Category: Theory (Support Vector Machines)
Presentation Preference: oral
Correspondence: N. Oliver

Natural Regularization in SVMs

Nuria Oliver[*,†]
nuria@media.mit.edu

Bernhard Schölkopf[*]
bs@first.gmd.de

Alex Smola[*]
smola@first.gmd.de

[*]**GMD FIRST.IDA**
Rudower Chaussee 5
12489 Berlin

[†]**Media Arts and Sciences Laboratory**
MIT, 20 Ames Street, E15-384C
Cambridge, MA 02139

Abstract

We provide a regularization-theoretic analysis of a class of SV kernels based on generative models with density $p(x|\theta)$ such as the Fisher kernel proposed by Jaakkola and Haussler. In particular, we show that the latter corresponds to a regularization operator (prior) penalizing the $L_2(p)$-norm of the estimated function. Moreover, we compute the corresponding eigendecompositions and show that the Fisher kernel isotropically maps the input space into feature space.

1 INTRODUCTION

Two significant developments have shaped learning theory over the last decade. First, generative probability models (such as HMMs, dynamic graphical models or mixtures of experts) have provided a mathematically principled framework for dealing with missing and incomplete data, uncertainty or variable length sequences, for example. Second, Support Vector (SV) and kernel methods have bridged the gap between abstract VC-theory and applied machine learning technology, leading to benchmark record results in a variety of domains. However, until recently, these two strands have been largely separated.

A promising approach to combine the strengths of both worlds has been made in the work of Jaakkola and Haussler [3]. They propose the use of a so called Fisher kernel to give a 'natural' similarity measure taking into account an underlying generative model. Defining a kernel function automatically implies assumptions about metric relations between the examples. In their paper they argue that these metric relations should be defined directly from a generative probability model $p(x|\theta)$.

Besides experimental proof they justify their choice from two perspectives: that of enhancing the discriminative power of the model and from an attempt to find a 'natural' comparison between examples induced by the generative model. What was missing so far, however, is a regularization theoretic analysis of the resulting kernel, namely to which regularization operator (and thus also which prior) [6] such a kernel corresponds to.

In the present paper we show that the Fisher kernel corresponds to a prior probability over the functions $f(\theta)$, $p(f) = \exp(-\|f\|_p^2)$, where $\|\cdot\|_p^2$ is the norm of the $L_2(p)$

space of functions square integrable with respect to the measure corresponding to $p(x|\theta)$. An analogous analysis is given for a large class of other priors. Furthermore, we derive its eigenvalue decomposition, with the goal to describe the image of input space in feature space, whose shape has consequences for the generalization behavior of the associated kernel method [11]. An analysis of some of the ambiguities that may arise from kernels based on generative models and a discussion of the results obtained concludes the paper.

2 KERNEL METHODS AND NATURAL KERNELS

Kernel Methods Many learning algorithms can be formulated in terms of dot products between input patterns $x_i \cdot x_j$ in $\mathbb{R}^N$, such as separating hyperplanes for pattern recognition. Using a suitable kernel k instead of a dot product in $\mathbb{R}^N$ corresponds to mapping the data into a possibly high-dimensional dot product space F by a (usually nonlinear) feature map $\Phi : \mathbb{R}^N \to F$, and taking the dot product there (cf. e.g. [1])

$$k(x, x') = (\Phi(x) \cdot \Phi(x')). \tag{1}$$

Any linear algorithm which can be carried out in terms of dot products can be made nonlinear by substituting the input patterns with features computed by an a priori chosen kernel. Examples of such algorithms include SV Machines and kernel PCA [1, 5]. The solutions of kernel algorithms are kernel expansions

$$f(x) = \sum_i \alpha_i k(x, x_i). \tag{2}$$

Since all the computations are done in terms of dot products, all information used for training, based on patterns $X := \{x_1, \ldots, x_\ell\} \subset \mathcal{X}$ (where for instance $\mathcal{X} \in \mathbb{R}^N$) resides in the Gram matrix

$$K_{ij} := k(x_i, x_j), \tag{3}$$

and in target values which might be provided additionally.

2.1 Natural Kernels

A class of kernels can be derived from generative models by employing techniques from information theory:

A measure class $\mathcal{M}$ is an affine space under its log-likelihood structure, and taking covariances of translations produces a natural Riemannian metric of sorts. The vector space of translations of $\mathcal{M}$ is a space of random variables R_Ω.

