Learning Bounds for Support Vector Machines with Learned Kernels

Nathan Srebro
Toyota Technological Institute at Chicago
nati@uchicago.edu

Shai Ben-David
University of Waterloo School of Computer Science
shai@cs.uwaterloo.ca

Abstract

Consider the problem of learning a kernel for use in SVM classification. We bound the estimation error of a large margin classifier when the kernel, relative to which this margin is defined, is chosen from a family of kernels based on the training sample. For a kernel family with *pseudodimension* d_ϕ, we present a bound of $\sqrt{\tilde{\mathcal{O}}(d_\phi + 1/\gamma^2)/n}$ on the estimation error for SVMs with margin γ. This is the first bound in which the relation between the margin term and the family-of-kernels term is **additive** rather then multiplicative. The pseudodimension of families of linear combinations of base kernels is the number of base kernels. Unlike in previous (multiplicative) bounds, there is no non-negativity requirement on the coefficients of the linear combinations. We also give simple bounds on the pseudodimension for families of Gaussian kernels.

This is an extended abstract of a paper presented at the 19th Annual Conference on Learning Theory (COLT), June 2006.

1 Introduction

In support vector machines (SVMs), as well as other similar methods, prior knowledge is represented through a *kernel function* specifying the inner products between an implicit representation of input points in some Hilbert space. A large margin linear classifier is then sought in this implicit Hilbert space. Using a “good” kernel function, appropriate for the problem, is crucial for successful learning: The kernel function essentially specifies the permitted hypothesis class, or at least which hypotheses are preferred.

In the standard SVM framework, one commits to a fixed kernel function apriori, and then searches for a large margin classifier with respect to this kernel. If it turns out that this fixed kernel in inappropriate for the data, it might be impossible to find a good large margin classifier. Instead, one can search for a data-appropriate kernel function, from some class of allowed kernels, permitting large margin classification. That is, search for both a kernel *and* a large margin classifier with respect to the kernel. In this paper we develop bounds for the sample complexity cost of allowing such kernel adaptation.

2 Learning the Kernel

As in standard hypothesis learning, the process of learning a kernel is guided by some family of potential kernels. A popular type of kernel family consists of kernels that are a linear, or convex, combinations of several base kernels [1, 2, 3][1]:

$$\mathcal{K}_{\text{linear}}(K_1,\ldots,K_k) \overset{\text{def}}{=} \left\{ K_{\vec{\lambda}} = \sum_{i=1}^{k} \lambda_i K_i \mid K_{\vec{\lambda}} \succcurlyeq 0 \text{ and } \sum_{i=1}^{k} \lambda_i = 1 \right\} \tag{1}$$

$$\mathcal{K}_{\text{convex}}(K_1,\ldots,K_k) \overset{\text{def}}{=} \left\{ K_{\vec{\lambda}} = \sum_{i=1}^{k} \lambda_i K_i \mid \lambda_i \geq 0 \text{ and } \sum_{i=1}^{k} \lambda_i = 1 \right\} \tag{2}$$

Such kernel families are useful for integrating several sources of information, each encoded in a different kernel, and are especially popular in bioninformatics applications [4, 5, 6, and others].

Another common approach is learning (or "tuning") parameters of a parameterized kernel, such as the covariance matrix of a Gaussian kernel, based on training data [7, 8, 9, 10, and others]. This amounts to learning a kernel from a parametric family, such as the family of Gaussian kernels:

$$\mathcal{K}^{\ell}_{\text{Gaussian}} \overset{\text{def}}{=} \left\{ K_A : (x_1,x_2) \mapsto e^{-(x_1-x_2)' A (x_1-x_2)} \mid A \in \mathbb{R}^{\ell\times\ell} \ , \ A \succcurlyeq 0 \right\} \tag{3}$$

Infinite-dimensional kernel families have also been considered, either through *hyperkernels* [11] or as convex combinations of a continuum of base kernels (e.g. convex combinations of Gaussian kernels) [12, 13]. In this paper we focus on finite-dimensional kernel families, such as those defined by equations (1)–(3).

Learning the kernel matrix allows for greater flexibility in matching the target function, but this of course comes at the cost of higher estimation error, i.e. a looser bound on the expected error of the learned classifier in terms of its empirical error. Bounding this estimation gap is essential for building theoretical support for kernel learning, and this is the focus of this paper.

3 Learning Bounds with Learned Kernels—Previous Work

For standard SVM learning, with a fixed kernel, one can show that, with high probability, the estimation error (gap between the expected error and empirical error) of a learned classifier with margin γ is bounded by $\sqrt{\tilde{\mathcal{O}}(1/\gamma^2)/n}$ where n is the sample size and the $\tilde{\mathcal{O}}()$ notation hides logarithmic factors in its argument, the sample size and the allowed failure probability. That is, the number of samples needed for learning is $\tilde{\mathcal{O}}\big(1/\gamma^2\big)$.

