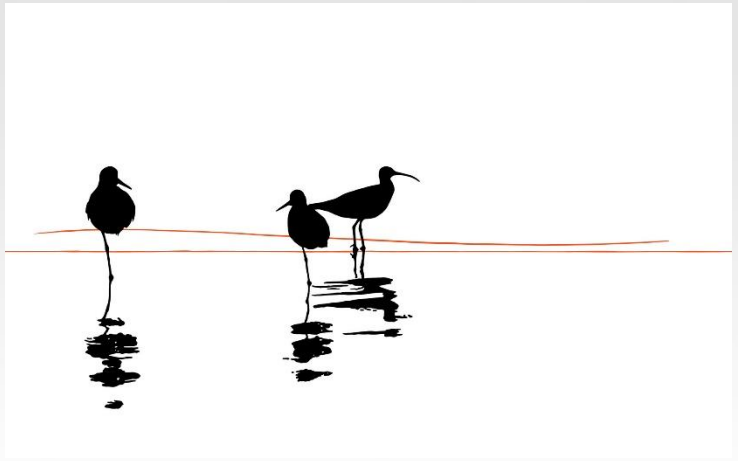
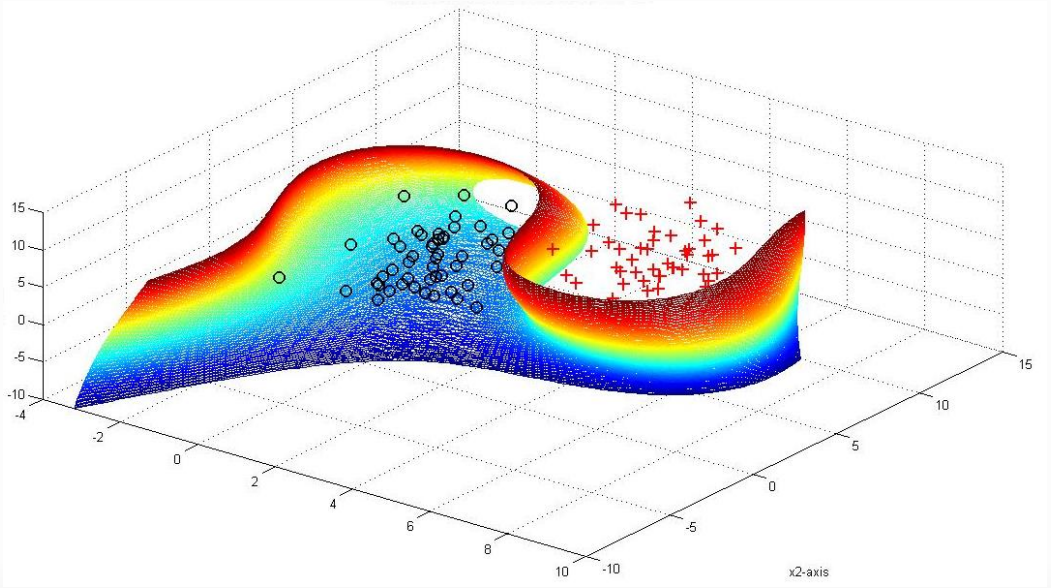


Functional Analysis through Applications: Kernel Methods in Data Mining

Hagen Knaf, Hochschule RheinMain

Outline

5. What is Data Mining?
6. Discriminant Analysis – an Overview
7. Kernel Fisher Discriminant Analysis
8. The Kernel Method in general



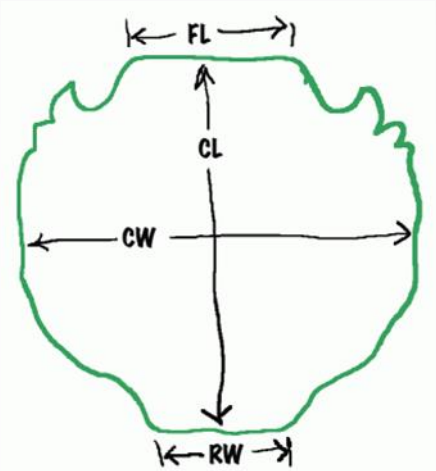
What is Data Mining?

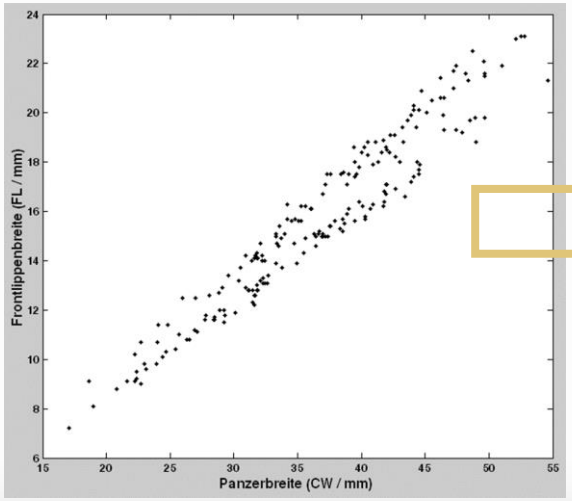
»Definitions«

- *Data Mining*: any activity aiming at the extraction of potentially useful information from given data.
- *Data*: a finite set of objects (*samples*) of the same type/structure. Typically an object consists of an ordered collection (*tupel*) of numbers, texts, images etc. The components of such a tupel are called *attributes*.
  - Example: (First name, Surname, Sex, Age, Weight)
- *Information*: description of the data in terms of relations between the attribute values.

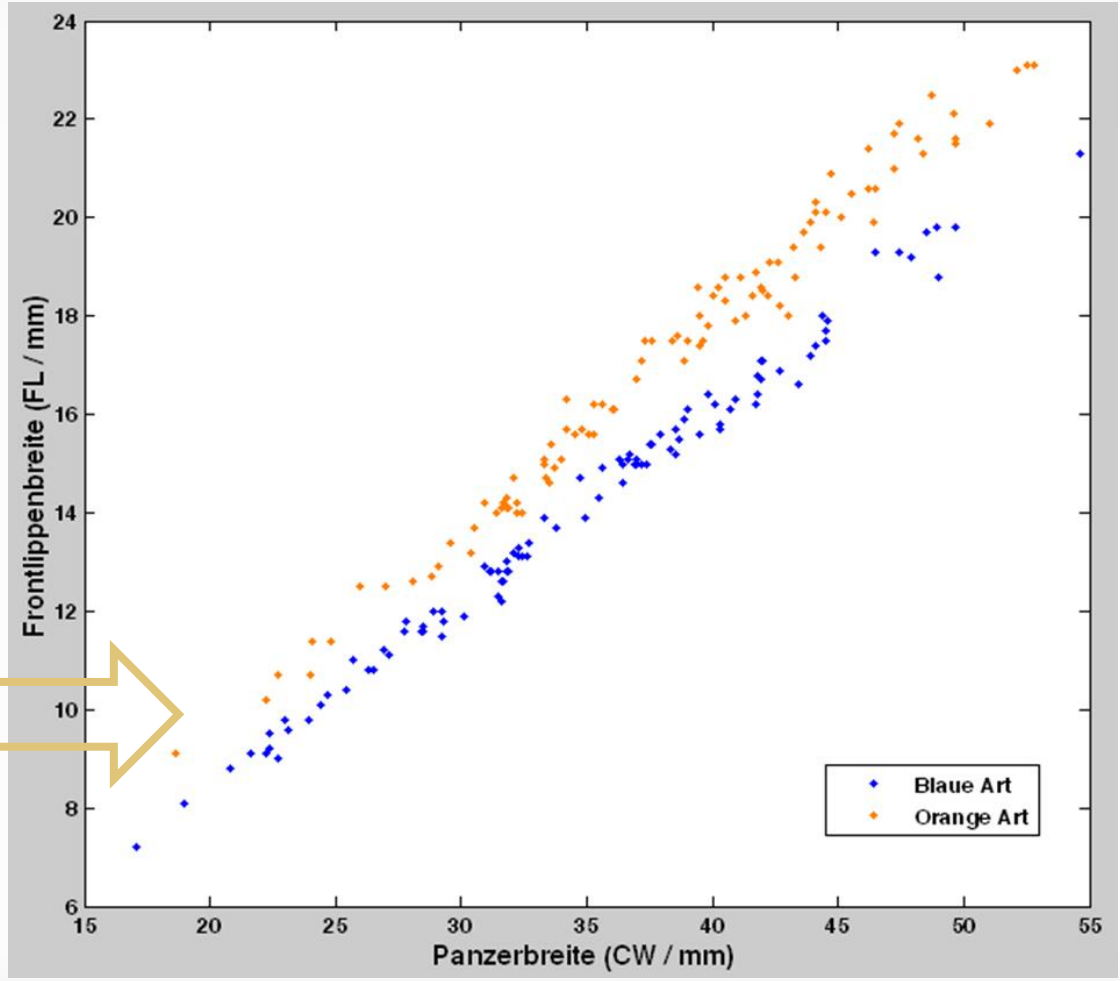
What is Data Mining?

