Kern-Fisher-Diskriminanzanalyse aus analytischer Sicht

ITWM

Fraunhofer Institut
Techno- und
Wirtschaftsmathematik

Dr. Hagen Knaf

Abteilung *Adaptive Systeme*

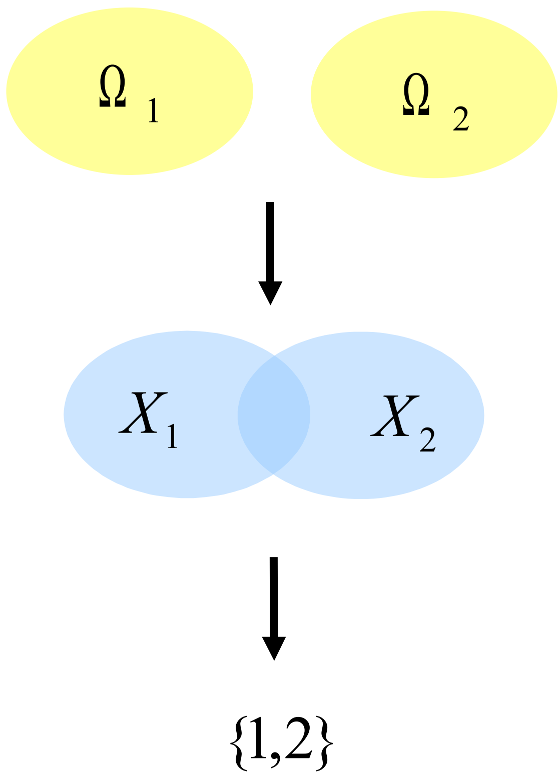
Schwerpunkt *Entscheidungsunterstützung in Medizin und Technik*

Überblick

1. Diskriminanzanalyse – Problem und Lösungsansatz
2. Lineare Diskriminanzanalyse nach Fisher
3. Reproduzierende Kerne funktionaler Hilberträume
4. Kern-Fisher-Diskriminanzanalyse
5. Der Algorithmus von S. Mika et. al.
6. Beispiele

Sir Ronald Aylmer Fisher, *1890, London, England, † 1962, Adelaide, Australien

Diskriminanzanalyse: Das allgemeine Problem



Man betrachtet eine in g>1 paarweise disjunkte Gruppen unterteilte Grundgesamtheit:

$$\Omega = \Omega_1 \cup \ldots \cup \Omega_g$$

Jedem Objekt $\omega \in \Omega$ sind bestimmte Merkmale

$$x(\omega) \in X$$

zugeordnet.

Finde eine Entscheidungsfunktion

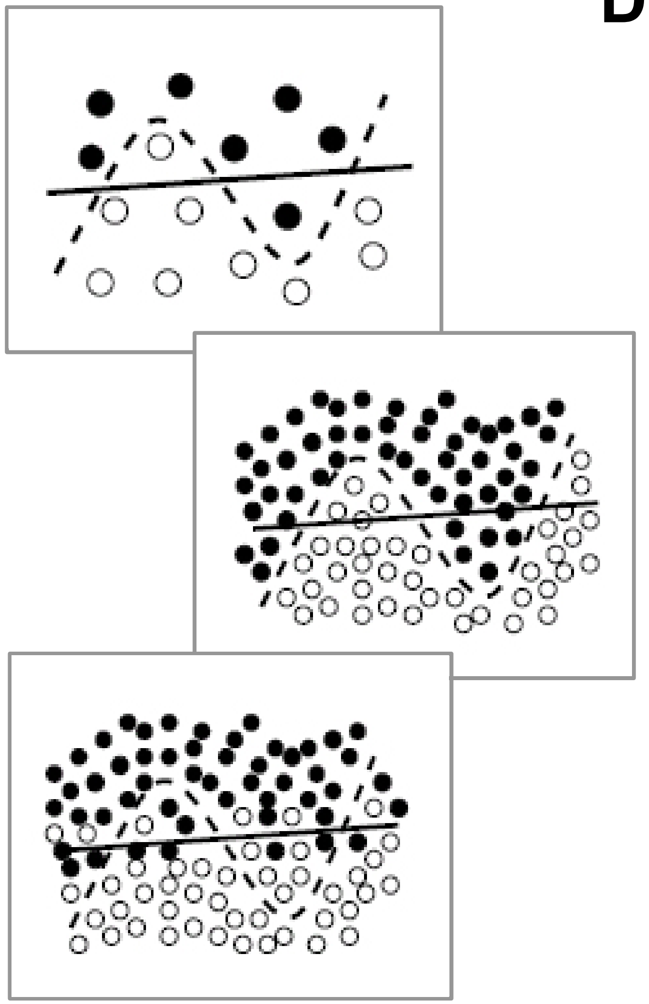
$$e : X \to \{1, \ldots, g\},$$

die jedem $\omega \in \Omega$ eine vermutete Gruppenzugehörigkeit

$$e(x(\omega))$$

so zuordnet, daß »möglichst wenige« Fehlzuordnungen auftreten.

Diskriminanzanalyse: Das allgemeine Problem



Aus der unbekannten Grundgesamtheit zieht man eine (endliche) Stichprobe

$$\Lambda = \Lambda_1 \cup \ldots \cup \Lambda_g \subseteq \Omega$$

Basierend auf der Merkmalsmenge

$$Y := \{x(\omega) : \omega \in \Lambda\}.$$

wird eine geschätzte Entscheidungsfunktion

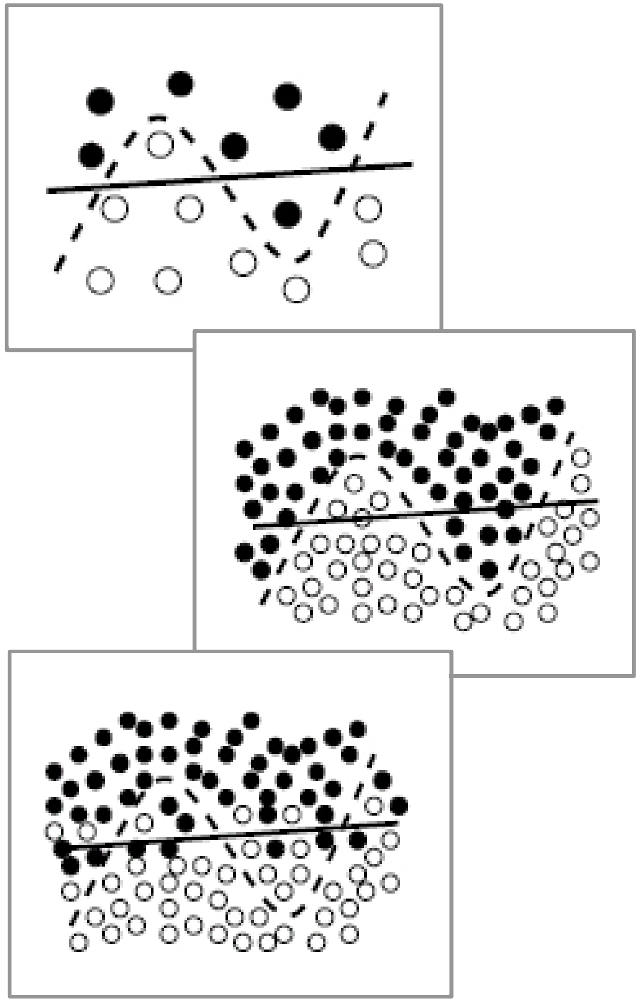
$$\hat{e} : X \to \{1, \ldots, g\}$$

ermittelt.