A statistical manifold P is a subset of $\wp$, the space of probability measures in a measure class $\mathcal{M}$ on a sample space Ω, which is also a manifold and satisfies two conditions:

1. Choosing any origin μ for $\mathcal{M}$ the log-likelihood function $l : P \to R_\Omega, p = e^f \mu \to f$ is a smooth function.
2. For any point p and any coordinates θ about p the random variables $\frac{\partial l}{\partial \theta^i}(p)$ are linearly independent.

The derivative map is called usually the *score map* $U_\theta : \mathcal{X} \to \mathbb{R}^r$ with

$$U_\theta(x) := (\partial_{\theta^1} l(x, \theta), \ldots, \partial_{\theta^r} l(x, \theta)) = (\partial_{\theta^1} \log p(x|\theta), \ldots, \partial_{\theta^r} \log p(x|\theta)) \tag{4}$$

whose coordinates are usually called a 'natural' basis of tangent vectors. This basis represents the direction in which the value of the ith coordinate increases while the others are fixed. Since the manifold is Riemannian, there is an inner product defined in its tangent space T_p whose metric tensor is given by the *Fisher information matrix* I.

$$I(p) := E_p\left[U_\theta(x)U_\theta(x)^\top\right] \text{ i.e. } I_{ij}(p) = E_p\left[\partial_{\theta^i}\log p(x|\theta)\partial_{\theta^j}\log p(x|\theta)\right] \qquad (5)$$

This metric is called the *Fisher information metric* and induces a 'natural' distance in the manifold. Thus this is the metric to use when measuring the difference in the generative process between a pair of examples x_i and x_j.

Assume now that we have a generative probability density function $p(x|\theta)$ parametrized by θ. This sort of generative probability models is interesting because it provides a principled way of treating missing information by estimating the probability of the data X given the parameters θ, i.e. computing $p(x|\theta)$.

The previous definitions enable us to define the class of *natural kernels* as

$$k_M^{\text{nat}}(x, x') := U_\theta(x)^\top M^{-1} U_\theta(x') = \nabla_\theta \log p(x|\theta)^\top M^{-1} \nabla_\theta \log p(x'|\theta), \qquad (6)$$

where M is a positive definite matrix. In [3] Haussler and Jaakkola set $M = I$ to obtain the so called 'Fisher kernel'. In some of the experiments, however, they simply use the identity matrix rather than I in the kernel. In the next section we will give a complete regularization theoretic analysis of the class of natural kernels, in particular also of k_I^{nat} and k_1^{nat}.

3 THE NATURAL REGULARIZATION OPERATOR

In regularization networks and SV machines, the weighted sum of empirical risk functional $R_{emp}[f]$ and regularization term $||Pf||^2$ is minimized (for some $\lambda > 0$):

$$R_{reg}[f] = R_{emp} + \lambda Q[f] = R_{emp} + \lambda ||Pf||^2 \qquad (7)$$

For instance, by choosing a suitable operator that penalizes large variations of f one can reduce overfitting. Another typical setting is to use an operator P mapping from $L_2(R^d)$ into some Reproducing Kernel Hilbert Space [9].

A Bayesian regularization risk is given by:

$$R_{reg}[f] = -\sum_i \log p(\mathbf{x}_i|\theta) + \lambda ||Pf||^2 = -\sum_i \log p(\mathbf{x}_i|\theta) - \lambda log(p(f(\theta))) \qquad (8)$$

One can show [6, 10, 2] that for f as defined in (2)

$$-\log p(f) \propto ||Pf||^2 = \boldsymbol{\alpha}^\top K \boldsymbol{\alpha}. \qquad (9)$$

Here P denotes a regularization operator in the sense of [7], i.e. P is a positive semidefinite operator mapping from the Hilbert Space $\mathcal{H}$ of functions f under consideration to a dot product space D. The second equality in (9) holds if k is a kernel of a corresponding Reproducing Kernel Hilbert Space RKHS [9], i.e. if

$$k(x, x') = \langle P_z k(x, z), P_z k(x', z)\rangle. \qquad (10)$$

This is in particular the case if k is a Greens function of P^*P [6]. In fact, in SV machines, the weights $\boldsymbol{\alpha}$ are essentially computed by finding a constrained maximum of a quadratic form (9) (additional linear terms depend on the specific setting). Therefore to analyze the properties of natural kernels, we exploit this connection between kernels and regularization operators by finding the operator P_M^{nat} such that (10) holds. To do so, we finally need to define a proper dot product. In our case we use the standard dot product in $L_2(p)$ space, i.e. $\langle f, g\rangle := \int f(x)g(x)p(x|\theta)dx$ (any other measure would also have been possible, however the operator P would have taken on a different form). Substituting (6) into (10) yields:

$$\begin{aligned} k(x,x') &= \nabla_\theta \log p(x|\theta)^\top M^{-1} \nabla_\theta \log p(x'|\theta) && (11) \\ &= \left\langle P_z k_M^{\mathrm{nat}}(x,z), P_z k_M^{\mathrm{nat}}(x',z) \right\rangle \\ &= \int \nabla_\theta \log p(x|\theta)^\top M^{-1} \left[P_z \nabla_\theta \log p(z|\theta) \right] \cdot \\ &\qquad \left[P_z \nabla_\theta \log p(z|\theta)^\top M^{-1} \right] \nabla_\theta \log p(x'|\theta) p(z|\theta) dz && (12) \end{aligned}$$

To ensure that the rhs (11) equals (12) we collect all terms in the integral independent of z outside and obtain the necessary condition

$$M = \int \left[P_z \nabla_\theta \log p(z|\theta) \right] \left[P_z \nabla_\theta \log p(z|\theta) \right]^\top p(z|\theta) dz. \quad (13)$$

Let us consider the two special cases pointed out in [3].

The Fisher Kernel ($M = I$): If M is the Fisher information, then (13) is satisfied when $Pf(x) = f(x)$. Therefore the regularization term in equation 7 is

$$\|Pf\|^2 = \|f\|^2_{L_2(p)}, \quad (14)$$

i.e. the $L_2(p)$-*norm* we used as the default norm in our analysis. We have thus proven the following result.

Proposition 1 (Fisher Regularization Operator) *The Fisher Kernel induced by a generative probability model with density p corresponds to a regularization term equal to the $L_2(p)$-norm of the estimated function.*

To understand intuitively what this regularizer does, let us spell it out explicitly. The solution of SV regression using the Fisher kernel has the form $f(x) = \sum_{i=1}^{\ell} \alpha_i k_I^{nat}(x, x_i)$, where the x_i are the SVs, and $\boldsymbol{\alpha}$ is the solution of the SV programming problem. Applied to this function, the regularization term $\|f\|^2_{L_2(p)} = \|\sqrt{p} f\|^2$ reads

$$\int \left(\sum_{i=1}^{\ell} \alpha_i \nabla_\theta \log p(x|\theta) I^{-1} \nabla_\theta \log p(x_i|\theta) \right)^2 p(x|\theta) dx. \quad (15)$$

To understand this term, first recall that what we actually minimize is (8), the sum of (15) and the empirical risk given by the negative log likelihood. Minimizing (8) without regularization would lead to overfitting. How does (15) prevent this from happening?

Let us assume we are given two parameter vectors θ_1 and θ_2 which both lead to the same high likelihood. In that case, the regularizer will pick the parameter vector with the property that *perturbing* it will (on average) lead to a smaller change in the likelihood, for in that case $\nabla_\theta \log p(x|\theta)$ will be smaller. Consequently, the regularizer will favor the solution which is more stable. Figure 3 illustrates the 'flatness' on parameter space of the natural regularizer.

Note, however, that the validity of this intuitive explanation is somewhat limited since some effects can compensate each other as the α_i come with different signs. Finally, we remark that the regularization operator of the conformal transformation $\sqrt{p(x|\theta)}\sqrt{p(x'|\theta)} k_I^{\mathrm{nat}}(x, x')$ of the Fisher kernel is the identity in L_2 space with respect to the uniform measure [**?**].

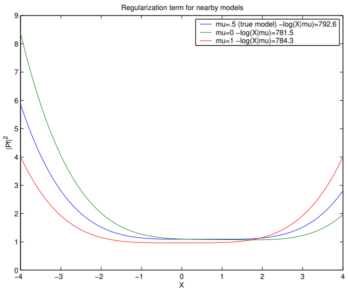

Figure 1: Flatness of the natural regularizer for a Normal generative pdf N(.5, 3)

