Kernel combination in SVMs for classification purposes: Geometry and Information

Javier González Hernández
Alberto Muñoz García
Universidad Carlos III de Madrid
Department of Statistics

Leaning problem

LEARNING

Automatic cloud identification

Space X
Satellite data

LEARNING

Space Y
Type of cloud

To find f: X ⟶ Y

SVM History

- Mercer theorem: Mercer, 1909.
- Geometrical interpretation of kernels: Aizerman et al., 1964.
- Hyperplane in an non parametric context: Vapnik and Chervonenkis, 1964.
- SVM origin: Boser, Guyon y Vapnik, 1992.
- SVMs as regularization problem: Wahba, 1999.
- SVM review and open problems: Moguerza and Muñoz, 2006.

Ill-posed problems

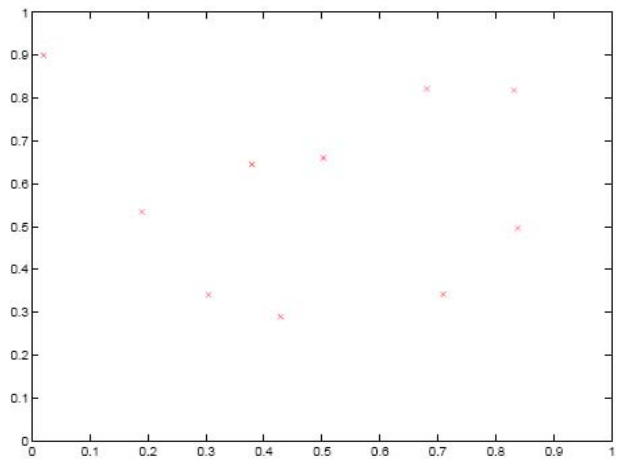
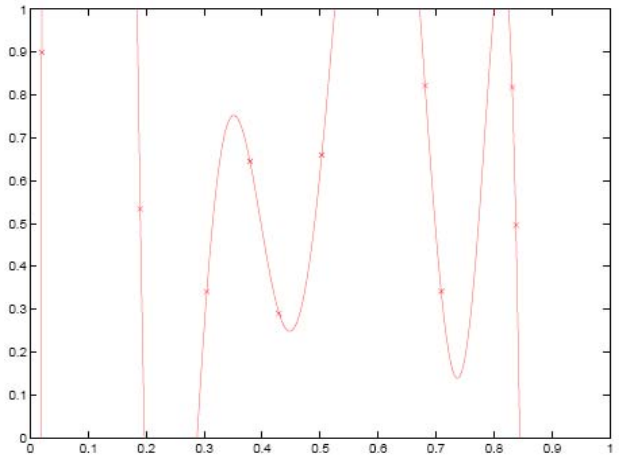
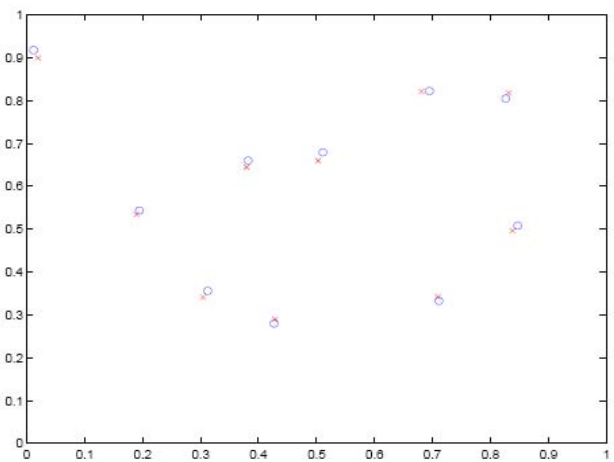
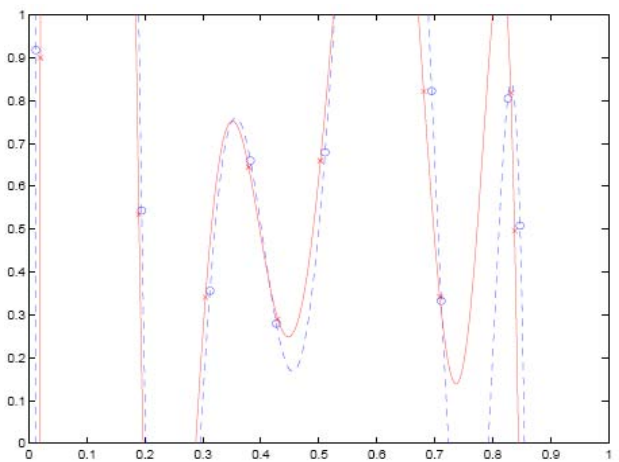
Well-posed problems (Hadamard)

- A solution exists.
- The solution is unique.
- The solution depends continuously on the data.