Gütecharakteristika und Gegenspieler:

- hohe Trefferquote of Y,
- verallgemeinerungsfähig auf neue Samples.

Lösungsansatz: Diskriminanzfunktionen



Vereinfachende Annahmen: $g = 2, X \subseteq R^m$.

Betrachte geschätzte Entscheidungsfunktionen der Form:

$$\hat{e} : X \to \{-1, +1\}, x \mapsto \operatorname{sgn}(d(x)),$$

wobei sgn die Signumfunktion und

$$d : X \to R$$

eine reellwertige Funktion ist – die *Diskriminanzfunktion*.

Die Hyperfläche

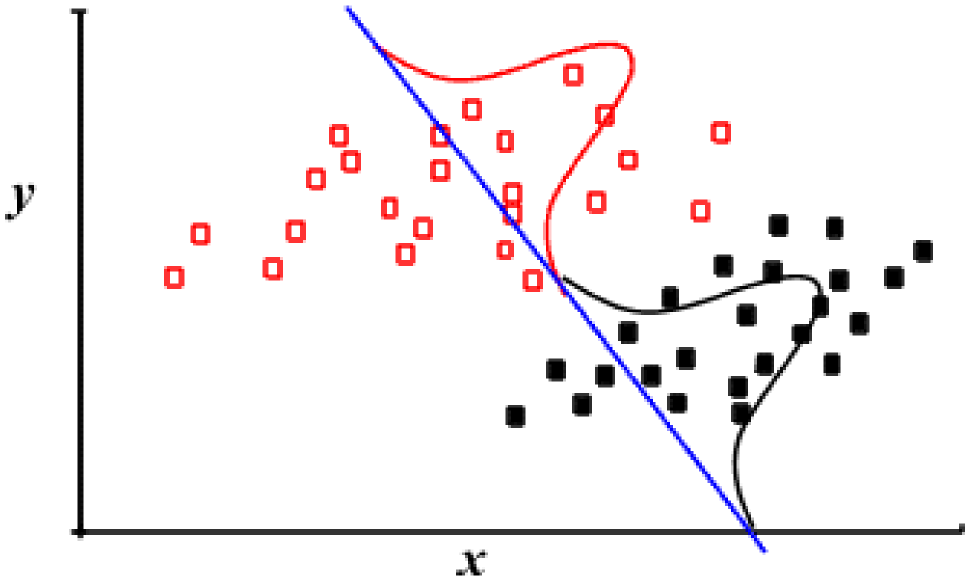
$$H := \{x \in R^m : d(x) = 0\}$$

trennt die beiden Gruppen; ihre Elemente bleiben bei strikter Sicht unklassifiziert.

Die Funktion d wird innerhalb einer Funktionenklasse F basierend auf den Daten Y gemäß der Gütekriterien für die Entscheidungsfunktion bestimmt.

Lineare Diskriminanzanalyse: Fisher‘s Ansatz

Annahmen: $g = 2$ (nicht notwendig), $X \subseteq R^m$ (notwendig), $Y = Y_{-1} \cup Y_{+1}, Y_{-1} \cap Y_{+1} = \{\}$ (notwendig).



Wähle ein lineares Funktional

$$f : R^m \to R, x \mapsto \langle a, x \rangle$$

so, daß die *Fisherdiskriminante*

$$Q(a) := \frac{\left(\langle a, \bar{y}_{-1} \rangle - \langle a, \bar{y}_{+1} \rangle\right)^2}{s_{-1}^{\;2} + s_{+1}^{\;2}}$$

maximal wird, wobei $\bar{y}_k$ Mittelwert von Y_k ist und:

$$s_k^{\;2} = \sum_{y \in Y_k} \left(\langle a, y \rangle - \langle a, \bar{y}_k \rangle\right)^2 .$$

Lineare Diskriminanzfunktion

- Fisher's Ansatz liefert die lineare Diskriminanzfunktion:

$$d : R^m \to R, x \mapsto \langle a, x \rangle + b,$$

$$b = \frac{1}{2}\left(\langle a, \bar{y}_{-1} \rangle + \langle a, \bar{y}_{+1} \rangle\right).$$

$$a = W^{-1}(\bar{y}_{-1} - \bar{y}_{+1})$$

$$W = \sum_{k=1}^{2} \sum_{y \in Y_k} \langle y - \bar{y}_k, y - \bar{y}_k \rangle$$

- Dieses Ergebnis hat auch eine statistische Interpretation, wenn man geeignete Verteilungs-annahmen über die Stichprobendaten macht.

Nichtlineare Diskriminanzfunktionen

Parametrischer Ansatz: Betrachte eine parametrisierte Klasse F von Diskriminanz-funktionen

$$d_p : X \to R, p \in P$$

und schätze p aus den Daten.

Nichtparametrische Methoden:

- Nächste-Nachbar-Methode
- Neuronale Netze
- Klassifikationsbäume
- …
- Kernmethoden

Funktionale Hilberträume

Im Folgenden werden Hilberträume über dem Körper κ der reellen oder komplexen Zahlen betrachtet.

Definition: Der Hilbertraum H heißt **funktional über der Menge X**, falls er Untervektorraum des Abbildungsraums $\mathrm{Abb}(X,\kappa)$ (versehen mit der punktweisen Addition und skalaren Multiplikation) ist.

- Für einen funktionalen Hilbertraum H existieren die linearen Auswertungsfunktionale

$$\varphi_x : H \to \kappa, \; f \mapsto f(x) \quad (x \in X).$$

- Die Auswertungsfunktionale sind im unendlich-dimensionalen Fall nicht notwendig stetig.

Struktursatz: Ein Hilbertraum ist isomorph zu $\ell^2(T)$ für geeignetes T und damit funktional über T.

Der reproduzierende Kern

Situation: Der Hilbertraum H sei funktional über der Menge X und besitze stetige Auswertungsfunktionale.

Riesz'scher Darstellungssatz: $\forall x \in X : \exists K_x \in H : \varphi_x = \langle \circ, K_x \rangle$.

Definition: Die Abbildung $K : X \times X \to \mathbb{K}$

$$(x,y) \mapsto \langle K_y, K_x \rangle = K_y(x)$$

heißt **reproduzierender Kern von H**.

Sie induziert eine Abbildung

$$\Phi : X \to H,\ x \mapsto K(\circ, x) = K_x$$

die injektiv ist, falls H die Punkte von X trennt: $x_1, x_2 \in X : x_1 \neq x_2 \Rightarrow \exists h \in H : h(x_1) \neq h(x_2)$.

Es gilt die **Schlüsselidentität**: $\langle \Phi x, \Phi y \rangle = K(y,x)$.