Data: sizes of crap carpaces





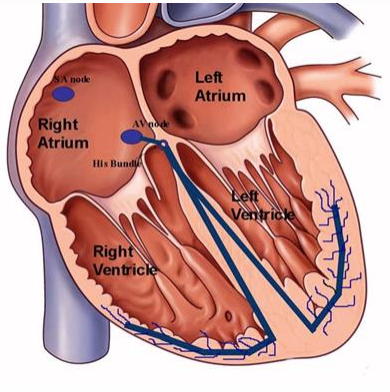
Information: subdivision of the data into two groups

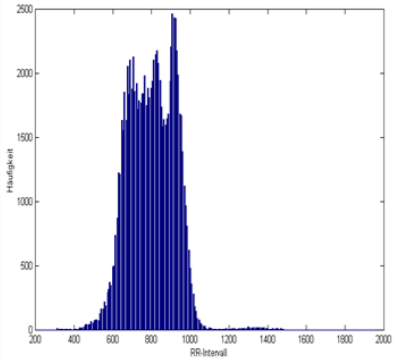


What is Data Mining?

Data acquisition

- Data can be collected via specifically prepared measurements, data base searches, polls etc.
- Data from different sources are often mixed together.
- The process of data acquisition is frequently considered/modelled as a stochastic process (measurement errors, inherent random-ness).





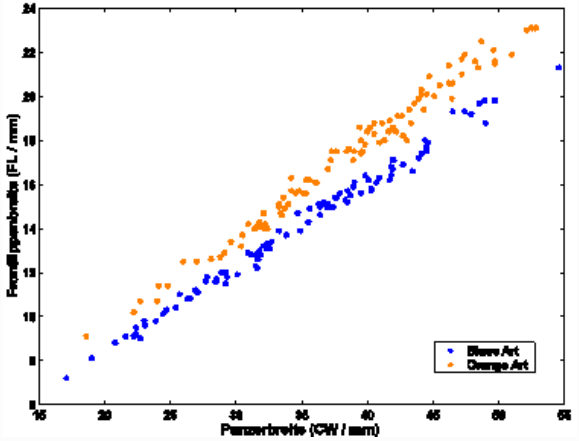
What is Data Mining?

Principal goals of data mining

- Find new previously unknown information (*unsupervised learning*).
- Confirm hypotheses and specify them precisely (*supervised learning*).

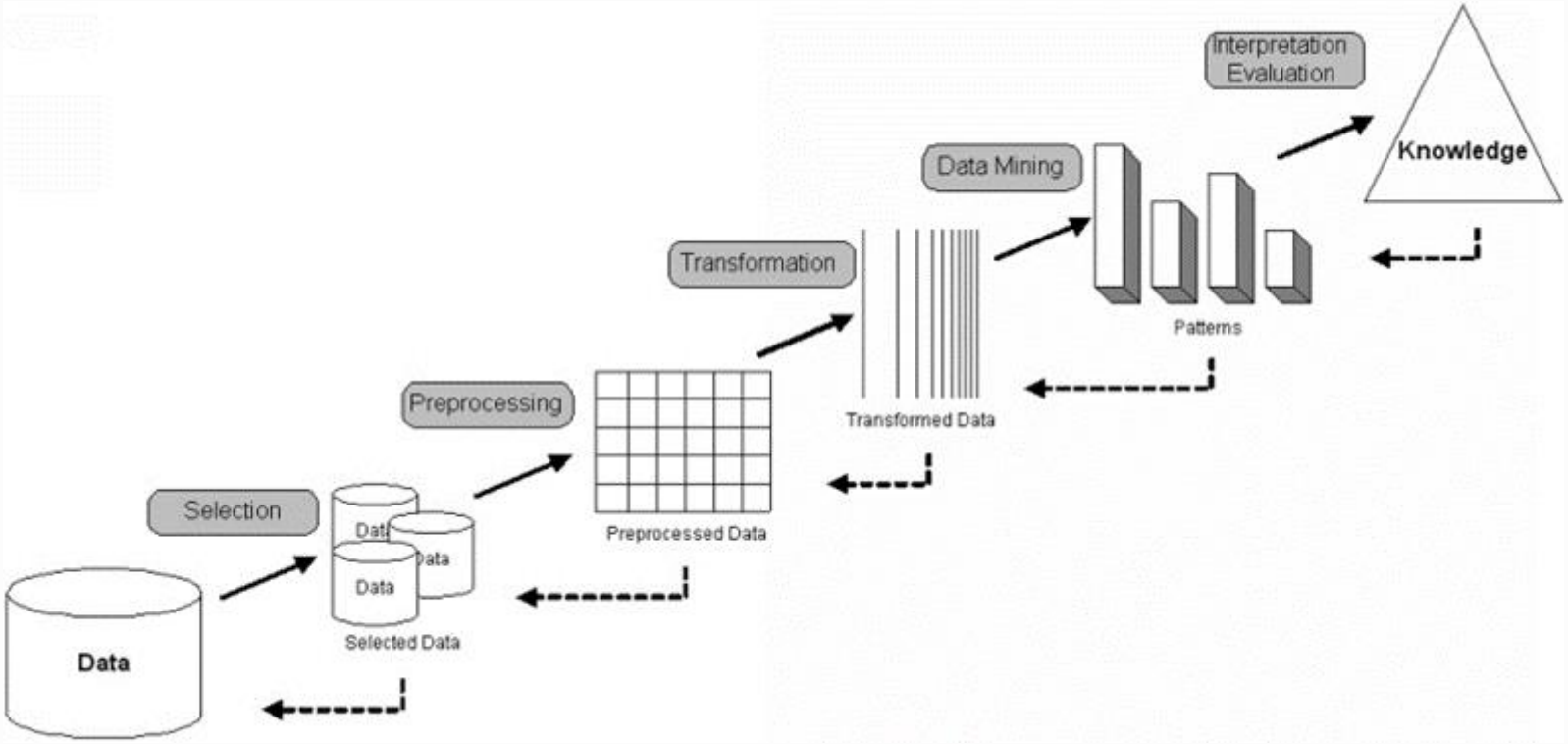
Mathematical methods in data mining

- Multivariate Statistics
- Geometry (e.g. metric spaces)
- Analysis



What is Data Mining?

Data mining as a standardised process



Process model of U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, 1995

Discriminant Analysis – an Overview

The problem addressed by discriminant analysis

- Given: a finite set $X \subset \mathbb{R}^m$ subdivided into $r \geq 2$ pairwise disjoint subsets $X = X_1 \cup X_2 \cup \ldots \cup X_r$.

  It is assumed that X is randomly drawn from a *population* $\Omega \subseteq \mathbb{R}^m$ subdivided into r pairwise disjoint subsets $\Omega_i \supseteq X_i$ called *classes*.

- Given: a class $\mathcal{F}$ of functions $f : \mathbb{R}^m \to \mathbb{R}$.

- Find functions $d_1, \ldots, d_r \in \mathcal{F}$ (so-called *discriminant functions*) such that the *classification rule*

  ***C**: $x \in \mathbb{R}^m$ is assigned to the class Ω_k if and only if*

  $$d_k(x) = \max(d_i(x) : i = 1, \ldots, r).$$

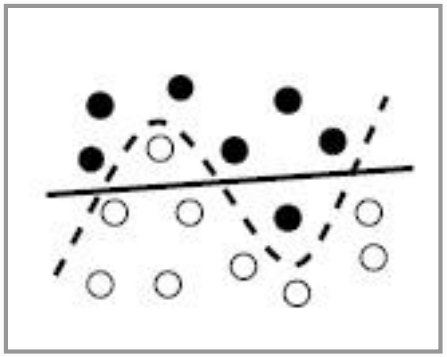
  is »sufficiently good«.

