Kernel combination in support vector machines for classification purposes

Javier González Hernández
Alberto Muñoz García

Universidad Carlos III de Madrid
Departament of Statistics

March 22, 2015

Motivation and objetive I

SVMs and kernel combinations

- Support Vector Machines (SVM) have proven to be a successful method for the solution of a wide range of classification problems.
- Kernel combinations in support vector machines for classification purposes is a successful area of research.

Motivation and objetive II

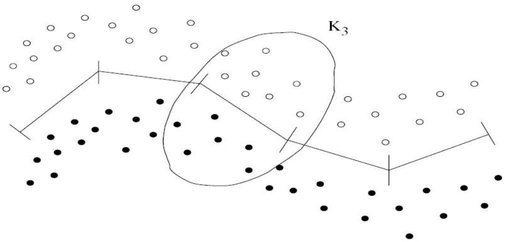
SVMs and the Bayes Risk

- 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).

Motivation and objetive III

Objective

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

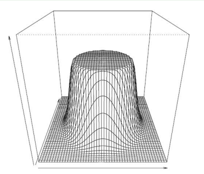


Indicator functions

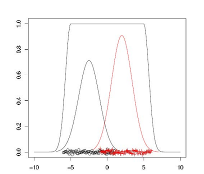
Indicator kernel functions. (a) 2D case. (b) 1D case.

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

Example



(a)

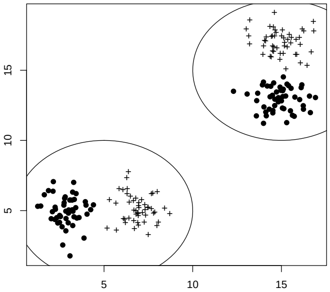


(b)

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:

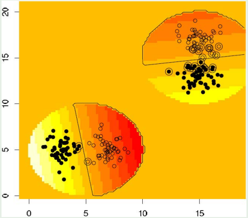
$$K_R(x, y) = H_1(x, y)K_1(x, y) + H_2(x, y)K_2(x, y) \, .$$

Solution

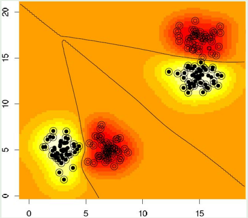
$$f(x) = \sum_{x_i \in A_1} \alpha_i K_1(x, x_i) + \sum_{x_j \in A_2} \alpha_j K_2(x, x_j) + b$$

Solution for the two areas problem

Example



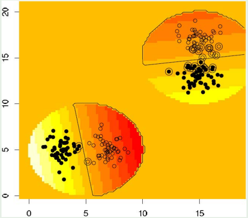
(c) Railway kernel solution



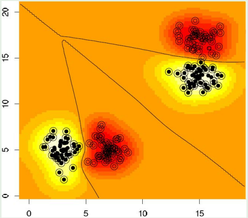
(d) RBF kernel solution

Solution for the two areas problem

Example



(e) Railway kernel solution



(f) RBF kernel solution

Good properties of the railway kernel

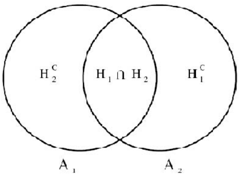
- Non tuning parameter dependence.
- Simple solution (locally optimal).
- Low dimension of the feature space.
- Small number of support vector (high generalization capability).

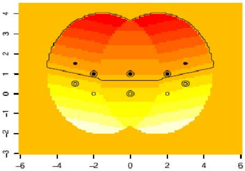
Intersections

Intersections

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

Average of the kernels





Itersections

Analytical expressions

$$K_R(x,y) = \begin{cases} K_1(x,y) & \text{if} \quad x,y \in A_1 \cap A_2^c\,, \\ K_2(x,y) & \text{if} \quad x,y \in A_1^c \cap A_2\,, \\ \frac{1}{2}(K_1(x,y) + K_2(x,y)) & \text{if} \quad x,y \in A_1 \cap A_2\,, \\ 0 & \text{otherwise}\,, \end{cases}$$

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.

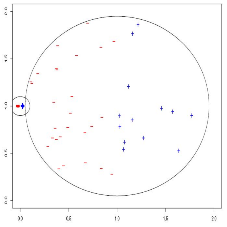
Data description

- 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)

Danke