Eigenschaften des reproduzierenden Kerns

Der reproduzierende Kern K eines funktionalen Hilbertraums H über X besitzt folgende Eigenschaften:

1. $\forall\, x, y \in X : K(x,y) = \overline{K(y,x)}.$

2. Für jede Wahl von paarweise verschiedenen Punkten $\{x_1,\ldots,x_n\} \subseteq X$ ist die Matrix $A := (K(x_k, x_l))_{k,l\in\{1,\ldots,n\}}$ positiv semi-definit: $\forall\, z \in K^n : z^t A \bar{z} \geq 0.$

   Beweis

3. Der von den Funktionen $\{K_x : x \in X\}$ erzeugte Untervektorraum von H liegt dort dicht.

   Beweis

Kernfunktionen

Definition: Eine Abbildung $K : X \times X \to \kappa$ mit den Eigenschaften (1) und (2) der vorangegangenen Folie heißt **Kernfunktion**.

Ist X ein topologischer Raum, so nennt man eine *stetige* Kernfunktion **Mercer-Kern (*)**.

Eigenschaften: Sind $K_i : X \times X \to \kappa, i = 1,2,$ Kernfunktionen / Mercer-Kerne, so auch:

- die punktweise Summe $K_1 + K_2$,
- positive Vielfache $\alpha K_1, \alpha > 0$,
- die Funktion $K(x, y) := f(x) K_1(x, y) \overline{f(y)}$ für [stetiges] $f : X \to \kappa$,
- das punktweise Produkt $K_1 \cdot K_2$.

J. Mercer, *1883 Liverpool, †1932 London.

Kernfunktionen

Beispiele von Mercer-Kernen auf R^n **:**

- Polynomkern vom Grad d: $K(x,y) := (\langle x,y\rangle + c)^d, c > 0,$
- Gaußkern der Bandweite h: $K(x,y) := e^{-\frac{\|x-y\|^2}{h^2}}.$
- **Keine** Kernfunktion ist: $K(x,y) := \tanh(\langle x,y\rangle + c).$

Das Theorem von Aronzajn-Moore (*)

Theorem: Zu jeder Kernfunktion $K : X \times X \to \kappa$ existiert ein über X funktionaler Hilbertraum, dessen reproduzierender Kern gleich K ist.

Sind H_1, H_2 über X funktionale Hilberträume mit reproduzierenden Kernen K_1, K_2, so gilt:

$$(K_1 = K_2) \Rightarrow (H_1 = H_2).$$

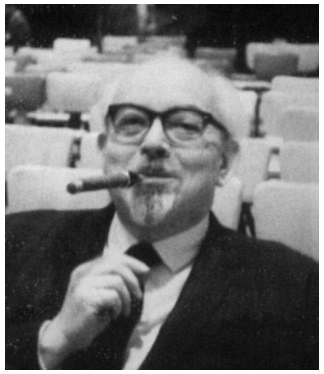
Bemerkungen:

1. Der durch die Kernfunktion K eindeutig bestimmte Hilbertraum wird mit H(K) bezeichnet.
2. Dies ist die Grundlage einer Kategorienäquivalenz zwischen der Kategorie der Kernfunktionen und der Kategorie funktionaler Hilberträume.
3. Anstelle von $\mathrm{Abb}(X,\kappa)$ können auch interessante Unterräume des vollen Abbildungsraums betrachtet werden, wie zum Beispiel r-fach stetig differenzierbare oder glatte Funktionen.

Beweisskizze

N. Aronzajn, *1907 Polen(?), †1980 USA(?).

R.L. Moore, *1882 Dallas, †1974 Austin.



Bemerkungen zum Theorem von Aronzajn-Moore

- Der Hilbertraum H(K) ist endlichdimensional, falls X endlich ist.
- Für jeden Polynomkern auf R^n ist H(K) endlichdimensional.
- Für den Gaußkern auf R^n ist H(K) unendlichdimensional.

- Um nachzuweisen, daß eine gegebene Funktion K eine Kernfunktion ist, kann die Konstruktion eines funktionalen Hilbertraums H, dessen reproduzierender Kern K ist, einfacher / konzeptioneller als ein direkter Beweis sein.

Beispiel: Die Tatsache, daß das Produkt KK' von Kernfunktionen K,K' auf einer Menge X eine ebensolche ist,

... kann direkt durch Benutzung der Eigenschaften des Schurprodukts von Matrizen beweisen werden,
... oder indem man zeigt, daß es der reproduzierende Kern des Tensorprodukts der Hilberträume H(K) und H(K') ist.

Situation: Gegeben ist eine endliche Stichprobe $Y = Y_{-1} \cup Y_{+1} \subset R^m,\ Y_{-1} \cap Y_{+1} = \{\}$ einer in zwei Gruppen eingeteilten Grundgesamtheit.

$K : R^m \times R^m \to R$ sei eine Kernfunktion auf R^m.

- Der große Definitionsbereich wird zur Klassifikation neuer Samples benötigt.

Im Folgenden arbeitet man in den funktionalen Hilberträumen

$$H(K|_{Y\times Y}) \subseteq H(K)$$

- $H(K|_{Y\times Y})$ wird von den Funktionen $K_y := K(\circ, y),\ y \in Y$ erzeugt.

Man betrachtet weiter die zugehörige Einbettung der Daten

$$\Phi : R^m \to H(K),\ y \mapsto K_y,$$

und setzt Injektivität von $\Phi|_Y$ voraus.

Fakt 1: Für die eingebettete Stichprobe $\Phi Y = \Phi Y_{-1} \cup \Phi Y_{+1} \subset H$ läßt sich mittels Fisher‘s Ansatz eine lineare Diskriminanzfunktion

$$d : H(K_{Y\times Y}) \to R, x \mapsto \langle a, x\rangle + b$$

bestimmen. Der Pullback

$$d \circ \Phi : R^m \to R, \; y \mapsto \langle a, \Phi y\rangle + b$$

ist dann eine Diskriminanzfunktion für die ursprüngliche Stichprobe.

Subtiler aber wichtiger Punkt: Das Skalarprodukt $\langle a, x\rangle$ kann via der natürlichen Inklusion $H(K|_{Y\times Y}) \subseteq H(K)$ als Skalarprodukt in $H(K)$ aufgefaßt werden. Die Diskriminanz-funktion d ist also auf dem eventuell unendlich-dimensionalen Raum $H(K)$ definiert. Damit können mittels $d \circ \Phi$ auch neue Samples klassifiziert werden.

Fraunhofer Institut Techno- und Wirtschaftsmathematik

Fakt 2: Die Diskriminanzfunktion $d \circ \Phi$ lässt sich vollständig in Termini der Kernfunktion K und der Stichprobendaten Y ausdrücken.

(In der Literatur wird diese Tatsache häufig als »Kerntrick« bezeichnet.)

Das Element $a \in H(K|_{Y \times Y})$ hat die Form

$$a = \sum_{y \in Y} \alpha_y \Phi(y), \quad \alpha_y \in R$$

und löst

$$\max_{h \in H(K|_{Y \times Y})} \left(Q(h) := \frac{\left(\langle h, \bar{h}_{-1} \rangle - \langle h, \bar{h}_{+1} \rangle \right)^2}{s_{-1}^2 + s_{+1}^2} \right), \quad s_k^2 = \sum_{y \in Y_k} \left(\langle h, \Phi y \rangle - \langle h, \bar{h}_k \rangle \right)^2,$$

$\bar{h}_{-1}, \bar{h}_{+1}$ die Mittelwerte der Gruppen $\Phi Y_{-1}, \Phi Y_{+1}$.

