Similarity-Based Theoretical Foundation for Sparse Parzen Window Prediction

Maria-Florina Balcan NINAMF@CS.CMU.EDU
Avrim Blum AVRIM@CS.CMU.EDU
Computer Science Department, Carnegie Mellon University

Nathan Srebro NATI@UCHICAGO.EDU
Toyota Technological Institute at Chicago, 1427 East 60th Street, Chicago IL 60637, USA

Extended Abstract

1. Sparse Parzen Window Prediction

We are concerned here with predictors of the form:

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x, x'_i) \tag{1}$$

where $x'_1, \ldots, x'_n \in \mathcal{X}$ are *landmark* instances in our instance space $\mathcal{X}$ and $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a function encoding the relationship between instances. These type of predictors are fairly common and natural and are obtained by various learning rules. Support Vector Machines (SVM) learn a predictor of the form (1) (binary labels y'_i are often included explicitly in the predictor, but can also be encoded as the sign of α_i) by minimizing an objective related to a dual large-margin problem. SVMs enjoy performance guarantees based on interpreting $K(\cdot, \cdot)$ as an inner product in an implicit Hilbert space, and also tend to yield sparse predictors, i.e. with few non-zero α_is. Parzen window prediction (aka soft nearest neighbor prediction) corresponds to a predictor of the form (1), with $\alpha_i = y'_i$. There is no need to interpret K as an inner product nor to require that it be positive semidefinite—we can simply think of K as specifying similarity. Still thinking of K as encoding similarity, and perhaps also dissimilarity, we might prefer to learn a sparse predictor, with $\alpha_i = 0$ for many landmarks x'_i as in SVMs, instead of simply fixing $\alpha_i = y'_i$. A more direct way of doing so is by minimizing a loss (here the hinge loss) with an explicit constraint on the L_1-norm of the coefficients α_i:

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{m} [1 - y_i f(x_i)]_+ \\ \text{s.t.} \quad & \sum_{j=1}^{n} |\alpha_j| \leq M \end{aligned} \tag{2}$$

where $(x_i, y_i) \in \mathcal{X} \times \{\pm 1\}$ are labeled training examples, which might, or might not, be the same as the landmarks x'_i (note that unlabeled examples can also be used as landmarks), and $[1 - yx]_+ = max(1 - yx, 0)$ is the hinge loss. This is a linear program and can be solved efficiently.

We view K as a similarity function, and provide a natural condition on K, that does not require K be positive semidefinite, and justifies the learning rule (2). Our condition guarantees the success of learning rule (2) and provides bounds on the required number of landmarks and training examples.

Furthermore, we show that any similarity function that is good as a kernel, i.e. can ensure SVM learning, also satisfies our condition and can thus also ensure learning using the learning rule (2) (though possibly with some deterioration of the learning guarantees). These arguments can be used to justify (2) as an alternative to SVMs.

2. Prior Work

The learning rule (2), usually with the same set of points both as training examples and landmarks, and variations of this rule, have been suggested by various authors and is fairly common in practice. Such a learning rule is typically discussed as an alternative to SVMs: Tipping (2001) suggested the Relevance Vector Machine (RVM) as a Bayesian alternative to SVMs. The MAP estimate of the RVM is given by an optimization problem similar to (2), though with a loss function different from the hinge loss (the hinge-loss cannot be obtained as a log-likelihood). Similarly, Singer (2000) suggests Norm-Penalized Leveraging Procedures as a boosting-like approach that mimics SVMs. Again, although the specific loss functions

studied by Singer are different from the hinge-loss, the method (with a norm exponent of 1, as in Singer's experiments) otherwise corresponds to a coordinate-descent minimization of (2). Other authors do use the hinge loss and discuss the learning rule (2) as given here, with the express intent of achieving sparsity more directly by minimizing the L_1 norm of the coefficients (Bennett & Campbell, 2000; Roth, 2001; Guigue et al., 2005).

