Kernel Fisher discriminant functions – a concise and rigorous introduction

Hagen Knaf

September 2006

Abstract

In the article the application of kernel functions – the so-called »kernel trick« – in the context of Fisher's approach to linear discriminant analysis is described for data sets subdivided into two groups and having real attributes. The relevant facts about functional Hilbert spaces and kernel functions including their proofs are presented. The approximative algorithm published in [Mik3] to compute a discriminant function given the data and a kernel function is briefly reviewed. As an illustration of the technique an artificial data set is analysed using the algorithm just mentioned.

Keywords: discriminant analysis, functional Hilbert space, kernel function, reproducing kernel.

MSC: 46E22, 46N30, 62-07.

Introduction

A fundamental problem in data mining consists of separating two disjoint, finite subsets X_{-1} and X_{+1} of the euclidean space $\mathbb{R}^m$ through a hypersurface $S \subset \mathbb{R}^m$. More precisely and in analytic terms this amounts to choosing a function

$$d_{\theta^*} : \mathbb{R}^m \to \mathbb{R}$$

within a given parametrized family

$$F = \{d_\theta : \mathbb{R}^m \to \mathbb{R} \mid \theta \in \Theta\},\ \Theta \subseteq \mathbb{R}^p,$$

of functionals, such that

$$X_k \subseteq S_k,\ k \in \{-1, +1\}, \tag{1}$$

where $S_k := \{x \in \mathbb{R}^m \mid \operatorname{sign}(d_{\theta^*}(x)) = k\}$ are the half-spaces defined by $S := d_{\theta^*}^{-1}(0)$. Of course depending on F such a function d_{θ^*} needs not exist. One therefore replaces the requirement (1) by the weaker

$$\text{maximize } \operatorname{obj}(X_{-1} \cap S_{-1}, X_{+1} \cap S_{+1}), \tag{2}$$

where obj is an objective function like for example the »total hitrate«

$$\operatorname{obj}(X_{-1} \cap S_{-1}, X_{+1} \cap S_{+1}) := \frac{|X_{-1} \cap S_{-1}| + |X_{+1} \cap S_{+1})|}{|X_{-1}| + |X_{+1}|}.$$

In [Fis] Fisher showed that within the family

$$F = \{d_{(a,c)}(x) := \langle x, a \rangle + c \mid a \in \mathbb{R}^m, c \in \mathbb{R}\} \tag{3}$$

there exists a function $d_{(a^*,c^*)}$ that maximizes the objective function

$$\operatorname{obj}(a, c) := \frac{||p(\overline{x}_{-1}) - p(\overline{x}_{+1})||^2}{\overline{s}_{-1}^2 + \overline{s}_{+1}^2}, \tag{4}$$

where $p(\overline{x}_i)$ and $\overline{s}_i^2$ are the mean values and variances of the sets X_{-1} and X_{+1} after orthogonal projection to a line perpendicular to the hyperplane $d_{(a,c)}^{-1}(0)$. The optimal parameters (a^*, c^*) can be determined analytically and are unique if the set $X := X_{-1} \cup X_{+1}$ is in a »sufficiently general« position.

Only some years ago it was realized that the particular form of Fisher's objective function (4) allows to apply the so-called »kernel trick« to obtain a non-linear version of Fisher's discriminant function $d_{(a^*,c^*)}$: instead of (3) one considers families of the form

$$F = \{\sum_{y\in X} a_y K(x,y) + c \mid a_y, c \in \mathbb{R}\},$$

where $K : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ is a so-called kernel function. The particular properties of K allow to construct a Hilbert space H consisting of functions $h : X \to \mathbb{R}$ such that H is generated by the elements of $Z := \{K(\cdot, y) \mid y \in X\}$. Solving the optimization problem (2) using Fisher's objective function (4) in the space H and for the data set $Z = Z_{-1} \cup Z_{+1}$, where $Z_k := \{K(\cdot, y) \mid y \in X_k\}$, yields a in general nonlinear discriminant function

$$d^*(x) = \sum_{y\in X} a_y^* K(x,y) + c^*.$$

In the present article a concise and (hopefully) rigorous presentation of this whole process is provided, taking an analytic point of view.

1 Discriminant Analysis

In this section a brief review of the general problem of discriminant analysis is given, followed by a summary of relevant facts about Fisher's linear discriminant function.

1.1 The problem

Let Ω be a set that is decomposed into $g \geq 2$ pairwise disjoint subsets Ω_i, in the sequel called **groups**:

$$\Omega = \Omega_1 \cup \ldots \cup \Omega_g. \tag{5}$$

This decomposition yields the **label map**

$$\ell : \Omega \to \{1, \ldots, g\},\ (\omega \in \Omega_k) \mapsto k, \tag{6}$$

that assigns to an object $\omega \in \Omega$ its group label.

Every object $\omega \in \Omega$ comes equipped with a finite set of **attributes** $x(\omega)$. To keep the overview focussed on the facts relevant for the present article it is assumed that these attributes are real numbers, thus giving the **attribute map**

$$\Omega \to A \subseteq \mathbb{R}^m,\ \omega \mapsto x(\omega). \tag{7}$$

Based on the known attributes and group labels of a set of samples $\Lambda \subset \Omega$ randomly drawn from Ω in discriminant analysis one seeks to determine an estimate

$$\hat{\ell} : \Omega \to \{1, \ldots, g\} \tag{8}$$

of the unknown label function ℓ. Since by assumption the estimated label $\hat{\ell}(\omega)$ depends on $x(\omega)$ only, determining $\hat{\ell}$ amounts to determine a **decision function**

$$\delta : A \to \{1, \ldots, g\} \tag{9}$$

such that $\hat{\ell}(\omega) = \delta(x(\omega))$.

A typical example is the medical diagnosis of a specific disease based on a blood test say. Here the set Ω consists of a population of human beings, the attributes $x(\omega)$ are the values of the parameters measured during the blood test and the label $\ell(\omega)$ encodes whether the human ω is suffering of the considered disease ($\ell(\omega) = 1$ say) or not ($\ell(\omega) = 2$). The estimated label $\hat{\ell}(\omega)$ represents the diagnosis based on the blood test. It can be correct or false. Note that in this situation the map $\omega \mapsto x(\omega)$ needs not be injective, so that the diagnostic performance is limited already by the choice of the particular parameters measured during the blood test.

From now on only the case $g = 2$ is considered. Problems with $g > 2$ can be treated by successively splitting off one group after the other. Using the algorithm presented in Section 3 this can be done effectively.

For reasons that will become clear in the next paragraph it is assumed that the group labels for the two groups are -1 and $+1$. Moreover a decision function is allowed to take the value 0 with the meaning that for an object $\omega \in \Omega$ with the property $\delta(x(\omega)) = 0$ the actual group label $\ell(\omega)$ cannot be estimated using the decision function δ.

An intuitive geometric way to determine a (modified) decision function

$$\delta : A \to \{-1, 0, 1\} \tag{10}$$

is to fix a hypersurface $S \subset \mathbb{R}^m$ such that most of the samples in the set $X_{-1} := \{x(\omega) \mid \omega \in \Lambda \cap \Omega_{-1}\}$ »lie on one side of S«, while most of the samples in the set $X_{+1} := \{x(\omega) \mid \omega \in \Lambda \cap \Omega_{+1}\}$ »lie on the other side«. This approach can be made more precise in the following way: a function $d : \mathbb{R}^m \to \mathbb{R}$ yields a hypersurface through

$$S := d^{-1}(0)$$

provided it for example satisfies the prerequisites of the Implicit Function Theorem. The decision function associated to that hypersurface in the sense vaguely described above is then given by

$$\delta : A \to \{-1, 0, 1\},\ x \mapsto \operatorname{sign}(d(x)), \tag{11}$$

where sign denotes the function that maps a real number to its sign. The function d is called a **discriminant function for the grouping** $X_{-1} \cup X_{+1}$.

A strategy to choose a discriminant function is to consider a sufficiently general family

$$F := \{d_\theta : \mathbb{R}^m \to \mathbb{R} \mid \theta \in \Theta\},\ \Theta \subseteq \mathbb{R}^p, \tag{12}$$

of functionals possessing the necessary properties to define hypersurfaces via $S := d_\theta^{-1}(0)$ and to choose a parameter θ^* optimal with respect to some objective function obj derived from the known restriction $\ell|_\Lambda$ of the label map.

1.2 Fisher's linear discriminant function

Let Λ be a finite set of samples drawn from the population $\Omega = \Omega_{-1} \cup \Omega_{+1}$ grouped into two groups with labels -1 and $+1$. The attribute map then yields

$$x(\Lambda) =: X = X_{-1} \cup X_{+1} \subset \mathbb{R}^m,\ X_k := x(\Lambda \cap \Omega_k),\ k \in \{-1, +1\}. \tag{13}$$

From now on and throughout the whole article it is assumed that the attribute map is injective. Thus (13) is a subdivision of X into two disjoint groups.