Discriminant Analysis – an Overview

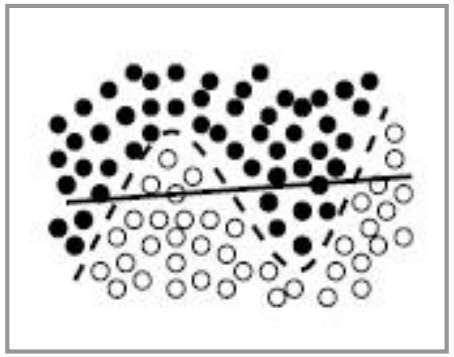
The quality of a classification rule

The quality of the classification rule $\mathbf{C}$ is estimated based on the following requirements:

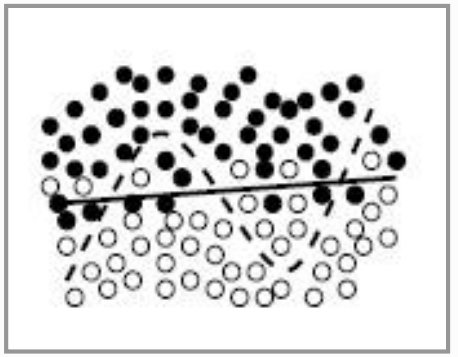
- Its hitrate (either total or a weighted mean of the class-wise hitrates) on the partition $X = X_1 \cup X_2 \cup \ldots \cup X_r$ is high.
- The likelihood for correct assignment of new samples $x \in \Omega$ is high.



Two classification rules
In the case m=2, r=2



Truth A



Truth B

Discriminant Analysis – an Overview

Special cases of discriminant analysis

- In the case of $r = 2$ classes $\Omega = \Omega_1 \cup \Omega_2$ in the population it suffices to determine *one* discriminant function instead of two: using the function $d(x) := d_2(x) - d_1(x)$ the original classification rule becomes

  $\boldsymbol{C}$*:* $x \in \mathbb{R}^m$ *is assigned to the class* Ω_2 *if and only if* $d(x) > 0$.

  $x \in \mathbb{R}^m$ *is assigned to the class* Ω_1 *if and only if* $d(x) < 0$.

- The set $S := \{x \in \mathbb{R}^m : d(x) = 0\}$ is a hypersurface separating the two classes.

- If $\mathcal{F}$ consists of all affine functions $f : \mathbb{R}^m \to \mathbb{R}$, that is $f(x) = \langle a, x\rangle_2 + b$, $a \in \mathbb{R}^m$, $b \in \mathbb{R}$ one speaks of *linear discriminant analysis.*

  If in addition $r = 2$ the separating hypersurface is a hyperplane, and the vector a is orthogonal to the vectors in that hyperplane.

Discriminant Analysis – an Overview

Linear discriminant analysis: Fisher's approach for $r = 2$

- Situation: $\mathbb{R}^m \supset X = X_1 \cup X_2,\ X_1 \cap X_2 = \emptyset,\ \sum_{x \in X} \mathbb{R}x = \mathbb{R}^m,$

$$\mathcal{F} = \{f(x) = \langle a, x \rangle_2 + b : a \in \mathbb{R}, \|a\|_2 = 1, b \in \mathbb{R}\}.$$

- Approach: determine $d(x) = \langle a^*, x \rangle_2 + b^* \in \mathcal{F}$ such that the separating hyperplane $S := \{x \in \mathbb{R}^m : d(x) = 0\}$ has the properties:

  1. the orthogonal projections of the sets X_1, X_2 to the line $g = \mathbb{R}a^*$, possess small spread.
  2. the centers of these orthogonal projections lie far apart.

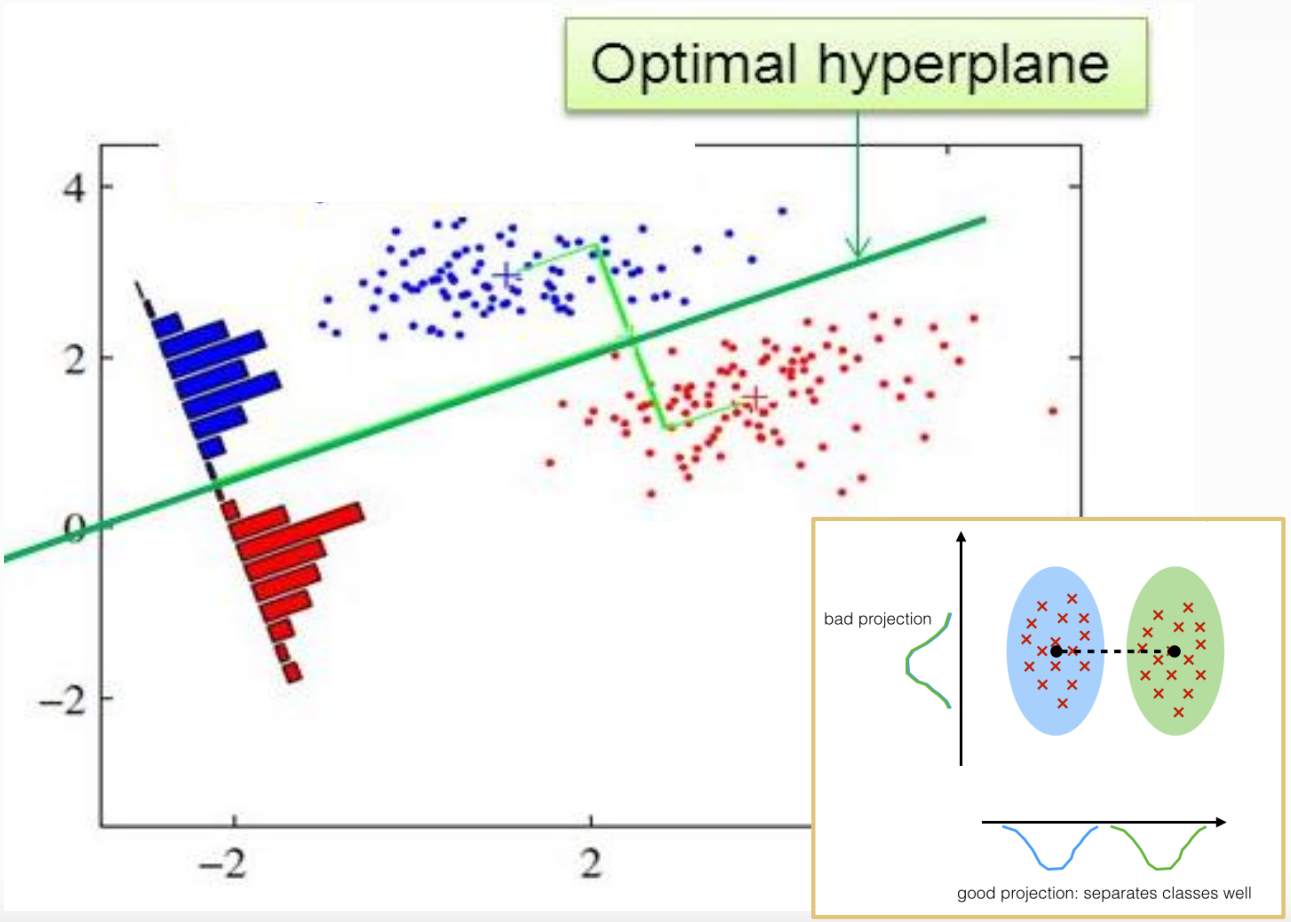
- Note that the orthogonal projection $p : \mathbb{R}^m \to \mathbb{R}a$ for $\|a\|_2 = 1$ is given by $p(x) = \langle a, x \rangle_2 \, a$.

- The position of the hyperplane S is determined only up to translations.

Discriminant Analysis – an Overview

Linear discriminant analysis: Fisher's approach for $r = 2$

Sir R. A. Fisher
1890 – 1962



Discriminant Analysis – an Overview

An analytic solution to Fisher's approach for $r = 2$

- For $\overline{x}_i := \frac{1}{|X_i|} \sum_{x \in X_i} x$ the element $\langle a, \overline{x}_i \rangle_2 \, a$ is the center of $p(X_i)$.
- The spread of $p(X_i)$ is quantified through $s_i^2 := \sum_{x \in X_i} (\langle a, x \rangle_2 - \langle a, \overline{x}_i \rangle_2)^2$.
- In $d(x) = \langle a^*, x \rangle_2 + b^*$ let $a^* \in \mathbb{R}^m$ be a solution of the optimisation problem

$$\max\left(\frac{(\langle a, \overline{x}_1 \rangle_2 - \langle a, \overline{x}_2 \rangle_2)^2}{s_1^2 + s_2^2} : a \in \mathbb{R}^m, \, \|a\|_2 = 1\right).$$

- To solve this problem one brings it into matrix form by expressing a as a linear combination of a system of generators of $\mathbb{R}^m$.
- Normally one would use a basis of $\mathbb{R}^m$, but in the present context it is essential to use $X = \{x_1, \ldots, x_n\}$ itself as generators.

