur Entscheidungsebene



Die Diskriminantenfunktion $g(x)$ mißt die Distanz von x zu H:

$$x = x_p + r\frac{w}{\|w\|}$$

Da $g(x_p) = 0$ bzw. $w^\top x_p = -w_0$, ist

$$\begin{aligned} g(x) &= w^\top x + w_0 \\ &= w^\top x_p + w_0 - r\frac{w^\top w}{\|w\|} = r\|w\| \end{aligned}$$

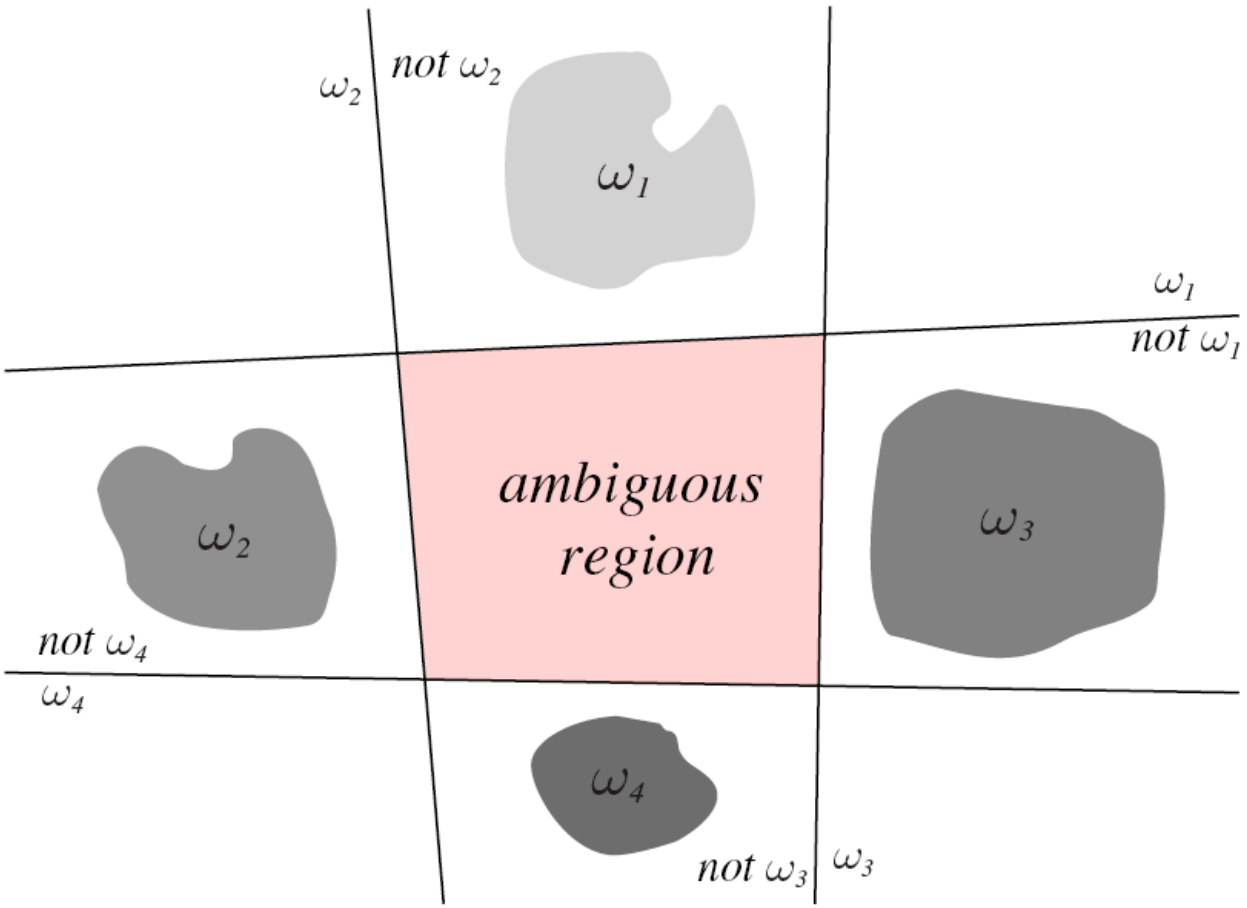
oder

$$r = \frac{g(x)}{\|w\|}$$

Abstand zum Ursprung: $g(x) = w_0 \quad \Rightarrow \quad r_0 = w_0/\|w\|.$

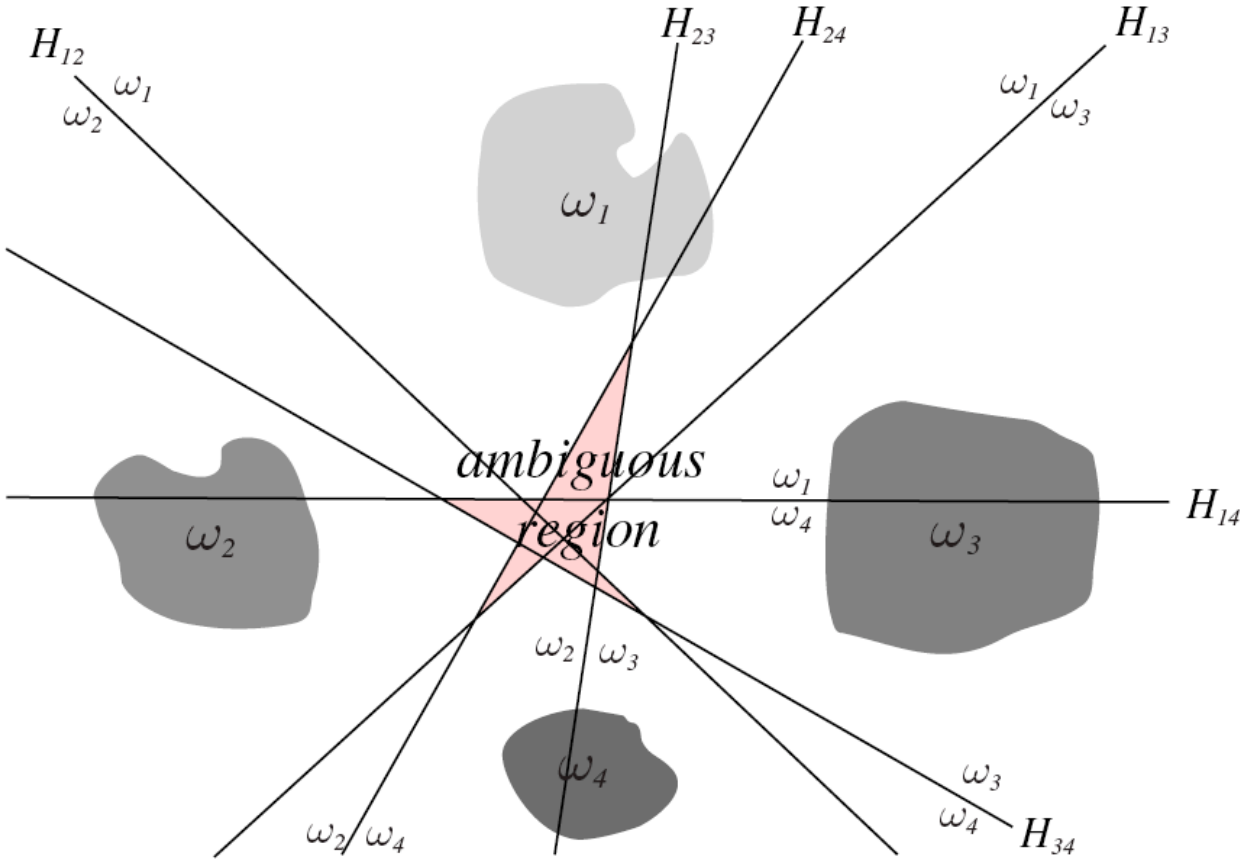
Polychotomien (1)

One-vs.-All-Architektur:



Polychotomien (2)

One-vs.-One-Architektur:

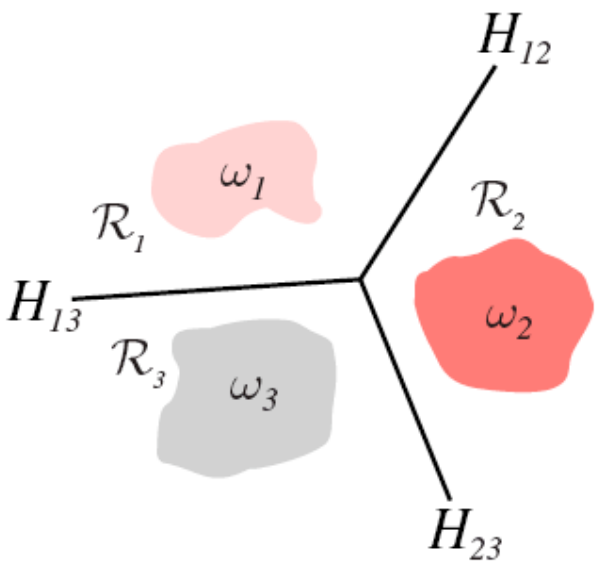


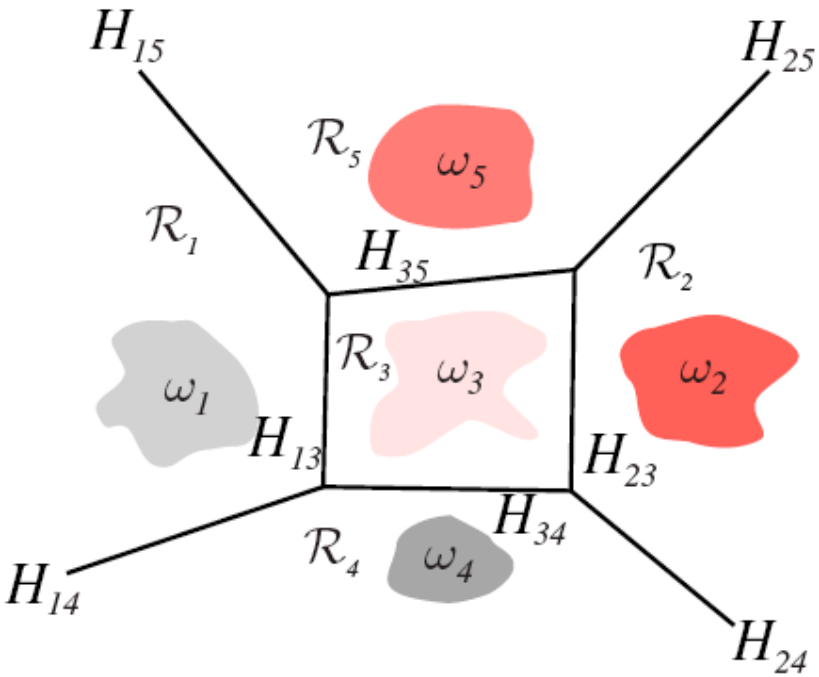
Polychotomien: Lineare Maschinen

Für c Klassen werden c lineare Diskriminantenfunktionen

$$g_i(x) = w_i^\top x + w_{0i}$$

benutzt. Entscheidung für ω_i, wenn $g_i(x) > g_j(x)$ für alle $j \neq i$.





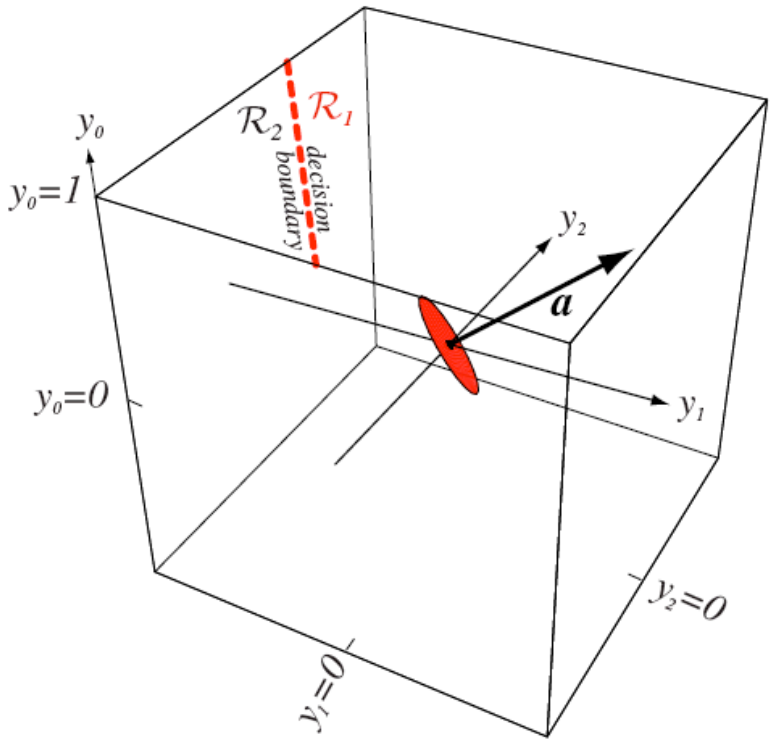
Entscheidungsgrenzen in linearen Maschinen

- Zwei angrenzende Entscheidungsregionen $\mathcal{R}_i$ und $\mathcal{R}_j$ werden durch einen Abschnitt H_{ij} einer Hyperebene getrennt:

$$g_i(x) = g_j(x) \quad \text{bzw.} \quad (w_i - w_j)^\top x + (w_{i0} - w_{j0}) = 0,$$

  d.h. der Normalenvektor von H_{ij} ist $(w_i - w_j)$, der (vorzeichenbehaftete) Abstand von x zu H_{ij} ist $(g_i(x) - g_j(x))/\|w_i - w_j\|$.
- Jede Entscheidungsregion ist einfach zusammenhängend.
- Lineare Maschinen sind daher besonders für Probleme mit unimodalen A-posteriori-Verteilungene geeignet.
- Aber: Es gibt multimodale Probleme, bei denen lineare Diskriminanten hervorragend abschneiden, und unimodale Probleme, bei denen sie schlecht funktionieren (s. Gauß)

Erweiterte Merkmalsvektoren



Einführung einer zusätzlichen konstanten Koordinate $x_0 = 1$ vereinfacht Notation.

$$g(x) = w_0 + \sum_{i=1}^{d} w_i x_i = \sum_{i=0}^{d} w_i x_i$$

Erweiterter Merkmalsvektor:

$$y = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{bmatrix} = \begin{bmatrix} 1 \\ x \end{bmatrix}$$

Erweiterter Gewichtsvektor:

$$a = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix} = \begin{bmatrix} w_0 \\ w \end{bmatrix}$$