Among the various ways to choose an **affine discriminant function**

$$d(x) = f(x) + c,\ f \in \operatorname{Hom}(\mathbb{R}^m, \mathbb{R}),\ c \in \mathbb{R}, \tag{14}$$

to separate the two groups in (13) R. A. Fisher's approach is known to be robust and has the advantage of leading to an analytic solution: for a one-dimensional sub-vector space $L \subseteq \mathbb{R}^m$ consider the orthogonal projection $p_L : \mathbb{R}^m \to L$ and define the **Fisher discriminant of** $X_{-1} \cup X_{+1}$ **with respect to** L as:

$$\mathcal{F}(L) := \frac{||p_L(\overline{x}_{-1}) - p_L(\overline{x}_{+1})||^2}{\overline{s}_{-1}^2 + \overline{s}_{+1}^2}, \tag{15}$$

where the

$$\overline{x}_k := \frac{1}{|X_k|} \sum_{x \in X_k} x$$

are the **group centroids** and

$$\overline{s}_k^2 := \sum_{x \in X_k} ||p_L(x) - p_L(\overline{x}_k)||^2, \tag{16}$$

up to a factor are the variances of the sets $p_L(X_k)$. Fisher proposes to consider the discriminant function

$$d^*(x) = p_{L^*}(x) + c^*, \tag{17}$$

where L^* is a solution of the optimization problem

$$\max(\mathcal{F}(L) \mid L \subseteq \mathbb{R}^m \text{ a one-dimensional sub-vector space}) \tag{18}$$

and

$$c^* := -\frac{1}{2}(p_{L^*}(\overline{x}_{-1}) + p_{L^*}(\overline{x}_{+1})). \tag{19}$$

Note that $-c^*$ is the mean value of the centroids of the projected groups $p_{L^*}(X_k)$.

In Section 3 a non-linear variant of Fisher's discriminant function is constructed using the so-called »kernel trick«. This is possible because a solution L^* of the optimization problem (18) can be obtained from quantities, that can be expressed solely in terms of scalar products between the elements $x \in X$. To verify this property one has to make the reasonable assumption

$$\sum_{x \in X} \mathbb{R}x = \mathbb{R}^m. \tag{20}$$

The orthogonal projection p_L to the one-dimensional sub-vector space $L \subseteq \mathbb{R}^m$ can be expressed as $p_L(x) = \langle a, x \rangle a$ with some $a \in \mathbb{R}^m$ having the property $||a|| = 1$. Using (20) one can write

$$a = \sum_{j=1}^{n} a_j x_j,\ a_j \in \mathbb{R},$$

where $\{x_1, \ldots, x_n\} = X$ is some numbering of the elements of X. Defining the vectors

$$\alpha := (a_1, \ldots, a_n) \in \mathbb{R}^n$$

and

$$m_k := (\langle x_1, \overline{x}_k \rangle, \ldots, \langle x_n, \overline{x}_k \rangle),\ k \in \{-1, +1\}, \tag{21}$$

one gets

$$\begin{aligned} ||p_L(\overline{x}_{-1}) - p_L(\overline{x}_{+1})||^2 &= \langle \sum_{j=1}^{n} a_j x_j, \overline{x}_{-1} - \overline{x}_{+1} \rangle^2 \\ &= (\sum_{j=1}^{n} a_j \langle x_j, \overline{x}_{-1} - \overline{x}_{+1} \rangle)^2 \\ &= (\alpha^t (m_{-1} - m_{+1}))^2 \end{aligned}$$

and thus the relation

$$||p_L(\overline{x}_{-1}) - p_L(\overline{x}_{+1})||^2 = \alpha^t (m_{-1} - m_{+1})(m_{-1} - m_{+1})^t \alpha \tag{22}$$

for the numerator of (15). The entries of the matrix

$$B := (m_{-1} - m_{+1})(m_{-1} - m_{+1})^t \in \mathbb{R}^{n \times n} \tag{23}$$

are sums of products of scalar products between elements of X. The denominator of (15) can be written in a similar way:

$$\begin{aligned} \overline{s}_{-1}^2 + \overline{s}_{+1}^2 &= \sum_{k \in \{-1,+1\}} \sum_{x \in X_k} \langle a, x \rangle^2 + \langle a, \overline{x}_k \rangle^2 - 2\langle a, x \rangle \langle a, \overline{x}_k \rangle \\ &= (\sum_{j=1}^{n} \langle a, x_j \rangle^2) - |X_{-1}| \langle a, \overline{x}_{-1} \rangle^2 - |X_{+1}| \langle a, \overline{x}_{+1} \rangle^2 \\ &= (\sum_{j=1}^{n} \langle a, x_j \rangle^2) - |X_{-1}| \alpha^t m_{-1} m_{-1}^t \alpha - |X_{+1}| \alpha^t m_{+1} m_{+1}^t \alpha. \end{aligned}$$

Using the matrix

$$K := (\langle x_i, x_j \rangle)_{i,j \in \{1,\ldots,n\}} \in \mathbb{R}^{n \times n} \tag{24}$$

the first summand on the right hand side can also be written as a matrix product:

$$\sum_{j=1}^{n} \langle a, x_j \rangle^2 = \alpha^t K K^t \alpha.$$

The optimization problem (18) can now be reformulated: the solutions L^* of (18) correspond to the solutions $\alpha^* = (a_1^*, \ldots, a_n^*)$ of the optimization problem

$$\max\left(\frac{\alpha^t B \alpha}{\alpha^t W \alpha} \mid \alpha \in \mathbb{R}^n \setminus 0\right) \tag{25}$$

through setting $p_{L^*}(x) := \langle a^*, x \rangle a^*$, $a^* := \sum_{j=1}^{n} a_j^* x_j$. Here the matrix W is defined as:

$$W := KK^t - |X_{-1}| m_{-1} m_{-1}^t - |X_{+1}| m_{+1} m_{+1}^t \in \mathbb{R}^{n \times n}. \tag{26}$$

It is well-known that (25) possesses the unique solution

$$\alpha^* := W^{-1}(m_{-1} - m_{+1}) \tag{27}$$

provided that W is invertible.

The constant term c^* can be expressed in terms of scalar products between the elements of X and the solution α^*:

$$c^* = -\frac{1}{2}(\alpha^*)^t (m_{-1} + m_{+1}). \tag{28}$$

2 Hilbert spaces from kernel functions

In this section an exposition of those parts of the theory of functional Hilbert spaces is given, that lead to the following statement sometimes called the »kernel trick«:

every real- or complex-valued function $K : X \times X \to \mathbb{K}$ *with the properties*

- $\forall x_1, x_2 \in X : \; K(x_1, x_2) = \overline{K(x_2, x_1)}$,
- $\forall n \in \mathbb{N}, \; \forall (x_1, \ldots, x_n) \in X^n : \; (K(x_i, x_j))_{i,j \in \{1,\ldots,n\}} \in \mathbb{K}^{n \times n}$ *is positive semi-definite,*

gives rise to a Hilbert space $H(K)$ *consisting of functions* $h : X \to \mathbb{K}$*, and to a map* $\phi : X \to H(K)$ *such that*

$$\forall x_1, x_2 \in X : \langle \phi(x_1), \phi(x_2) \rangle = K(x_2, x_1),$$

where $\langle \cdot, \cdot \rangle$ *denotes the scalar product of* $H(K)$*.*

2.1 Functional Hilbert spaces

For a set X the set $\mathrm{Map}\,(X, \mathbb{K})$ of all maps $h : X \to \mathbb{K}$ into a field $\mathbb{K}$ becomes a vector space over $\mathbb{K}$ if addition and scalar multiplication are defined pointwise. The maps $h_x \in \mathrm{Map}\,(X, \mathbb{K})$ defined by $h_x(x) = 1$ and $h_x(y) = 0$ for $x \neq y$ are linearly independent and form a basis of $\mathrm{Map}\,(X, \mathbb{K})$ if X is finite. Hence $\mathrm{Map}\,(X, \mathbb{K})$ has finite dimension if and only if X is finite.

Every $x \in X$ gives rise to the $\mathbb{K}$-linear **evaluation functional**

$$e_x : \mathrm{Map}\,(X, \mathbb{K}) \to \mathbb{K}, \; h \mapsto h(x). \tag{29}$$

These maps play a particular role in the sequel.

From now on let $\mathbb{K}$ be either the field of reals or the field of complex numbers.

A Hilbert space H over $\mathbb{K}$ is called **functional over** X if it is a sub-vector space of $\mathrm{Map}\,(X, \mathbb{K})$. If H has infinite dimension, the evaluation functionals (29) may not be continous. However from now on only functional Hilbert spaces with the property that *all* evaluation functionals are continous are considered. In particular norm convergence in H implies pointwise convergence:

$$(\lim_{k \to \infty} h_k = h) \;\; \Rightarrow \;\; (\forall x \in X : \; \lim_{k \to \infty} h_k(x) = h(x)). \tag{30}$$

The principal result of the structure theory of Hilbert spaces states that every Hilbert space H is isomorphic to a space $\ell^2(X)$ of square-summable maps

$X \to \mathbb{K}$. In particular it follows that every Hilbert space is functional over some set X, even with continous evaluation functionals.

