Learning with Multiple Similarity Functions

Maria-Florina Balcan
Computer Science Department
Carnegie Mellon University
ninamf@cs.cmu.edu

Avrim Blum
Computer Science Department
Carnegie Mellon University
avrim@cs.cmu.edu

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

1 Introduction

Kernel functions have become an extremely popular tool in machine learning, with many applications and an attractive theory [1, 12, 10]. There has also been substantial work on learning kernel functions from data [7, 11, 2]. A sufficient condition for a kernel to allow for good generalization on a given learning problem is that it induce a large margin of separation between positive and negative classes in its implicit space. In recent work [4, 5, 3] we have developed a theory that more broadly holds for general similarity functions that are not necessarily legal kernel functions. In particular, we have introduced a notion of a good similarity function for a given learning problem that (a) is fairly natural and intuitive (it does not require an implicit space and allows for functions that are not positive semi-definite), (b) is a sufficient condition for learning well, and (c) strictly generalizes the notion of a large-margin kernel function in that any such kernel is also a good similarity function, though not necessarily vice-versa. We also have partial progress on extending the theory of learning with *multiple* kernel functions to this more general notion. In this note, we describe the main definitions and results of [4], give our results on learning with multiple similarity functions, and present several open questions.

2 Good Similarity Functions

We consider classification problems specified by a joint distribution P over labeled examples (x, ℓ), where $\ell \in \{-1, 1\}$. 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 notion of a good similarity function K is summarized in the following definition. This definition is based on the intuitive idea that ideally we would like most examples to be substantially more similar on average to examples of their own label than to examples of the opposite label. However, this is too strong a requirement to capture all large-margin kernel function. So, instead we relax the condition to instead ask for existence of a non-negligble set R of "representative points" such that most examples are on average more similar to those in R of their own label than to those in R of the opposite label (this set R need not be known in advance). More formally,

Definition 1 *A similarity function $K : X \times X \to [-1, 1]$ is an (ϵ, γ, τ)-**good similarity function** for a learning problem P if there* exists *a subset R of the domain such that the following conditions hold:*

1. *A $1 - \epsilon$ probability mass of examples (x, ℓ) satisfy*

$$\mathbf{E}_{(x',\ell')\sim P}[\ell\ell' K(x, x') \mid x' \in R] \geq \gamma \tag{1}$$

2. $\Pr[R] \geq \tau$.

where formally we allow membership in R to be probabilistic ($R(x)$ is a probabilistic indicator function, and all probabilities are over both P and any randomization in R). We say K is (ϵ, γ, τ)-
good in hinge loss *if we can replace (1) with*

$1'$. $\mathbf{E}_{(x,\ell)\sim P}\Big[\,[1 - \ell g(x)/\gamma]_+\,\Big] \leq \epsilon$, *where* $g(x) = \mathbf{E}_{(x',\ell')\sim P,R}[\ell' K(x,x') \mid x' \in R]$.

That is, rather than defining ϵ to be the probability mass of examples that fail, ϵ is the expected degree of failure.

If the set R of "representative points" is $50/50$ positive and negative, we can interpret condition (1) as stating that most examples x are on average 2γ more similar to random representative examples x' of their own label than to random representative examples x' of the other label. Condition (2) is that at least a τ fraction of the points should be representative.

These definitions have two important properties: (a) they are sufficient for learning, and (b) they are are satisfied—albeit with some loss in parameters—by any large-margin kernel function. We begin with sufficiency for learning: in particular, we show that by choosing a set of "landmark" points at random and then using similarity to these as explicit features, we can convert our learning problem to one of learning a linear separator of good L_1-margin.

Theorem 1 *Let K be an (ϵ, γ, τ)-good similarity function for a learning problem P. Let $S = \{x'_1, x'_2, \ldots, x'_d\}$ be a (potentially unlabeled) sample of*

$$d = \frac{2}{\tau}\left(\log(2/\delta) + 8\frac{\log(2/\delta)}{\gamma^2}\right)$$

points ("landmarks") drawn from P. Consider the mapping $\phi^S : X \to \mathbb{R}^d$ defined as follows: ${\phi^S}_i(x) = K(x, x'_i)$, $i \in \{1, \ldots, d\}$. Then, with probability at least $1 - \delta$ over the random sample S, the induced distribution $\phi^S(P)$ in R^d has a separator of error at most $\epsilon + \delta$ relative to L_1 margin at least $\gamma/2$, and all examples have L_∞ norm at most 1.

Proof Sketch: First, note that since $|K(x,x')| \leq 1$ for all x, x', we have $\left\|\phi^S(x)\right\|_\infty \leq 1$.

