Knowledge-Enhanced RBF Kernels

Kernel Methods for Prior-Knowledge Incorporation into SVMs

Student: Antoine Veillard
Supervisors: Dr. Stéphane Bressan, Dr. Daniel Racoceanu

School of Computing/Image and Pervasive Access Lab

May 23, 2012

Outline

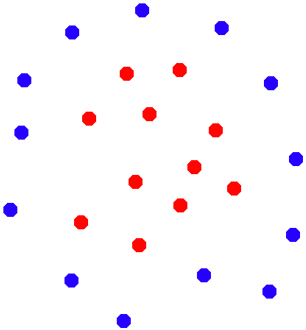
Knowledge-Enhanced RBF framework

Set of 3 kernel methods (ξRBF, pRBF, gRBF) for the incorporation of prior-knowledge into SVMs.

- Wide range of task-specific prior-knowledge
- Effective and practical
- Enables learning with very small and strongly biased training sets

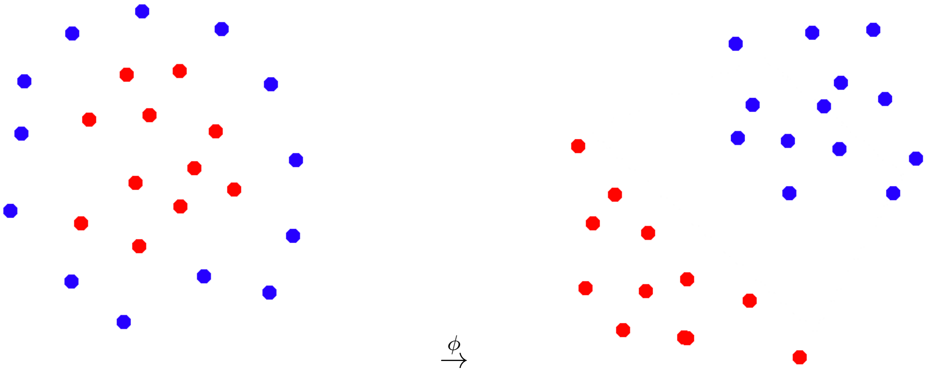
Contents

1. Support vector methods
2. Prior-knowledge incorporation into SVMs
3. Knowledge-Enhanced RBF kernels (ξRBF, pRBF, gRBF)
4. Application: MICO project



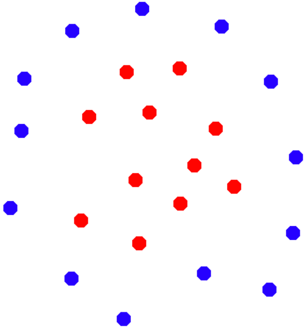
Original space $\mathcal{X}$

SVMs in a nutshell II

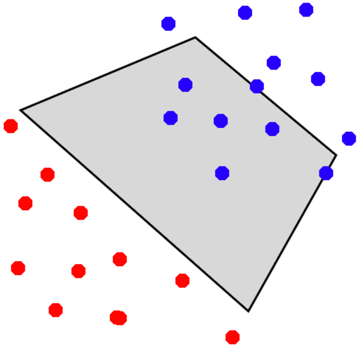


Original space $\mathcal{X}$ Hilbert space $\mathcal{H}$

SVMs in a nutshell III



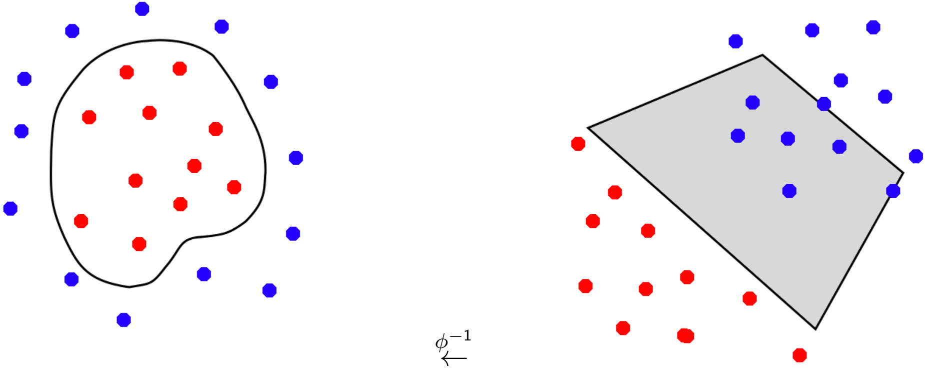
Original space $\mathcal{X}$



$$f(x) = \langle \sum_{i=1}^{N} \Phi(x_i), \Phi(x) \rangle_{\mathcal{H}}$$

Hilbert space $\mathcal{H}$

SVMs in a nutshell IV



$$f(x) = \sum_{i=1}^{N} K(x_i, x)$$

Original space $\mathcal{X}$

$$f(x) = \langle \sum_{i=1}^{N} \Phi(x_i), \Phi(x) \rangle_{\mathcal{H}}$$

Hilbert space $\mathcal{H}$

SVMs in a nutshell V

Key features

- Classification and regression
- Mapping $\Phi : \mathcal{X} \rightarrow \mathcal{H}$ can be implicit
- Only need positive-definite kernels:
  $K(x_1, x_2) = \langle \Phi(x_1), \Phi(x_2) \rangle_{\mathcal{H}}$