Proposition 2.1 *For a Hilbert space H functional over X and possessing continous evaluation functionals there exists a unique function $K : X \times X \to \mathbb{K}$ with the properties:*

$$\begin{aligned} \forall x \in X : \qquad & K_x := K(\cdot, x) \in H, \\ & e_x = \langle \cdot, K_x \rangle. \end{aligned} \tag{31}$$

*The function K is called the **reproducing kernel of** H. It has the following properties:*

(1) $\forall x_1, x_2 \in X : K(x_1, x_2) = \overline{K(x_2, x_1)}$,

(2) $\forall n \in \mathbb{N},\ \forall (x_1, \ldots, x_n) \in X^n :\ K(x_1, \ldots, x_n) := (K(x_i, x_j))_{i,j \in \{1,\ldots,n\}}$ *is positive semi-definite, that is*

$$\forall z \in \mathbb{K}^n :\ z^t K(x_1, \ldots, x_n) \overline{z} \geq 0.$$

Proof. By the Representation Theorem of Riesz for every $x \in X$ there exists a unique element $K_x \in H$ such that $e_x = \langle \cdot, K_x \rangle$. Thus one can define the map K through

$$K(x, y) := K_y(x). \tag{32}$$

The uniqueness of K now also follows.

By definition K satisfies

$$K(x, y) = K_y(x) = \langle K_y, K_x \rangle, \tag{33}$$

which implies the symmetry of K. As for positive semi-definiteness this relation yields

$$\begin{aligned} z^t K(x_1, \ldots, x_n) \overline{z} &= \sum_{j=1}^{n} \sum_{k=1}^{n} z_j K(x_j, x_k) \overline{z_k} \\ &= \langle \sum_{k=1}^{n} \overline{z_k} K_{x_k}, \sum_{j=1}^{n} \overline{z_j} K_{x_j} \rangle \geq 0, \end{aligned}$$

where $z^t = (z_1, \ldots, z_n) \in \mathbb{K}^n$. □

As a consequence of Proposition 2.1 for a Hilbert space H functional over X and having continous evaluation functionals one can define a map

$$\phi : X \to H,\ x \mapsto K_x \tag{34}$$

that, having the applications in mind, unfortunately is not necessarily injective:

Lemma 2.2 *The map ϕ is injective if and only if for every pair of distinct points $x_1, x_2 \in X$ there exists an $h \in H$ such that $h(x_1) \neq h(x_2)$.*

This follows immediately from the equation $e_x = \langle \cdot, K_x \rangle$.

Utilizing the map ϕ one arrives at a key identity with respect to applications, namely

$$\forall x_1, x_2 \in X : \langle \phi(x_1), \phi(x_2) \rangle = K(x_2, x_1). \tag{35}$$

It follows directly from the definitions of K and ϕ.

Although the family $(K_x)_{x \in X}$ in general does not form a basis of H it generates H in the topological sense:

Proposition 2.3 *For a Hilbert space H functional over X and possessing continous evaluation functionals the subspace*

$$U := \sum_{x \in X} \mathbb{K} K_x$$

lies dense in H.

Proof. It suffices to prove that only the zero-function is perpendicular to every function K_x. Indeed $\langle h, K_x \rangle = 0$ for all $x \in X$ by (31) is equivalent to $h = 0$. □

As a consequence of Proposition 2.3 the inclusion $H \subseteq \mathrm{Map}\,(X, \mathbb{K})$ can be described using the functions $u \in U$ (and thus the functions K_x) only: every function $h \in H$ can be expressed as a limit $\lim\limits_{k \to \infty} u_k$, $u_k \in U$. The continuity of the evaluation functionals (30) thus yields

$$h(x) = \lim_{k \to \infty} u_k(x). \tag{36}$$

This fact can be used to reduce analytic to algebraic statements.

2.2 Kernel functions

Let X be a set. Motivated by Proposition 2.1 one calls a function

$$K : X \times X \to \mathbb{K}$$

with the properties 1 and 2 stated in that proposition a **kernel function (on X)**. Slightly abusing language property 1 is called **symmetry of** K while property 2 is named **positive semi-definiteness**. If for all $n \in \mathbb{N}$ and for pairwise distinct $x_1, \ldots, x_n \in X$ the matrices $K(x_1, \ldots, x_n)$ are positive definite, then K is called **positive definite**.

The set of kernel functions on X to the field $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ carries a rich algebraic and order structure that can be utilized effectively in applications. Again the focus in the present section is put onto the facts relevant for the applications in discriminant analysis.

Theorem 2.4 *The set $\mathcal{K}(X)$ of kernel functions on X is closed under pointwise addition, multiplication and scalar multiplication with non-negative scalars $c \in \mathbb{R}$, that is it forms a (linear) cone. In particular $\mathcal{K}(X)$ forms a commutative semi-ring taking pointwise addition and multiplication as compositions.*

Proof. Once the closedness of $\mathcal{K}(X)$ under the operations mentioned in the theorem is verified, the validity of all other ring axioms is obvious since the operations are all defined pointwise.

Addition: for a family $(x_1, \ldots, x_n) \in X^n$, $K_1, K_2 \in \mathcal{K}(X)$ and $K := K_1 + K_2$ one has

$$K_1(x_1, \ldots, x_n) + K_2(x_1, \ldots, x_n) = K(x_1, \ldots, x_n).$$

Hence for $z \in \mathbb{K}^n$ one gets

$$z^t K(x_1, \ldots, x_n)\overline{z} = z^t K_1(x_1, \ldots, x_n)\overline{z} + z^t K_2(x_1, \ldots, x_n)\overline{z} \geq 0.$$

Scalar multiplication: for a family $(x_1, \ldots, x_n) \in X^n$, $K \in \mathcal{K}(X)$ and positive $c \in \mathbb{R}$ one has

$$z^t(cK(x_1, \ldots, x_n))\overline{z} = c(z^t K(x_1, \ldots, x_n)\overline{z}) \geq 0.$$

Multiplication: the proof is based on considering the tensor product of bi-/sesquilinear forms. using the fact that if $\mathcal{B} := (b_1, \ldots, b_n)$ is a basis of V, then $(b_i \otimes b_j \mid i, j \in \{1, \ldots, n\})$ is a basis of $V \otimes V$, for bi-/sesquilinear forms α, β on some vector space V of finite dimension one can define the tensor product

$$\alpha \otimes \beta : (V \otimes V) \times (V \otimes V) \to \mathbb{K}$$

through setting

$$(\alpha \otimes \beta)((v_1 \otimes v_2), (w_1 \otimes w_2)) := \alpha(v_1, w_1) \cdot \beta(v_2, w_2). \tag{37}$$

$\alpha \otimes \beta$ is a bi-/sesquilinear form that is symmetric (and positive semi-definite) if α and β are symmetric (and positive semi-definite).

From now on assume that α and β are symmetric and positive semi-definite; then $(\alpha \otimes \beta)|_{U \times U}$ is symmetric and positive semi-definite for every sub-vector space $U \subseteq V \otimes V$. Take U to be the subspace spanned by the symmetric tensors $\mathcal{C} := (b_i \otimes b_i \mid i = 1, \ldots n)$; note that $\mathcal{C}$ is a basis of U.

Let $A = (a_{ij})_{i,j \in \{1,\ldots,n\}}$ and $B = (b_{ij})_{i,j \in \{1,\ldots,n\}}$ be the matrices of α and β with respect to the basis $\mathcal{B}$. It is then straightforward to check that the matrix $C \in \mathbb{K}^{n \times n}$ of $(\alpha \otimes \beta)|_{U \times U}$ with respect to the basis $\mathcal{C}$ of U looks like

$$C = \begin{pmatrix} a_{11}b_{11} & a_{12}b_{12} & \ldots & a_{1n}b_{1n} \\ a_{21}b_{21} & a_{22}b_{22} & \ldots & a_{2n}b_{2n} \\ \cdot & \cdot & \ldots & \cdot \\ \cdot & \cdot & \ldots & \cdot \\ a_{n1}b_{n1} & a_{n2}b_{n2} & \ldots & a_{nn}b_{nn} \end{pmatrix}. \tag{38}$$

Consequently the matrix C is symmetric and positive semi-definite. To denote the particular relationship between A, B and C one writes $C = A \odot B$. The matrix C is frequently called the **Hadamard or Schur product** of A and B. One has thus shown that the Hadamard product of symmetric, positive semi-definite matrices is symmetric and positive semi-definite.

