

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

Abstract. The success of Support Vector Machine (SVM) gave rise to the development of a new class of theoretically elegant learning machines which use a central concept of kernels and the associated reproducing kernel Hilbert space (r.k.h.s.). Exponential families, a standard tool in statistics, can be used to unify many existing machine learning algorithms based on kernels (such as SVM) and to invent novel ones quite effortlessly. In this paper we will discuss how exponential families, a standard tool in statistics, can be used with great success in machine learning to unify many existing algorithms and to invent novel ones quite effortlessly. A new derivation of the novelty detection algorithm based on the one class SVM is proposed to illustrate the power of the exponential family model in a r.k.h.s.

1 Introduction

Machine learning is proving increasingly important tools in many fields such as text processing, machine vision, speech to name just a few. Among these new tools, kernel based algorithms have demonstrated their efficiency on many practical problems. These algorithms performed function estimation, and the functional framework behind these algorithm is now well known [1]. But still too little is known about the relation between these learning algorithms and more classical statistical tools such as likelihood, likelihood ratio, estimation and test theory. A key model to understand this relation is the generalized or non parametric exponential family. This exponential family is a generic way to represent any probability distribution since any distribution can be well approximated by an exponential distribution. The idea here is to retrieve learning algorithm by using the exponential family model with classical statistical principle such as the maximum penalized likelihood estimator or the generalized likelihood ratio test. To do so the paper (following [2]) is organized as follows. First section presents the functional frameworks and reproducing kernel Hilbert space. Then the exponential family on a r.k.h.s. is introduced and classification as well as density

*National ICT Australia is funded through the Australian Government's *Backing Australia's Ability* initiative, in part through the Australian Research Council. This work was supported by grants of the ARC and by the IST Programme of the European Community, under the Pascal Network of Excellence, IST-2002-506778.

estimation and regression kernels based algorithms such as SVM are derived. In a final section new material is presented establishing the link between the kernel based one class SVM novelty detection algorithm and classical test theory. It is shown how this novelty detection can be seen as an approximation of a generalized likelihood ratio thus optimal test.

2 Functional framework

Definition 1 (reproducing kernel Hilbert space (r.k.h.s.)) *A Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ is a r.k.h.s. if it is defined on $\mathbb{R}^{\mathcal{X}}$ (pointwise defined functions) and if the evaluation functional is continuous on \mathcal{H} .*

For instance \mathbb{R}^n , the set \mathcal{P}_k of polynomials of order k , as any finite dimensional set of genuine functions are r.k.h.s.. The set of sequences ℓ^2 is also a r.k.h.s.. Usual L^2 (with Lebesgue measure) is not because it is not a set of pointwise functions.

Definition 2 (positive kernel) *A function from $\mathcal{X} \times \mathcal{X}$ to \mathbb{R} is a positive kernel if it is symmetric and if for any finite subset $\{x_i\}, i = 1, n$ of \mathcal{X} and any sequence of scalar $\{\alpha_i\}, i = 1, n$*

$$\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j K(x_i, y_j) \geq 0$$

This definition is equivalent to Aronszajn definition of positive kernel.

Proposition 1 (bijection between r.k.h.s. and Kernel) *Corollary of proposition 23 in [3] and theorem 1.1.1 in [4]. There is a bijection between the set of all possible r.k.h.s. and the set of all positive kernels.*

Thus Mercer kernels are a particular case of a more general situation since every Mercer kernel is positive in the Aronszajn sense (definition 2) while the converse is false. One of the key property to be used here after is the reproducing ability in the r.k.h.s.. It is closely related with the fact that in r.k.h.s. functions are pointwise defined and the evaluation functional is continuous. Thus, because of this continuity Riesz theorem can be stated as follows

$$\forall f \in \mathcal{H}, \quad x \in \mathcal{X}, \quad f(x) = \langle f(\cdot), k(x, \cdot) \rangle_{\mathcal{H}} \quad (1)$$

In the remaining of the paper the reproducing kernel Hilbert space, its dot product and its kernel k will be assumed to be given. In this case the so called feature space is given by the kernel and the dot product considered is the one of the r.k.h.s..