Radial basis function kernel

$$K_{\text{rbf}}(x_1, x_2) = \exp(-\gamma \|x_1 - x_2\|_2^2)$$

- Nonlinear
- Invariant by rotation and translation
- Bandwidth parameter γ to control over-fitting

$\Longrightarrow$ SVM+RBF combination = general-purpose learning tool

Data and prior-knowledge

SVMs

- Learning black-boxes
- Requires a large amount of high-quality training data

Real-world problems

- Data hard to obtain (cost, time, ethical reasons. . .)
- Seldom black-boxes: general and/or specific knowledge often available.

$\Longrightarrow$ need methods for the incorporation of PN into SVMs

PN incorporation into SVMs: the state-of-the-art

	Domain-specific	Data-specific	Problem-specific
Sample-based	• Virtual samples • π-SVM		• Knowledge initialization
Kernel-based	• Jittering kernels • Tangent distance kernels • Tangent vector kernels • Haar integration kernels • Kernels for finite sets • Local alignment kernel	• Weighted samples • Knowledge-driven kernel selection	
Optimization-based	• π-SVM • Semi-definite programming machines • Invariant hyperplanes	• Weighted samples • Transductive SVM	• KBSVM • Extensional KBSVM • Simpler KBSVM • Online KBSVM

KE-RBF: motivation

In real-life problems, specific information relevant to the task is often available. A few examples:

- In climatology, measurements have known pseudo-periods: seasonal and diurnal (dominant frequencies).
- In anatomy, the weight of a specimen increases *w.r.t.* its dimensions and the increase is cubic (monotonicity, correlation patterns).
- In oncology, small and regular cells are typical while large and irregular cells are atypical (regions of the feature space).

KE-RBF kernels provide a way to leverage on such prior-knowledge.

KE-RBF: framework I

KE-RBF framework

3 original kernel methods (ξRBF, pRBF and gRBF) based on adaptation of the pervasive RBF kernel for the incorporation of prior-knowledge into SVMs.

Main features:

- Deals with a *wide variety of prior-knowledge* that is *problem-specific*.
- Compensates for *small or biased training sets*.
- Preserves the *versatility* of the RBF kernel.
- *Ease of use:* just apply the kernel trick.

KE-RBF: framework II

		ξRBF	pRBF	gRBF
semi-global	unlabeled regions	$\times$		
	labeled regions			$\times$
global	monotonicity		$\times$	
	pseudo-periodicity	$\times$		
	frequency decomposition	$\times$		
	explicit correlation		$\times$	

ξRBF I

ξRBF kernel

$$K_a(x_1, x_2) = (\lambda + \mu\xi(x_1, x_2))K_{\text{rbf}}(x_1, x_2)$$

where $\xi : \mathcal{X}^2 \to \mathbb{R}$ contains the prior-knowledge and $\mu = 1 - \lambda \in [0, 1]$ controls the the amount of prior-knowledge.

Motivation

Induce appropriate modifications to the kernel distance according to the prior-knowledge.

Types of prior-knowledge

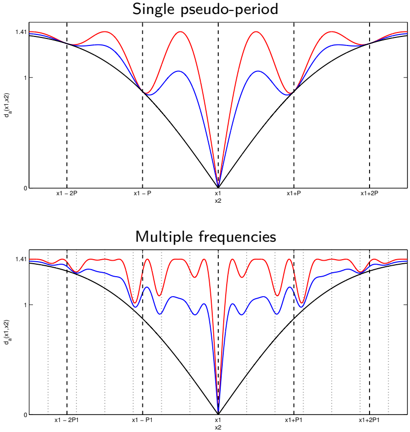
- Unlabeled sets (similarity)
- Frequency decomposition

Frequency decomposition

$$K_a(x_1, x_2) = \left(\lambda + \mu \prod_{i=1}^{N_0} \xi_i(x_1, x_2) \right) K_{\text{rbf}}(x_1, x_2)$$

with

$$\xi_i(x_1, x_2) = \frac{\cos\left(\frac{2\pi}{P_i}(x_{1,j} - x_{2,j})\right) + 1}{2} = \frac{\cos(2\pi f_i(x_{1,j} - x_{2,j})) + 1}{2}$$



black $\mu = 0$, blue $\mu = 0.5$ and red $\mu = 1$

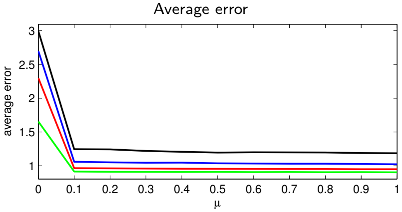
ξRBF III

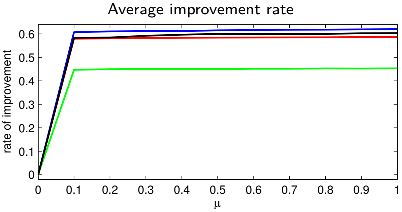
Application: meteorological predictions

- Prediction of daily temperatures in UK from 1914 to 2006.
- Publicly available from "UK Climate Projections" database.

