Reproducing Kernel Hilbert Spaces

Steffen Dähne
Wolfgang Karl Härdle
Dedy Dwi Prastyo

Ladislaus von Bortkiewicz Chair of Statistics
C.A.S.E. – Center for Applied Statistics
and Economics
Humboldt–Universität zu Berlin
http://lvb.wiwi.hu-berlin.de
http://www.case.hu-berlin.de

RKHS – what are they?

- They are ideal space for smooth objects, i.e. Hilbert spaces limited to smooth functions
- Main field of application – Support Vector Machines (SVM)
- This method of linear discrimination can be generalised to smooth nonlinear contexts
- The generalisation involves kernel and thus, Reproducing Kernel Hilbert Spaces (Aronszajn, 1950)

Recap SVM

- SVM provide a modern and powerful statistical tool for classification
- Training data $(x_1, y_1), \ldots, (x_m, y_m) \in \mathcal{X} \times \{\pm 1\}$
- From training data construct a classifier function $f : \mathcal{X} \to \{\pm 1\}$
- Focus on efficiency and speed
- Example: classify from accounting data in bankrupt and non-bankrupt companies

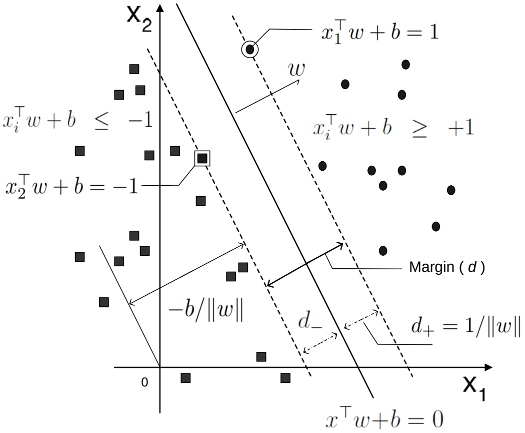


Figure 1: Linearly separable case

Let $x_1 \in \{x | \langle w, x \rangle + b = 1\}$ and $x_2 \in \{x | \langle w, x \rangle + b = -1\}$. Then,

$$\begin{aligned} \langle w, (x_1 - x_2) \rangle &= 2 \\ \langle \frac{w}{\|w\|}, (x_1 - x_2) \rangle &= \frac{2}{\|w\|} \end{aligned}$$

From projecting two points from the two classes being closest to each other on the separating hyperplane's normal vector $\frac{w}{\|w\|}$ it is clear that the margin equals $\frac{2}{\|w\|}$

Margin maximisation

Maximise the margin in order to separate the points from both classes with the highest "safest" distance (margin) between them. Maximising the margin is equivalent to minimising the norm of w.

$$\begin{aligned} \min_{w \in \mathcal{X}, b \in \mathbb{R}} \quad & \frac{1}{2}\|w\|^2 \\ s.t. \quad & y_i(\langle w, x_i \rangle + b) \geq 1, \forall i = 1, \ldots, m \end{aligned}$$

Corresponding Lagrangian:

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{m} \alpha_i \left\{ y_i(\langle w, x_i \rangle + b) - 1 \right\}$$

Dual problem

Due to Wolfe (1961) this optimisation program is equivalent to the dual one (see ▸ appendix for details):

$$\begin{aligned} \max_{\alpha \in \mathbb{R}^m} \quad & \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \\ s.t. \quad & \alpha_i \geq 0, \forall i = 1, \ldots, m \\ & \sum_{i=1}^{m} \alpha_i y_i = 0 \end{aligned} \tag{1}$$

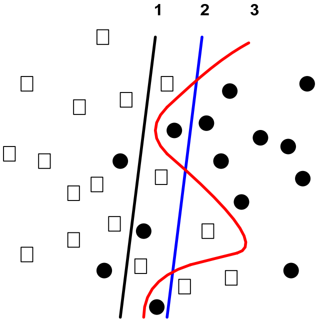


Figure 2: Classifier

Nonlinear classifier

Real data is not linearly separable!
Idea – *Elevate* data point via a **feature map** Ψ into RKHS $\mathcal{H}$ and then solve a linear classification problem in $\mathcal{H}$

$$\begin{array}{rcl} \Psi : \mathcal{X} & \to & \mathcal{H} \\ x & \mapsto & \overset{\text{def}}{=} \Psi(x) \end{array}$$