To prove closedness under pointwise multiplication take $(x_1, \ldots, x_n) \in X^n$, $K_1, K_2 \in \mathcal{K}(X)$ and let $K := K_1 \cdot K_2$. By definition one then has

$$K(x_1, \ldots, x_n) = K_1(x_1, \ldots, x_n) \odot K_2(x_1, \ldots, x_n),$$

hence positive semi-definiteness of $K(x_1, \ldots, x_n)$. The symmetry of K is obvious. □

From the previous proof one can extract the following

Corollary 2.5 *The set* $\mathrm{SPSD}(n)$ *of symmetric, positive semi-definite* $n \times n$*-matrices becomes a commutative semi-ring with* 1 *if one takes the ordinary matrix addition as addition and the Hadamard product as multiplication.*

For a finite set $X = \{x_1, \ldots, x_n\}$ *with* n *elements the map*

$$\mathcal{K}(X) \to \mathrm{SPSD}(n),\ K \mapsto K(x_1, \ldots, x_n)$$

is an isomorphism of semi-rings.

In the context of Corollary 2.5 note that if $K(x_1, \ldots, x_n)$ is positive semi-definite, then for every permutation $(x_{s(1)}, \ldots, x_{s(n)})$ of the elements x_k and for every subfamily $(x_{k_1}, \ldots, x_{k_r})$ the matrices $K(x_{s(1)}, \ldots, x_{s(n)})$, $K(x_{k_1}, \ldots, x_{k_r})$ are positive semi-definite.

Corollary 2.6 *Let* K *be a kernel function on* X *and let* $p = \sum_{k=0}^{d} a_k X^k \in \mathbb{R}[X]$ *be a polynomial with positive coefficients, then the function*

$$p(K) = \sum_{k=0}^{d} a_k K^k$$

understood pointwise is a kernel function on X*.*

Note that $K^0 := 1$, the function $X \times X \to \mathbb{K}$ that maps everything to 1. Taking the standard scalar product $\langle \cdot, \cdot \rangle$ for K in Corollary 2.6 yields the widely used **polynomial kernel of degree** d:

$$K(x_1, x_2) := (\langle x_1, x_2 \rangle + c)^d,\ c > 0,\ d \in \mathbb{N}. \tag{39}$$

It is not difficult to replace the polynomial in Corollary 2.6 by a power series:

Proposition 2.7 *Let $(K_i)_{i\in\mathbb{N}}$ be a sequence of kernel functions on X such that the limit*

$$\lim_{i\to\infty} K_i(x_1, x_2)$$

exists for all $x_1, x_2 \in X$. Then the pointwise limit $K := \lim_{i\to\infty} K_i$ is a kernel function on X.

Proof. Symmetry of K is straightforward to check. For $(x_1, \ldots, x_n) \in X^n$ and $z \in \mathbb{K}^n$ one gets

$$z^t K(x_1, \ldots, x_n)\overline{z} = \lim_{i\to\infty} z^t K_i(x_1, \ldots, x_n)\overline{z} \geq 0,$$

since the entries in the matrix $K(x_1, \ldots, x_n)$ are the limits of the corresponding entries in the matrices $K_i(x_1, \ldots, x_n)$. □

Corollary 2.8 *Let $f : U \to \mathbb{R}$, $U \subseteq \mathbb{K}$, be a function defined through a (convergent) power series with positive coefficients:*

$$f(u) = \sum_{i=0}^{\infty} a_i u^i.$$

Let K be a kernel function on X such that $K(x_1, x_2) \in U$ for all $x_1, x_2 \in X$. Then the pointwise limit $f(K)$ defined through

$$f(K)(x_1, x_2) := \sum_{i=0}^{\infty} a_i K(x_1, x_2)^i$$

is a kernel function on X.

Proof. By Proposition 2.6 the functions

$$K_j := \sum_{i=0}^{j} a_i K^i$$

are kernel functions. By assumption $f(K) = \lim_{j\to\infty} K_j$, hence Proposition 2.7 yields the assertion. □

Corollary 2.8 yields the fact that Gauß' error function is a kernel function. The subsequent result is used in the proof but is also relevant in other contexts:

Proposition 2.9 *For every $K \in \mathcal{K}(X)$ and every non-zero function $f : X \to \mathbb{K}$ the map*

$$K' : X \times X \to \mathbb{K},\ (x_1, x_2) \mapsto f(x_1)K(x_1, x_2)\overline{f(x_2)}$$

is a kernel function on X.

Proof. Symmetry is immediate to check. For $(x_1, \ldots, x_n) \subseteq X^n$ and $z \in \mathbb{K}^m$ one gets

$$\begin{aligned} z^t K'(x_1, \ldots, x_n)\overline{z} &= \sum_{i=1}^{n}\sum_{j=1}^{n} z_i f(x_i) K(x_i, x_j)\overline{f(x_j)}\overline{z_j} \\ &= w^t K(x_1, \ldots, x_n)\overline{w} \geq 0, \end{aligned}$$

where $w = (z_1 f(x_1), \ldots, z_n f(x_n))$. □

Corollary 2.10 *For every $h \in \mathbb{R}$ the function*

$$K(x_1, x_2) := e^{-\frac{||x_1 - x_2||^2}{h^2}}, \tag{40}$$

*is a kernel function on $\mathbb{K}^m$. It is called the **Gauß kernel of bandwidth** h.*

Proof. One first has to rewrite K in an appropriate way – the naive approach of applying Corollary 2.8 directly does not work since the exponent $-\frac{||x_1 - x_2||^2}{h^2}$ is no kernel function:

$$e^{-\frac{||x_1 - x_2||^2}{h^2}} = [e^{-\frac{\langle x_1, x_1\rangle}{h^2}} e^{-\frac{\langle x_2, x_2\rangle}{h^2}}](e^{\frac{\langle x_1, x_2\rangle}{h^2}})^2. \tag{41}$$

The factor in squared brackets on the right side of (41) is a kernel function according to Proposition 2.9. Since the scalar product of $\mathbb{K}^m$ is a kernel function Corollary 2.8 yields that $e^{\frac{\langle x_1, x_2\rangle}{h^2}}$ is a kernel function. Consequently K itself is a kernel function due to Theorem 2.4. □

2.3 The Theorem of Aronszajn-Moore

A functional Hilbert space H having continous evaluation functionals gives rise to a kernel function K, namely the reproducing kernel of H. From a theoretical point of view it is natural to ask whether every kernel function K arises as the reproducing kernel of some Hilbert space $H(K)$. At the same time the positive answer to this question forms the basis for many applications of kernel functions:

Theorem 2.11 (N. Aronszajn, R. L. Moore) *For every kernel function $K : X \times X \to \mathbb{K}$ there exists a Hilbert space H functional over X and possessing continous evaluation functionals, such that K is the reproducing kernel of H. This Hilbert space $H = H(K)$ is uniquely determined by K.*

Proof. Let $K_y := K(\cdot, y) \in \mathrm{Map}\,(X, \mathbb{K})$ and consider the subvector space $U \subseteq \mathrm{Map}\,(X, \mathbb{K})$ generated by the family $(K_y \mid y \in X)$. On U one can define a

symmetric bi- respectively sesquilinear form through

$$\beta : \ U \times U \to \mathbb{K}, \ (\sum_{x \in X} a_x K_x, \sum_{y \in X} b_y K_y) \mapsto \sum_{x \in X} \sum_{y \in X} a_x K(y,x) \overline{b_y} \ . \tag{42}$$

Since the family $(K_y \mid y \in X)$ in general need not be linearly independent it remains to show that β is well-defined: to this end one has to verify that for every linear combination

$$u := \sum_{x \in X} a_x K_x = 0$$

and every function K_y, $y \in X$, the equation

$$\sum_{x \in X} a_x K(y,x) = 0$$

holds and similarly for switched roles of the two variables of β. Indeed

$$\sum_{x \in X} a_x K(y,x) = u(y) = 0$$

by assumption about u.

Claim: β is positive definite.

β is positive semi-definite since K is a kernel function. Hence the assertion follows from the Cauchy-Schwarz inequality $|\beta(u_1,u_2)|^2 \leq \beta(u_1,u_1)\beta(u_2,u_2)$ once one knows that $\beta(u_1,u_2) = 0$ for all $u_2 \in U$ implies $u_1 = 0$. Indeed $0 = \beta(u_1, K_y)$ for all $y \in X$ implies $u_1(y) = 0$ for all $y \in X$.

Let H be the completion of U with respect to the norm $||u||^2 := \beta(u,u)$. It is well-known that H is a Hilbert space taking the unique continous extension of β to H as the scalar product.

Claim: H is a subvector space of $\mathrm{Map}\,(X, \mathbb{K})$.

Express $h \in H$ as a limit $h = \lim\limits_{i \to \infty} u_i$, $u_i \in U$. For every $x \in X$ the Cauchy-Schwarz inequality then yields

$$|u_i(x) - u_j(x)|^2 = |\beta(u_i - u_j, K_x)|^2 \leq \beta(u_i - u_j, u_i - u_j) K(x,x), \tag{43}$$

hence $(u_i(x))_{i \in \mathbb{N}}$ is a Cauchy sequence in $\mathbb{K}$. If $h = \lim\limits_{i \to \infty} v_i$, $v_i \in U$, is a second representation of h, then $(u_i - v_i)_{i \in \mathbb{N}}$ is a zero-sequence. Consequently by (43) applied to $u_i - v_i$ the sequence $(u_i(x) - v_i(x))_{i \in \mathbb{N}}$ is a zero-sequence too. Therefore one can interpret h as a function $X \to \mathbb{K}$ through setting

$$h(x) := \lim_{i \to \infty} u_i(x).$$

This interpretation leads to a linear map $H \to \mathrm{Map}\,(X, \mathbb{K})$.