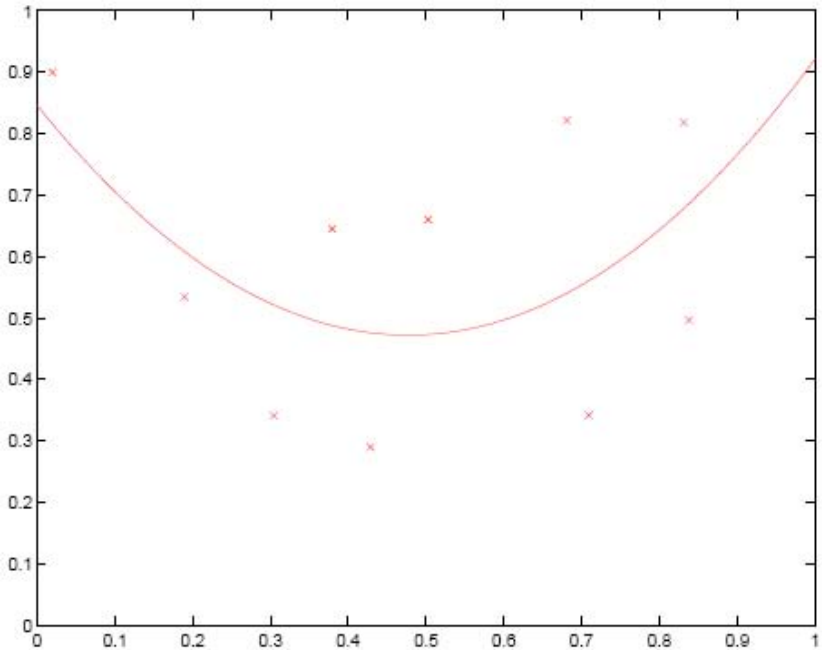
Examples of ill-posed problems

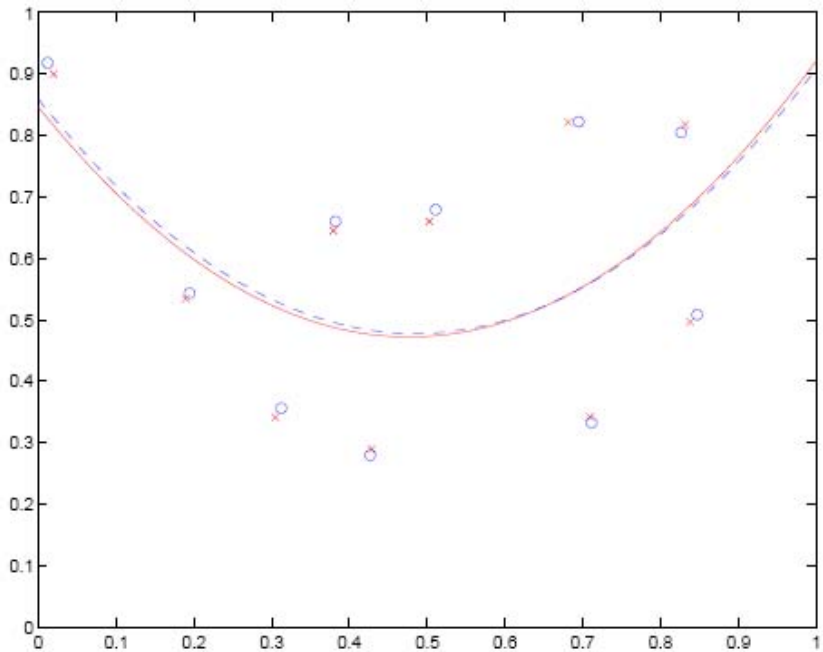
- Density estimation.
- *Classification* problems.
- Regression problems.

Example of ill-posed problem

Example of well-posed problem





Elements of the problem and notation

Elements of the problem

- There exists $f : X \rightarrow Y$.
- A probability measure p over $X \times Y$. $E[y|\mathbf{x}] = f(\mathbf{x})$.
- X a compact domain or manifold in an Euclidean space.
- $L(f(\mathbf{x}), y)$ a generic loss function.

Objetive

To find the best approximation to $f : X \rightarrow Y$ given a sample $M = \{(\mathbf{x}_i, y_\mathbf{i}) \in X \times Y\}_{i=1}^{n}$.

Hypothesis space

Structure of the hypothesis space

Let $C(X)$ a Banach space of the continuos functions over X with the norm

$$\|f\|_\infty = \sup_{\mathbf{x}\in X} |f(\mathbf{x})|.$$

¿Where to search for f?

In a compact subspace $\mathcal{H}$ de $C(X) \Rightarrow$ hypothesis space.