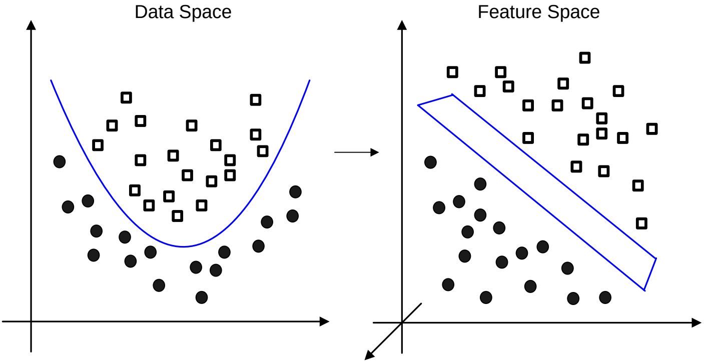


Figure 3: Feature map $\Psi(x) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$

The reproducing property

- We need to calculate the inner product in the feature space, see equation (1). This is computationally intensive due to the high dimensionality of $\mathcal{H}$
- Cover's Theorem (Cover, 1965) – number of linear separations increases with the dimensionality

Here, the reproducing property helps (the so called kernel trick)

$$x^\top x = \langle \Psi(x), \Psi(x') \rangle = k(x, x')$$

Outline

1. Motivation ✓
2. Definition
3. Construction
4. RKHS – definitions and properties
5. Illustrations
6. References
7. Appendix

The merit of the Reproducing Property

Thanks to the reproducing property ($\langle f, k(x,.)\rangle_{\mathcal{H}} = f(x)$), one can evaluate inner products in the feature space without explicitely calculating the mapping $\Psi(x_i)$.
Hilbert spaces with this reproducing property are called RKHS.

Hilbert space

1. Vector space (i.e. vector space axioms are valid)
2. Normed space (i.e. a norm exist)
3. Complete (i.e. every Cauchy sequence's limit is inside)
4. The norm is defined by an inner product (i.e. $\|v\|^2 = \langle v, v \rangle, \forall v$)

Inner product

1. Mapping $\mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is linear in each component (bilinear)
2. Symmetric: $\forall x, y \in \mathcal{H}, \langle x, y \rangle = \langle y, x \rangle$
3. Strictly positive: $\forall x \in \mathcal{H} \backslash \left\{ 0 \right\}, \langle x, x \rangle$ greater than zero

What is a (positive definite) kernel?

- A kernel k is positive definite *iff* for $x_1, \ldots, x_m \in \mathcal{X}$ the Gram matrix K defined by $K_{ij} \overset{\text{def}}{=} k(x_i, x_j)$ is positive definite
- Kernel are symmetric, i.e. $k(x_i, x_j) = k(x_j, x_i)$
- Kernel help constructing generalised inner products
- An RKHS possesses the Reproducing Property wrt one **unique** kernel

Map to a feature space

Map observations $\{x_i\}_{i=1}^{m}$ to a feature space such that the kernel is an inner product in that space. More precisely

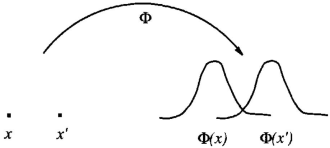
$$\Psi : x \mapsto k(x, .)$$

A particular form of the kernel could be:

$$k(x, .) = \exp\left\{ -\frac{1}{2\sigma_2} \|x - .\|^2 \right\}$$

(Gaussian Radial Basis Function)

Map to a feature space



$k(x,.)$ can be interpreted as a *similarity* measure of x to all other elements of $\mathcal{X}$.
Here, $\Psi : \mathcal{X} \mapsto \mathbb{R}^{\mathcal{X}}$, where $\mathbb{R}^{\mathcal{X}} \overset{\text{def}}{=} \{f : \mathcal{X} \mapsto \mathbb{R}\}$ is (possibly infinite dimensional) function space