Discriminant Analysis – an Overview

An analytic solution to Fisher's approach for $r = 2$

This leads to the following matrix form of the optimisation problem

$$\max\left(\frac{\alpha^t B \alpha}{\alpha^t W \alpha} : \alpha \in \mathbb{R}^n \setminus 0\right),$$

where:

- $\alpha := (\alpha_1, \ldots, \alpha_n)^t$ mit $\sum_{i=1}^{n} \alpha_i x_i = a$,
- $B := (m_1 - m_2)(m_1 - m_2)^t \in \mathbb{R}^{n \times n}$, $m_k := (\langle x_1, \overline{x}_k \rangle, \ldots, := (\langle x_n, \overline{x}_k \rangle)^t$,
- $W := (\langle x_i, x_j \rangle)_{i,j} (\langle x_i, x_j \rangle)_{i,j}^t - |X_1| m_1 m_1^t - |X_2| m_2 m_2^t \in \mathbb{R}^{n \times n}$.

Discriminant Analysis – an Overview

An analytic solution to Fisher's approach for $r = 2$

- The solution of the optimisation problem $\max(\frac{\alpha^t B \alpha}{\alpha^t W \alpha} : \alpha \in \mathbb{R}^n \setminus 0)$ is

$$\alpha^* = W^{-1}(m_1 - m_2)$$

  provided that W is invertible and $m_1 \neq m_2$.

- For the general case Sebastian Mika has provided an algorithm to find sparse solutions of the optimisation problem, that is solutions $\alpha = (\alpha_1, \ldots, \alpha_n)$ possessing many components having small absolute value.

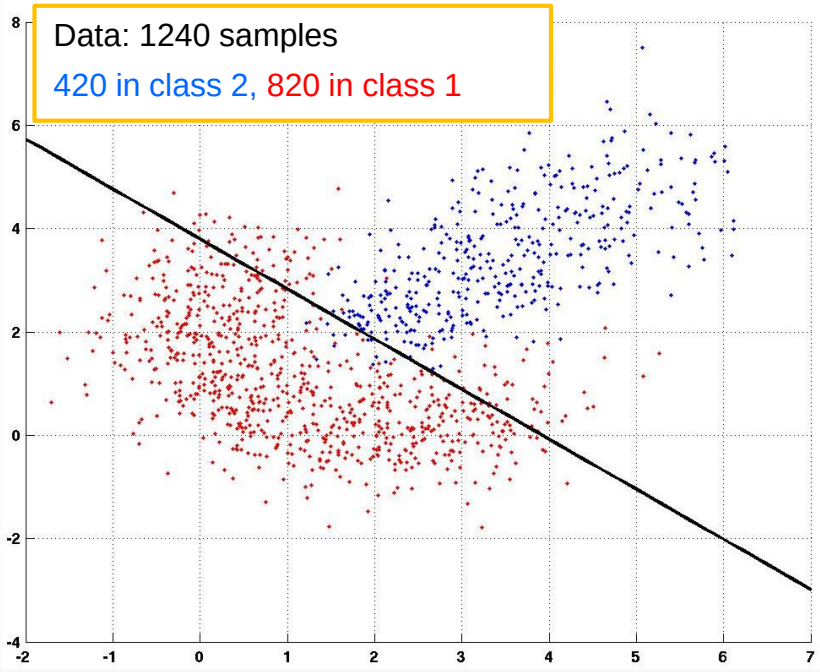
- Note that the optimisation problem is formulated entirely in terms of scalar products $\langle x_i, x_j \rangle$ between data points.

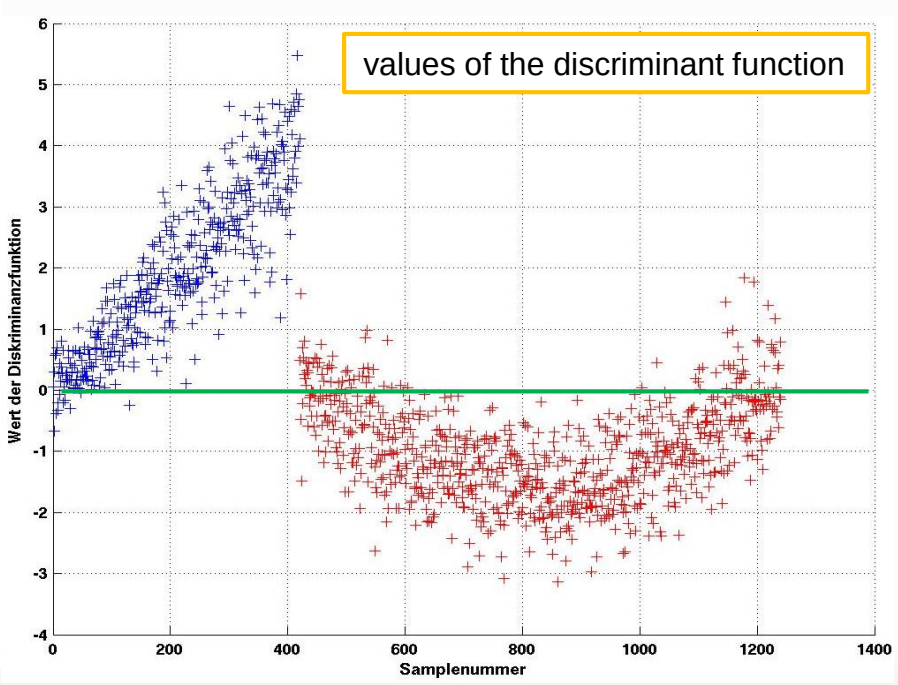
- $b^* := -\frac{1}{2}(\alpha^*)^t(m_1 + m_2)$ provided that the separating hyperplane S lies in the middle between the centers $\overline{x}_1$ and $\overline{x}_2$.

Discriminant Analysis – an Overview

An unsatisfying case of Fisher's discriminant analysis (Example 1)

- Hitrates are sufficient … at least for the »blue class«.
- The separating straight line however doesn't look convincing – a curved line seems to be more appropriate.





Kernel Fisher Discriminant Analysis

The approach

- Situation: $\mathbb{R}^m \supset \{x_1, x_2, \ldots, x_n\} = X = X_1 \cup X_2$, $\sum_{i=1}^{n} \mathbb{R}x = \mathbb{R}^m$,

  K_θ, $\theta \in \Theta$, a family of kernel functions on $\mathbb{R}^m$,

  such that the maps $\phi_\theta : \mathbb{R}^m \to H(K_\theta)$ are injective.

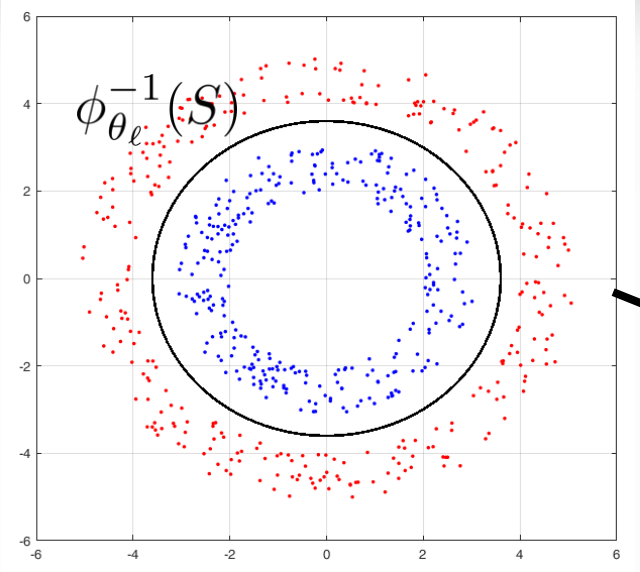
- Approach: Choose parameters $\theta_1, \ldots, \theta_r \in \Theta$.