Generalization Error

Theoretical criteria for searching f

Minimize the risk functional $R(f) : C(X) \rightarrow \mathbb{R}$ (generalization error)

$$R(f) = \int_{X \times Y} L(f(\mathbf{x}) - y) p(\mathbf{x}, y) d\mathbf{x} dy.$$

Existence of f

- The existance of f^* is guaranteed due to the compactness of $\mathcal{H}$ an the continuity of $R(f)$.
- If $\mathcal{H}$ is convex f^* is unique $\Rightarrow$ well-posed problem.

Empirical Error

Practical criteria for searching f

To minimize the functional

$$R_{emp}(f) = \frac{1}{n}\sum_{i=1}^{n} L(f(\mathbf{x}_i), y_i),$$

Error for f on the sample M.

¿Makes sense?

Yes,

- If $\mathcal{H}$ is compact $\Rightarrow$ the problem is well posed.
- The convergence of the empirical error to the generalization error (for the SVM loss function) is guaranteed.

Compactness of the hypothesis space

Compactness imposition

Through the Tikhonov regularization. To minimize on H the functional risk

$$F(f) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(\mathbf{x}_i)) + \mu\Omega(f)\,,$$

- $\mu > 0$.
- H is an appropriate space of functions.
- $\Omega(f)$ is a functional convex and positive.

Mercer kernels

Definition

Let $K : X \times X \to \mathbb{R}$ a continuous and symmetric function. Let assume that K is positive definite, that is, given a set $\{x_1, ..., x_n\} \subset X$ the matrix $K[x]$ with components $K(x_i, x_j)$ is positive definite. Then K is a Mercer kernel.

Mercer theorem

Theorem

Let X a compact domain or manifold, ν a Borel measure over X and $K : X \times X \longrightarrow \mathbb{R}$ un Kernel de Mercer. Sea λ_k the k-th eigenvalue of L_K and $\{\phi_k\}_{k \geq 1}$ the corresponding eigenvector. Then, for all $x, y \in X$

$$K(x, y) = \sum_{k=1}^{\infty} \lambda_k \phi_k(x) \phi_k(y)$$

where the where the convergence is absolute (for each $(x, y) \in X \times X$) and uniform (on $(x, y) \in X \times X$).

Interpretation of the Mercer theorem

Geometrical interpretation

$K(\mathbf{x}, \mathbf{y})$ can be interpreted as a scalar product in the transformed space by $\Phi(\mathbf{x}) = (\sqrt{\lambda_1}\phi_1(\mathbf{x}), \sqrt{\lambda_2}\phi_2(\mathbf{x}), ...)$.

$$K(\mathbf{x}, \mathbf{y}) = < \Phi(\mathbf{x}), \Phi(\mathbf{y}) >$$

Examples

- Linear kernel $K(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T\mathbf{y}$.
- Polinomial kernel $K(\mathbf{x}, \mathbf{y}) = (a + \mathbf{x}^T\mathbf{y})^b$.
- RBF kernel $K(\mathbf{x}, \mathbf{y}) = e^{-\frac{1}{2\sigma^2}\|\mathbf{x}-\mathbf{y}\|^2}$.

Reproducing Kernel Hilbert Spaces I

Construction of RKHS

By the completion of the space generated by the linear combinations of $K(\mathbf{x}, \mathbf{x}_i)$:

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

Hyperplanes on the RKHS

$$f(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i k(\mathbf{x}, \mathbf{x}_i) = \sum_{i=1}^{m} \alpha_i \Phi(\mathbf{x})\Phi(\mathbf{x}_i) = \mathbf{w}^T \Phi(x),$$

$f(\mathbf{x}) = 0 \Rightarrow$ Hiperplane on the transformed space.

SVM as regularization method I

The SVM minimize the risk functional

$$F(f) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(\mathbf{x}_i)) + \mu\Omega(f)\,,$$

given a loss function and a hypothesis space:

SVM

- Loss function: hinge loss: $L(f(\mathbf{x}_i), y_i) = (1 - y_i f(\mathbf{x}_i))_+$, with $(\mathbf{x})_+ = max(\mathbf{x}, 0)$.
- Hypothesis space: RKHS of reproducing kernel K.

SVM as regularization method II

Problem to solve

$$\min_{f \in H_k} \frac{1}{n} \sum_{i=1}^{n} (1 - y_i f(\mathbf{x}_i))_+ + \mu \|f\|_K^2 .$$

¿Everything works?

$\{f \in H_K : \|f\|_K^2 \leq (sup_{y \in Y} L(y, 0))/\mu\}$

Compact hypothesis space.

Solution to the regularization problem I

Solution to the regularization problem

By the Representer theorem,

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i K(\mathbf{x}_i, \mathbf{x}) + b,$$

where the constant b can be added without lost of generality.