Prior-knowledge

Cycle of seasons: pseudo-period of 365.25 days.





black $N = 50$, blue $N = 100$, red $N = 200$ and green $N = 400$

Definition

$K_a = K_{\text{rbf}} \otimes K$

PD kernel!

Prior-knowledge

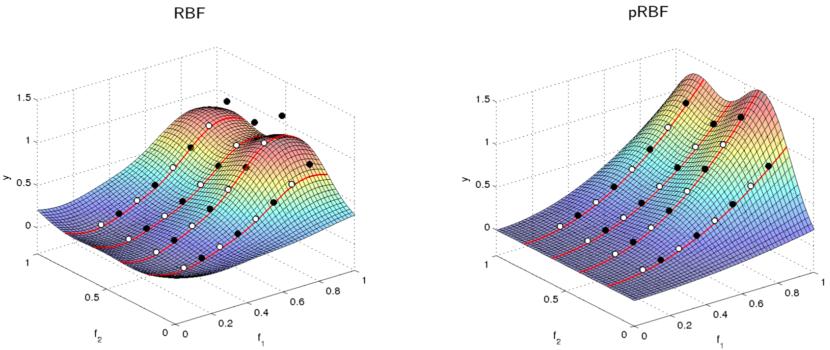
- Correlation patterns *w.r.t.* features.
- Monotonicity *w.r.t.* features.

There are restrictions on K.

Theorem (sketch)

Let E be a real vector space.
If $\{K_x | x \in \mathcal{X}\} \subset E$ then $\hat{f} \in E$.

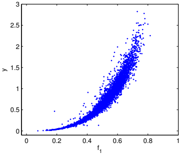
Illustration: quadratic correlation.

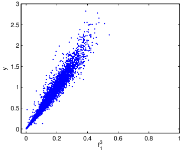


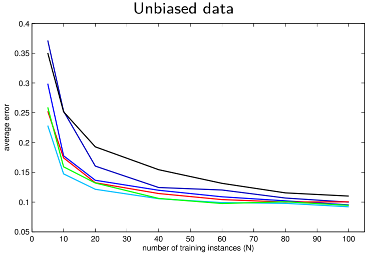
pRBF III

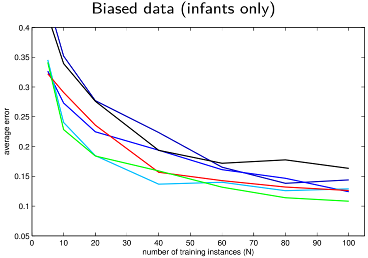
Application: Anatomy of abalones

- Predict weight of abalones (y) from morphological parameters including length (f_1), width (f_2), height (f_3) and other features.
- From public "UCI abalone" dataset.
- A priori correlation between dimensions and weight.









Black RBF, d. blue f_1, blue f_1^2, l. blue f_1^3, red $f_1 f_2$ and green $f_1 f_2 f_3$.

Generalization of the RBF kernel from points to arbitrary sets.

Definition

$$\begin{aligned} K_{\mathrm{grbf}} : \quad \mathfrak{P}(\mathbb{R}^n)^2 &\rightarrow \mathbb{R} \\ (\mathcal{A}, \mathcal{B}) &\mapsto \exp(-\gamma d(\mathcal{A}, \mathcal{B})^2) \end{aligned}$$

with

$$d(\mathcal{A}, \mathcal{B}) = \begin{cases} \inf_{a \in \mathcal{A} \wedge b \in \mathcal{B}} \|a - b\|_2 & \text{if } \mathcal{A} \neq \emptyset \text{ and } \mathcal{B} \neq \emptyset \\ \infty & \text{otherwise} \end{cases}$$

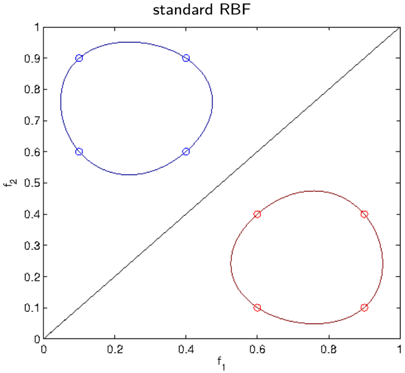
NOT PD!