Claim: K is the reproducing kernel of H.

Indeed by definition of β for every $x \in X$ one has

$$\begin{aligned} e_x(h) = h(x) &= \lim_{i\to\infty} u_i(x) \\ &= \lim_{i\to\infty} \beta(u_i, K_x) \\ &= \beta(\lim_{i\to\infty} u_i, K_x) \\ &= \beta(h, K_x). \end{aligned}$$

Note that this also shows that e_x is continous.

Finally assume that H_i, $i = 1, 2$, are functional Hilbert spaces both possessing $K : X \times X \to \mathbb{K}$ as their reproducing kernel. Let U_i be the subvector space of H_i spanned by the functions $K_x = K(\cdot, x)$. Note that for $h \in U_i$ one has $h(x) = \sum_{j=1}^{r} a_{x_j} K(x, x_j)$ so that $U_1 = U_2$ as subvector spaces of $\mathrm{Map}\,(X, \mathbb{K})$ holds. Similarly for the restrictions $(||\cdot||_i)|_{U_i}$ of the norms of the H_i one deduces

$$||h||_1^2 = \sum_{j=1}^{r}\sum_{k=1}^{r} a_{x_j} K(x_k, x_j)\overline{a_{x_k}} = ||h||_2^2.$$

Consequently using the remark (36) following Proposition 2.3 one concludes $H_1 = H_2$. □

2.4 Dimension

For a thorough analysis of the results one gets by using Kernel Fisher discriminant functions it is necessary to have some idea about the dimension of the Hilbert space $H(K)$ associated to a kernel function $K : X \times X \to \mathbb{K}$. Since $H(K)$ is functional over X one gets

$$\dim H(K) \leq |X| \tag{44}$$

if X is a finite set. Moreover as a related general result one should mention:

Proposition 2.12 *For a kernel function $K \in \mathcal{K}(X)$ the functions $(K_x \mid x \in X)$ are $\mathbb{K}$-linearly independent if and only if K is positive definite.*

Proof. The linear relation

$$\sum_{i=1}^{n} a_i K_{x_i} = 0$$

is equivalent with

$$\begin{aligned} 0 &= \langle \sum_{i=1}^{n} a_i K_{x_i}, \sum_{j=1}^{n} a_j K_{x_j} \rangle \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} a_i \overline{a_j} K(x_j, x_i). \end{aligned}$$

□

Corollary 2.13 *For a positive definite kernel function $K \in \mathcal{K}(X)$ the Hilbert space $H(K)$ has finite dimension if and only if X is finite; then the family $(K_x \mid x \in X)$ forms a basis of $H(K)$.*

The proof of Proposition 2.12 yields a criterion for the injectivity of the map $\phi : X \to H(K)$ – a fundamental property in the context of discriminant functions.

Corollary 2.14 *If the kernel function $K \in \mathcal{K}(X)$ has the property*

$$\forall x_1, x_2 \in X,\ x_1 \neq x_2 : \quad K(x_1, x_2) \text{ is positive definite},$$

then $H(K)$ separates the points of X and thus ϕ is injective.

Proof. By the argument used in the proof of Proposition 2.12 the functions K_{x_1}, K_{x_2} for $x_1 \neq x_2$ are linearly independent hence $K_{x_1} \neq K_{x_2}$. □

For a polynomial kernel the functions $(K_x \mid x \in X)$ can be linearly dependent as the case of degree one shows. Moreover the dimension of $H(K)$ depends on X and on the degree of the kernel. For »large« X one gets:

Proposition 2.15 *For a polynomial kernel $K(x_1, x_2) := (\langle x_1, x_2 \rangle + c)^d$ the following properties are valid:*

(1) if X contains a non-empty open set of $\mathbb{K}^m$, then the map $\phi :\ X \to H(K),\ x \mapsto K_x$ is injective,

(2) $\dim H(K) \leq \binom{m+d}{d}$ with equality for $X = \mathbb{K}^m$.

Proof. 1.: Assume that $K_{x_1} = K_{x_2}$ on X for some $x_1, x_2 \in X$. Then

$$\left(\frac{\langle x, x_1 \rangle + c}{\langle x, x_2 \rangle + c} \right)^d = 1,$$

for all $x \in X$ at which the rational function on the left-hand side is defined; denote that set by U. Since X contains a non-empty open set and $\{x \in \mathbb{K}^m \mid \langle x, x_2 \rangle + c = 0\}$ is a hyperplane in $\mathbb{K}^m$, the set U contains a

non-empty open set of $\mathbb{K}^m$. On U the rational function

$$f(x) := \frac{\langle x, x_1\rangle + c}{\langle x, x_2\rangle + c}$$

attains only finitely many values, namely a subset of the d-th roots of unity. Consequently f is constant, that is $\langle x, x_1\rangle + c = a(\langle x, x_2\rangle + c)$, which implies $a = 1$ and thus $x_1 = x_2$ as asserted.

2.: The Hilbert space $H(K)$ is a subvector space of the vector space $P(X, d)$ of polynomial functions on X in m variables and of degree d. The restriction yields a surjective linear map $P(d) \to P(X, d)$, where $P(d)$ denotes the vector space of polynomial functions on $\mathbb{K}^m$ in m variables and of degree d. Therefore the well-known formula $\dim P(d) = \binom{m+d}{d}$ completes the proof. □

Gauß kernels behave much nicer than polynomial kernels:

Lemma 2.16 *For pairwise distinct numbers $x_1, \ldots, x_n \in \mathbb{C}$ the functions*

$$e^{-||x-x_k||^2},\ k = 1, \ldots, n,$$

are linearly independent over $\mathbb{C}$.

Proof. Take a linear relation

$$\sum_{k=1}^{n} a_k e^{-||x-x_k||^2} = 0,\ a_k \in \mathbb{C},$$

between functions as stated in the lemma and rewrite it in the form

$$0 = \sum_{k=1}^{n} (a_k e^{-||x_k||^2}) e^{-||x||^2} e^{2\mathrm{Re}\,\langle x, x_k\rangle} = \sum_{k=1}^{n} b_k e^{-||x||^2} e^{2\mathrm{Re}\,\langle x, x_k\rangle},$$

where $b_k := a_k e^{-||x_k||^2}$; then:

$$\sum_{k=1}^{n} b_k e^{2\mathrm{Re}\,\langle x, x_k\rangle} = 0$$

and since this equation holds for all $x \in \mathbb{R}^m$ one gets the homogenous system

$$\sum_{k=1}^{n} b_k e^{2\mathrm{Re}\,\langle jx, x_k\rangle} = 0,\ j = 0, \ldots, n-1, \tag{45}$$

for fixed $x \in \mathbb{R}^m$. Next choose $x \in \mathbb{R}^m$ such that the values $\operatorname{Re}\langle x, x_k\rangle$, $k = 1, \ldots, n$, are pairwise distinct. Then the coefficient matrix

$$A := \begin{pmatrix} 1 & 1 & \ldots & 1 \\ e^{2\operatorname{Re}\langle x, x_1\rangle} & e^{2\operatorname{Re}\langle x, x_2\rangle} & \ldots & e^{2\operatorname{Re}\langle x, x_n\rangle} \\ e^{2\operatorname{Re}\langle 2x, x_1\rangle} & e^{2\operatorname{Re}\langle 2x, x_2 \operatorname{Re}\rangle} & \ldots & e^{2\operatorname{Re}\langle 2x, x_n\rangle} \\ \cdot & \cdot & \ldots & \cdot \\ \cdot & \cdot & \ldots & \cdot \\ e^{2\operatorname{Re}\langle (n-1)x, x_1\rangle} & e^{2\operatorname{Re}\langle (n-1)x, x_2\rangle} & \ldots & e^{2\operatorname{Re}\langle (n-1)x, x_n\rangle} \end{pmatrix}$$

of (45) is a Vandermonde matrix. By the choice of the values $\operatorname{Re}\langle x, x_k\rangle$ and the injectivity of the real exponential function A is invertible. Consequently the system (45) only has the trivial solution $b_1 = \ldots = b_n = 0$, which implies $a_1 = \ldots = a_n = 0$ thus completing the proof. □

Lemma 2.16 combined with Proposition 2.12 yields:

Proposition 2.17 *The Gauß kernel $K : X \times X \to \mathbb{K}$ on some set $X \subseteq \mathbb{K}^m$ is positive definite. Consequently the Hilbert space $H(K)$ has finite dimension if and only if X is finite; then the equation*

$$\dim H(K) = |X|$$

holds.