Other kernels

Anisotropic Gaussian Kernel

$$k(x,.) = \exp\left\{\frac{1}{2}\left(x_i - .\right)^\top r^{-2}\Sigma^{-1}\left(x_i - .\right)\right\}$$

Polynomial Kernel

$$k(x,.) = \langle x,.\rangle^d$$

Hyperbolic Tangent Kernel

$$k(x,.) = \tanh\left(\nu\langle x,.\rangle +\right)$$

Element of the feature space

The elements of the feature space are defined as:

$$f(.) = \sum_{i=1}^{m} \alpha_i k(., x_i)$$

where $m \in \mathbb{N}, \alpha_i \in \mathbb{R}$ and arbitrary $x_1, \ldots, x_m \in \mathcal{X}$.

Inner product on the feature space

For two elements f and g

$$f(.) = \sum_{i=1}^{m} \alpha_i k(., x_i), \quad g(.) = \sum_{j=1}^{m'} \beta_j k(., x_j')$$

define

$$\langle f, g \rangle \overset{\text{def}}{=} \sum_{i=1}^{m} \sum_{j=1}^{m'} \alpha_i \beta_j k(x_i, x_j')$$

Recall that we aim at:

$$\langle k(x, .), f \rangle = f(x)$$
$$\langle \Psi(x_i), \Psi(x_j) \rangle = k(x_i, x_j)$$

Is it an inner product?

- ⊡ Symmetric: $\langle f, g\rangle = \langle g, f\rangle$
- ⊡ Biliner: $\langle f, g\rangle = \sum_{j=1}^{m'} \beta_j f(x_j') = \sum_{i=1}^{m'} \alpha_i g(x_i)$
- ⊡ Positive definite: $\langle f, f\rangle = \sum_{i,j=1}^{m} \alpha_i \alpha_j k(x_i, x_j) \geq 0$
- ⊡ $\langle f, f\rangle = 0 \Rightarrow f = 0$

Consequently, the completion of the feature space is now a Hilbert space!

Is this space an RKHS?

Choose m finite for simplicity

$$\begin{aligned}\langle k(.,x), f\rangle &= \langle k(.,x), \sum_{i=1}^{m} \alpha_i k(.,x_i)\rangle \\ &= \sum_{i=1}^{m} \alpha_i k(x,x_i) \\ &= f(x)\end{aligned}$$

By definition of $\langle .,.\rangle$ (with $m' = 1$ and $\beta_1 = 1$) and $f(.)$

The "kernel trick"

Let

$$f(.) = k(x, .)$$

Then

$$\begin{aligned} \langle k(x,.), k(x',.) \rangle &= \langle f, k(x',.) \rangle \\ &= f(x') \\ &= k(x, x') \\ \overset{\Psi: x \mapsto k(x,.)}{\Longleftrightarrow} \langle \Psi(x), \Psi(x') \rangle &= k(x, x') \end{aligned}$$

i.e. the **inner product** in the feature space is equivalent to the value of the **kernel** in the input space

A different point of view: Riesz's theorem

1. $\mathcal{H}$ Hilbert space of functions $f : \mathcal{X} \mapsto \mathbb{R}$ and corresponding inner product $\langle .,. \rangle$
2. ξ_x linear, continuous functional

$$\begin{aligned} \xi_x : \mathcal{H} &\mapsto \mathbb{R} \\ f &\mapsto f(x), \forall x \in \mathcal{X} \\ \xi_x(f) &= f(x) \end{aligned}$$

Then $\exists ! y \in \mathcal{H}, \quad \forall f \in \mathcal{H}, \quad \xi_x(f) = \langle y, f \rangle$.
Hence $f(x) = \langle k(x,.), f \rangle$ and $y = k(x,.) \in \mathcal{H}$.