Die Kern-Fisher-Diskriminanzfunktion

Äquivalent hierzu löst der Vektor $\alpha := (\alpha_y)_{y \in Y} \in R^n, \; n := |Y|$ das Problem:

$$\max_{\beta \in R^n} \left(Q(\beta) := \frac{\beta^t M \beta}{\beta^t N \beta} \right)$$

$$M = (M_{-1} - M_{+1})(M_{-1} - M_{+1})^t$$

$$(M_i)_y = \frac{1}{n_i} \sum_{y' \in Y_i} K(y, y'), \quad (i = -1, +1)$$

$$N = KK^t - n_{-1} M_{-1} M_{-1}^{\ t} - n_{+1} M_{+1} M_{+1}^{\ t}$$

$$\alpha = N^{-1}(M_{+1} - M_{-1})$$

ITWM
Fraunhofer Institut
Techno- und
Wirtschaftsmathematik

Die explizite Kern-Fisher-Diskriminanzfunktion

Hat das Element $a \in H(K|_{Y \times Y})$ die Form $a = \sum_{y \in Y} \alpha_y \Phi(y)$ so gilt für die zugehörige Diskriminanzfunktion:

$$d \circ \Phi : R^m \to R, \; x \mapsto \sum_{y \in Y} \alpha_y K(y, x) + b.$$

Probleme

1. Die Berechnung von α, b aus n Datenpunkten im n-dimensionalen Raum ist ein schlecht gestelltes Problem.
2. In realen Data Mining Problemen kann n sehr groß sein (z.B. Bildverarbeitung).

Der Algorithmus von S.Mika et. al.

- Umformulierung in ein konvexes, quadratisches Optimierungsproblem mit näherungsweise gleicher Lösung:

$$\left.\begin{array}{c}\min\limits_{\alpha,\xi\in R^n, b\in R}\left(\|\xi\|^2 + C\|\alpha\|^2\right)\\ K\alpha + \mathbf{1}b = L + \xi\\ \langle \mathbf{1}_+,\xi\rangle = 0 = \langle \mathbf{1}_-,\xi\rangle\end{array}\right\}\quad (P)\quad \left\{\begin{array}{l}\alpha\ :\text{Koordinaten der Lösung}\\ b:\text{konstanter Koeffizient der Fisherfunktion}\\ K := (K(y_i,y_j))_{i,j}\in R^{n\times n},\ Y = \{y_1,\dots,y_n\}\\ \mathbf{1} := (1,\dots,1)\in R^n\\ L\in\{-1,+1\}^n\ \text{Vektor der Gruppenlabels}\\ \mathbf{1}_+ := \max(L,0),\ \mathbf{1}_- := \max(-L,0)\\ C:\text{Regularisierungskonstante}\end{array}\right.$$

- Kann mit Standardalgorithmen gelöst werden, was aber je nach Stichprobenumfang zeitaufwendig ist.

Der Algorithmus von S.Mika et. al.

- Die optimale Regularisierung »Anzahl der von 0 verschiedenen Koeffizienten« ist analytisch nicht zu handhaben …
- … und liefert die Motivation für einen Algorithmus, der die Anzahl in der Lösung verwendeter Kernfunktionen schrittweise erhöht:

Start mit der leeren Menge S={} von Kernsamples.

Wahl: Wähle ein Sample $y_k \in Y \setminus S$ mit der Optimalitätseigenschaft:
Die Lösung $\alpha_{S \cup \{y_k\}}, b_{S \cup \{y_k\}}$ des Optimierungsproblems (P) zu der Menge $S \cup \{y_k\}$ führt zum stärksten Abfall der primären Zielfunktion $\|\xi\|^2 + C\|\alpha\|^2$ im Vergleich aller Lösungen $\alpha_{S \cup \{y_l\}}, b_{S \cup \{y_l\}},\ y_l \in Y \setminus S.$

Abbruch?: Setze $S := S \cup \{y_k\}$ und breche ab, falls die Lösungsgüte ausreichend ist, oder gehe zu **Wahl**.

Fraunhofer Institut Techno- und Wirtschaftsmathematik

Der Algorithmus von S.Mika et. al.

- Weshalb geht das alles »schnell«?
  - Numerisch sind im Wesentlichen Rang-1-Updates von Matrixinversen und Maxima von quadratischen Polynomen in 2 Variablen zu bestimmen:

$$\left.\begin{array}{l}\min\limits_{a\in R^{n+1}}\left(\frac{1}{2}a^t Ha - c^t a + \frac{n}{2}\right) \\ A_+^t a - n_+ = 0 = A_-^t a - n_-\end{array}\right\} \quad (P)$$

$n_+, n_- :=$ Samplezahl von Y_+, Y_-

$$a := \begin{bmatrix} \alpha \\ b \end{bmatrix}$$

$$H := \begin{bmatrix} n & \mathbf{1}^t K \\ K^t \mathbf{1} & K^t K + CE \end{bmatrix}$$