3 Kernel Fisher discriminant functions

Building on the theoretical basis layed in Section 2 in the present section explicit formulae for kernel Fisher discriminant functions are derived. Next an algorithm to estimate the parameters in these formulae using classified data is described. Finally an example is presented.

3.1 Explicit formulae

Let $X = X_{-1} \cup X_{+1} \subset \mathbb{R}^m$ be a finite set (of attributes) divided into two disjoint groups. Fix a kernel function

$$K : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$$

and consider the functional Hilbert space $H(K|_{X\times X})$ along with the associated map

$$X \xrightarrow{\phi} H(K|_{X\times X})$$

defined in (34). Throughout the whole section it is assumed that ϕ is injective; consequently $\phi(X_{-1}) \cap \phi(X_{+1}) = \emptyset$.

The Hilbert space $H(K|_{X\times X})$ has finite dimension (44). Consequently Fisher's approach as formulated in Subsection 1.2 can be used to separate the sets $\phi(X_{-1})$ and $\phi(X_{+1})$ using a hyperplane: note first that due to Proposition 2.3 the family $(\phi(x) \mid x \in X)$ generates the space $H(K|_{X\times X})$. Choosing a numbering $X = \{x_1, \ldots, x_n\}$ a Fisher discriminant function

$$d : H(K|_{X\times X}) \to \mathbb{R}, h \mapsto \langle a^*, h\rangle + c^*, \tag{46}$$

can thus be defined by taking a solution $\alpha^* = (a_1^*, \ldots, a_n^*) \in \mathbb{R}^n$ of the optimization problem (25) replacing the elements x_i by the elements $\phi(x_i)$. In that way one arrives at

$$a^* = \sum_{j=1}^{n} a_j^* \phi(x_j)$$

and

$$c^* = -\frac{1}{2}(\alpha^*)^t(m_{-1} + m_{+1}),$$

where

$$m_k = \frac{1}{|X_k|} \sum_{x\in X_k} (\langle\phi(x_1), \phi(x)\rangle, \ldots, \langle\phi(x_n), \phi(x)\rangle) \in \mathbb{R}^n$$

for $k \in \{-1, +1\}$ – see (19) and (21). The matrices B and W in the current setting of the optimization problem (25) are defined through

$$B = (m_{-1} - m_{+1})(m_{-1} - m_{+1})^t$$

and

$$W := KK^t - |X_{-1}|m_{-1}m_{-1}^t - |X_{+1}|m_{+1}m_{+1}^t$$

where

$$K := (\langle \phi x_i, \phi x_j \rangle)_{i,j \in \{1,\ldots,n\}} \in \mathbb{R}^{n \times n}, \tag{47}$$

see (23), (26) and (24). At that point the essence of the kernel method becomes clearly visible: the solutions α^* of the optimization problem (25) for the grouping $\phi(X) = \phi(X_{-1}) \cup \phi(X_{+1})$ are determined by quantities – namely the matrices B and W – that themselves depend on scalar products of the form $\langle \phi(x), \phi(x') \rangle$, $x, x' \in X$ only. Hence due to the key identity $\langle \phi(x), \phi(x') \rangle = K(x', x)$ (35) these solutions can be computed working in the original sample space $\mathbb{R}^m$ instead of $H(K|_{X \times X})$. Note that the matrix K defined in (47) is indeed »equal« (in an obvious sense) to the kernel function $K : X \times X \to \mathbb{R}$.

The discriminant function to separate the groups X_{-1} and X_{+1} is given by $d \circ \phi : X \to \mathbb{R}$, which can be expressed explicitely using α^*, the kernel function K and the data X: for $x \in X$ we get

$$\begin{aligned} (d \circ \phi)(x) &= \langle \sum_{j=1}^{n} a_j^* \phi(x_j), \phi(x) \rangle + c^* \\ &= \sum_{j=1}^{n} a_j^* K(x, x_j) + c^*. \end{aligned} \tag{48}$$

New samples $x \in \mathbb{R}^m$ are classified by plugging in x into (48) and using the decision function $\operatorname{sign}(d(\phi(x)))$. Formally this is possible because the kernel function K is defined on the whole of $\mathbb{R}^m$. Conceptually this procedure can be justified in the usual way: if the data set X sufficiently faithfully represents the distributions underlying the groups Ω_{-1} and Ω_{+1}, then a discriminant function derived from these data is supposed to generalize well.

3.2 Computation

The computation of the kernel Fisher discriminant function (48) from the given data and a kernel function requires to handle the in general huge kernel matrix K (47) as well as the derived matrices B and W. One therefore has to carefully choose among the existing methods for solving the optimization problem (25) in the present context. In this subsection an algorithm for the approximative solution of (25) proposed in [Mik3] is described without going into the details or giving proofs.

Reformulation of the optimization problem: given a grouping $X = X_{-1} \cup X_{+1}$ and a kernel function $K \in \mathcal{K}(\mathbb{R}^m)$ in the article [Mik2] the problem of finding a discriminant function on $\mathbb{R}^m$ of the form

$$d(x) = \sum_{j=1}^{n} a_j K(x, x_j) + c \tag{49}$$

is considered as a regression problem for the label function $\ell : X \to \{-1, +1\}$. Let $X = \{x_1, \ldots, x_n\}$ be some numbering of X and define

$$\xi(\alpha, c) := (d(x_j) - \ell(x_j))_{j=1,\ldots,n}$$

to be the vector of deviations of the discriminant values $d(x_j)$ from the labels; here $\alpha = (a_1, \ldots, a_n)$. Moreover let

$$y := (\ell(x_j))_{j=1,\ldots,n}$$

be the vector of labels, let K be the kernel matrix (47) and define

$$\mathbf{1} := (1, \ldots, 1) \in \mathbb{R}^n,\ \mathbf{1}_{+1} = \frac{1}{2}(y + \mathbf{1}),\ \mathbf{1}_{-1} = \mathbf{1} - \mathbf{1}_{+1}.$$

Finally let $P : \mathbb{R}^n \to \mathbb{R}$ be a function that penalizes solutions α with »many« non-zero coordinates with a penalty weight $R \geq 0$. Then according to [Mik2]:

Proposition 3.1 *For every penalty function P and constant $R > 0$ the optimization problems*

$$\begin{aligned} \min(||\xi(\alpha, c)||^2 &+ RP(\alpha) \mid \alpha \in \mathbb{R}^n, c \in \mathbb{R}) && (50)\\ K\alpha + c\mathbf{1} &= y + \xi && (51)\\ \langle \mathbf{1}_i, \xi \rangle &= 0,\ i \in \{-1, +1\} && (52) \end{aligned}$$

and

$$\begin{aligned} \min(\alpha^t W \alpha + RP(\alpha) &\mid \alpha \in \mathbb{R}^n) && (53)\\ \langle \alpha, (m_{+1} - m_{-1}) \rangle &= 2 && (54) \end{aligned}$$

have the same solutions. For the second optimization problem the constant c is computed using the formula (28).

Note that the second optimization problem is just a reformulation of (25): the fraction

$$\frac{\alpha^t B \alpha}{\alpha^t W \alpha}$$

does not change its value when replacing α by a multiple $\lambda\alpha$, $\lambda \neq 0$. This substitution corresponds to multiplying the associated discriminant function by the factor λ. We can thus assume (54) and consequently that $\alpha^t B \alpha$ has a fixed

(positive) value, which in turn yields that it suffices to minimize the denominator $\alpha^t W \alpha$. In that way one arrives at the second optimization problem appearing in Proposition 3.1. Taking into account that the matrix W can and frequently will be singular, the penalty term $RP(\alpha)$ can also be understood as a form of regularization.

A greedy algorithm ... to solve the optimization problem (50) and thus (53) approximatively. It is proposed in the article [Mik3] and is based on the idea of building up the discriminant function (48) stepwise by adding one suitable term $K(x, x_i)$ at each step and stopping the process once the performance of the discriminant function is sufficient. In this way a sparse approximate solution of (50) can be efficiently computed. In principal, choosing the number of non-zero coordinates of α as the penalty function $P(\alpha)$, sparse solutions could also be obtained avoiding the stepwise approach but their computation is not efficient. In the greedy algorithm presented here the penalty term $RP(\alpha)$ should better be considered as a regularization.

To give an explicit description of the crucial step of adding a new term to the discriminant function assume that an approximative solution of (50) involving only $r \geq 1$ terms has already been determined and let $I = \{i_1, \ldots, i_r\}$ be the indices of the samples $x_i \in X$ such that $K(x, x_i)$ appears in this solution. Indeed only the terms $K(x, x_i)$, $i \in I$, are of interest here. The associated coefficients a_i will be recomputed at each step.