  Perform Fisher's discriminant analysis of $\phi_{\theta_i}(X) = \phi_{\theta_i}(X_1) \cup \phi_{\theta_i}(X_2)$

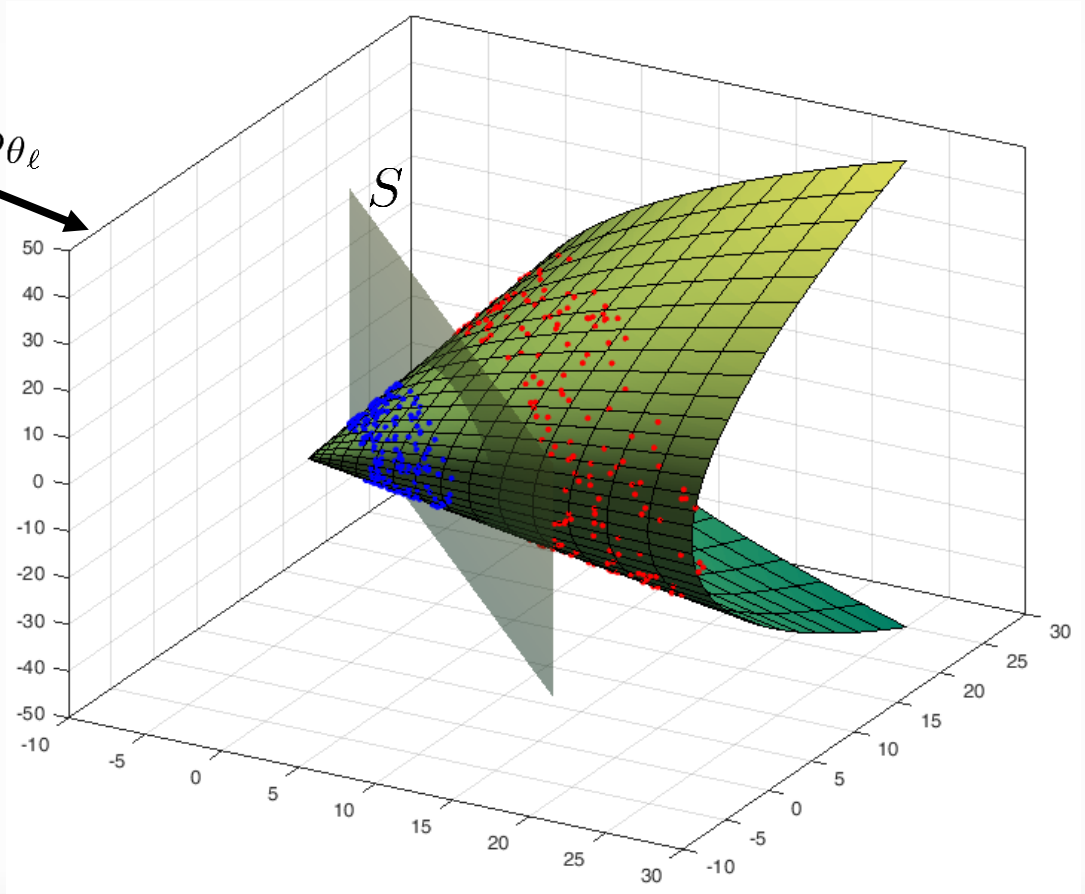
  in $U_i := \sum_{y \in X} \mathbb{R}\phi_{\theta_i}(y)$ to obtain discriminant functions $d_i : U_i \to \mathbb{R}$.

  Select a discriminant function d_ℓ of highest quality and take the pullback $d : \mathbb{R}^m \to \mathbb{R}$, $x \mapsto d_\ell(\phi_{\theta_\ell}(x))$ as a discriminant function for $X = X_1 \cup X_2$.

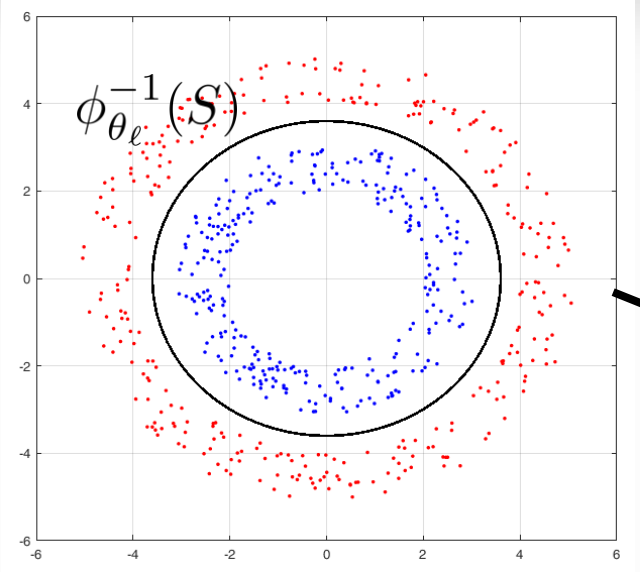
Kernel Fisher Discriminant Analysis



Using the kernel function $K_{\theta_\ell}(x) = \langle x, y \rangle^2$.

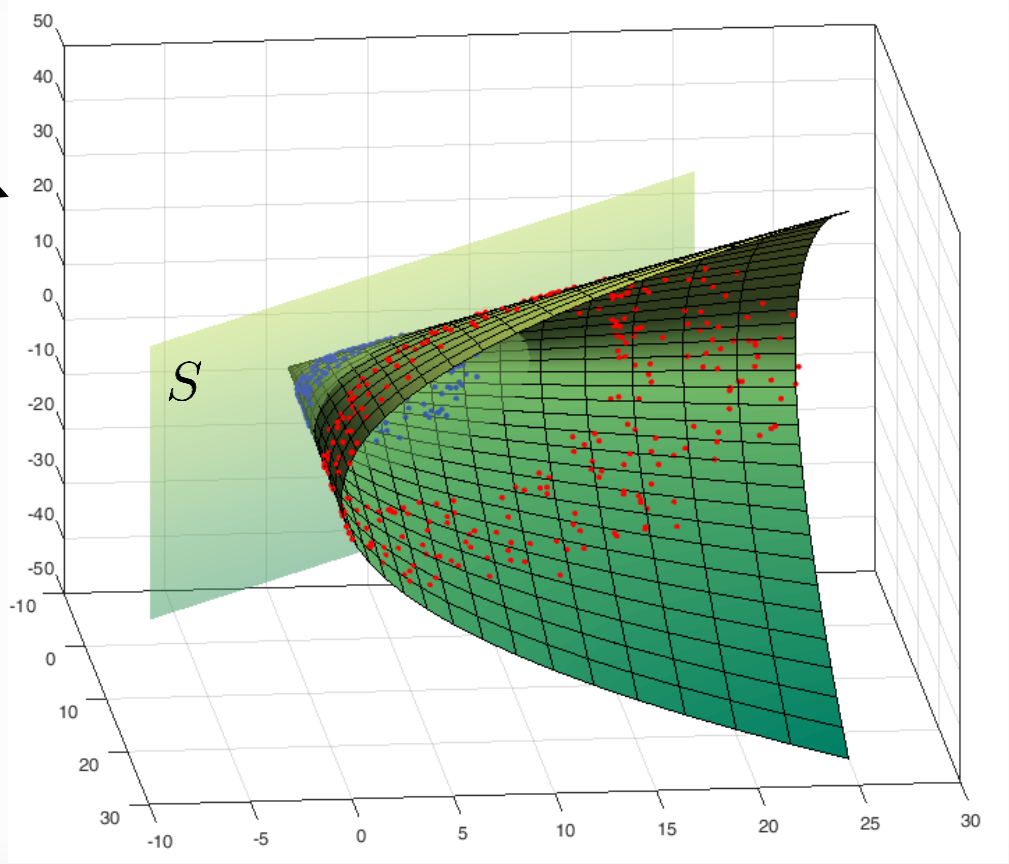


Kernel Fisher Discriminant Analysis



ϕ_{θ_ℓ}

Using the kernel function $K_{\theta_\ell}(x) = \langle x, y \rangle^2$.



Kernel Fisher Discriminant Analysis

Finding solutions

- Situation: $\mathbb{R}^m \supset \{x_1, x_2, \ldots, x_n\} = X = X_1 \cup X_2$, $\sum_{i=1}^{n} \mathbb{R}x_i = \mathbb{R}^m$,

  K a kernel function on $\mathbb{R}^m$, $\phi : \mathbb{R}^m \to H(K)$ is injective.