Consider the linear separator $\alpha \cdot \phi^S(x) \geq 0$ for $\alpha \in \mathbb{R}^d$ given by $\alpha_i = \ell(x'_i)R(x'_i)/d_1$ where $d_1 = \sum_i R(x'_i)$ is the number of landmarks with $R(x') = 1$. This normalization ensures $\|\alpha\|_1 = 1$. Moreover, the dot-product $\ell(x)\alpha \cdot \phi^S(x)$ is an empirical estimate (over S) of the conditional expectation in condition (1) of Definition 1. Our sample size d is sufficiently large so that with probability at least $1 - \delta/2$, the number of representative points d_1 in S is large enough so that for any given x, by Hoeffding bounds, the empirical estimate is within $\gamma/2$ of the expectation in (1) with probability at least $1 - \delta^2/2$. This in turn, by Markov's inequality, implies that with probability at least $1 - \delta$, at most a δ fraction of the points x satisfying condition (1) have $\phi^S(x)$ with margin less than $\gamma/2$. Therefore, at most an $\epsilon + \delta$ fraction of points overall have margin less than $\gamma/2$, as desired. ∎

Similarly, for hinge-loss, we can show the following.

Theorem 2 *Let K be an (ϵ, γ, τ)-good similarity function in hinge-loss for a learning problem P. For any $\epsilon_1 > 0$ and $0 < \lambda < \gamma\epsilon_1/4$ let $S = \{x'_1, x'_2, \ldots, x'_d\}$ be a sample of size $d = \frac{2}{\tau}\left(\log(2/\delta) + 16\log(2/\delta)/(\epsilon_1\gamma)^2\right)$ drawn from P. With probability at least $1 - \delta$ over the random sample S, the induced distribution $\phi^S(P)$ in R^d, for ϕ^S as defined in Theorem 1, has a separator achieving hinge-loss at most $\epsilon + \epsilon_1$ at L_1 margin γ.*

In particular, the above theorem implies that we can use algorithms designed for learning linear separators of good L_1 margin (for example, Winnow), and get generalization bounds that are logarithmic in the dimension $d = |S|$ and quadratic in $1/\gamma$ [8].

Turning to the connection to large-margin kernels, if a similarity function is positive semidefinite and "good" in the traditional kernel sense, then we can show it also satisfies Definition 1, though

with some loss in parameters. 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 (ϵ, γ)-kernel good in hinge-loss 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 3 *If a positive semidefinite K is an (ϵ_0, γ)-good kernel in hinge loss for learning problem (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.*

Note that Theorem 3 implies that the learning procedure given by Theorem 2 can also be applied to good kernel functions, with roughly a quadratic increase in sample complexity. Despite the increase in sample complexity, this can be useful for other purposes. For example, inspired by results in our earlier work [6] this approach has been used for transfer learning by Quattoni, Collins, and Darrell et al. [9]. In addition, the intuitiveness of Definition 1 may make it easier to design good kernel functions for a given learning problem.

3 Learning with Multiple Similarity Functions

We consider here as in Lanckriet et al. [7] the problem of learning with multiple similarity functions. In particular, suppose that rather than having a single similarity function, we were instead given n functions $K_1, ..., K_n$, and our hope is that some *convex combination* of them will satisfy Definition 1. Is this sufficient to be able to learn well, and how does performance degrade with n? The following generalization of Theorem 1 shows that indeed we can still learn well and that the degradation with n is only slight. (The analog of Theorem 2 can be derived similarly.)

Theorem 4 *Suppose $K_1, \ldots, K_n$ are similarity functions such that some (unknown) convex combination of them is (ϵ, γ, τ)-good. For any $\delta > 0$, let $S = \{x'_1, x'_2, \ldots, x'_d\}$ be a sample of size $d = 16\frac{\log(1/\delta)}{\tau\gamma^2}$ drawn from P. Consider the mapping $\phi^S : X \to \mathbb{R}^{nd}$ defined as follows:*

$$\phi^S{}_i(x) = (K_1(x, x'_1), \ldots, K_n(x, x'_1), \ldots, K_1(x, x'_d), \ldots, K_n(x, x'_d)).$$

With probability at least $1 - \delta$ over the random sample S, the induced distribution $\phi^S(P)$ in R^{nd} has a separator of error at most $\epsilon + \delta$ at L_1, L_∞ margin at least $\gamma/2$.

Before proving Theorem 4, notice that the margin achieved is identical to that in Theorem 1 when we had just a single similarity function. The only degradation is in the dimension, which has increased by a factor of n. However, because algorithms for learning linear separators with good L_1 margin have sample-size bounds that are only *logarithmic* in dimension, this causes only an $O(\log n)$ multiplicative penalty in the number of labeled examples needed to achieve good generalization (though running time would be impacted linearly with n).