$$a = H^{-1}(c - \lambda_+ A_+ - \lambda_- A_-)$$

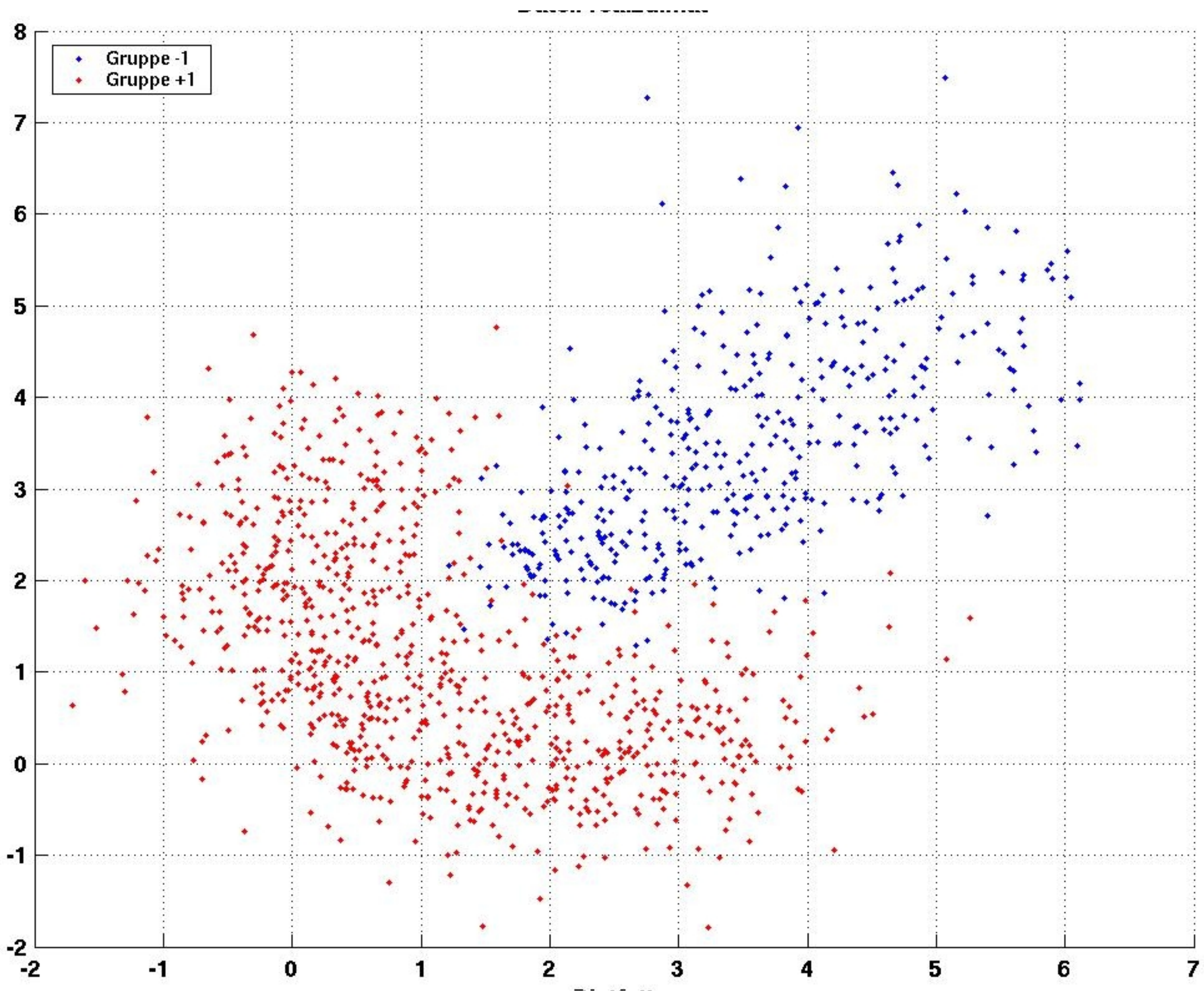
λ_+, λ_- : Lagrangemultiplikatoren von (P)

$$c := \begin{bmatrix} n_+ - n_- \\ K^t L \end{bmatrix}, \quad A_{+/-} := \begin{bmatrix} n_{+/-} \\ K^t \mathbf{1}_{+/-} \end{bmatrix}$$

Wahl des Kerns / der Kernparameter

- Der Kernfunktionstyp bzw. die Kernfunktionsparameter werden durch Mika's Algorithmus nicht festgelegt.
- Falls außer der Stichprobe keine Informationen vorliegen …
  - Schätzung der zu erwartenden Trefferquote (oder eines anderen Gütekriteriums) für eine konkrete Kernfunktion z.B. mittels Kreuzvalidierung,
  - Auswahl derjenigen Kernfunktion mit der höchsten Trefferquote (der höchsten Güte).
- Gefahr des Overfitting ist z.B. bei Gaußkernen mit kleinen Bandbreiten hoch …
- … daher Anzahl der Kernfunktionen in einer potentiellen Lösung vorab beschränken.
- Es existieren Algorithmen zur simultanen Kernwahl und Diskriminanzfunktionsbestimmung.

Datensatz



- 1240 Samples
- 420 in Gruppe -1
- 820 in Gruppe +1

Erwartete Trefferquoten (10-fache Kreuzvalidierung)

	Lineare Fisher-DA	Kern-Fisher-DA Gaußkern	Kern-Fisher-DA Quadratischer Polynomkern
Gruppe -1	97%	95%	98%
Gruppe +1	89%	99%	97%