Problem y and solution of the SVM I

Regularization problem

$$\min_{f \in H_k} \frac{1}{n} \sum_{i=1}^{n} (1 - y_i f(\mathbf{x}_i))_+ + \mu \|f\|_K^2 .$$

Reformulated problem

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{i=1}^{n} \xi_i \\ s.a. \quad & y_i \left(\mathbf{w}^T \Phi(\mathbf{x}_i) + b \right) \geq 1 - \xi_i, \quad \forall i = 1, \ldots, n, \\ & \xi_i \geq 0 \qquad\qquad\qquad\qquad \forall i = 1, \ldots, n, \end{aligned}$$

Problem y and solution of the SVM II

Dual of the reformulated problem

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \\ s.a. \quad & 0 \leq \alpha_i \leq C \\ & \sum_{i=1}^{n} \alpha_i y_i = 0 \,. \end{aligned}$$

Solution to the dual problem

$$D^*(\mathbf{x}) = (\mathbf{w}^*)^T \Phi(\mathbf{x}) + b^* = \sum_{i=1}^{n} \lambda_i^* y_i K(\mathbf{x}, \mathbf{x}_i) + b^* .$$

Geometrical Idea

Steps of the SVM

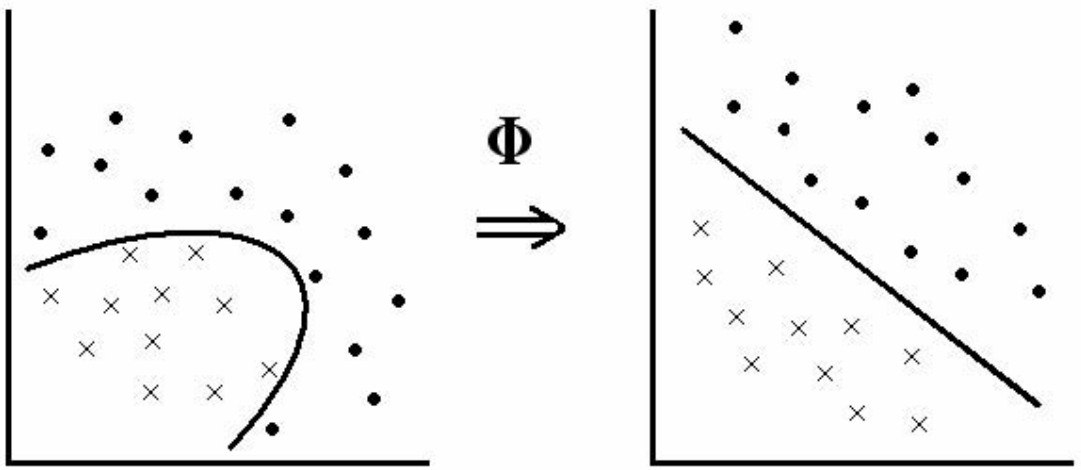
1. Data transformation onto a high dimensional space by the use of kernels.
2. Solution to the problem by the maximization of the margin between the classes.

Data transformation

Searching for linearity

First, the data are mapped into an space (generally of high dimension) bye the use of a kernel.

$$K(\mathbf{x}, \mathbf{y}) = < \Phi(\mathbf{x}), \Phi(\mathbf{y}) >$$



Hyperplane of maximum separation between the classes

Criteria for searching the hyperplane

Infinite feasible hyperplanes $\Rightarrow$ To maximize the margin between the classes.

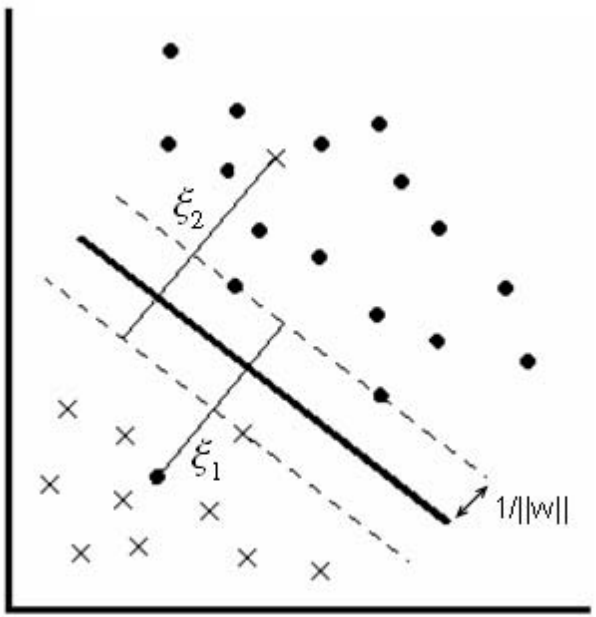
Problem to solve

$$\begin{array}{ll} \underset{\mathbf{w},b}{min} & \frac{1}{2}\left\|\mathbf{w}\right\|_2^2 \\ s.a. & y_i\left(\mathbf{w}^T\Phi(\mathbf{x}_i)+b\right)\geq 1 \quad \forall i=1,\ldots,n\,. \end{array}$$

Maximun separating hyperplane

Problem

If under Φ the problem does not become linearly separable $\Rightarrow$ Penalization of the misclassified observations.