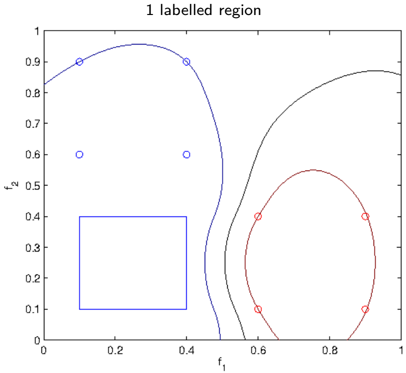
Prior-Knowledge

Labelled regions of $\mathcal{X}$.

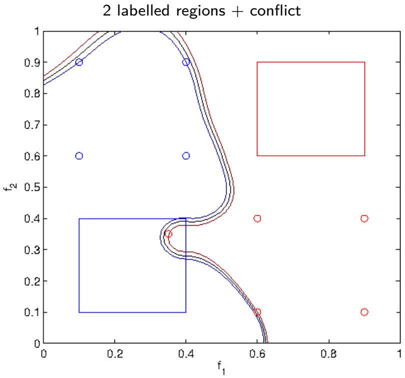
gRBF II

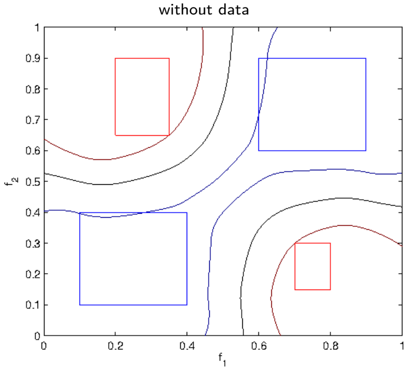
Examples (classification)



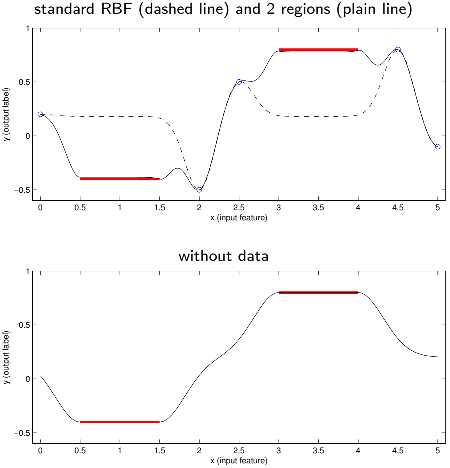


Examples (classification)

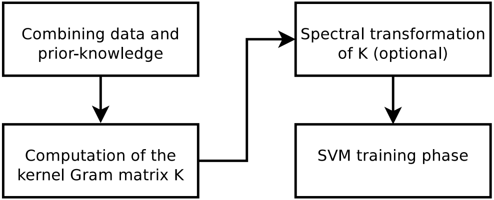




Examples (regression)



gRBF V



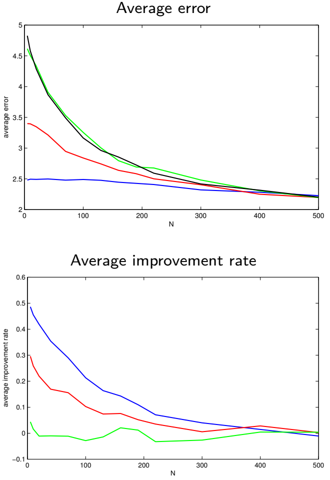
Computational challenges

- Dealing with non-PD kernels: flipping and shifting.
- Computing the set distance: balls, orthotopes, convex polytopes.
- Dealing with conflicts between data and prior-knowledge.
- Managing the computational complexity.

gRBF VI

Application: daily meteorological predictions using averages

- Daily temperatures for 10 years at 100 locations.
- Prior-knowledge: yearly, seasonal, monthly averages.
- Data publicly available from "UK Climate Projections" database.



Black RBF, blue monthly, red seasonal and green yearly

KE-RBF: conclusion

- Effective: drastic improvement of results by the incorporation of PN
- Efficient: computational complexity comparable to RBF
- Enables training with much smaller training sets
- Enables training with strongly biased training sets

Application: MICO I

Cognitive Microscope Project

- 3 years ANR project
- Partners: IPAL, LIP6, Thales, AGFA, TRIBVN, GHU-PS
- Automatic breast cancer grading (BCG): diagnosis/prognosis of breast cancer from surgical biopsies

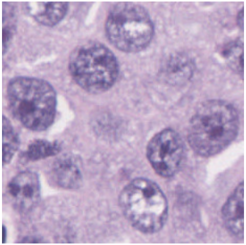
Assessment of cytonuclear atypiae (CNA)

- Central component in BCG
- Based on the morphology of cell nuclei
- Requires accurate extraction of the cell nuclei