Lanckriet *et al.* [1] showed that when a kernel is chosen from a convex combination of k base kernels, the estimation error of the learned classifier is bounded by $\sqrt{\tilde{\mathcal{O}}(k/\gamma^2)/n}$ where γ is the margin of the learned classifier under the learned kernel. Note the multiplicative interaction between the margin complexity term $1/\gamma^2$ and the number of base kernels k. More recently, Micchelli *et al.* [14] derived bounds for the family of Gaussian kernels of equation (3). The dependence of these bounds on the margin and the complexity of the kernel family is also multiplicative—the estimation error is bounded by $\sqrt{\tilde{\mathcal{O}}(C_\ell/\gamma^2)/n}$, where C_ℓ is a constant that depends on the input dimensionality ℓ.

The multiplicative interaction between the margin and the complexity measure of the kernel class is disappointing. It suggests that learning even a few kernel parameters (e.g. the coefficients λ) leads to a multiplicative increase in the required sample size. It is important to understand whether such a multiplicative increase in the number of training samples is in fact necessary.

Bousquet and Herrmann [2, Theorem 2] and Lanckriet *et al.* [1] also discuss bounds for families of convex and linear combinations of kernels that appear to be independent of the number of base kernels. However, we show that these bounds are meaningless: The bound on the expected error is never less than one. We are not aware of any previous work describing meaningful explicit bounds for the family of linear combinations of kernels given in equation (1).

[1] Lanckriet *et al.* [1] impose a bound on the trace of the Gram matrix of K_λ—this is equivalent to bounding $\sum \lambda_i$ when the base kernels are normalized.

4 New, Additive, Learning Bounds

We bound the estimation error, when the kernel is chosen from a kernel family $\mathcal{K}$, by $\sqrt{\tilde{\mathcal{O}}(d_\phi + 1/\gamma^2)/n}$, where d_ϕ is the *pseudodimension* of the family $\mathcal{K}$. This establishes that the bound on the required sample size, $\tilde{\mathcal{O}}\big(d_\phi + 1/\gamma^2\big)$ grows only **additively** with the dimensionality of the allowed kernel family (up to logarithmic factors). This is a much more reasonable price to pay for not committing to a single kernel apriori.

The pseudodimension of most kernel families matches our intuitive notion of the dimensionality of the family, and in particular:

- The pseudodimension of a family of linear, or convex, combinations of k base kernels (equations 1,2) is at most k.
- The pseudodimension of the family $\mathcal{K}^\ell_{\text{Gaussian}}$ of Gaussian kernels (equation 3) for inputs $x \in \mathbb{R}^\ell$, is at most $\ell(\ell+1)/2$. If only diagonal covariances are allowed, the pseudodimension is ℓ. If the covariances (and therefore A) are constrained to be of rank at most k, the pseudodimension is at most $k\ell \log_2(22k\ell)$.

5 Plan of Attack

For a fixed kernel, it is well known that, with probability at least $1-\delta$, the estimation error of all margin-γ classifiers is at most $\sqrt{\mathcal{O}(1/\gamma^2 - \log\delta)/n}$ [15]. To obtain a bound that holds for all margin-γ classifiers with respect to *any* kernel K in some *finite* kernel family $\mathcal{K}$, consider a union bound over the $|\mathcal{K}|$ events "the estimation error is large for some margin-γ classifier with respect to K" for each $K \in \mathcal{K}$. Using the above bound with δ scaled by the cardinality $|\mathcal{K}|$, the union bound ensures us that with probability at least $1-\delta$, the estimation error will be bounded by $\sqrt{\mathcal{O}(\log|\mathcal{K}| + 1/\gamma^2 - \log\delta)/n}$ for all margin-γ classifiers with respect to any kernel in the family.

In order to extend this type of result also to infinite-cardinality families, we employ the standard notion of ϵ-nets: Roughly speaking, even though a continuous family $\mathcal{K}$ might be infinite, many kernels in it will be very similar and it will not matter which one we use. Instead of taking a union bound over all kernels in $\mathcal{K}$, we only take a union bound over "essentially different" kernels. We apply standard results in order to show that the number of "essentially different" kernels in a family grows exponentially only with the dimensionality of the family, yielding an additive term (almost) proportional to the dimensionality.