Despite the interest in the learning rule (2), none of the above works suggest learning guarantees. In the case of SVMs, we have an established theory that ensures us that when K is positive semidefinite and is a "good kernel" for the learning problem (i.e. corresponds to an implicit Hilbert space where the problem is mostly separable with large margin), then the SVM learning rule is guaranteed to find a predictor of the form (1) with small generalization error. However, to the best of our knowledge, no such theory has been previously suggested for the learning rule (2). Even when the SVM pre-conditions hold, and the SVM learning-rule would work, we do not know of a previous guarantee for the alternate learning rule (2). Furthermore, since the learning rule (2) does not require K to be positive semidefinite, nor refer to an implied Hilbert space, one might hope for a more direct condition on K, that does not require it be positive semidefinite, and is sufficient to guarantee the success of the learning rule (2).

In fact, in order to enjoy the SVM guarantees while using L_1 regularization to obtain sparsity, some authors suggest regularizing both the L_1 norm $\|\alpha\|_1$ of the coefficient vector α (as in (2)), and the norm $\|\beta\|$ of the corresponding predictor $\beta = \sum_j \alpha_j \phi(x'_j)$ in the Hilbert space implied by K, where $K(x, x') = \langle \phi(x), \phi(x') \rangle$, as when using a SVM with K as a kernel (Osuna & Girosi, 1999; Gunn & Kandola, 2002).

3. Our Guarantees

We consider learning problems specified by a joint distribution P over labeled examples (x, y). We consider learning a predictor based on both labeled examples drawn from this distribution, as well as unlabeled examples drawn from the marginal over x. Our goal is to obtain a predictor with low expected error with respect to P.

Our main condition for a similarity function K is summarized in the following definition:

Definition 1 *A similarity function K is an (ϵ, γ, τ)-**good similarity function** for a learning problem P if there* exists *a (probabilistic) set R of "reasonable points" (one may think of R as a random indicator function) such that the following conditions hold:*

1. *We have*

$$\mathbf{E}_{(x,y)\sim P}\Big[\,[1 - yg(x)/\gamma]_+\,\Big] \leq \epsilon, \qquad (3)$$

   where $g(x) = \mathbf{E}_{(x',y',R(x'))}[y'K(x,x') \mid R(x')]$.

2. $\Pr_{x'}[R(x')] \geq \tau$.

That is, we require that at least τ fraction of the points are "reasonable" (in expectation), and that most points can be predicted according to the reasonable points similar, or dis-similar, to them (or rather, that the expected hinge loss of using this prediction is low).

If a similarity function is good under Definition 1, then we can guarantee there is a predictor $f(x)$ of the form (1) with low L_1-norm $|\alpha|_1$ achieving low expected hinge loss. This in turn yields a learning guarantee for the learning rule (2):

Theorem 1 *Let K be an (ϵ, γ, τ)-good similarity function for a learning problem P. For any $\delta, \epsilon_1 > 0$, let $x'_1, \ldots, x'_n$ be a (potentially unlabeled) sample of*

$$n = \frac{2}{\tau}\left(\log(2/\delta) + 16\frac{\log(2/\delta)}{\epsilon_1^2\gamma^2}\right)$$

landmarks drawn from P. Then with probability at least $1-\delta$, there exists a predictor of the form (1) with

$$|\alpha|_1 = \sum_{i=1}^{n} |\alpha_i| \leq 1/\gamma$$

and expected hinge loss

$$\mathbf{E}_{(x,y)\sim P}\Big[\,[1 - yf(x)]_+\,\Big] \leq \epsilon + \epsilon_1.$$

Corollary 1 *Let K be an (ϵ, γ, τ)-good similarity function for a learning problem P. For any $\delta, \epsilon_1 > 0$, with probability at least $1-\delta$ the predictor obtain from learning rule (2), with*

$$n = O\left(\frac{\log(1/\delta)}{\tau\gamma^2\epsilon_1^2}\right)$$

(unlabeled) landmarks and

$$m = \tilde{\mathcal{O}}\left(\frac{\log n \log(1/\delta)}{\gamma^2\epsilon_1^2}\right)$$

labeled training examples, has expected hinge loss at most $\epsilon + \epsilon_1$.

