

Kernel methods and the exponential family

Stéphane Canu¹ and Alex J. Smola²

1- PSI - FRE CNRS 2645
INSA de Rouen, France
St Etienne du Rouvray, France
Stephane.Canu@insa-rouen.fr

2- Statistical Machine Learning Program
National ICT Australia and ANU
Alex.Smola@nicta.com.au

ESANN 2005



two streams in statistical learning theory

	one stream	an other one
Model	model probabilities	model target function
Optimization	use likelihood	minimize some contrast
Example	mixture model	SVM
pro & cons pro & cons	all formal slow	focus on what is needed formal limits (P ?)

Question

Can we build a bridge between these two frameworks?

1 Theoretical framework

- hypothesis: RKHS
- probabilities: exponential family

2 The bridge for the 2 class problem

- conditional exponential family in a RKHS
- parameter estimation: from functions to vectors

3 The bridge for novelty detection

- detection of abrupt changes
- generalized likelihood ratio

In the beginnig there was the kernel

Definition

- positive function: symmetric & $\sum_{i=1}^n \sum_{j=1}^m \alpha_i \alpha_j k(x_i, x_j) > 0$
- kernel: any positive function $k(x, y) \quad \forall x, y \in \Omega$

- $\mathcal{H}_0 = \{f \in \mathbb{R}^\Omega \mid f(x) = \sum_{i=1}^n \alpha_i k(x, x_i), n \in \mathbf{N}, \alpha_i \in \mathbb{R}, x_i \in \Omega\}$

- scalar product on \mathcal{H}_0

$$\langle f(\cdot), g(\cdot) \rangle_{\mathcal{H}_0} = \sum_{i=1}^n \sum_{j=1}^m \alpha_i \beta_j k(x_i, x_j)$$

properties of the scalar product on \mathcal{H}_0

- $\langle f(\cdot), k(x, \cdot) \rangle_{\mathcal{H}_0} = f(x)$ evaluation functional
- $\langle k(x, \cdot), k(y, \cdot) \rangle_{\mathcal{H}_0} = k(x, y)$ reproducing property

hypothesis: RKHS

From Kernel to hypothesis and back

the Hilbert space has to be complete

$$\mathcal{H} = \overline{\mathcal{H}_0}$$

definition: RKHS - reproducing kernel Hilbert space

- a Hilbert space
- $\mathcal{H} \subset \mathbb{R}^\Omega$
- there exists some kernel k such that

$$\forall f \in \mathcal{H}, \quad \langle f(\cdot), k(x, \cdot) \rangle_{\mathcal{H}} = f(x)$$

...and back

if the **evaluation functional is continuous** in some hilbertian sub space then it is a RKHS

proof: there exists some k (Riesz theorem)

Exponential family

definition: Exponential family

$$\mathbf{P}(x|\theta) = \mu(x) \exp^{\theta^\top \varphi(x) - g(\theta)}$$

- $\mu(x)$ the support measure
- θ the “natural” parameter
- $\varphi(x)$ the sufficient statistic

$g(\theta)$ the Log partition function is **CONVEX**

$$\int \mathbf{P}(x) dx = 1 \quad \Leftrightarrow \quad g(\theta) = \log \int \exp^{\theta^\top \varphi(x)} \mu(x) dx$$

definition: Domain

$$\Theta = \left\{ \theta \in \mathbf{R}^k \mid g(\theta) < \infty \right\}$$

probabilities: exponential family

Examples: Bernoulli, Normal, power exponential family, even the uniform!

the Bernoulli probability distribution

- $Y \sim \mathcal{B}(p) \quad p \in [0, 1]$
- $\mathbf{P}(Y = y) = p^y(1 - p)^{1-y}, \quad \forall y \in \{0, 1\}$
- $\mathbf{P}(Y = y) = \exp^{y \log p + (1-y) \log(1-p)},$
- $\mathbf{P}(Y = y) = \exp^{\theta y - g(\theta)},$
- $\theta = \log \frac{p}{1-p}, \quad g(\theta) = \log(1 + \exp^{\theta})$ it is CONVEX

likelihood maximization is a CONVEX problem

$$\ell(x, \theta) = -\log \mathbf{P}(Y = y)$$

$$\ell(x, \theta) = -\theta y + g(\theta) \quad \text{it is CONVEX}$$

It is a matter of parameterization to get convexity

Generalized exponential family

definition: Exponential family in a RKHS

- \mathcal{H} is a RKHS with kernel k (...or the other way round)
- the “natural” parameter $\theta \in \mathcal{H}$
- the sufficient statistic is $\varphi(x) = k(x, \cdot)$

$$\mathbf{P}(x) = \mu(x) \exp^{\langle \theta(\cdot), k(x, \cdot) \rangle_{\mathcal{H}} - g(\theta)}$$