Definition RKHS

Let $\mathcal{X}$ be a nonempty set and $\mathcal{H}$ a Hilbert space of functions $f : \mathcal{X} \mapsto \mathbb{R}$. Then $\mathcal{H}$ is called an RKHS endowed with the inner product $\langle ., . \rangle$ if there is exist a function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ with the following properties:

1. k has the reproducing property $\langle f, k(x, .) \rangle = f(x), \quad \forall f \in \mathcal{H}$
2. k spans $\mathcal{H}$, i.e. $\mathcal{H} = \overline{span\,\{k(x, .) | x \in \mathcal{X}\}}$

Intuition RKHS

- Benefit from the richness of Hilbert spaces (inner product, norm, projections) . . .
- . . . and restrict them to ensure the Reproducing property
- Mapping to the feature space *flattens out non-linearities* due to the Reproducing property

Mercer’s theorem

- Provide conditions under which a RKHS associated with a certain kernel exists
- RKHS exist for every kernel that is continuous, symmetric and positive definite
- Represenation of RKHS in a standardised basis

Mercer's theorem

Suppose $k \in L_\infty(\mathcal{X}^2)$ is a real-valued function such that

$$\begin{array}{rcl} T_k : L_2(\mathcal{X}) & \mapsto & L_2(\mathcal{X}) \\ f(.) & \mapsto & (T_k f)(.) \\ (T_k f)(x) & \overset{\text{def}}{=} & \int_{\mathcal{X}} k(x,x')f(x')d\mu(x') \end{array}$$

is positive definite. Let $\phi_j \in L_2(\mathcal{X})$ be normalised orthogonal eigenfunctions of T_k associated with eigenvalues $\lambda_j > 0$. Then

1. $\lambda_j \in \ell_1$
2. Eigenfunction expansion $k(x,x') = \sum_{j=1}^{N_{\mathcal{H}}} \lambda_j \phi_j(x)\phi_j(x')$ holds for almost all (x,x') ($N_{\mathcal{H}}$ is the possibly infnite number of dimension of $\mathcal{H}$)

Implied *Mercer* map

- Somewhat standardised feature map implied by Mercer's theorem:

$$\begin{array}{rcl}\Psi : \mathcal{X} & \mapsto & \ell_2^{N_{\mathcal{H}}} \\ x & \mapsto & \left\{\sqrt{\lambda_j}\phi_j(x)\right\}_{j=1,\ldots,N_{\mathcal{H}}}\end{array}$$

- Elements of $\mathcal{H}$:

$$f(x) = \sum_{i=1}^{\infty} \alpha_i k(x, x_i) = \sum_{i=1}^{\infty} \alpha_i \sum_{j=1}^{N_{\mathcal{H}}} \lambda_j \phi_j(x) \phi_j(x_i)$$

- Bi-linearity implies:

$$\langle f, k(.,x') \rangle_{\mathcal{H}} = \sum_{i=1}^{\infty} \alpha_i \sum_{j=1}^{N_{\mathcal{H}}} \sum_{m=1}^{N_{\mathcal{H}}} \lambda_j \phi_j(x_i) \langle \phi_j, \phi_m \rangle \lambda_m \phi_m(x')$$

- Now, choose $\langle ., . \rangle$ such that

$$\langle \phi_j, \phi_m \rangle = \frac{\delta_{jm}}{\lambda_j}$$

- Then, obtain the Reproducing property

$$\langle f, k(.,x') \rangle = \sum_{i=1}^{\infty} \alpha_i \sum_{j=1}^{N_{\mathcal{H}}} \lambda_j \phi_j(x) \phi_j(x') \overset{\text{per def}}{=} f(x')$$

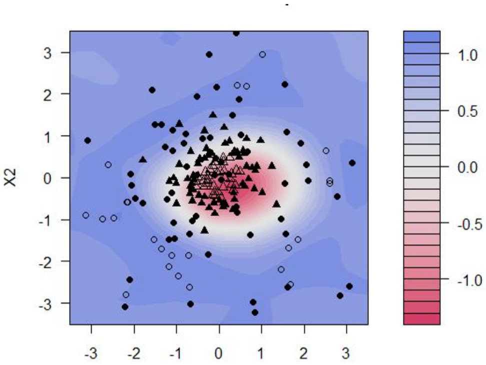


Figure 4: SVM for orange peel data, $n = 200$, $d = 2$, $n_{-1} = n_{+1} = 100$, $x_{+1,i} \sim \mathrm{N}((0,0)^\top, 2^2\mathcal{I})$, $x_{-1,i} \sim \mathrm{N}((0,0)^\top, 0.5^2\mathcal{I})$ with SVM parameters $r = 0.5$ and $C = 20/200$. The solid circle and triangle are observations that are used as support vector. MVAsvmOrangePeel.R