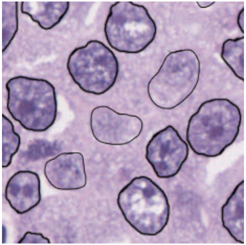
Application: MICO II

Challenges

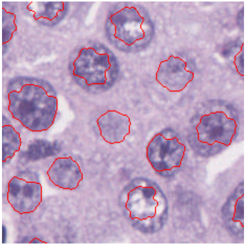
- Inhomogeneous objects in inhomogeneous background
- Low object-background contrast
- Frequent overlaps between nuclei
- Existing methods based on pixel intensities perform poorly



Original image



Manual segmentation

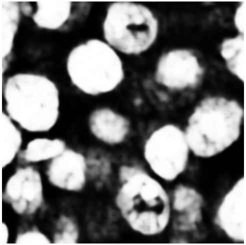


Automatic segmentation

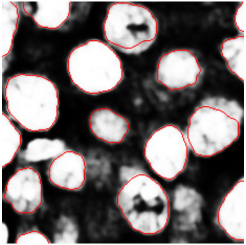
Application: MICO III

Solution

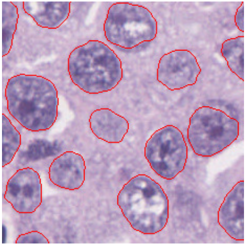
- Use SVMs with KE-RBF kernels to create a new modality from the original image using color, texture, scale and shape priors.
- The new modality is a probability map where objects and backgrounds are smoothed out.
- Apply the segmentation algorithms on the new modality



Probability map



Segmentation on the probability map



Results on original image

Student's publications

- A. Veillard, D. Racoceanu, and S. Bressan "pRBF Kernels: A Framework for the Incorporation of Task-Specific Properties into Support Vector Methods", submitted.
- A. Veillard, M. S. Kulikova, and D. Racoceanu, "Cell Nuclei Extraction from Breast Cancer Histopathology Images Using Color, Texture, Scale and Shape Information", TP2012.
- M. S. Kulikova, A. Veillard, L. Roux, and D. Racoceanu, "Nuclei extraction from histopathological images using a marked point process approach", SPIE medical imaging 2012.
- A. Veillard, D. Racoceanu, and S. Bressan, "Incorporating Prior-Knowledge in Support Vector Machines by Kernel Adaptation", ICTAI2011.
- C-H. Huang, A. Veillard, L. Roux, N. Lomenie, and D. Racoceanu, "Time-efficient sparse analysis of histopathological Whole Slide Images", CMIG vol 35 (2011).
- A. Veillard, N. Lomenie, and D. Racoceanu, "An Exploration Scheme for Large Images: Application to Breast Cancer Grading", ICPR2010.
- A. Veillard, E. Melissa, C. Theodora, and S. Bressan, "Learning to Rank Indonesian-English Machine Translations", MALINDO2010.
- A. Veillard, E. Melissa, C. Theodora, D. Racoceanu, and S. Bressan, "Support Vector Methods for Sentence Level Machine Translation Evaluation", ICTAI2010.
- L. Roux, A E. Tutac, A. Veillard, J-R. Dalle, D. Racoceanu, N. Lomenie, and J. Klossa, "A Cognitive Approach to Microscopy Analysis Applied to Automatic Breast Cancer Grading", ECP2009.
- L. Roux, A E. Tutac, N. Lomenie, D. Balensi, D. Racoceanu, A. Veillard, W-K. Leow, J. Klossa, and T C. Putti, "A cognitive virtual microscopic framework for knowlege-based exploration of large microscopic images in breast cancer histopathology", EMBC2009.
- D. Racoceanu, A E. Tutac, W. Xiong, J-R. Dalle, C-H. Huang, L. Roux, W-K. Leow, A. Veillard, J-H. Lim, T C. Putti, et al., "A virtual microscope framework for breast cancer grading", A-STAR CCO workshop 2009.

APPENDIX

- PD kernels
- Kernel trick
- Statistical learning
- Structural risk minimization
- SVMs: a statistical approach
- Learning bounds in RKHS
- Representer theorem
- Graphical interpretation of SVMs
- C-SVM
- ξRBF: unlabeled sets
- pRBF main theorem
- Dealing with indefinite kernels
- gRBF: managing conflicts
- Application: machine translation evaluation
- Application: exploration of very large images

Positive definite kernels

PD kernels

$K : \mathcal{X}^2 \to \mathbb{R}$ is a PD kernel if:

1. $\forall (x_1, x_2) \in \mathcal{X}^2,\ K(x_1, x_2) = K(x_2, x_1)$ (K symmetric)
2. $\forall (x_1, \ldots, x_N) \in \mathcal{X}^N,\ \forall (v_1, \ldots, v_N) \in \mathbb{R}^N,$
   $\sum_{i=1}^N \sum_j j = 1^N v_i v_j K(x_i, x_j) \geq 0$ (the Gram matrix is PSD)

Aronszajn (1950)

The following assertions are equivalent:

1. $K : \mathcal{X}^2 \to \mathbb{R}$ is a PD kernel
2. There is a Hilbert space $\mathcal{H}$ and $\Phi : \mathcal{X} \to \mathcal{H}$ such that:
   $\forall (x_1, x_2) \in \mathcal{X}^2,\ K(x_1, x_2) = \langle \Phi(x_1), \Phi(x_2) \rangle_{\mathcal{H}}$

$\Longrightarrow$ *A PD kernel is a generalization of the "dot" product in* $\mathbb{R}^n$.