- Performing Fisher's discriminant analysis of $\phi(X) = \phi(X_1) \cup \phi(X_2)$ in $U := \sum_{j=1}^{n} \mathbb{R}\phi(x_j) = \sum_{j=1}^{n} \mathbb{R}K(\cdot, x_j)$ yields a discriminant function $d : U \to \mathbb{R},\ u \mapsto \langle a^*, u\rangle + b^*$, where $a^* = \sum_{i=1}^{n} \alpha_i^* K(\cdot, x_i)$.

- For the pullback $d \circ \phi$ die identity $\langle \phi(x_i), \phi(x)\rangle = K(x, x_i)$ hence gives

$$(d \circ \phi)(x) = \langle \sum_{i=1}^{n} \alpha_i^* K(\cdot, x_i), K(\cdot, x)\rangle + b^* = \sum_{i=1}^{n} \alpha_i^* K(x, x_i) + b^*.$$

Kernel Fisher Discriminant Analysis

Finding solutions

- Note that the coefficients $\alpha^* = (\alpha_1^*, \ldots, \alpha_n^*)$ of the pullback

$$a^* = \sum_{i=1}^{n} \alpha_i^* K(\cdot, x_i)$$

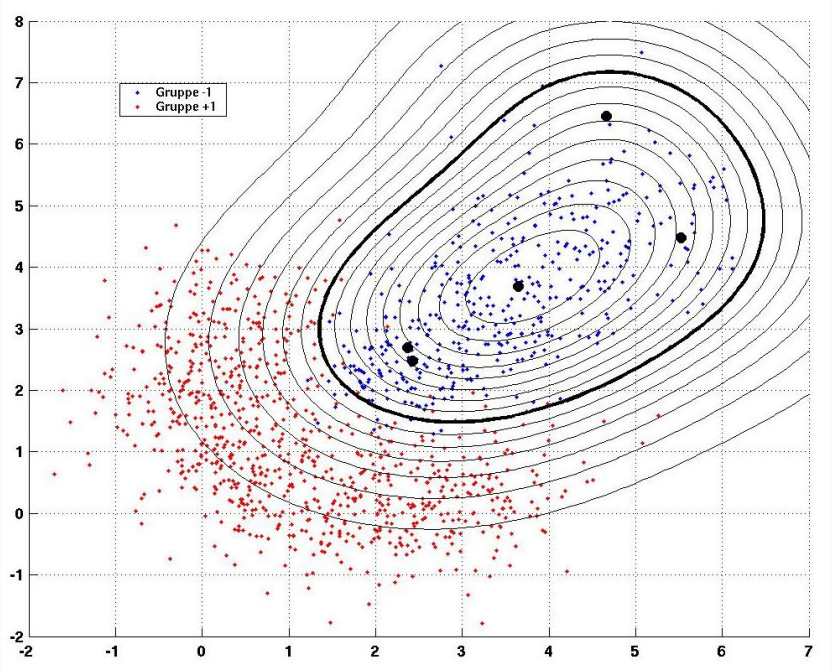
can be computed using scalar products $\langle \phi(x_i), \phi(x_j) \rangle$ thus the values $K(x_i, x_j)$ of the kernel function K only.