Proof: Let $K = \alpha_1 K_1 + \ldots + \alpha_n K_n$ be an (ϵ, γ, τ)-good convex-combination of the K_i. By Theorem 1, had we instead performed the mapping: $\tilde{\phi}^S : X \to R^d$ defined as

$$\tilde{\phi}^S(x) = (K(x, x'_1), \ldots, K(x, x'_d)),$$

then with probability $1 - \delta$, the induced distribution $\tilde{\phi}^S(P)$ in R^d would have a separator of error at most $\epsilon + \delta$ at margin at least $\gamma/2$. Let $\hat{\beta}$ be the vector corresponding to such a separator in that space. Now, let us convert $\hat{\beta}$ into a vector in R^{nd} by replacing each coordinate $\hat{\beta}_j$ with the n values $(\alpha_1\hat{\beta}_j, \ldots, \alpha_n\hat{\beta}_j)$. Call the resulting vector $\tilde{\beta}$. Notice that by design, for any x we have $\left\langle \tilde{\beta}, \phi^S(x) \right\rangle = \left\langle \hat{\beta}, \tilde{\phi}^S(x) \right\rangle$. Furthermore, $\left\| \tilde{\beta} \right\|_1 = \left\| \hat{\beta} \right\|_1$. Thus, the vector $\tilde{\beta}$ under distribution $\phi^S(P)$ has the same properties as the vector $\hat{\beta}$ under $\tilde{\phi}^S(P)$. This implies the desired result. ∎

This result can be viewed as analogous to the idea of learning a kernel matrix studied by [7]. However, rather than explicitly learning the best convex combination, we are simply folding the learning process into the second stage of the algorithm. Of particular interest is that due to the use of L_1 margins, the number of labeled examples needed for good generalization grows only *logarithmically* with the number of similarity functions at hand. It is interesting to compare this with the results of [11] who give general sample-complexity bounds for learned kernels. Whereas our approach gives an $O(\log n)$ multiplicative penalty, the results of [11] on convex combinations of kernels give an $O(n)$ additive penalty. Thus, in a sense the results are incomparable. On the other hand, the logarithmic penalty of Theorem 4 means that one could even use more similarity functions than we have labeled data!

4 Open Questions and Discussion

While the result given in Theorem 4 is in a sense analogous to results of [7] and [11] on learning a convex combination of kernel functions, we do not actually produce a convex combination of similarity functions; rather, we just use its existence to learn a linear separator in the landmark-feature space. This suggests the following natural open question: Can bounds comparable to those resulting from Theorem 4 for learning the target function be obtained for algorithms that explicitly output a convex combination of similarity functions satisfying some approximation of Definition 1?

More broadly, Srebro and Ben-David [11] discuss the problem of learning over a general space $\mathcal{K}$ of kernel functions. One could obtain roughly comparable bounds for spaces of similarity functions by explicitly listing similarity functions making up an ϵ-net of $\mathcal{K}$ and then performing a mapping using this explicit list as in Theorem 4. However, this would be incredibly computationally inefficient. A second open question is whether for natural classes of similarity functions one can combine learning of the target function with learning the best similarity function in the class into a single, and ideally computationally efficient, optimization problem.

The approach of Theorem 4 may also have applications to multi-task or transfer learning. Suppose one has a set of similarity functions and a family of classification tasks, and for each classification task there is some (perhaps different) convex combination of the similarity functions that is good for it. Then one could reuse the same embedding and same landmarks for all of them. If furthermore one believes that the sets of "representative points" for the different tasks have high overlap, then following the approach of [9] one could use a joint regularization penalty when learning all of them. It would be interesting to better understand when this can be expected to perform well and more generally to explore transfer learning in this setting.

References

[1] *http://www.kernel-machines.org/*.

[2] A. Argyriou, R. Hauser, C.A. Micchellio, and M. Pontil. A DC algorithm for kernel selection. In *ICML*, 2006.

[3] M.-F. Balcan and A. Blum. On a theory of learning with similarity functions. In *ICML*, 2006.

[4] M.-F. Balcan, A. Blum, and N. Srebro. Improved guarantees for learning via similarity functions. In *COLT*, 2008.

[5] M.-F. Balcan, A. Blum, and N. Srebro. A theory of learning with similarity functions. *Machine Learning*, 2008.

[6] M.-F. Balcan, A. Blum, and S. Vempala. Kernels as features: On kernels, margins, and low-dimensional mappings. *Machine Learning*, 65(1):79 – 94, 2006.

[7] G. R. G. Lanckriet, N. Cristianini, P. L. Bartlett, L. El Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

[8] N. Littlestone. From online to batch learning. In *COLT*, pages 269–284, 1989.

[9] A. Quattoni, M. Collins, and T. Darrell. Transfer learning for image classification with sparse prototype representations. In *CVPR*, 2008.

[10] A. J. Smola and B. Schölkopf. *Learning with Kernels*. MIT Press, 2002.

[11] N. Srebro and S. Ben-David. Learning bounds for support vector machines with learned kernels. In *COLT*, 2006.

[12] V. N. Vapnik. *Statistical Learning Theory*. John Wiley and Sons Inc., 1998.