The kernel trick

Let $K_x(x') = K(x, x')$ ("sections" of K).

Reproducing Kernel Hilbert Space (RKHS)

- $\Phi_K : x \mapsto K_x$
- $\mathcal{H}_k = \mathrm{span}\{K_x | x \in \mathbb{R}\}$

are realizations of Φ and $\mathcal{H}$ from Aronszajn's theorem.

Generally, explicit computations in $\mathcal{H}$ is not practical or even feasible. Instead, projections are handled through evaluations of the kernel product.

Induced metric

$$\|\Phi(x_1) - \Phi(x_2)\|^2_{\mathcal{H}} = \langle \Phi(x_1) - \Phi(x_2), \Phi(x_1) - \Phi(x_2)\rangle_{\mathcal{H}}$$
$$= \langle \Phi(x_1), \Phi(x_1)\rangle_{\mathcal{H}} + \langle \Phi(x_2), \Phi(x_2)\rangle_{\mathcal{H}} - 2\langle \Phi(x_1), \Phi(x_2)\rangle_{\mathcal{H}}$$
$$= K(x_1, x_1) + K(x_2, x_2) - 2K(x_1, x_2) \text{ (Aronszajn)}$$

Kernel "trick"!

Let:

- $\mathscr{P}$ probability distribution with values in $\mathcal{X} \times \mathcal{Y}$ ($\mathcal{Y} \subset \mathbb{R}$) *a.k.a.* the *problem*.
- $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ set of labeling models *a.k.a. hypothesis*.
- $\mathcal{S}_N = (x_i, y_i)_{i=1}^N$ a *training set i.i.d.* according to $\mathscr{P}$.
- $\Lambda : \mathcal{X} \times \mathcal{Y} \times \mathcal{H} \to \mathbb{R}$ a *loss* function.

Find a labeling model $f \in \mathcal{H}$ minimizing:

Theoretical risk minimization

$R(f) = \mathbb{E}_{(X,Y)\sim\mathscr{P}}(\Lambda(X, Y, f))$

Problem: R is unknown in practice.

Empirical risk minimization

$R^*(f) = \frac{1}{N}\sum_{i=1}^N \Lambda(x_i, y_i, f)$

Problem: Prone to overfitting.

Structural risk minimization

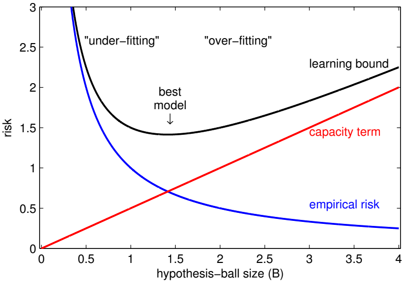
Adapted from the work by Vapnik and Chervonenkis (1974).

Learning bounds

Under certain conditions ($\mathcal{H}$ must be a RKHS!) and "high-probability":

$$R(f) \leq R^*(f) + \frac{\kappa B}{\sqrt{N}}$$

for some constant $\kappa > 0$ and $B \geq \|f\|_{\mathcal{H}}$.



$\Longrightarrow$ tradeoff between minimization of $R^*(f)$ and $\|f\|_{\mathcal{H}}$.

SVMs: a statistical approach

SVMs are a direct implementation of the SRM principle into an optimization problem.

SVM problem

$$\underset{f \in \mathcal{H}}{\operatorname{argmin}} \; R^*(f) + \lambda \|f\|_{\mathcal{H}}^2$$

The tradeoff parameter $\lambda \geq 0$ is usually adjusted with a tuning method such as grid search.

Solution space

By the *representer theorem*, the optimal solution $\hat{f}$ has the following form:

$$\hat{f} = \sum_{i=1}^{N} \alpha_i K_x \text{ with } \forall i, \; \alpha_i \in \mathbb{R}$$

which makes the problem convex and efficiently solvable.