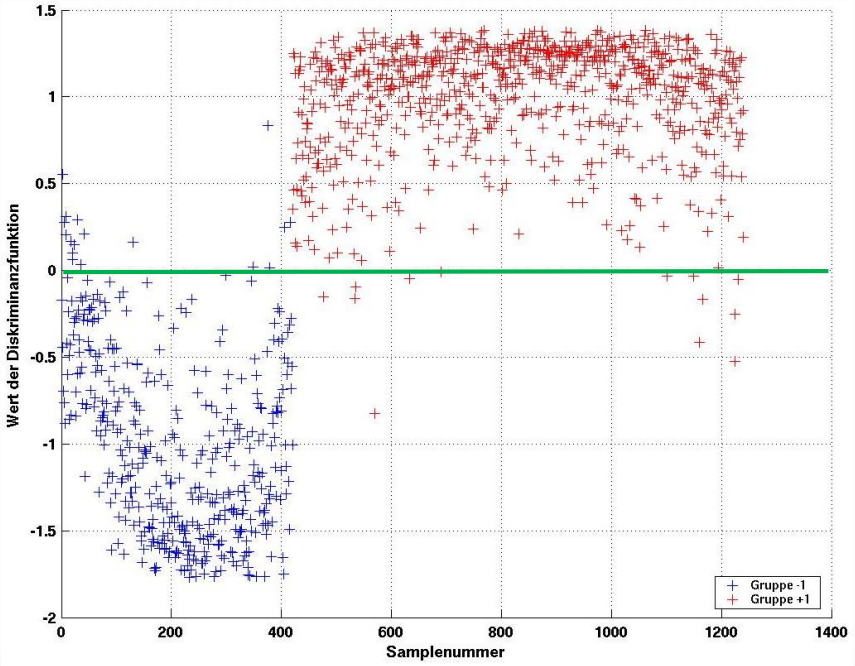
Kernel Fisher Discriminant Analysis

Example 1 using the Gauß kernel

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

values of the discriminant function



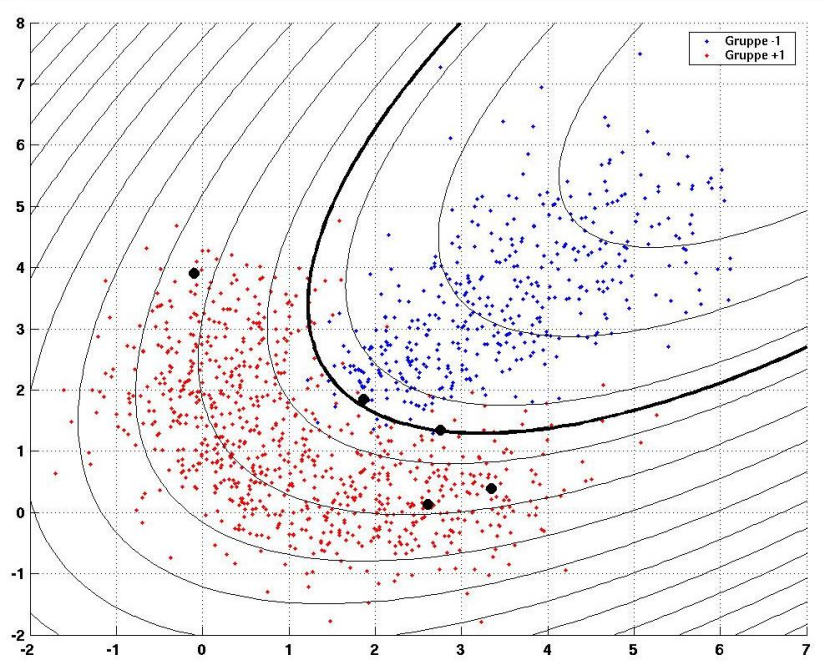


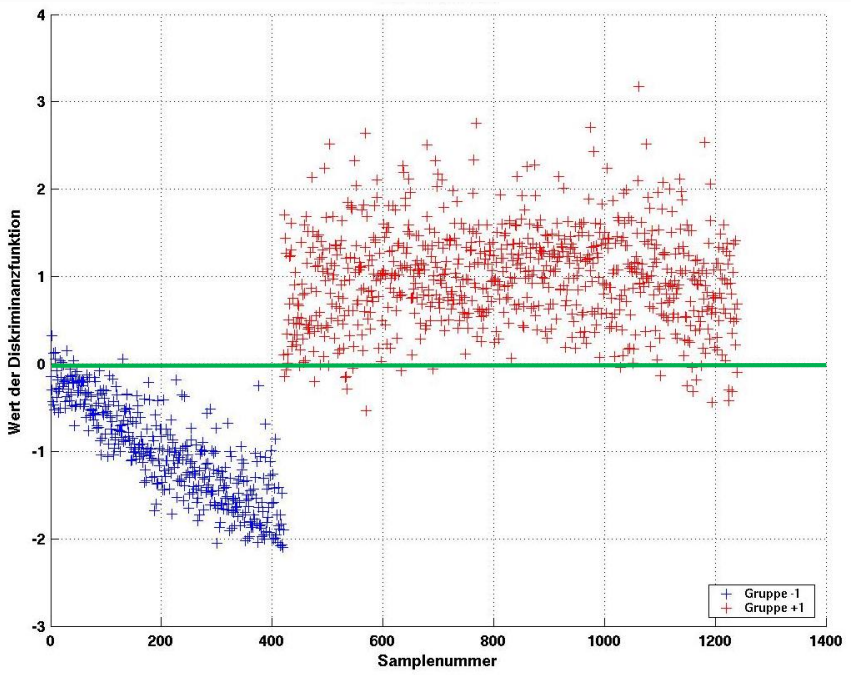
Kernel Fisher Discriminant Analysis

Example 1 using the polynomial kernel of degree 2

$$f(x) = \sum_{k=1}^{5} \alpha_k y_k k(x_k, x) - b$$

values of the discriminant function

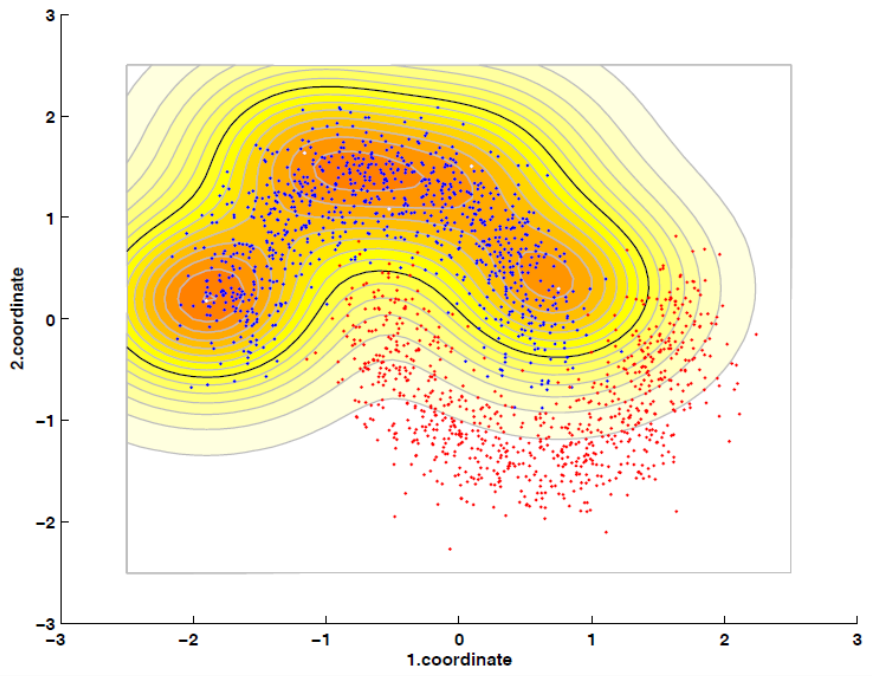


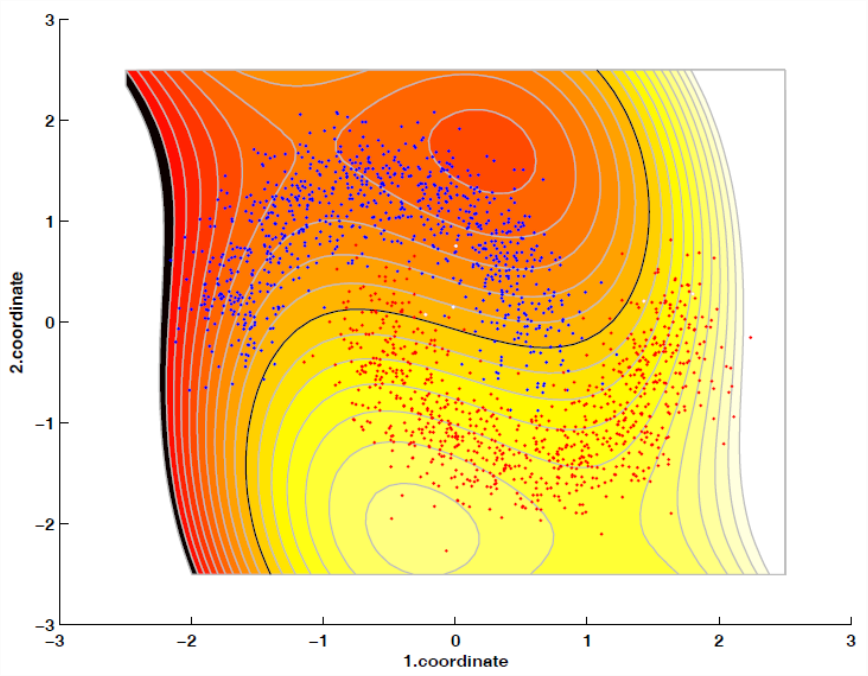


Kernel Fisher Discriminant Analysis

Example 2: Gauß kernel versus polynomial kernel of degree 3

- 5 summands, bandwidth h=0.858, c=0.75

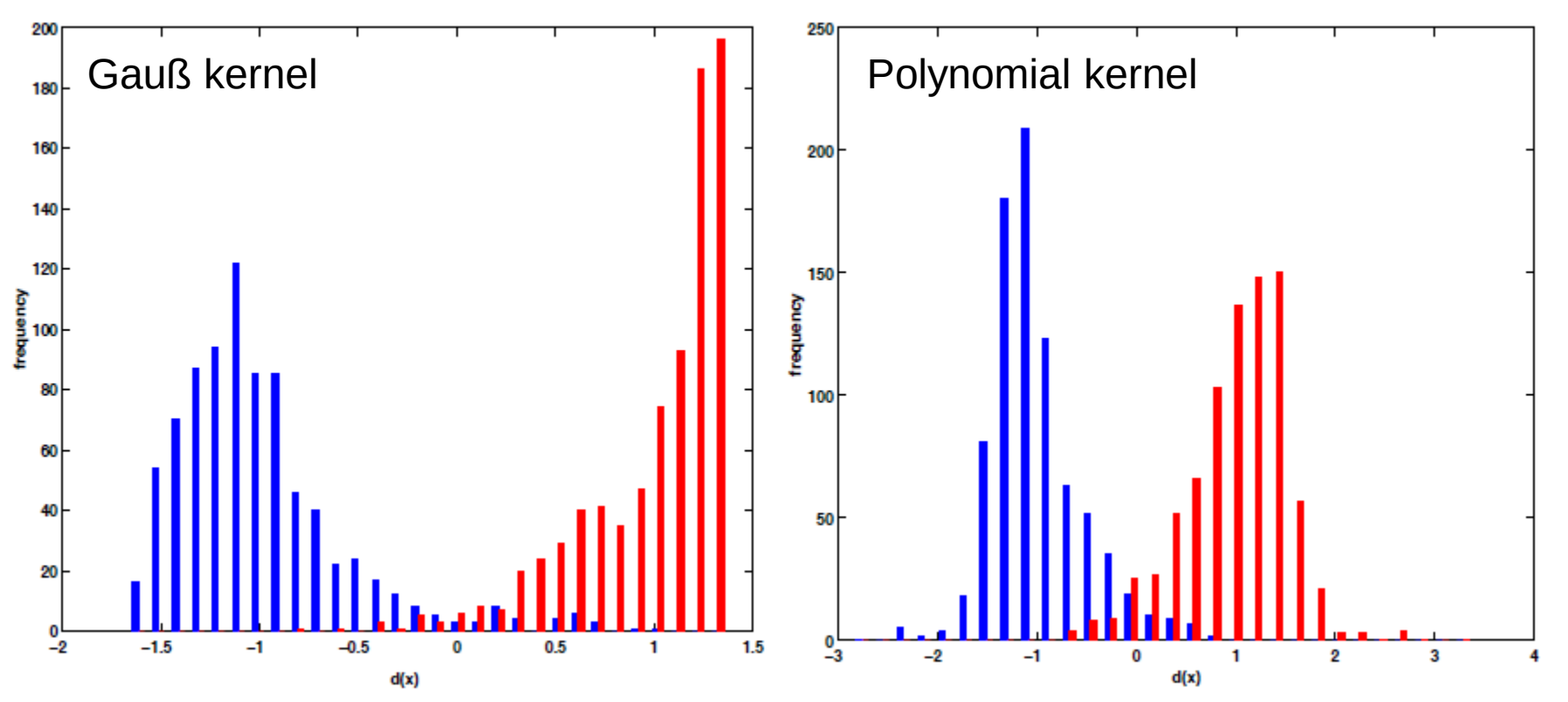




Kernel Fisher Discriminant Analysis

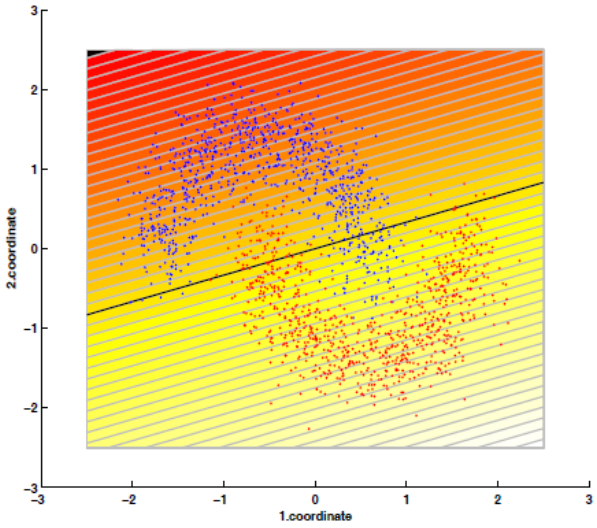
Example 2: Gauß kernel versus polynomial kernel of degree 3

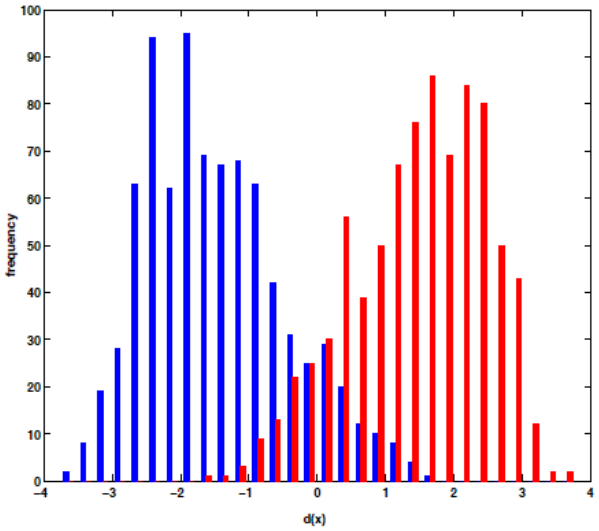
- Distribution of the discriminant function values d(x)



Kernel Fisher Discriminant Analysis

Example 2: Gauß kernel versus polynomial kernel of degree 3





Hitrates

Linear Fisher:	**89,5% , 91,1%**
Gauß kernel:	**96,1% , 98,1%**
Polynomial kernel:	**96,1% , 95%**

The Kernel Method in general

What can we learn from Kernel Fisher discriminant analysis concerning general applications of reproducing kernel Hilbert spaces?

We have performed the following steps:

- Embed a nonlinear problem / task into a RKHS H(K).
- Solve the linear version of the problem / task in H(K).
- Pull the solution back to the original space.