3 Kernel approaches for the exponential family

3.1 Exponential Families

We begin by reviewing some basic facts of exponential families. Assume there exists a reproducing kernel Hilbert space \mathcal{H} embedded with the dot product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and with a reproducing kernel k such that kernel $k(x, \cdot)$ is a sufficient statistics of x . Then in exponential families the density $\mathbb{P}(x; \theta)$ is given by

$$\mathbb{P}(x; \theta) = \exp(\langle \theta(\cdot), k(x, \cdot) \rangle_{\mathcal{H}} - g(\theta)), \quad g(\theta) = \log \int_{\mathcal{X}} \exp(\langle \theta(\cdot), k(x, \cdot) \rangle_{\mathcal{H}}) dx.$$

Here θ is the natural parameter and $g(\theta)$ is the log-partition function, often also called the moment generating function. When we are concerned with estimating conditional probabilities, the exponential families framework can be extended to conditional densities:

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

and $g(\theta|x) := \log \int_{\mathcal{Y}} \exp(\langle \theta(\cdot), k(x, y, \cdot) \rangle_{\mathcal{H}}) dy.$

$g(\theta|x)$ is commonly referred to as the conditional log-partition function. Both $g(\theta)$ and $g(\theta|x)$ are convex C^∞ functions in θ and they can be used to compute moments of a distribution:

$$\partial_\theta g(\theta) = \mathbf{E}_{p(x;\theta)}[k(x)] \quad \partial_\theta g(\theta|x) = \mathbf{E}_{p(x,y;\theta)}[k(x,y)|x] \quad \text{Mean} \quad (2)$$

$$\partial_\theta^2 g(\theta) = \text{Var}_{p(x;\theta)}[k(x)] \quad \partial_\theta^2 g(\theta|x) = \text{Var}_{p(x,y;\theta)}[k(x,y)|x] \quad \text{Variance} \quad (3)$$

We will also assume there exists some prior on parameter θ defined by

$$\mathbb{P}(\theta) = \frac{1}{Z} \exp(\langle \theta(\cdot), \theta(\cdot) \rangle_{\mathcal{H}} / 2\sigma^2)$$

where Z is a normalizing constant. In this case, the posterior density can be written as $\mathbb{P}(\theta|x) \propto \mathbb{P}(x|\theta)\mathbb{P}(\theta)$.

3.2 Kernel logistic regression and Gaussian process

Assume we observe some training data $x_i, y_i, i = 1, n$. The binary classification problem is when $y_i \in \{-1, +1\}$. In this case we can use the conditional exponential family to model $\mathbb{P}(Y = y|x)$. The estimation of its parameter θ using the maximum a posteriori (MAP) principle aims at minimizing the following cost function:

$$-\log \mathbb{P}(\theta|\text{data}) = - \sum_{i=1}^n \langle \theta(\cdot), k(x_i, y_i, \cdot) \rangle_{\mathcal{H}} + g(\theta, x_i) + \langle \theta(\cdot), \theta(\cdot) \rangle_{\mathcal{H}} / 2\sigma^2 + C$$

where C is some constant term. Note that this can be seen also as a penalized likelihood cost function and thus connected to maximum description length principle.

Since $y \in \{-1, +1\}$ the kernel can be simplified and written as $k(x_i, y_i, x, y) = k(x_i, x)y_i y$. based on that and using the reproducing property equation 1 ($\theta(x_i) = \langle \theta(\cdot), k(x_i, \cdot) \rangle_{\mathcal{H}}$) we have:

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

Then after some algebra the MAP estimator can be found by minimizing:

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

On this minimization problem, the representer theorem (see [5] for more details) gives us:

$$\theta(\cdot) = \sum_{i=1}^n y_i \alpha_i k(x_i, \cdot)$$

The associated optimization problem can be rewritten in terms of α :

$$\min_{\alpha \in \mathbf{R}^n} \sum_{i=1}^n \log \left(1 + \exp \left(2 \sum_{j=1}^n y_j \alpha_j k(x_i, x_j) \right) \right) + \frac{1}{2\sigma^2} \sum_{i=1}^n \sum_{j=1}^n y_j y_i \alpha_i \alpha_j k(x_i, x_j)$$

It is non linear and can be solved using Newton method. The connection is made with the kernel logistic regression since we have in our framework:

$$\log \frac{\mathbb{P}(Y = 1|x)}{\mathbb{P}(Y = -1|x)} = \sum_{i=1}^n y_i \alpha_{i,y} k(x_i, x) + b$$

and thus the decision of classifying a new data x only depends on the sign of the kernel term. Note that the multiclass problem can be solve by using the same kind of derivations assuming that $k(x_i, y_i, x, y) = k(x_i, x)\delta_{y_i y}$.

3.3 2 class Support vector machines

Instead of the MAP estimate take the robust minimization criteria:

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

Together with the exponential family model, the minimization of this criterion leads to the maximum margin classifier. Here again this can be easily generalized to the multiclass problem.

3.4 1 class Support vector machines

The one class SVM algorithm has been design to estimate some quantile from sample data. This is closely reated but simpler than estimating the whole density. It is also more relevant when the target application is novelty detection.