As is standard in obtaining such bounds, our notion of "essentially different" refers to a specific sample and so symmetrization arguments are required in order to make the above conceptual arguments concrete. To do so cleanly and cheaply, we use an ϵ-net of *kernels* to construct an ϵ-net of *classifiers* with respect to the kernels, noting that the size of the ϵ-net increases only multiplicatively relative to the size of an ϵ-net for any one kernel. An important component of this construction is the observation that kernels that are close as real-valued functions also yield similar classes of classifiers. Using our constructed ϵ-net, we can apply standard results bounding the estimation error in terms of the log-size of ϵ-nets, without needing to invoke symmetrization arguments directly.

6 Summary

We establish the first generalization error bounds for kernel-learning SVMs where the margin complexity term and the dimensionality of the kernel family interact *additively* rather then *multiplicatively* (up to log factors). The additive interaction yields stronger bounds. We believe that the implied additive bounds on the sample complexity represent its correct behavior (up to log factors), although this remains to be proved.

The results we present significantly improve on previous results for convex combinations of base kernels, for which the only previously known bound had a multiplicative interaction [1], and for Gaussian kernels with a learned covariance matrix, for which only a bound with a multiplicative interaction and an unspecified dependence on the input dimensionality was previously shown [14].

We also provide the first explicit non-trivial bound for linear combinations of base kernels—a bound that depends only on the (relative) margin and the number of base kernels. The techniques we introduce for obtaining bounds based on the pseudodimension of the class of kernels should readily apply to straightforward derivation of bounds for many other classes.

We note that previous attempts at establishing bounds for this setting [1, 2, 14] relied on bounding the Rademacher complexity [15] of the class $\mathcal{F}_{\mathcal{K}}$. However, generalization error bounds derived solely from the Rademacher complexity $\boldsymbol{\mathcal{R}}[\mathcal{F}_{\mathcal{K}}]$ of the class $\mathcal{F}_{\mathcal{K}}$ *must* have a multiplicative dependence on $\sqrt{B}/\gamma$. As we show, an analyzis based on covering number bounds yields significantly tighter bounds.

References

[1] Gert R.G. Lanckriet, Nello Cristianini, Peter Bartlett, Laurent El Ghaoui, and Michael I. Jordan. Learning the kernel matrix with semidefinite programming. *J Mach Learn Res*, 5:27–72, 2004.

[2] Olivier Bousquet and Daniel J. L. Herrmann. On the complexity of learning the kernel matrix. In *Adv. in Neural Information Processing Systems 15*, 2003.

[3] Koby Crammer, Joseph Keshet, and Yoram Singer. Kernel design using boosting. In *Advances in Neural Information Processing Systems 15*, 2003.

[4] G R G Lanckriet, T De Bie, N Cristianini, M I Jordan, and W S Noble. A statistical framework for genomic data fusion. *Bioinformatics*, 20, 2004.

[5] S Sonnenburg, G Rätsch, and C Schafer. Learning interpretable SVMs for biological sequence classification. In *Research in Computational Molecular Biology*, 2005.

[6] A Ben-Hur and W S Noble. Kernel methods for predicting protein-protein interactions. *Bioinformatics*, 21, 2005.

[7] Nello Cristianini, Cohn Campbell, and John Shawe-Taylor. Dynamically adapting kernels in support vector machines. In *Adv. in Neural Information Proceedings Systems 11*, 1999.

[8] O Chapelle, V Vapnik, O Bousquet, and S Makhuerjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46(1):131–159, 2002.

[9] S. Sathiya Keerthi. Efficient tuning of SVM hyperparameters using radius/margin bound and iterative algorithms. *IEEE Tran. on Neural Networks*, 13(5):1225–1229, 2002.

[10] Tobias Glasmachers and Christian Igel. Gradient-based adaptation of general gaussian kernels. *Neural Comput.*, 17:2099–2105, 2005.

[11] Cheng Soon Ong, Alexander J. Smola, and Robert C. Williamson. Learning the kernel with hyperkernels. *J. Mach. Learn. Res.*, 6, 2005.

[12] Charles A. Micchelli and Massimiliano Pontil. Learning the kernel function via regularization. *J. Mach. Learn. Res.*, 6, 2005.

[13] Andreas Argyriou, Charles A. Micchelli, and Massimiliano Pontil. Learning convex combinations of continuously parameterized basic kernels. In *18th Annual Conf. on Learning Theory*, 2005.

[14] Charles A. Micchelli, Massimiliano Pontil, Qian Wu, and Ding-Xuan Zhou. Error bounds for learning the kernel. Research Note RN/05/09, University College London Dept. of Computer Science, June 2005.

[15] Vladimir Koltchinskii and Dmitry Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Ann. Statist.*, 30(1), 2002.