- the domain: $\theta \in \Theta \subseteq \mathcal{H}$

Bayesian framework: prior distribution on θ

$$\mathbf{P}(\theta) = \frac{1}{Z} \exp^{-\frac{\|\theta\|_{\mathcal{H}}^2}{2\sigma^2}}$$

MAP

MAP principle

$$\max_{\theta \in \Theta} \mathbf{P}(\theta | \text{data}) \Leftrightarrow \min_{\theta \in \Theta} -\log \mathbf{P}(\text{data} | \theta) - \log \mathbf{P}(\theta)$$

$$\min_{\theta \in \Theta} \mathcal{J}(\theta) = - \sum_{i=1}^n \theta(x_i) + g(\theta) + \frac{1}{2\sigma^2} \|\theta\|_{\mathcal{H}}^2 \quad \text{it is CONVEX}$$

recall $\theta(x_i) = \langle \theta(\cdot), k(x_i, \cdot) \rangle_{\mathcal{H}}$

ok, nice, but this is a **functional** minimization problem...

conditional for classification

$x \in \Omega$ and $y \in \{-1, 1\}$

The exponential conditional densities:

$$\mathbb{P}(y|x; \tilde{\theta}) = \exp(\langle \tilde{\theta}(\cdot, \cdot), \tilde{k}(x, y, \cdot, \cdot) \rangle_{\tilde{\mathcal{H}}} - g(\tilde{\theta}(x)))$$

$$g(\tilde{\theta}(x)) = \log \sum_y \exp(\langle \tilde{\theta}(\cdot, \cdot), \tilde{k}(x, y, \cdot, \cdot) \rangle_{\tilde{\mathcal{H}}})$$

find out the kernel

$$\tilde{\theta}(x, y) = y\theta(x) \Rightarrow \langle \theta(\cdot), \tilde{k}(x, y, \cdot, \cdot) \rangle_{\tilde{\mathcal{H}}} = y \langle \theta(\cdot), k(x, \cdot) \rangle_{\mathcal{H}}$$

Conditional Rademacher

$$\mathbb{P}(y|x; \theta) = \exp^{y\theta(x) - g(\theta(x))}$$

$$g(\theta(x)) = \log \left(\exp^{\theta(x)} + \exp^{-\theta(x)} \right)$$

parameter estimation: from functions to vectors

MAP for classification

$$\mathbf{P}(y|x; \theta) = \exp^{y\theta(x) - g(\theta|x)} \quad \mathbf{P}(\theta) = \frac{1}{Z} \exp^{-\frac{\|\theta\|_{\mathcal{H}}^2}{2\sigma^2}}$$

$$J(\theta) = -\sum_{i=1}^n y_i \theta(x_i) + \sum_{i=1}^n g(\theta(x_i)) + \frac{1}{2\sigma^2} \|\theta\|_{\mathcal{H}}^2$$

Gateau differential

$$\nabla_{\theta} J(\theta) = -\sum_{i=1}^n y_i k(x_i, \cdot) + \sum_{i=1}^n g'(\theta(x_i)) k(x_i, \cdot) + \frac{1}{\sigma^2} \theta(\cdot)$$

The representer theorem

$$\nabla_{\theta} J(\theta) = 0 \quad \Leftrightarrow \quad \theta(\cdot) = \sum_{i=1}^n \underbrace{\sigma^2 (y_i - g'(\theta(x_i)))}_{\alpha_i} k(x_i, \cdot)$$

From functions to vectors

$$\min_{\theta \in H} J(\theta) = - \sum_{i=1}^n y_i \theta(x_i) + \sum_{i=1}^n \log \left(\exp^{\theta(x_i)} + \exp^{-\theta(x_i)} \right) + \frac{1}{2\sigma^2} \|\theta\|^2$$

$$\begin{aligned} \theta(x_i) &= \sum_{j=1}^n \alpha_j k(x_i, x_j) \\ \theta &= K\alpha \quad \in \mathbb{R}^n \end{aligned}$$

$$\min_{\alpha \in \mathbb{R}^n} J(\alpha) = -\mathbf{y}^\top K\alpha + \mathbb{1}^\top \log \left(\exp^{K\alpha} + \exp^{-K\alpha} \right) + \frac{1}{2\sigma^2} \alpha^\top K\alpha$$

the optimization problem in \mathbb{R}^n is tractable

Kernel logistic regression¹

How to make a decision for a given x

Bayes classifier takes the sign of the odd ratio

$$\log \frac{\mathbf{P}(Y = 1|x)}{\mathbf{P}(Y = -1|x)} = 2 \sum_{i=1}^n \alpha_i k(x_i, x)$$