Problema

$$
\begin{aligned}
\min_{\mathbf{w},b} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{n}\xi_i \\
s.a. \quad & y_i\left(\mathbf{w}^T\Phi(\mathbf{x}_i) + b\right) \geq 1 - \xi_i, \\
& \forall i = 1, \ldots, n, \\
& \xi_i \geq 0 \\
& \forall i = 1, \ldots, n,
\end{aligned}
$$

Open Problems

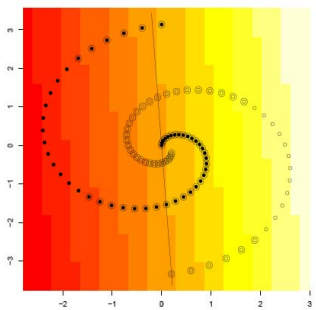
Open Problems

- Kernel election.
- Parameter tuning.

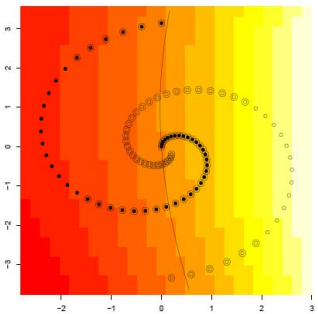
Objetive

Study and selection of the best kernel in classification problems.

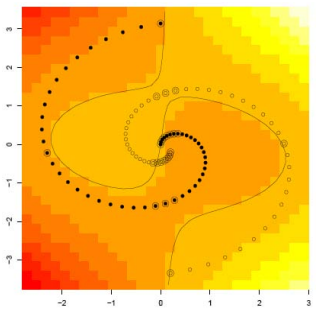
Spirals



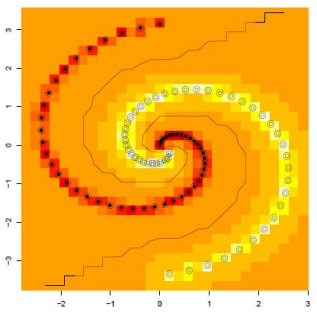
(e) Linear.



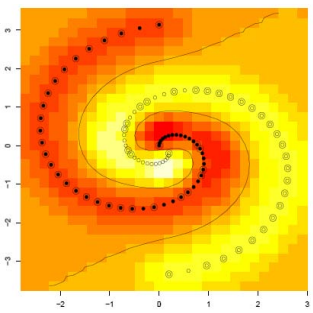
(f) Pol. gr. 2.



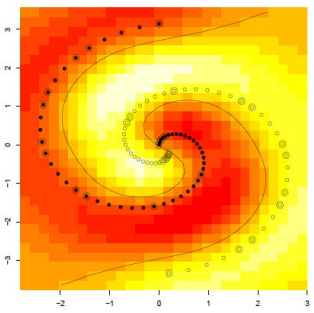
(g) Pol. gr. 3.



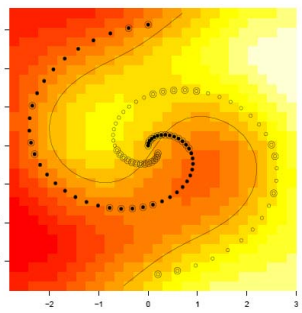
(h) $RBF_{\sigma=0.1}$.



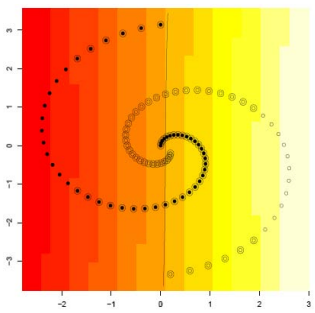
(i) $RBF_{\sigma=0.5}$.



(j) $RBF_{\sigma=1}$.



(k) $RBF_{\sigma=2}$.



(l) $RBF_{\sigma=10}$.

Why kernel combinations?

Since...

Any definite positive matrix is a Mercer kernel and can be used for training a SVM.

...then

It make sense to work with kernel combinations when the final matrix is positive definite.