As a matter of fact, any point not belonging to the support of a density can be seen a novel.

Back with our exponential family model for $\mathbb{P}(x)$, a robust approximation of maximum a posteriori (MAP) estimator for θ is the one maximizing:

$$\max_{\theta \in \mathcal{H}} \prod_{i=1}^n \min \left(\frac{\mathbb{P}_0(x_i|\theta)}{p_0}, 1 \right) \mathbb{P}(\theta)$$

with $p_0 = \exp(\rho - g(\theta))$. After some tedious algebra, this problem can be rewritten as:

$$\min_{\alpha \in \mathbb{R}^n} \sum_{i=1}^n \max(\rho - \langle \theta(\cdot), k(x_i, \cdot) \rangle_{\mathcal{H}}, 0) + \frac{1}{2\sigma^2} \|\theta\|_{\mathcal{H}}^2 \quad (4)$$

On this problem again the representer theorem gives us the existence of some coefficient α_i such that:

$$\theta(\cdot) = \sum_{i=1}^n \alpha_i k(x_i, \cdot)$$

and thus the estimator has the following form:

$$\hat{\mathbb{P}}(x) = \exp \left(\sum_{i=1}^n \alpha_i k(x_i, \cdot) - b \right)$$

where coefficients α are determined by solving the one class SVM problem (4). Parameter b represents the value of the log partition function and thus the normalization factor. It can be hard to compute it but it is possible to do without it in our applications.

Here again the one class SVM algorithm can be derived using the exponential family on a r.k.h.s. and a relevant cost function to be minimized.

3.5 Regression

It is possible to see the problem as a generalization of the classification case to continuous y . But in this case, a generalized version of the representer theorem shows that parameters α are no longer scalar but functions, leading to intractable optimization problems. Some additional hypothesis have to be made about the nature of the unknown distribution. One way to do is to use the conditional gaussian representation with its natural parameters:

$$\mathbb{P}(y|x) = \exp(y \theta_1(x) + y^2 \theta_2(x) - g(\theta_1(x), \theta_2(x)))$$

with $\theta_1(x) = m(x)/\sigma^2(x)$ and $\theta_2(x) = -1/2\sigma^2(x)$ where $m(x)$ is the conditional expectation of y given x and $\sigma^2(x)$ its conditional variance. The associated kernel can be written as follows:

$$k(x_i, y_i, x, y) = k_1(x_i, x)y + k_2(x_i, x)y^2$$

where k_1 and k_2 are two positive kernels. In this case the application of the represented theorem gives a heteroscedastic gaussian process (with non constant variance) as the model of the data, associated with a convex optimization problem.

4 Application to novelty detection

Let $X_i, i = 1, 2t$ be a sequence of random variables distributed according to some distribution \mathbb{P}_i . We are interested in finding whether or not a change has occurred at time t . To begin with a simple framework we will assume the sequence to be stationary from 1 to t and from $t+1$ to $2t$, *i.e.* there exists some distributions \mathbb{P}_0 and \mathbb{P}_1 such that $P_i = P_0, i \in [1, t]$ and $P_i = P_1, i \in [t+1, 2t]$. The question we are addressing can be seen as determining if $\mathbb{P}_0 = \mathbb{P}_1$ (no change has occurred) or else $\mathbb{P}_0 \neq \mathbb{P}_1$ (some change have occurred). This can be restated as the following statistical test:

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

In this case the likelihood ratio is the following:

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

since both densities are unknown the generalized likelihood ratio (GLR) has to be used:

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

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

Because we want our detection method to be universal, we want it to work for any possible density. Thus some approximations have to be done to clarify our framework. First we will assume both densities \mathbb{P}_0 and \mathbb{P}_1 belong to the generalized exponential family thus there exists a reproducing kernel Hilbert space \mathcal{H} embedded with the dot product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and with a reproducing kernel k such that:

$$\mathbb{P}_0(x) = \mu(x) \exp\langle \theta_0(\cdot), k(x, \cdot) \rangle_{\mathcal{H}} - g(\theta_0) \quad \text{and} \quad \mathbb{P}_1(x) = \mu(x) \exp\langle \theta_1(\cdot), k(x, \cdot) \rangle_{\mathcal{H}} - g(\theta_1)$$

where $g(\theta)$ is the so called log-partition function. Second hypothesis, the functional parameter θ_0 and θ_1 of these densities will be estimated on the data of respectively first and second half of the sample by using the one class SVM algorithm. By doing so we are following our initial assumption that before time t we know the distribution is constant and equal to some \mathbb{P}_0 . The one class SVM algorithm provides us with a good estimator of this density. The situation of $\widehat{\mathbb{P}}_1(x)$ is more simple. It is clearly a robust approximation of the maximum