Take $i_{r+1} \in \{1, \ldots, n\} \setminus I$ and set $I' := I \cup \{i_{r+1}\}$. The aim now is to efficiently compute an approximative solution of (50) involving the $r + 1$ terms $K(x, x_i)$, $i \in I'$. To this end one rewrites (50), (51) and (52) in matrix form. Setting

$$\beta := (c, a_{i_1}, \ldots, a_{i_{r+1}}) \in \mathbb{R}^{r+2}$$

and

$$K_{r+1} := \begin{pmatrix} K(x_1, x_{i_1}) & K(x_1, x_{i_2}) & \ldots & K(x_1, x_{i_{r+1}}) \\ K(x_2, x_{i_1}) & K(x_2, x_{i_2}) & \ldots & K(x_2, x_{i_{r+1}}) \\ \cdot & \cdot & \ldots & \cdot \\ \cdot & \cdot & \ldots & \cdot \\ K(x_n, x_{i_1}) & K(x_n, x_{i_2}) & \ldots & K(x_n, x_{i_{r+1}}) \end{pmatrix} \in \mathbb{R}^{n \times (r+1)}$$

the matrix form of the considered optimization problem is

$$\min(\beta^t H \beta - \frac{1}{2} \gamma^t \beta + n \quad | \quad \beta \in \mathbb{R}^{r+2}) \tag{55}$$

$$A_{-1}^t \beta + |X_{-1}| = 0 \tag{56}$$

$$A_{+1}^t \beta - |X_{+1}| = 0, \tag{57}$$

where

$$\begin{aligned}\gamma &:= (|X_{+1}| - |X_{-1}|, K_{r+1}^t y) \in \mathbb{R}^{r+2},\\ A_k &:= (|X_k|, K_{r+1}^t \mathbf{1}_k) \in \mathbb{R}^{r+2},\ k \in \{-1, +1\},\\ H &:= \begin{pmatrix} n & \mathbf{1}^t K_{r+1} \\ K_{r+1}^t \mathbf{1} & K_{r+1}^t K_{r+1} + R \end{pmatrix} \in \mathbb{R}^{(r+2)\times(r+2)}.\end{aligned}$$

The optimization problem (55), (56) and (57) can be solved using Lagrange multipliers thus obtaining:

$$\beta^* = H^{-1}(\gamma - \lambda_{+1}^* A_{+1} - \lambda_{-1}^* A_{-1}), \tag{58}$$

where the multipliers $(\lambda_{-1}^*, \lambda_{+1}^*)$ form a solution of the optimization problem

$$\max(-\frac{1}{2}\lambda^t Q \lambda + L\lambda - \frac{1}{2}\gamma^t H^{-1}\gamma + \frac{n}{2} \mid \lambda = (\lambda_{+1}, \lambda_{-1}) \in \mathbb{R}), \tag{59}$$

where

$$Q := \begin{pmatrix} A_{+1}^t H^{-1} A_{+1} & A_{+1}^t H^{-1} A_{-1} \\ A_{-1}^t H^{-1} A_{+1} & A_{-1}^t H^{-1} A_{-1} \end{pmatrix} \in \mathbb{R}^{2\times 2},$$

and

$$L := \big((-|X_{+1}| + \gamma^t H^{-1} A_{+1}) \ \ (|X_{-1}| + \gamma^t H^{-1} A_{-1})\big) \in \mathbb{R}^2.$$

Note that the optimization problem (59) can be solved analytically by taking the first derivative of the quadratic function involved.

Note further that due to the particular form of the matrices H one has an equation of the form

$$H_{r+1} = \begin{pmatrix} H_r & v \\ v^t & s \end{pmatrix},\ v \in \mathbb{R}^{r+1},\ s \in \mathbb{R},$$

relating the matrix H for r terms in the discriminant function with the matrix for $r+1$ terms. Consequently the inverse H^{-1} for $r+1$ terms can be computed efficiently from the inverse for r terms using the well-known Sherman-Woodbury-formula.

The algorithm can now be formulated as follows:

(1) Start with the empty set $I = \emptyset$ of indices $i \in \{1, \ldots, n\}$ such that $K(x, x_i)$ appears in the discriminant function.
(2) Compute the solutions $\beta^*(i')$ for all index sets $I \cup \{i'\}$, $i' \in \{1, \ldots, n\} \setminus I$, using the formula (58).
(3) Select an index $i_0 \in \{1, \ldots, n\} \setminus I$ such that the objective function (55) at the solution $\beta^*(i_0)$ is minimal among all of the objective function values at the solutions $\beta^*(i)$. Set $I := I \cup \{i_0\}$.

(4) Stop if $|I|$ exceeds a predefined maximal number of terms in the discriminant function or if the change of the value of the objective function (55) compared to the previous addition of a new term lies below a predefined bound. Otherwise go to step 2.

3.3 An example

In the sequel the results of an application of the greedy algorithm described in Subsection 3.2 to an artificial two-dimensional data set are presented with the aim to give a visual impression of Kernel Fisher discriminant functions in a concrete case (and not of the performance of the algorithm used).

The data set $X = X_{-1} \cup X_{+1}$ shown in Figure 1 consists of 1640 samples $x \in \mathbb{R}^2$, each of the groups X_{-1} and X_{+1} having 820 elements. It has been created by randomly distributing points about two semi-circles using a normal distribution with standard deviation $\sigma = 0.8$. The »center points« of the semi-circles lie at $(-0.5, 0)$ and $(0, 0.5)$, both having a »radius« equal to 1.

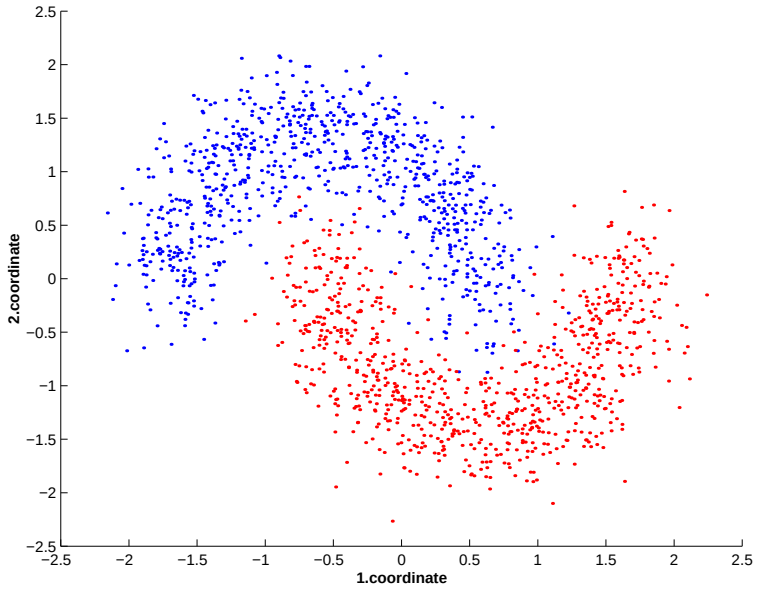


Figure 1 data set

First a Gauß kernel $K(x_1, x_2) = e^{-\frac{||x_1 - x_2||^2}{h^2}}$ (40) having bandwidth $h = 0.858$ was used. The value of h has been estimated from the data using 5-fold cross validation – knowing the way the data set has been created a reasonable result.

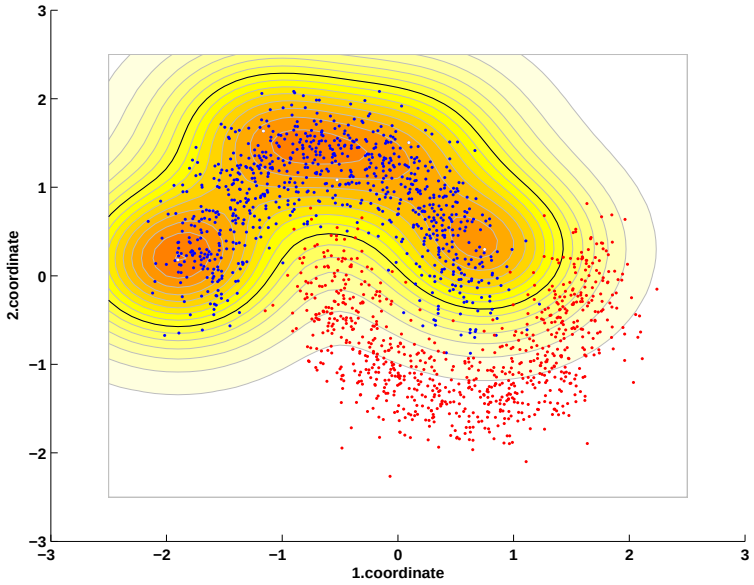
The algorithm gives the approximate discriminant function

$$\begin{aligned} d_{\text{Gauß}}(x) := \quad & -0.7591 \cdot K(x, x_1) - 2.8579 \cdot K(x, x_2) - 2.4329 \cdot K(x, x_3) \\ & -2.0455 \cdot K(x, x_4) - 1.6439 \cdot K(x, x_5) + 1.3246, \end{aligned} \tag{60}$$

with

$$\begin{aligned} x_1 &= (-0.5256, 1.0841), \quad & x_2 &= (-1.9011, 0.1797), \\ x_3 &= (\ \ 0.7500, 0.2981), \quad & x_4 &= (-1.1620, 1.6343), \\ x_5 &= (\ \ 0.0948, 1.5043). \end{aligned}$$

Figure 2 shows various level sets $d^{-1}_{\text{Gauß}}(l)$, $l \in \mathbb{R}$, of the function (60); the level curve $d^{-1}_{\text{Gauß}}(0)$, that is the separating curve, is drawn in black colour. In addition the level of the function values $d_{\text{Gauß}}(x)$ is depicted through a yellow-to-orange colour scale symbolising decreasing values of $d_{\text{Gauß}}(x)$. The locations of the points $x_1, \ldots, x_5$ are marked as white points.



level curves of $d_{\text{Gauß}}(x)$ Figure 2

Using the classification rule $\ell(x) = \text{sgn}(d_{\text{Gauß}}(x))$ (11) yields misclassification rates of 3.9% in the group X_{-1} and 1.9% in the group X_{+1}.