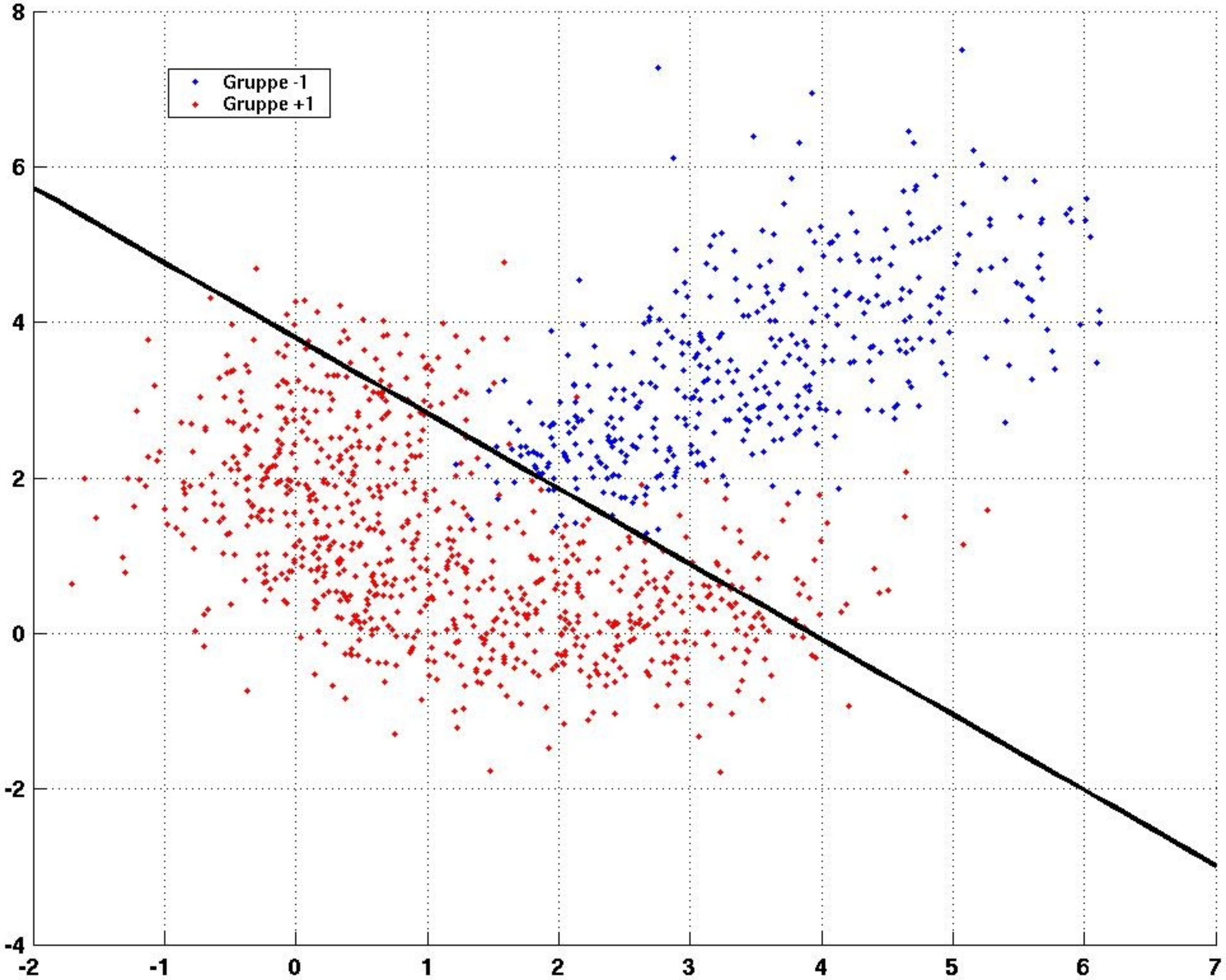
Grafik Grafik Grafik

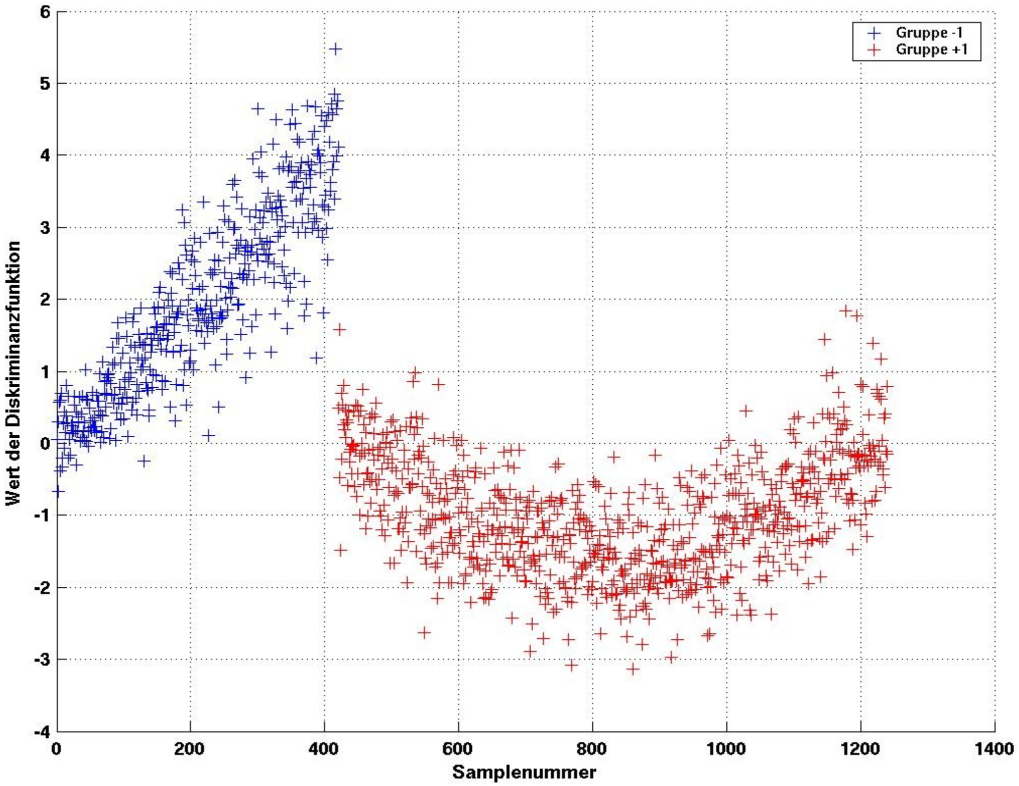
Zu den Gütekriterien für Diskriminanzfunktionen

- Kern-Fisher-DA maximiert die Fisherdiskriminante im funktionalen Hilbertraum …
- … nicht die gruppenweise / totale Trefferquote …
- … was auch nicht unbedingt wünschenswert ist (Generalisierung!) …
- … aber durch Nachjustieren der als Trennfläche verwendeten Niveaufläche der Diskriminanzfunktion angestrebt werden kann und wird.
- Angemessenere Gütekriterien werden aktuell diskutiert …
- … z.B. die Fläche unter der Receiver-Operating-Characteristic, ein globales Gütemaß, das die Trefferquoten zu den verschiedenen Trennflächen innerhalb einer 1-Parameter-familie erfaßt.
- □ Excellenzcluster Klassifikation