The Plain Kernel ($M = 1$): Using the identity matrix leaves us with the kernel $k_{\mathbf{1}}^{\text{nat}}$ that [3] have actually used in practice for simplicity. The regularization operator associated with this kernel, for the particular case of generative probability models belonging to the exponential family, is simply the gradient ∇_x. This can be seen by substituting $logp(x|\theta) = \theta \cdot x - K(\theta)$ (as is possible in the case of exponential families of densities) into the condition (13):

$$\begin{aligned} \int \left[\nabla_z \nabla_\theta \log p(z|\theta)\right]^\top \left[\nabla_z \nabla_\theta \log p(z|\theta)\right] p(z|\theta) dz &= \\ \int \left[\nabla_z (z - \nabla_\theta K(\theta))\right]^\top \left[\nabla_z (z - \nabla_\theta K(\theta))\right] p(z|\theta) dz &= \mathbf{1}. \end{aligned} \tag{16}$$

Arbitrary Operators: Often one is facing the opposite problem of identifying a kernel k_M^{nat} from its corresponding regularization operator. This problem can be solved in in a straightforward fashion by equating (13), i.e. computing the integral wrt. $p(x|\theta)$ for the appropriate class of operators. A possible choice would be a data dependent smoothness measure, i.e. the computation of Radon–Nikodym derivatives (e.g. $P = p^{-1}(x|\theta)\partial_\theta$ or powers thereof).

4 THE FEATURE MAP OF NATURAL KERNELS

The feature map in SVMs is usually constructed by utilizing Mercer's theorem, which establishes the conditions that a given symmetric kernel k needs to satisfy in order to correspond to a dot product in some feature space F. We state the version given in [11]. By "almost all" we mean except for sets of measure zero.

Theorem 2 (Mercer) *Let $(\mathcal{X}, \mu)$ be a finite measure space. Suppose $k \in L_\infty(\mathcal{X}^2)$ is a symmetric real-valued kernel such that the integral operator $T_k : L_2(\mathcal{X}) \to L_2(\mathcal{X})$, $(T_k f)(x) := \int_{\mathcal{X}} k(x,y) f(y) d\mu(y)$ is positive, i.e. for all $f \in L_2(\mathcal{X})$, we have $\int_{\mathcal{X}} k(x,y) f(x) f(y)\ d\mu(x) d\mu(y) \geq 0$. Let $\psi_j \in L_2(\mathcal{X})$ be the normalized eigenfunctions of T_k associated with the eigenvalues $\lambda_j > 0$, sorted in nonincreasing order. Then $\psi_j \in L_\infty(\mathcal{X})$ with $\sup_j \|\psi_j\|_{L_\infty} < \infty$, and the kernel admits an expansion $k(x,y) = \sum_{j=1}^{N_F} \lambda_j \psi_j(x) \psi_j(y)$ for almost all (x,y). Either $N_F \in \mathbb{N}$, or $N_F = \infty$; in the latter case, the series converges absolutely and uniformly for almost all (x,y).*

Consequently, for almost all $x, y \in \mathcal{X}$, $k(x,y)$ corresponds to a dot product in $l_2^{N_F}$, i.e. $k(x,y) = (\Phi(x) \cdot \Phi(y))$ with

$$\Phi : \mathcal{X} \to l_2^{N_F},\ x \mapsto (\sqrt{\lambda_j} \psi_j(x))_{j=1,\ldots,N_F}. \tag{17}$$

In order to obtain an eigendecomposition of the kernel in terms of eigenfunctions $\psi_i(x)$ and associated eigenvalues $(\lambda_M^{\mathrm{nat}})_i$, we consider k_M^{nat} as the kernel of an integral operator acting on $L_2(\mathcal{X}, p)$, and solve the eigenvalue equation.

$$\int k_M^{\mathrm{nat}}(x, x')\psi_i(x')p(x'|\theta)dx' \quad = \quad (\lambda_M^{\mathrm{nat}})_i \; \psi_i(x) \tag{18}$$

The following proposition holds:

Proposition 3 (Natural Feature Map) *Let $\lambda_i^M, s_i (i = 1, \ldots, r)$ be the eigenvalues and the eigenvectors of the matrix M defining the metric of the natural kernel k_M^{nat}, i.e. $M\, s_i = \lambda_i^M s_i$ (r is the number of dimensions of θ). Then an orthonormal system of eigenvectors of the natural kernel (with eigenvalues $\lambda_i = 1$) is given by*

$$\psi_i(x) = \frac{1}{\sqrt{\lambda_i^M}} P^* P s_i^\top \nabla_\theta \log(p(x|\theta)) \textit{ for all } 1 \leq i \leq r. \tag{19}$$

Note that the feature map (17) associated with these eigenfunctions maps into a feature space whose dimensionality equals that of the parameter space.