likelihood estimator. Using one class SVM algorithm and the exponential family model both estimate can be written as:

$$\widehat{\mathbb{P}}_0(x) = \mu(x) \exp \left(\sum_{i=1}^t \alpha_i^{(0)} k(x, x_i) - g(\theta_0) \right), \widehat{\mathbb{P}}_1(x) = \mu(x) \exp \left(\sum_{i=t+1}^{2t} \alpha_i^{(1)} k(x, x_i) - g(\theta_1) \right)$$

where $\alpha_i^{(0)}$ is determined by solving the one class SVM problem on the first half of the data (x_1 to x_t). while $\alpha_i^{(1)}$ is given by solving the one class SVM problem on the second half of the data (x_{t+1} to x_{2t}). Using these three hypothesis, the generalized likelihood ratio test is approximated as follows:

$$\begin{aligned} \Lambda(x_1, \dots, x_{2t}) > s &\Leftrightarrow \prod_{j=t+1}^{2t} \frac{\exp \sum_{i=t+1}^{2t} \alpha_i^{(1)} k(x_j, x_i) - g(\theta_1)}{\exp \sum_{i=1}^t \alpha_i^{(0)} k(x_j, x_i) - g(\theta_0)} > s \\ &\Leftrightarrow \sum_{j=t+1}^{2t} \left(\sum_{i=1}^t \alpha_i^{(0)} k(x_j, x_i) - \sum_{i=t+1}^{2t} \alpha_i^{(1)} k(x_j, x_i) \right) < s' \end{aligned}$$

where s' is a threshold to be fixed to have a given risk of the first kind a such that:

$$\mathbb{P}_0 \left(\sum_{j=t+1}^{2t} \left(\sum_{i=1}^t \alpha_i^{(0)} k(x_j, x_i) - \sum_{i=t+1}^{2t} \alpha_i^{(1)} k(x_j, x_i) \right) < s' \right) = a$$

It turns out what variation of $\sum_{i=t+1}^{2t} \alpha_i^{(1)} k(x_j, x_i)$ are very small in comparison to the one of $\sum_{i=1}^t \alpha_i^{(0)} k(x_j, x_i)$. Thus $\widehat{\mathbb{P}}_1(x)$ can be assumed to be constant, simplifying computations. In this case the test can be performed by only considering:

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

This is exactly the novelty detection algorithm as proposed in [6]. Thus we show how to derive it as a statistical test approximating a generalized likelihood ratio test, optimal under some condition in the Neyman Pearson framework.

5 Conclusion

In this paper we illustrates how powerful is the link made between kernel algorithms, reproducing kernel Hilbert space and the exponential family. A lot of learning algorithms can be revisited using this framework. We discuss here the logistic kernel regression, the SVM, the gaussian process for regression and the novelty detection using the one class SVM. This framework is applicable to many different cases and other derivations are possible: exponential family in a r.k.h.s. can be used to recover sequence annotation (via Conditional Random Fields) or boosting to name just a few. The exponential family framework is

powerful because it allows to connect, with almost no loss of generality, learning algorithm with usual statistical tools such as posterior densities and likelihood ratio. These links between statistics and learning were detailed in the case of novelty detection restated as a quasi optimal statical test based on a robust approximation of the generalized likelihood. Further works on this field regard the application of sequential analysis tools such as the CUSUM algorithm for real time novelty detection minimizing the expectation of the detection delay.

References

- [1] S. Canu, X. Mary, and A. Rakotomamonjy. *Advances in Learning Theory: Methods, Models and Applications NATO Science Series III: Computer and Systems Sciences*, chapter Functional learning through kernel. IOS Press, Amsterdam, 2003.
- [2] A Smola. Exponential families and kernels. Berder summer school, 2004. <http://users.rsise.anu.edu.au/~smola/teaching/summer2004/>.
- [3] Laurent Schwartz. Sous espaces hilbertiens d'espaces vectoriels topologiques et noyaux associés. *Journal d'Analyse Mathématique*, pages 115–256, 1964.
- [4] Grace Wahba. *Spline Models for Observational Data*. Series in Applied Mathematics, Vol. 59, SIAM, Philadelphia, 1990.
- [5] B. Scholkopf and A. Smola. *Learning with Kernels*. MIT Press, 2001.
- [6] B Scholkopf, R Williamson, A Smola, and J Shawe-Taylor. Support vector method for novelty detection. In SA Solla, TK Leen, and KR Muller, editors, *NIPS*, pages 582–588. MIT Press, 2000.