Lineare Fisher-DA

$$d(x) = \langle a, x \rangle + b$$

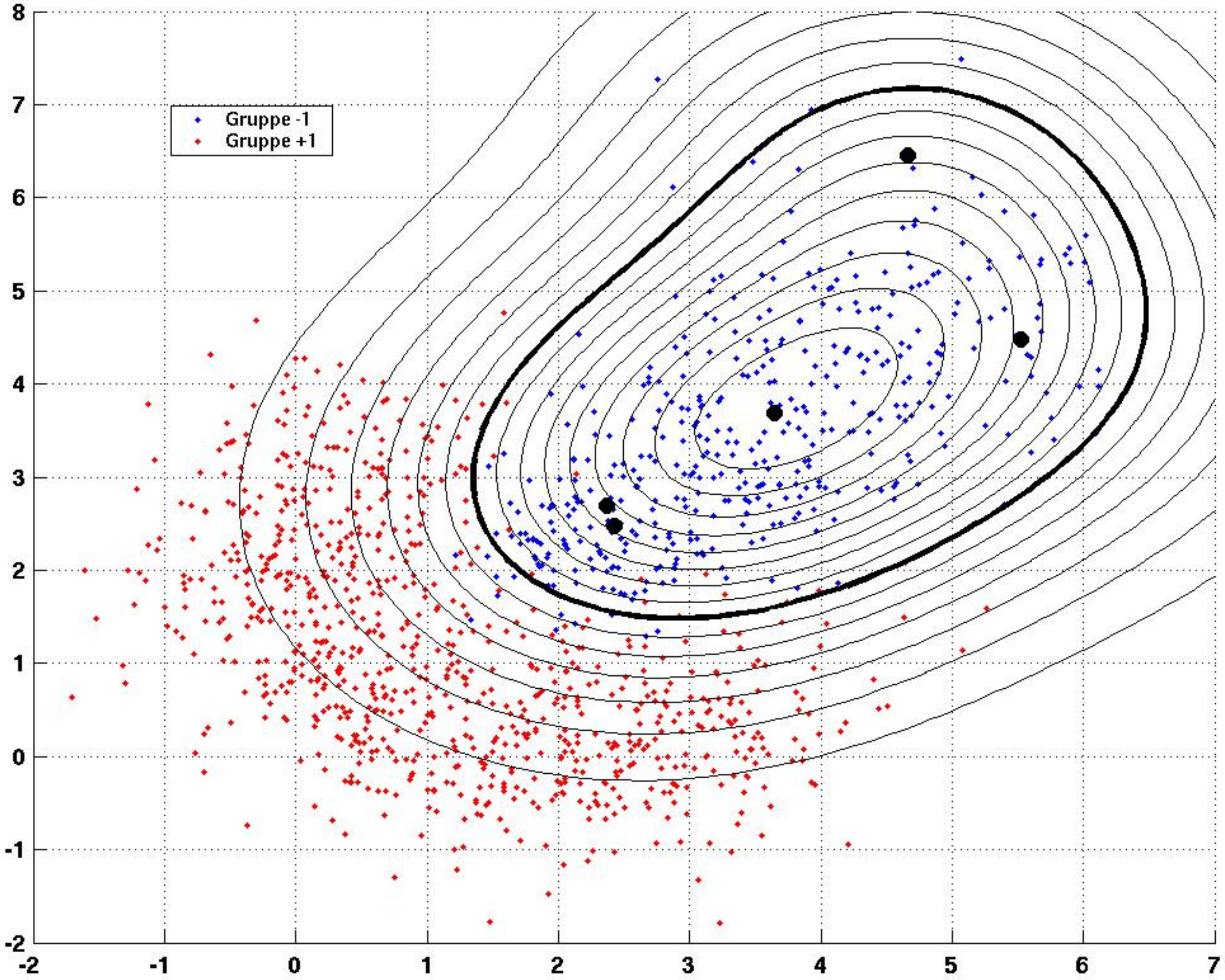




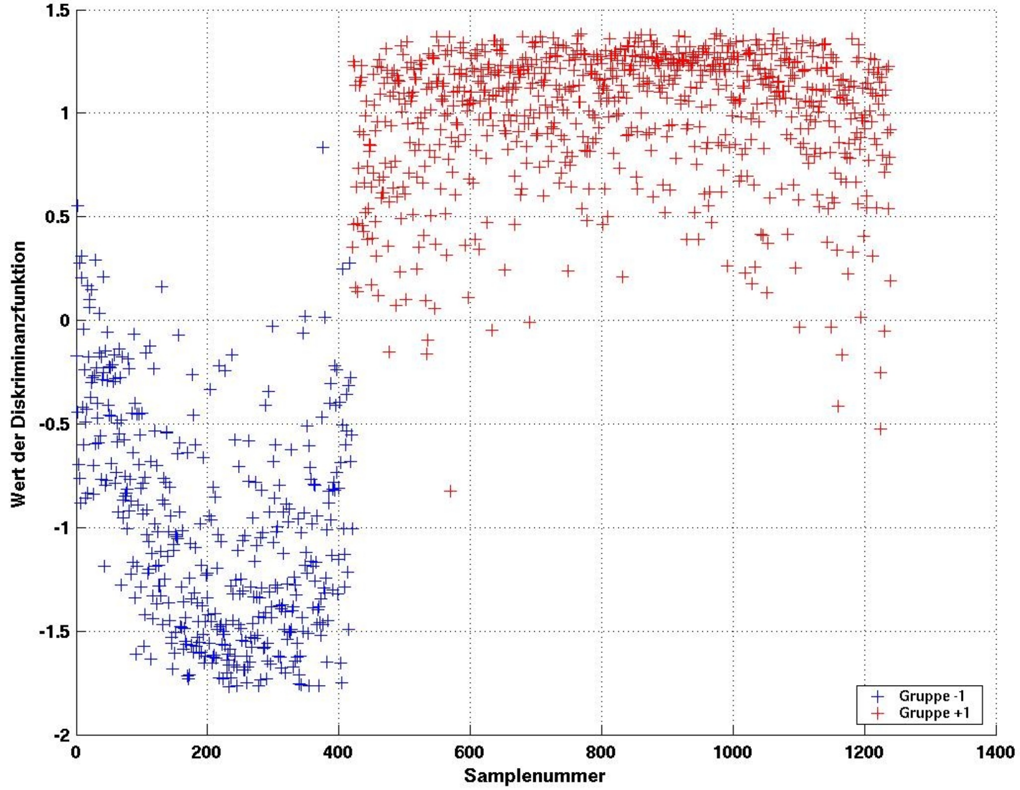
Fraunhofer Institut
Techno- und
Wirtschaftsmathematik