Proof We only have to show that for all $\psi_i(x)$ (18) is satisfied. Substituting (19) into (18) yields

$$\begin{aligned} \int \left[\nabla_\theta \log p(x|\theta)\right]^\top M^{-1} \left[\nabla_\theta \log p(x'|\theta)\right] P^* P s_i^\top \left[\nabla_\theta \log p(x'|\theta)\right] p(x'|\theta) dx' &= \\ \left[\nabla_\theta \log p(x|\theta)\right]^\top M^{-1} \int \left[P \nabla_\theta \log p(x'|\theta)\right] \left[P \nabla_\theta \log p(x'|\theta)\right]^\top s_i p(x'|\theta) dx' &= \\ \left[\nabla_\theta \log p(x|\theta)\right]^\top M^{-1} M s_i &= \psi_i(x) \end{aligned}$$

This transformation was possible since the vector s_i and the operator P comute. Moreover we used the definition of dot products in $L_2(p)$.

To show that this system of eigenfunctions is complete, one easily double-checks that indeed $\sum_{i=1}^r \psi_i(x)\psi_i(y) = k_I^{nat}(x, y)$. ■

The eigenvalues $(\lambda_M^{nat})_i$ of k_I^{nat} are all 1, reflecting the fact that the matrix M whitens the scores, $P\nabla_\theta \log(p(x|\theta))$. In the case of the Fisher Kernel, P is the identity operator and therefore (19) becomes $\psi_i(x) = \frac{1}{\sqrt{\lambda_i^M}} P^* P s_i^\top \nabla_\theta \log(p(x|\theta))$ for all $1 \leq i \leq r$.

Traditionally, people had used the standard VC dimension bounds [8] which state that the capacity of a linear classifier or regression algorithm is essentially given by $R^2 \cdot \Lambda^2$. Here, R is the radius of the smallest sphere containing the data (in feature space), and Λ is the maximal allowed length of the weight vector.

Recently, it has been shown that both the spectrum of an associated integral operator [11] and the spectrum of the Gram matrix (3) [4] can be used to formulate generalization error bounds. This was done by exploiting the fact that since $C := \sup_j \|\psi_j\|_{L_\infty}$ exists, (17) implies that $|\Phi_i(x)| = \sqrt{\lambda_i}|\psi_i(\mathbf{x})| \leq \sqrt{\lambda_i}C$, i.e. the mapped data live in some parallelepiped whose sidelengths are given by the square roots of the eigenvalues. New bounds improved upon these generic VC dimension bounds by taking into account this fact: due to the decay of the eigenvalues, the mapped data are not distributed isotropically. Therefore capturing the shape of the mapped data only by the radius of a sphere should be a rather rough approximation. On the other hand, taking into account the rate of decay of the eigenvalues allows one to formulate kernel-dependent bounds which are much more accurate than the standard VC-bounds.

Now in our case all λ_i are 1, therefore $|\Phi_i(\mathbf{x})| = |\psi_i(\mathbf{x})|$. Hence the upper bound simply states that the mapped data is contained in some box with equal sidelengths (hypercube). Moreover, the $L_2(p)$ normalization of the eigenfunctions ψ_i means that $\int \psi_i(x)^2 p(x)\, dx = 1$. Therefore, the squared averaged size of the feature map's ith coordinate is independent of i, implying that the the mapped data have the same range in all directions. This isotropy of the Fisher kernel suggests that the standard 'isotropic' VC bounds should be fairly precise here.

5 EXPERIMENTS AND DISCUSSION

The unit length of the eigenvalues of the natural kernels can be exploited as a way of performing *model selection*: given a set of $\mathcal{K}$ models $p(x|\theta^k), k = 1 \ldots \mathcal{K}$, select the model $\|$ such that the risk (8) is minimized. Given that $\lambda_i = 1$ for all i only in the case of the true model, one should select the model $\|$ such that the eigenvalues of its Gram Matrix are the closer to one. Figure 5 illustrates this procedure for selecting the optimal sufficient statistics (μ, σ) of a one dimensional Normal pdf $p(x|\theta) = N(\mu, \sigma)$ with 10 data points. In this case we computed the eigendecomposition of the empirical Gram Matrices of a set of different models. The figure contains the error bar plots of the ratio of the eigenvalues (note that in this case the parameter space is a 2 dimensional space). The minimum corresponds to the model to be selected. In our case: $\mu = .5$ and $\sigma =$.