use a threshold if necessary

iterative parameter optimization: conjugate gradient or Newton method

$$\nabla_{\alpha} J(\alpha) = -K\mathbf{y} + K \tanh(K\alpha) + \frac{1}{\sigma^2} K\alpha$$

$$\nabla_{\alpha}^2 J(\alpha) = K \left(\left(1 + \frac{1}{\sigma^2} \right) I - \tanh(K\alpha) \tanh(K\alpha)^{\top} \right)$$

summary of the 2 class example

- exponential family in a RKHS
- gaussian prior on θ
- optimize a criterion: MAP
- representer theorem: dual representation in \mathbb{R}^n

Generalize

- adapt to different problems (multiclass, density estimation, regression...)
- change the prior (conjugate, informational...)
- change the cost (MDL, penalized likelihood, robust one...)

summary of the 2 class example

- exponential family in a RKHS
- gaussian prior on θ
- optimize a criterion: MAP
- representer theorem: dual representation in \mathbb{R}^n

Generalize

- adapt to different problems (multiclass, density estimation, regression...)
- change the prior (conjugate, informational...)
- change the cost (MDL, penalized likelihood, robust one...)

the case of the SVM

SVM and the exponential family

$$MAP \quad - \sum_{i=1}^n \log \mathbf{P}(y_i | x_i, \theta) + \frac{1}{2\sigma^2} \|\theta\|_{\mathcal{H}}^2$$

for SVM use a robust pseudo-likelihood ratio

- introduce a margin ρ
- focus on the “worst” cases

$$\min_{\theta \in \Theta} \quad - \sum_{i=1}^n \max \left(\rho - \log \frac{\mathbf{P}(y_i | x_i, \theta)}{1 - \mathbf{P}(y_i | x_i, \theta)}, 0 \right) + \frac{1}{2} \|\theta\|_{\mathcal{H}}^2$$

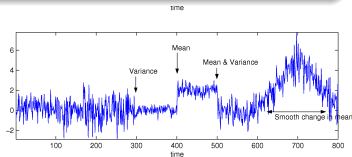
$\mathbf{P}(y_i | x_i, \theta)$ use conditional Rademacher

$$\min_{\theta \in \Theta} \quad - \sum_{i=1}^n \max (\rho - 2y_i \theta(x_i), 0) + \frac{1}{2} \|\theta\|_{\mathcal{H}}^2$$

The problem of novelty detection in a sequence²

the problem

detect a change in a sequence



$$\underbrace{X_1, X_2, \dots, X_{t-1}, X_t}_{P_0}, \underbrace{X_{t+1}, \dots, X_{2t}}_{P_1}$$

Model

$$\begin{cases} \mathcal{H}_0 & : \mathbf{P}_0 = \mathbf{P}_1 \\ \mathcal{H}_1 & : \mathbf{P}_0 \neq \mathbf{P}_1 \end{cases}$$

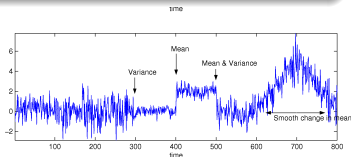
The problem of novelty detection in a sequence²

a good tool: **the likelihood ratio**

$$\Lambda_l(x_1, \dots, x_{2t}) = \frac{\prod_{i=1}^t \mathbf{P}_0(x_i) \prod_{i=t+1}^{2t} \mathbf{P}_1(x_i)}{\prod_{i=1}^{2t} \mathbf{P}_0(x_i)} = \prod_{i=t+1}^{2t} \frac{\mathbf{P}_1(x_i)}{\mathbf{P}_0(x_i)}$$

the problem

detect a change in a sequence



$$\underbrace{x_1, x_2, \dots, x_{t-1}, x_t}_{\mathbf{P}_0}, \underbrace{x_{t+1}, \dots, x_{2t}}_{\mathbf{P}_1}$$

Model

$$\begin{cases} \mathcal{H}_0 & : \mathbf{P}_0 = \mathbf{P}_1 \\ \mathcal{H}_1 & : \mathbf{P}_0 \neq \mathbf{P}_1 \end{cases}$$

Generalized likelihood ratio

$$\Lambda(x_1, \dots, x_{2t}) = \prod_{i=t+1}^{2t} \frac{\widehat{\mathbf{P}}_1(x_i)}{\widehat{\mathbf{P}}_0(x_i)}$$