Combining kernels

By using Semidefinite programming (Lanckriet)

$$\sum_{m=1}^{M} \mu_m K_m \, .$$

Martín, Muñoz, Moguerza kernel combinations

$$K^* = \bar{K} + \tau Y \sum_{i<j} g(K_i - K_j) Y \, .$$

$$K^* = \sum_i W_i \otimes K_i \, .$$

Comparative study

Comparative study for the cancer data set by using three kernels: lineal, polinomial y exponencial.

Method	Error Train	Error Test	% Support Vectors
Polinomial	0.1 (0.1)	7.8 (2.5)	8.3 (0.8)
RBF	0.0 (0.0)	10.8 (1.7)	65.6 (1.0)
Linear	2.6 (0.5)	3.7 (1.8)	7.1 (0.8)
AV	2.4 (0.3)	3.1 (1.3)	2.9 (0.4)
SDP	0.0 (0.0)	6.2 (1.6)	65.5 (1.9)

Combinations, is that all?

Kernel combination can be improved taking into account the geometrical structure of the problem:

Solution

- Local data structure
- Global data geometry

Motivation

SVMs and the Bayes Risk

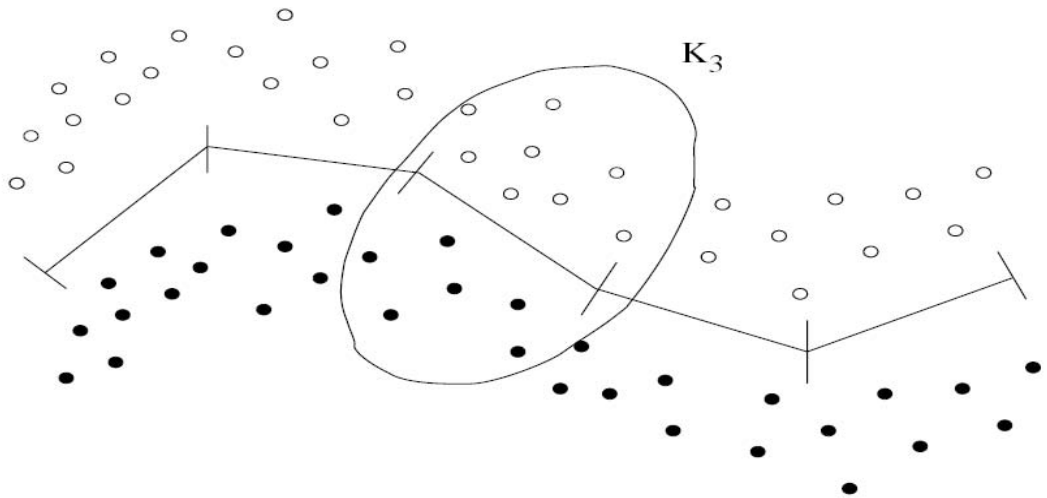
- Linear SVMs are optimal in the classical setting in which two normally distributed populations have to be separated.
- The support vector machine error converges to the optimal Bayes risk. and approaches the optimal Bayes rule (Lin, 2002), (Moguerza and Muñoz, 2006).

Local Linear Approximation for Kernel Methods: The Railway Kernel. Alberto Muñoz, Javier González and Isaac Martín de Diego. CIARP 2006: 936-944

Objetive

Objetive

To build a global kernel for general nonlinear classification problems that locally behaves as a linear (optimal) kernel.

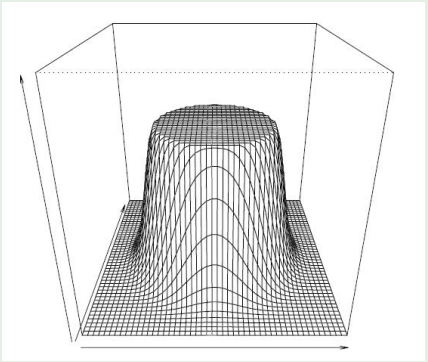


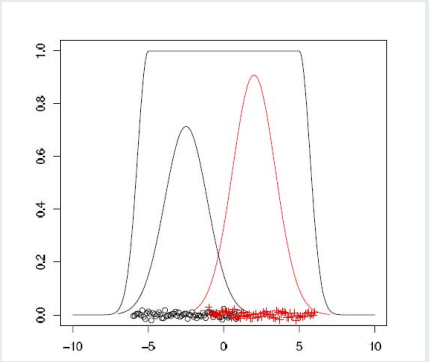
Indicator functions

Indicator kernel functions.

$$\lambda(x) = \begin{cases} 1 & \text{if} \|x-c\|^{1/2} \leq r \\ e^{-\gamma\left(\|x-c\|^2 - r^2\right)} & \text{if} \|x-c\|^{1/2} > r \end{cases}$$