In a second run a polynomial kernel $K(x_1, x_2) = (\langle x_1, x_2 \rangle + c)^3$ (39) of degree 3 with constant term $c = 0.75$ has been used, the value of c again being estimated using 5-fold crossvalidation. The resulting estimated discriminant function is:

$$\begin{aligned} d_{\text{poly}}(x) := \quad & -7.1287 \cdot K(x, x_1) + 0.3231 \cdot K(x, x_2) + 0.3612 \cdot K(x, x_3) \\ & -0.5870 \cdot K(x, x_4) + 0.9951 \cdot K(x, x_5) + 2.4521, \end{aligned} \tag{61}$$

with

$$\begin{aligned} x_1 &= (-0.0150, 0.1517), \quad & x_2 &= (0.0049, 0.7557), \\ x_3 &= (\ \ 1.8009, 0.2819), \quad & x_4 &= (1.4297, 0.2090), \\ x_5 &= (-0.2253, 0.0755). \end{aligned}$$

Figure 3 depicts the level sets of d_{poly}, the separating curve and the positions of the x_k using the same symbolism as in the previous example.

The classification rule $\ell(x) = \operatorname{sgn}(d_{\text{poly}}(x))$ this time yields a misclassification rate of 3.9% in the group X_{-1} and 5.0% in the group X_{+1}.

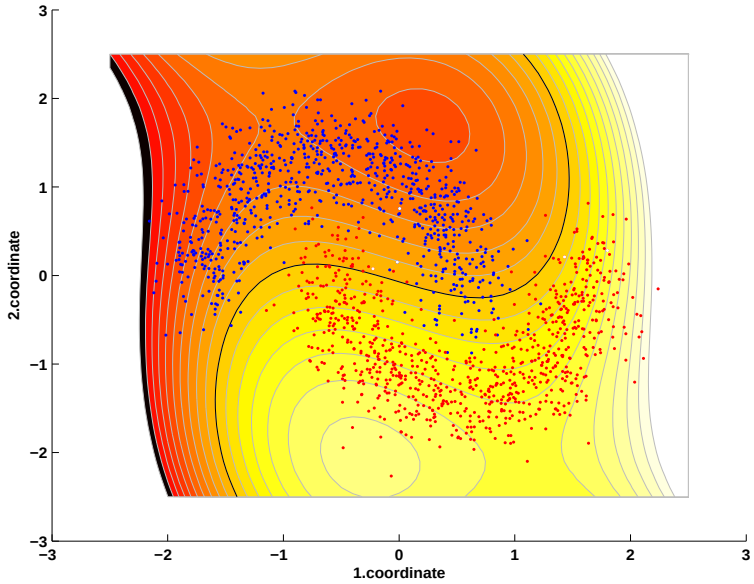
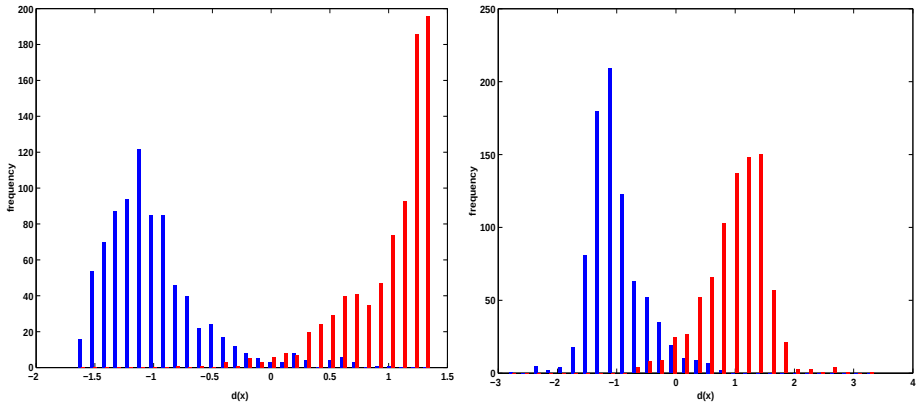


Figure 3 level curves of $d_{\text{poly}}(x)$

Assuming that the functions $d_{\text{Gauß}}$ and d_{poly} are good approximations to the respective optimal discriminant function (46), their function values $d_{\text{Gauß}}(x)$ and $d_{\text{poly}}(x)$, $x \in X$, up to a constant factor and a translation represent the orthogonal projections of the points $\phi(x)$ to the line $\mathbb{R}\alpha \subseteq H(K|_{X\times X})$, where α is the vector of coefficients of $d_{\text{Gauß}}$ respectively d_{poly} amended by zeros.

The left plot in Figure 4 shows a histogram of the distribution of $d_{\text{Gauß}}(X_{-1})$ plotted in blue and of $d_{\text{Gauß}}(X_{+1})$ plotted in red. Bars in different colour directly beside each other refer to one and the same bin of the histogram. The right plot in Figure 4 shows the distribution of $d_{\text{poly}}(X_{-1})$ and $d_{\text{poly}}(X_{+1})$ using the same colour code.

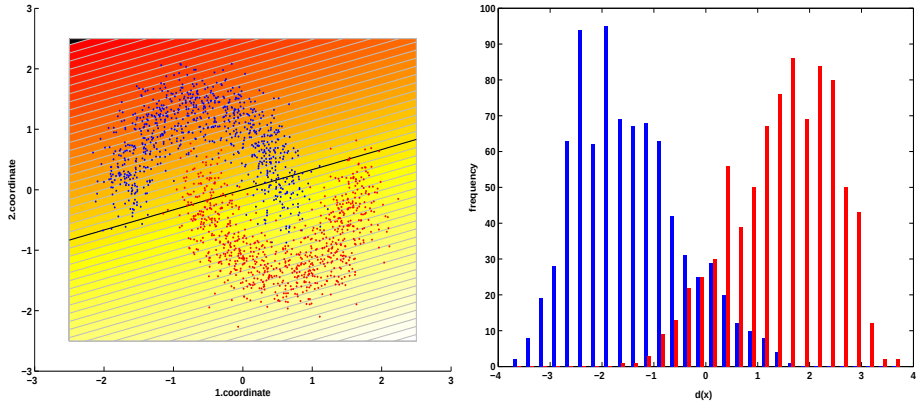


distribution of $d_{\text{Gauß}}(X)$(left) and $d_{\text{poly}}(X)$(right) Figure 4

Figure (5) shows the results of the classical Fisher approach. The discriminant function here is:

$$d_{\text{F}}(x) := \langle (0.4951, -1.4871), x \rangle \tag{62}$$

with a misclassification rate of 10.5% in group X_{-1} and 8.9% in group X_{+1}. Note that one can determine Fisher's linear discriminant function approximately by applying Mika's algorithm utilizing the kernel function $K(x_1, x_2) := \langle x_1, x_2 \rangle$. However here the usual approach has been choosen.



level curves of $d_{\text{F}}(x)$(left), distribution of $d_{\text{F}}(X)$ (right) Figure 5

References

[Aiz] M. Aizerman, E. Braverman, L. Rozonoer *Theoretical foundations of the potential function method in pattern recognition learning*, Automation and Remote Control **25** (1964), 821-837.

[Aro] N. Aronszajn, *Theory of reproducing kernels*, Trans. Amer. Math. Soc. **68** (1950), 337-404.

[Fis] R. A. Fisher, *The Use of Multiple Measurements in Taxonomic Problems*, Annals of Eugenics **7** (1936), 179-188.

[Mik1] S. Mika et. al., *Fisher Discriminant Analysis With Kernels*, Neural Networks for Signal Processing **IX** (1999), 41-48.

[Mik2] S. Mika et. al., *A Mathematical Programming Approach to the Kernel Fisher Algorithm*, Advances in Neural Information Processing Systems **13** (2001).

[Mik3] S. Mika et. al., *An Improved Training Algorithm for Kernel Fisher Discriminants*, Proceedings AISTATS 2001.

[Moo] E. H. Moore, *On properly positive Hermitian matrices*, Bull. Amer. Math. Soc. **23(59)** (1916), 66-67.

[ST-N] J. Shawe-Taylor, N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, 2004.