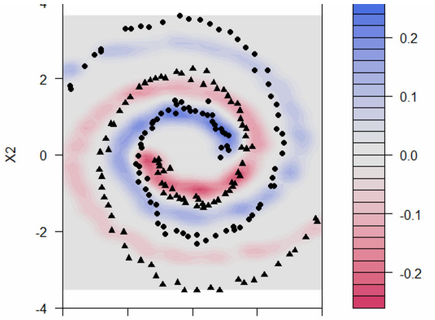
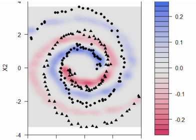


Figure 5: SVM for noisy spiral data. The spirals spread over 3π radian; the distance between the spirals equals 1.0. $d = 2$, $n_{-1} = n_{+1} = 100$, $n = 200$. The noise was injected with the parameters $\varepsilon_i \sim \mathrm{N}(0, 0.1^2\mathcal{I})$. The separation is perfect with SVM parameters $r = 0.1$ and $C = 10/200$. The solid circle and triangle are observations that are used as support vector. MVAsvmSpiral.R

Reproducing Kernel Hilbert Spaces

Steffen Dähne
Wolfgang Karl Härdle
Dedy Dwi Prastyo

Ladislaus von Bortkiewicz Chair of Statistics
C.A.S.E. – Center for Applied Statistics
and Economics
Humboldt–Universität zu Berlin
http://lvb.wiwi.hu-berlin.de
http://www.case.hu-berlin.de



References

Aronszajin, N.
Theory of reproducing kernels
Transaction of the American Mathematical Society, 1950, 68(3): 337-404

Cover, T.M.
Geometrical and statistical properties of system of linear inequalities with applications in pattern recognition
IEEE Transactions on Electronic Computers, 1965, 14: 326-334

Wolfe, P.
A duality theorem for nonlinear programming
Quarterly of Applied Mathematics, 1961, 19: 239-244

References

- Vapnik, V.
  The Nature of Statistical Learning Theory
  Springer Verlag, New York, 1995

Wolfe duality

▸ Dual Problem

Primal problem

$$\begin{aligned}\min_{w\in\mathcal{X},b\in\mathbb{R}} \quad & \frac{1}{2}\|w\|^2 \\ s.t. \quad & y_i(\langle w, x_i\rangle + b) \geq 1, \quad \forall i = 1, \ldots, m \\ L(w, b, \alpha) \quad = \quad & \frac{1}{2}\|w\|^2 - \sum_{i=1}^{m} \alpha_i \left\{ y_i(\langle w, x_i\rangle + b) - 1 \right\}\end{aligned}$$

Injecting the FOC into the Lagrangian yields the dual problem

$$\begin{aligned}
\frac{\partial}{\partial b} L(w, b, \alpha) = 0 \quad &\Longleftrightarrow \quad \sum_{i=1}^{m} \alpha_i y_i = 0 \\
\frac{\partial}{\partial w} L(w, b, \alpha) = 0 \quad &\Longleftrightarrow \quad w = \sum_{i=1}^{m} \alpha_i y_i x_i \\
G(\alpha) \quad &\overset{\text{def}}{=} \quad \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle
\end{aligned}$$

Wolfe (1961):

$$\max_{w \in \mathcal{X}, b \in \mathbb{R}} L \Longleftrightarrow \max_{\alpha \in \mathbb{R}^m} G$$

Obtain optimal w and b

Then obtain w by:

$$w = \sum_{i=1}^{m} \alpha_i y_i x_i$$

Choose b such that the optimality condition

$$y_i(\langle w, x_i \rangle + b) \geq 1$$

is verified for all $i = 1, \ldots, m$ (more complex algorithm needed)