Example

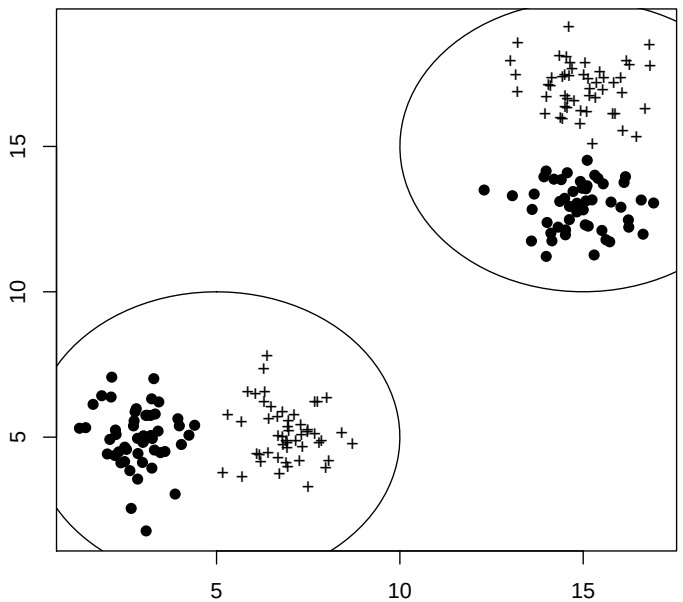




Two areas problem

Solution

Kernel K_1 solves the classification problem in area A_1 and so does K_2 in area A_2.



Kernel and solution

Railway kernel for a two areas problem

We define:

- $H_1(x, y) = \lambda_1(x)\lambda_1(y)$
- $H_2(x, y) = \lambda_2(x)\lambda_2(y)$

The global Railway Kernel K_R as follows:

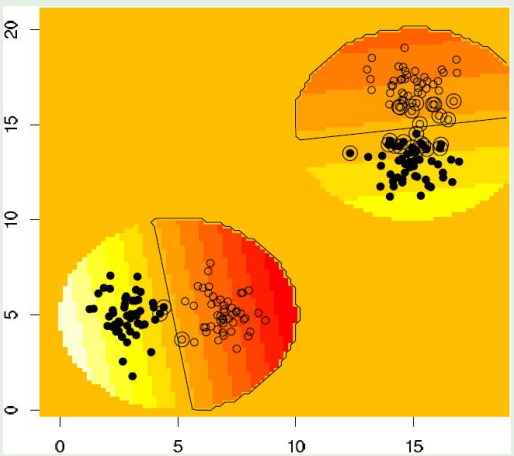
$$K_R(x, y) = H_1(x, y)K_1(x, y) + H_2(x, y)K_2(x, y)\,.$$

Solution

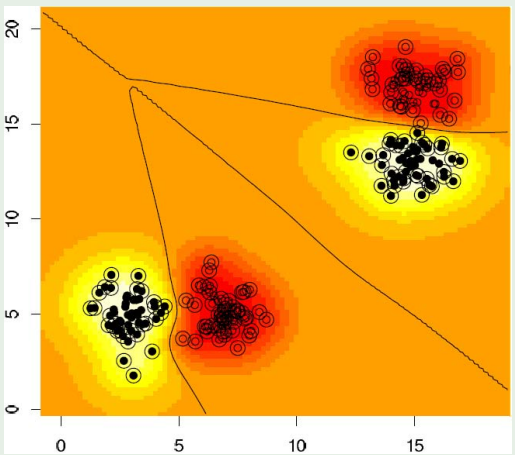
$$f(x) = \sum_{x_i \in A_1} \alpha_i K_1(x, x_i) + \sum_{x_j \in A_2} \alpha_j K_2(x, x_j) + b$$

Solution for the two areas problem

Example



(o) Railway kernel solution



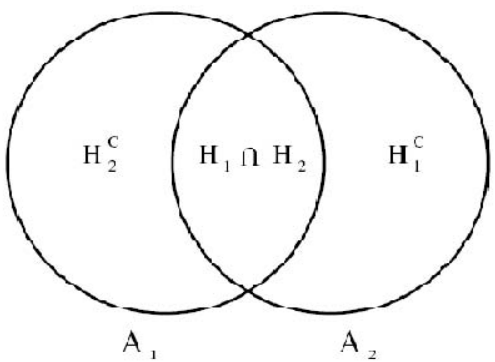
(p) RBF kernel solution

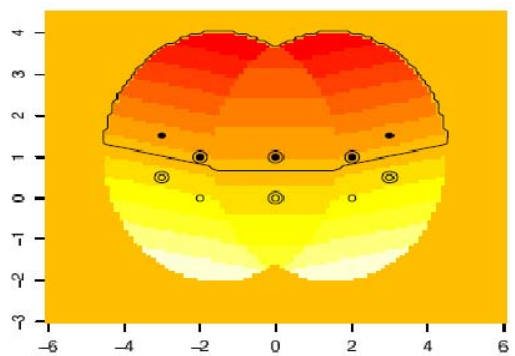
Intersections

Intersections

Areas where both kernels achieve the same performance, and should be equally weighted.