1 2 3 4 5 6

Kern-Fisher-DA / Gaußkern

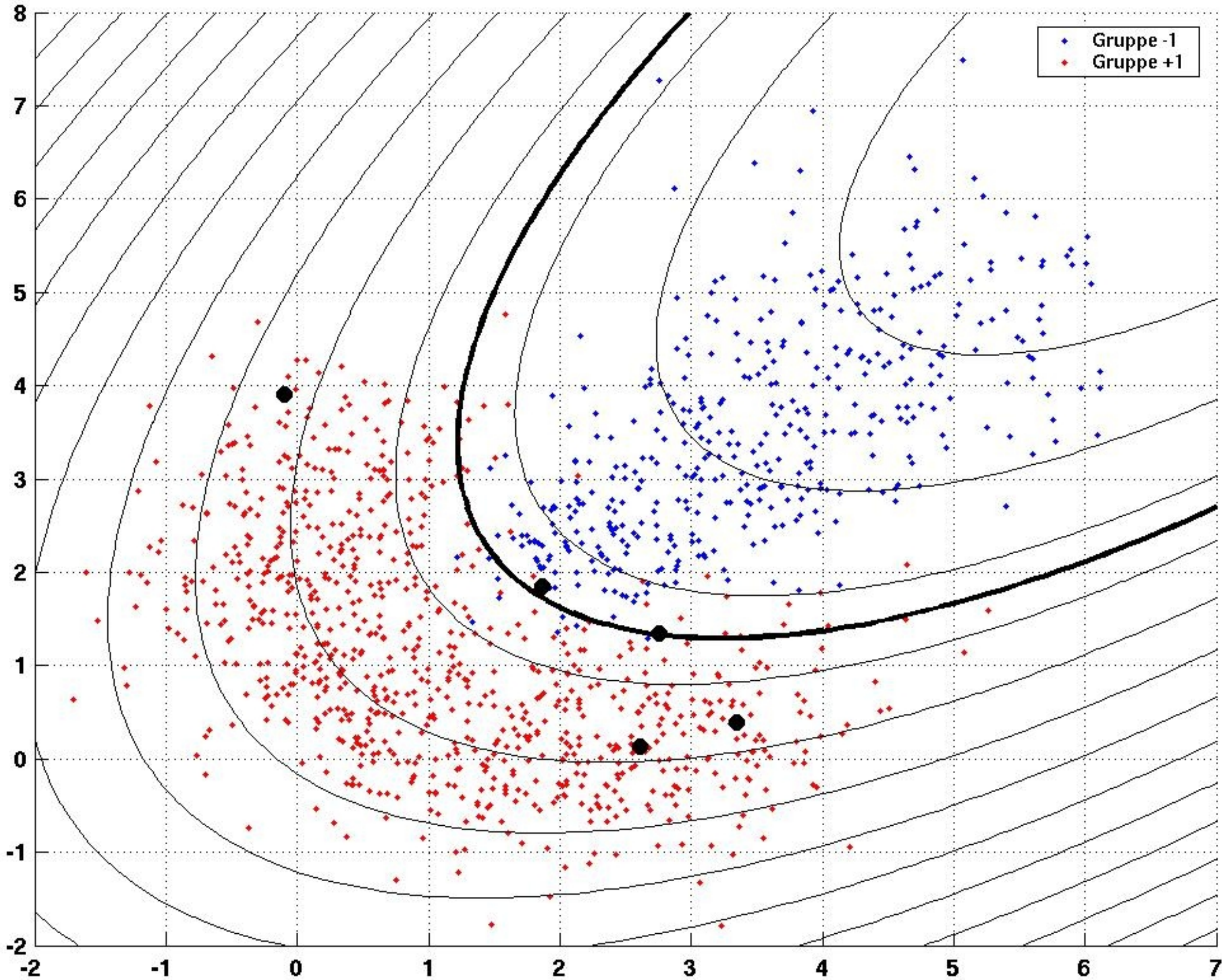


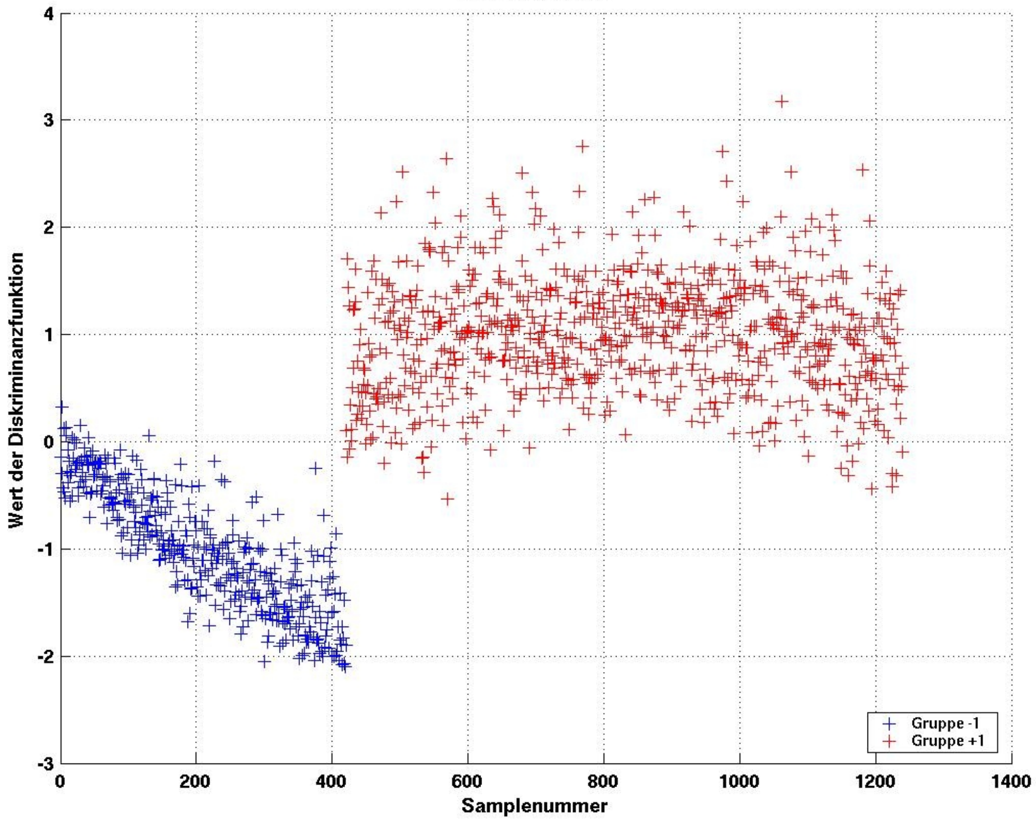
$$d(x) = \sum_{k=1}^{5} \alpha_k e^{-\frac{\|x - x_k\|^2}{0.86^2}}$$



Kern-Fisher-DA/Polynomkern

$$d(x) = \sum_{k=1}^{5} \alpha_k \left(\langle x, x_k \rangle + 0.3 \right)^2$$





Beweis für die Positiv-Semi-Definitheit des reproduzierenden Kerns

$$
\begin{aligned}
z^t A\bar{z} &= \sum_{i=1}^{n}\sum_{j=1}^{n} z_i K(x_i, x_j)\bar{z}_j \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} z_i \left\langle K_{x_j}, K_{x_i}\right\rangle \bar{z}_j \\
&= \left\langle \sum_{j=1}^{n} \bar{z}_j K_{x_j}, \sum_{i=1}^{n} \bar{z}_i K_{x_i} \right\rangle \geq 0
\end{aligned}
$$

Beweis für die Dichtheit des von den Kernfunktionen erzeugten Unterraums

Es genügt zu zeigen, daß einzig der Nullvektor senkrecht auf allen Kernfunktionen steht.

$$
\begin{aligned}
& \forall x \in X : \langle h, K_x \rangle = 0 \\
\Leftrightarrow\ & \forall x \in X : h(x) = 0 \\
\Leftrightarrow\ & h = 0
\end{aligned}
$$

Beweisskizze für das Theorem von Aronzajn-Moore / 1

Existenz

- Betrachte den Vektorraum $W := \sum_{y \in X} \kappa \cdot K_y \subseteq \mathrm{Abb}(X, \kappa), \quad K_y := K(\circ, y);$
- definiere eine Bi-/Sesquilinearform auf W durch:

$$b : W \times W \to \kappa, \quad b\Big(\sum_{y \in X} a_y \cdot K_y, \sum_{y' \in X} b_{y'} \cdot K_{y'}\Big) := \sum_{y \in X} \sum_{y' \in X} a_y \overline{b_{y'}} \cdot K(y', y) \ \text{(Summen endlich)}$$

- b ist positiv definit und erfüllt $b(h, K_y) = h(y)$.
- Definiere den Hilbertraum H als Komplettierung von (W,b),
- es gilt dann: $(h = \lim_{k \to \infty} w_k, \ w_k \in W) \Rightarrow (\forall x \in X : \lim_{k \to \infty} w_k(x) \text{ existiert}),$
- damit ist H via der Auffassung $h(x) := \lim_{k \to \infty} w_k(x)$ funktional über X,

Beweisskizze für das Theorem von Aronzajn-Moore / 2

- schließlich gilt: $(h = \lim_{k \to \infty} w_k,\ w_k \in W) \Rightarrow (\forall x \in X : \langle h, K_x \rangle = h(x)),$
- und damit sind die Auswertungsfunktionale stetig ...
- ... und der reproduzierende Kern von H ist gleich K.

Lemma

Sei H ein funktionaler Hilbertraum über der Menge X, mit stetigen Auswertungsfunktionalen.

Dann gilt: Normkonvergenz in H impliziert punktweise Konvergenz auf X.

Beweisskizze für das Theorem von Aronzajn-Moore / 3

Eindeutigkeit

- Betrachte die durch die Kernfunktionen in H_1, H_2 erzeugten Untervektorräume W_1, W_2 ,
- es gilt $\left(h = \sum_{k=1}^{n} a_{y_k} \cdot K_{y_k} \in W_i\right) \Rightarrow \left(h(x) = \sum_{k=1}^{n} a_{y_k} \cdot K_i(x, y_k)\right)$ $(i = 1{,}2)$,
- und damit $W_1 = W_2 =: W$.
- Weiter gilt: $\left(h = \sum_{k=1}^{n} a_{y_k} \cdot K_{y_k} \in W_i\right) \Rightarrow \left(\|h\|_i^2 = \sum_{j=1}^{n} \sum_{k=1}^{n} a_{y_j} \overline{a}_{y_k} \cdot K_i(y_k, y_j)\right)$ $(i = 1{,}2)$,
- und damit: $\|\circ\|_1 = \|\circ\|_2$ auf W.
- Jedes $h \in H_1$ ist Limes einer Cauchyfolge $(w_k)_k$ aus W; diese hat einen Grenzwert

  $g \in H_2$ womit nach dem Lemma h=g folgt.
- Es folgt $H_1 = H_2$ und $\|\circ\|_1 = \|\circ\|_2$ wegen der Dichtheit von W.