Learning bounds in RKHS

Hypothesis

- $\mathscr{P}$ be a problem *w.r.t.* $\mathcal{X}$ and $\mathcal{Y} = \{-1, +1\}$;
- Λ be a L_ϕ-Lipschitz ϕ-loss function;
- $\mathcal{H}_B \subset \mathbb{R}^{\mathcal{X}}$ a RKHS ball of models with radius B;
- $\mathcal{S}_n$ a set of n independent observations of $S = (X, Y) \sim \mathscr{P}$
- Λ is bounded by ψ_Λ for any observation from $\mathscr{P}$

Bound

With probability at least $1 - \delta$ (for any $\delta \in [0, 1]$):

$$R_{\Lambda,\mathscr{P}}(f) \leq R_{\mathrm{emp}_{\Lambda,\mathcal{S}_n}}(f) + 2BL_\phi \sqrt{\frac{\mathbb{E}_X\left[K(X, X)\right]}{n}} + \psi_\Lambda \sqrt{\frac{-\log \delta}{2n}}$$

Weak representer theorem

Let:

- $\mathcal{X}$ be a non-empty set
- $K : \mathcal{X}^2 \to \mathbb{R}$ be a PD kernel with RKHS $\mathcal{H}_K$.
- $\mathcal{S} = \{x_1, \ldots, x + n\} \subset \mathcal{X}$ be a finite subset of $\mathcal{X}$
- $\Lambda : \mathbb{R}^n \to \mathbb{R}$ be a "loss" function
- $\lambda > 0$
- $\Omega : \mathbb{R} \to \mathbb{R}$ be a strictly increasing function

If $\hat{f}$ is a solution of the optimization problem:

$$\hat{f} = \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \Lambda(f(x_1), \ldots, f(x_n)) + \lambda\Omega(\|f\|_{\mathcal{H}_K})$$

then $\hat{f}$ admits a solution of the form:

$$\hat{f} = \sum_{i=1}^{n} \alpha_i K_{x_i}$$

C-SVM

$$
\begin{aligned}
&\underset{(\beta_i)_{i=1,\dots,N}\in\mathbb{R}^N,\ b\in\mathbb{R}}{\text{minimize}} && C\sum_{i=1}^{N}\xi_i + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \beta_i \beta_j K(x_i, x_j) \\
&\text{subject to} && y_i\left(\sum_{j=1}^{N} y_j \beta_j K(x_i, x_j) + b\right) - 1 + \xi_i \geq 0, \quad i = 1, \dots, N \\
& && \xi_i \geq 0, \quad i = 1, \dots, N \\
& && 0 \leq \beta_i \leq C, \quad i = 1, \dots, N
\end{aligned}
$$

ξRBF: unlabeled sets I

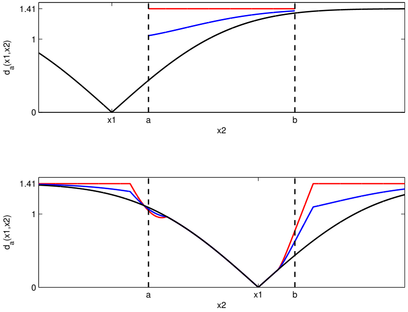
Unlabeled set $\mathcal{A}$ (crisp)

$$\chi(x) = \begin{cases} 1 & \text{if } x \in \mathcal{A} \\ -1 & \text{if } x \notin \mathcal{A} \end{cases}$$

Unlabeled set $\mathcal{A}$ (fuzzy)

$$\chi(x) \in [-1, 1]$$

K_a is PD.

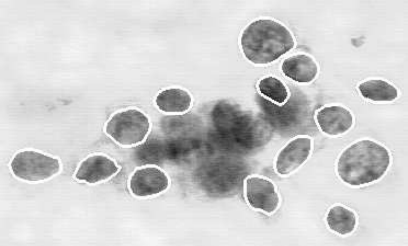


black $\mu = 0$, blue $\mu = 0.5$ and red $\mu = 1$

ξRBF: unlabeled sets II

Application: Breast cancer diagnosis from FNA

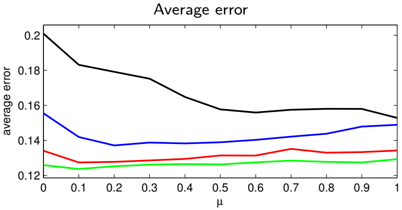
- Diagnose cancer from cell morphology.
- Publicly available "UCI Wisconsin Breast Cancer" dataset.

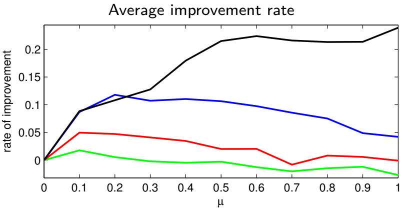


Prior-knowledge

Advice from pathologist:

- Cells with a smooth contour and a regular texture are typical of normal tissue.
- Cells with a rough contour and a irregular texture are atypical.





black $N = 8$, blue $N = 16$, red $N = 32$ and green

$N = 64$

pRBF main result

Let:

- E be a vector field over $\mathbb{R}$;
- K be a PD kernel over $\mathbb{R}^m$ such that $\{K_x | x \in \mathbb{R}^m\} \subset E$;
- $$\begin{array}{llll} K_a : & (\mathbb{R}^{n-m} \times \mathbb{R}^m)^2 & \to & \mathbb{R} \\ & ((x_{1,1}, x_{2,1}), (x_{1,2}, x_{2,2})) & \mapsto & K_{\text{rbf}}(x_{1,1}, x_{2,1}) K(x_{1,2}, x_{2,2}) \end{array}$$

  be a pRBF kernel over $\mathbb{R}^n$ ($m < n$) with $\mathcal{H}_a$ its RKHS;
- $\mathcal{S} = \{x_1, \dots x_N\} \in (\mathbb{R}^n)^N$ be a finite set;
- $\Omega : \mathbb{R} \to \mathbb{R}$ a strictly increasing function;
- $\lambda > 0$;
- $\Lambda : \mathbb{R}^N \to \mathbb{R}$ be any function.

If $\hat{f} : \mathbb{R}^{n-m} \times \mathbb{R}^m \to \mathbb{R}$ is a the solution of the optimization problem:

$$\underset{f \in \mathcal{H}_a}{\operatorname{argmin}} \Lambda(f(x_1), \dots, f(x_N)) + \lambda \Omega(\|f\|_{\mathcal{H}_a})$$

then $\forall x' \in \mathbb{R}^{n-m}$, $\hat{f}_{x'} \in E$ where:

$$\begin{array}{llll} \hat{f}_{x'} : & \mathbb{R}^m & \to & \mathbb{R} \\ & x & \mapsto & \hat{f}(x', x) \end{array}$$

Dealing with indefinite kernels

- The kernel Gram matrix K is symmetric, therefore:

$$K = U\mathrm{diag}(\lambda_1, \ldots, \lambda_N)U^T$$

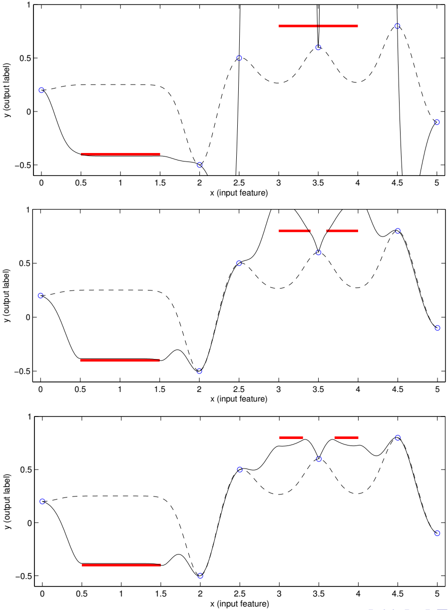
- **Flipping**

$$\mathrm{flip}(K) = U\mathrm{diag}(|\lambda_1|, \ldots, |\lambda_N|)U^T$$

- **Shifting**

$$mathrmshift(K) = U\mathrm{diag}(\lambda_1 + \eta, \ldots, \lambda_N + \eta)U^T$$

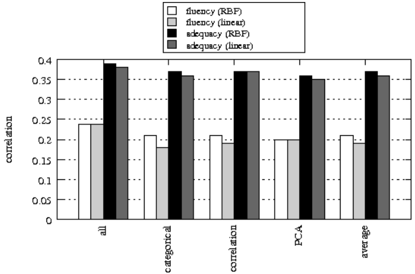
gRBF: managing conflicts



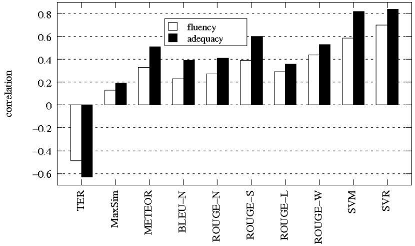
Machine translation evaluation

- Standard metrics for MTE: ROUGE, BLEU, NIST, METEOR. . .
- Metrics tend to perform poorly with less common languages and domains.
- ML-based approach using SVMs.
- Focus on feature modeling and learning machine.

Different feature models and SVMs



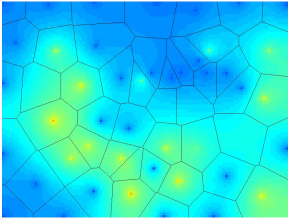
Our SVM-based approach outperforms previous works



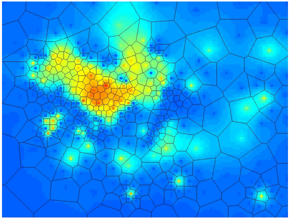
Overview

- A whole slide image typically consists of several thousands individual frames $\Longrightarrow$ an exhaustive analysis is not feasible.
- Selection of the highest scoring frames with a dynamic sampling algorithm based on computational geometry.

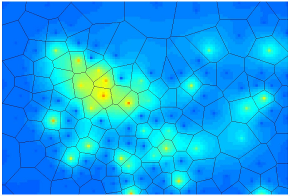
50 samples



400 samples



150 samples



Result