Average of the kernels





Good properties of the railway kernel

- Non tuning parameter dependence.
- Simple solution (locally optimal).
- Low dimension of the feature space.
- Small number of support vector (high generalization capability).

Areas Location

Before constructing the kernel $\Rightarrow$ Areas identification

A two steps algorithm is used:

1. Single labeled areas are created.
2. Final areas are obtained joining the nearest areas with different labels.

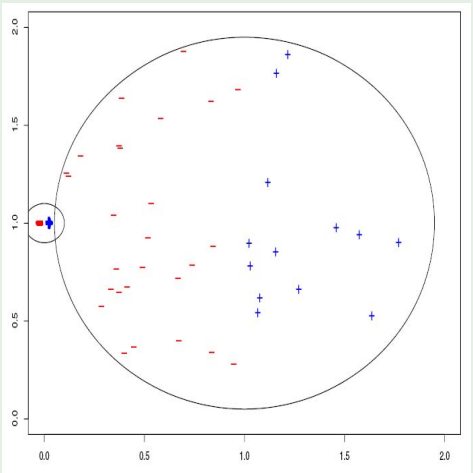
Data description

- The data set consists of 400 points in $\mathbb{R}^2$.
- Two main areas are created with different dispersion matrix.
- We use 80% of the data for training and 20% for testing.
- Several RBFs were compared with the railway kernel results.
- In SVM1 the parameter σ is chosen as a function of the data dimension For SVM2 we follow the heuristic proposed in (Keethi, 2003).

Simulated data

Simulated data representation

Two areas with different scattering matrices. The first area center is $(0, 1)$ and the second area center is $(1, 1)$. The areas do not coincide with the classes $\{-1, +1\}$.



Simulated data results

Results

Percentage of missclassified data and percentage of support vectors for the two different scattering data set: A_1 stands for the less scaterring group, A_2 stands for the most dispersive one.

Method	Train Error			Test Error			Support Vectors
	Total	A_1	A_2	Total	A_1	A_2	Total
RBF$_{\sigma=0.5}$	2.4	3.0	0.0	13.4	4.1	51.0	39.2
RBF$_{\sigma=5}$	4.6	5.8	0.0	13.6	8.6	35.0	82.6
RBF$_{\sigma=10}$	29.1	36.2	0.5	36.0	44.1	10.0	94.4
Railway Kernel	3.7	3.6	15.6	4.2	0.1	20.6	14.1
SVM$_1$	2.1	2.6	0.0	13.5	4.1	51.0	39.6
SVM$_2$	2.1	2.6	0.0	11.0	3.3	41.5	37.6

Experiment description

- The data set consists of 683 observations with 9 features each.
- We use 80% of the data for training and 20% for testing.
- Several RBFs were compared with the railway kernel results.

Breast Cancer data set

Example

Percentage of missclassified data, sensitivity (Sens.), specificity (Spec.) and percentage of support vectors for the cancer data set. Standard deviations in brackets.

Method	Train			Test			Support Vectors
	Error	Sens.	Spec.	Error	Sens.	Spec.	
Railway Kernel	2.5 (0.3)	0.979	0.974	2.9 (0.4)	0.975	0.876	18.6 (3.6)
SVM$_1$	0.1 (0.1)	1.000	0.999	4.2 (1.4)	0.989	0.942	49.2 (1.0)
SVM$_2$	0.0 (0.0)	1.000	0.999	2.9 (1.6)	0.963	0.975	49.2 (1.0)

New published Advances

Springer. ICANN 2007.

- *Spectral Measures for kernel matrices comparison*. Javier González and Alberto Muñoz.

New similarity measure for kernel matrices based on the definition on matrix pencils and simultaneous diagonalization.

Springer. CIARP 2007. (Submitted)

- *Joint Diagonalization of Kernels for Information Fusion.* Alberto Muñoz and Javier Gonzalez.

Analysis and solutions to possible redundances in the kernel fusion process.

References

- Support *Vector Machines with applications*. Javier Moguerza and Alberto Muñoz. Statistical Science. 2006, Vol. 21, No. 3, 322336. With Comments.
- *Estimation of High-Density Regions Using One-Class Neighbor Machines*. Alberto Muñoz, Javier M. Moguerza. IEEE Trans. Pattern Anal. Mach. Intell. 28(3): 476-480 (2006).
- Combining Kernel Information for Support Vector Classification. Isaac Martín de Diego, Javier M. Moguerza and Alberto Muñoz. Springer Multiple Classifier Systems 2004: 102-111.