To solve the linear version in H(K) it was actually not necessary to work in H(K) due to the identity

$$K(x, y) = \langle \phi(x), \phi(y) \rangle$$

that allowed to work in the original space using the kernel function K.

The Kernel Method in general

What can we learn from Kernel Fisher discriminant analysis concerning general applications of reproducing kernel Hilbert spaces?

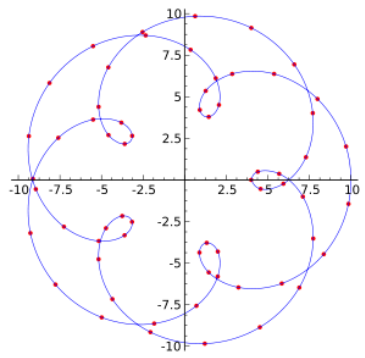
The method used to develop Kernel Fisher discriminant analysis works for problems / tasks possessing the following features:

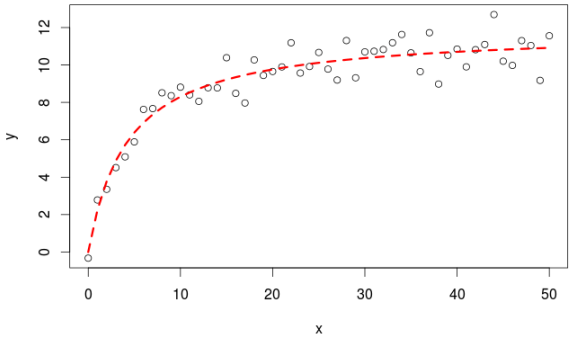
- There exists a linear version of the problem / task.
- There exists a solution / an algorithm for finding the solution that works with scalar products of the input data only.

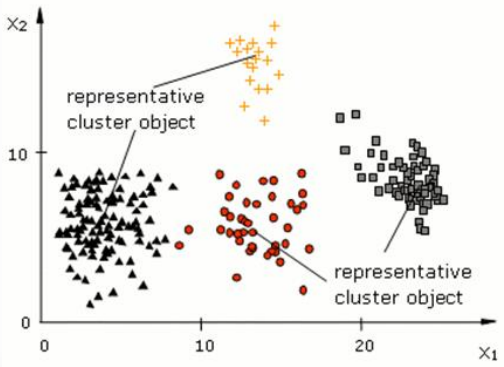
The Kernel Method in general

Further Examples of problems / tasks to which the kernel method can be applied:

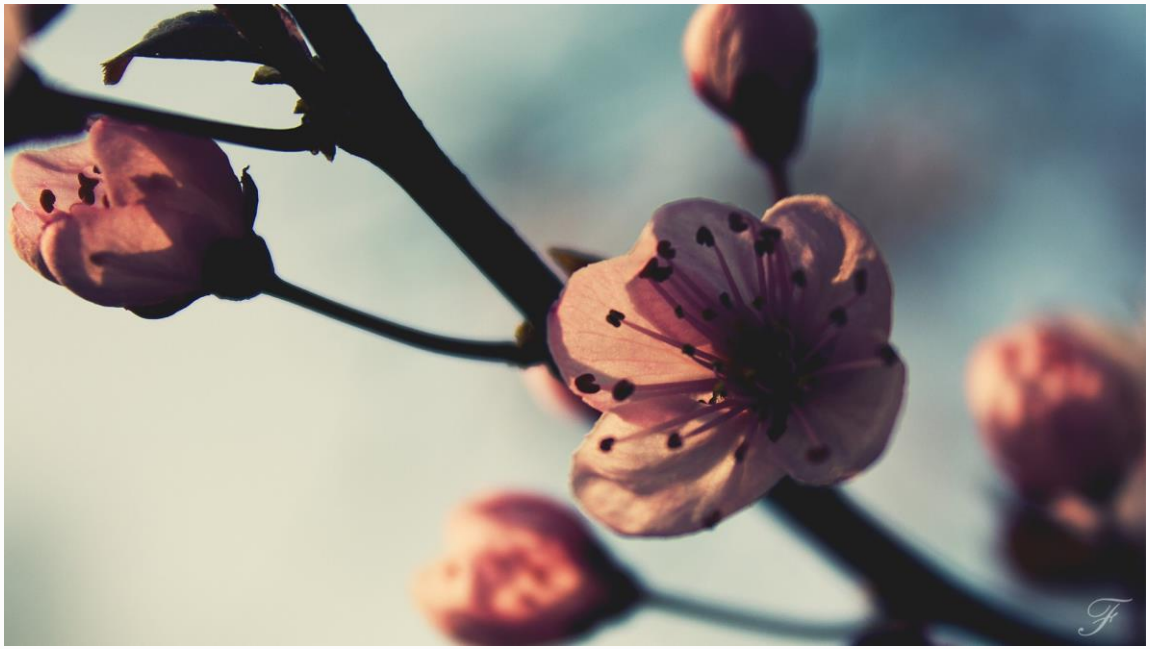
- Interpolation using functions, that are linear combinations of kernel functions,
- Regression using functions, that are linear combinations of kernel functions,
- Cluster analysis based on Euclidean distances.







Thank you for your attention.



Further Reading – scientific articles

- R. A. Fisher, *The Use of Multiple Measurements in Taxonomic Problems*, Annals of Eugenics **7** (1936), 179-188.
- S. Mika et. al., *A Mathematical Programming Approach to the Kernel Fisher Algorithm*, Advances in Neural Information Processing Systems **13** (2001).