As discussed earlier, we also establish that if a similarity function is positive semidefinite and "good" in the traditional kernel sense, then it also satisfies Definition 1, yielding a learning guarantee on the learning rule (2). Recall that a function $K : \mathcal{X} \times \mathcal{X}$ is *positive semidefinite* iff there exists a mapping $\phi : \mathcal{X} \to \mathcal{H}$ into a Hilbert space $\mathcal{H}$ such that $K(x, x') = \langle \phi(x), \phi(x') \rangle$. With this representation of K in mind:

Definition 2 *We say that a positive semidefinite K is an (ϵ, γ)-**good kernel** if there exists a vector $\beta \in \mathcal{H}$, $\|\beta\| \leq 1/\gamma$ such that*

$$\mathbf{E}_{(x,y)\sim P}[[1 - \ell\langle \beta, \phi(x) \rangle]_+] \leq \epsilon.$$

Theorem 2 *If a positive semidefinite K is an (ϵ_0, γ)-good kernel in hinge loss for learning problem P (with deterministic labels), then for any $\epsilon_1 > 0$ there exists $c > 1$ such that K is also a $(\epsilon_0 + \epsilon_1, \frac{c\gamma^2}{1+\epsilon_0/2\epsilon_1}, \frac{2\epsilon_1+\epsilon_0}{c})$-good similarity function in hinge loss.*

Corollary 2 *Let K be a positive semidefinite (ϵ, γ)-good kernel for a learning problem P. For any $\delta, \epsilon_1 > 0$, with probability at least $1 - \delta$ the predictor obtained from learning rule (2), with*

$$n = O\left(\frac{(1 + \epsilon/\epsilon_1)^2 \log(1/\delta)}{(\epsilon + \epsilon_1)\gamma^4 \epsilon_1^2}\right)$$

(unlabeled) landmarks and

$$m = \tilde{\mathcal{O}}\left(\frac{(1 + \epsilon/\epsilon_1)^2 \log n \log(1/\delta)}{\gamma^4 \epsilon_1^2}\right)$$

labeled training examples, has expected hinge loss at most $\epsilon + \epsilon_1$.

Note that if $\epsilon_1 = \Omega(\epsilon)$, e.g. if we aim for a fixed percentile increase over "optimal" error, or in the noiseless case $\epsilon = 0$, the sample sizes simplify to: $n = O\left(\frac{\log 1/\delta}{\gamma^4 \epsilon_1^3}\right)$ and $m = \tilde{\mathcal{O}}\left(\frac{\log n \log 1/\delta}{\gamma^4 \epsilon_1^2}\right)$.

Proofs of these theorems (in slightly different forms) appear in (Balcan et al., 2008), which focuses on generalizing the theory of learning with kernels to broader classes of pairwise similarity functions. Here, our focus is on how this extension can be used to provide formal guarantees for the common sparsity inducing learning rule given in equation (2).

References

Balcan, M., Blum, A., & Srebro, N. (2008). Improved Guarantees for Learning via Similarity Functions. *COLT*.

Bennett, K. P., & Campbell, C. (2000). Support vector machines: hype or hallelujah? *SIGKDD Explor. Newsl.*, *2*, 1–13.

Guigue, V., Rakotomamonjy, A., & Canu, S. (2005). Kernel basis pursuit. *Proceedings of the 16th European Conference on Machine Learning (ECML'05)*. Springer.

Gunn, S. R., & Kandola, J. S. (2002). Structural modelling with sparse kernels. *Mach. Learn.*, *48*, 137–163.

Osuna, E. E., & Girosi, F. (1999). Reducing the run-time complexity in support vector machines. In *Advances in kernel methods: support vector learning*, 271–283. Cambridge, MA, USA: MIT Press.

Roth, V. (2001). Sparse kernel regressors. *ICANN '01: Proceedings of the International Conference on Artificial Neural Networks* (pp. 339–346). London, UK: Springer-Verlag.

Singer, Y. (2000). Leveraged vector machines. *Advances in Neural International Proceedings System 12*.

Tipping, M. E. (2001). Sparse bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.*, *1*, 211–244.