$\widehat{\mathbf{P}}_0$ and $\widehat{\mathbf{P}}_1$ are the maximum likelihood estimates of the densities

the exponential family at the rescue

$$\mathbf{P}_0(x) = \mu(x) \exp\langle \theta_0(\cdot), k(x, \cdot) \rangle_{\mathcal{H}} - g(\theta_0)$$

$$\mathbf{P}_1(x) = \mu(x) \exp\langle \theta_1(\cdot), k(x, \cdot) \rangle_{\mathcal{H}} - g(\theta_1)$$

one class SVM

$$\widehat{\mathbf{P}}_0(x) = \mu(x) \exp\left(\sum_{i=1}^t \alpha_i^{(0)} k(x, x_i) - g(\theta_0)\right)$$

$$\widehat{\mathbf{P}}_1(x) = \mu(x) \exp\left(\sum_{i=t+1}^{2t} \alpha_i^{(1)} k(x, x_i) - g(\theta_1)\right)$$

$\alpha_i^{(0)}$ using the 1st half (x_1 to x_t) and $\alpha_i^{(1)}$ 2nd half (x_{t+1} to x_{2t})

the acceptance region of the test

$$\Lambda(x_1, \dots, x_{2t}) > s \Leftrightarrow \prod_{j=t+1}^{2t} \frac{\exp \sum_{i=t+1}^{2t} \alpha_i^{(1)} k(x_j, x_i) - g(\theta_1)}{\exp \sum_{i=1}^t \alpha_i^{(0)} k(x_j, x_i) - g(\theta_0)} > s$$

$$\Leftrightarrow \sum_{j=t+1}^{2t} \left(\sum_{i=1}^t \alpha_i^{(0)} k(x_j, x_i) - \sum_{i=t+1}^{2t} \alpha_i^{(1)} k(x_j, x_i) \right) < s + t(b_1 - b_0)$$

s is fixed for a given risk of the first kind

simplification: $(\sum_{i=t+1}^{2t} \alpha_i^{(1)} k(x_j, x_i))$ almost constant

$$\sum_{j=t+1}^{2t} \left(\sum_{i=1}^t \alpha_i^{(0)} k(x_j, x_i) \right) < s' + tg(\theta_0)$$

the acceptance region of the test

$$\Lambda(x_1, \dots, x_{2t}) > s \Leftrightarrow \prod_{j=t+1}^{2t} \frac{\exp \sum_{i=t+1}^{2t} \alpha_i^{(1)} k(x_j, x_i) - g(\theta_1)}{\exp \sum_{i=1}^t \alpha_i^{(0)} k(x_j, x_i) - g(\theta_0)} > s$$

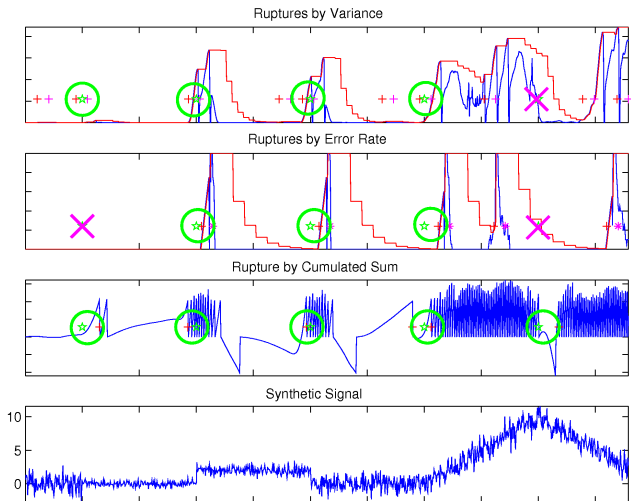
$$\Leftrightarrow \sum_{j=t+1}^{2t} \left(\sum_{i=1}^t \alpha_i^{(0)} k(x_j, x_i) - \sum_{i=t+1}^{2t} \alpha_i^{(1)} k(x_j, x_i) \right) < s + t(b_1 - b_0)$$

s is fixed for a given risk of the first kind

simplification: $(\sum_{i=t+1}^{2t} \alpha_i^{(1)} k(x_j, x_i))$ almost constant

$$\sum_{j=t+1}^{2t} \left(\sum_{i=1}^t \alpha_i^{(0)} k(x_j, x_i) \right) < s' + tg(\theta_0)$$

evaluate the future based on the past - **it is very fast!**



Concluding remarks

Summary

- exponential family is the bridge expected between probabilities and kernels based methods
- parameters change from functions to vectors
- applicable to many algorithms
- application to graphical models (CRF)

Open Problems

- more to come: e.g. temporal statistical tests (cusum)
- develop practical consequences (new methods)
- clarify some functional aspects of the exponential family

Questions?

Questions?

`http://users.rsize.anu.edu.au/~smola`

`http://asi.insa-rouen.fr/~scanu`