On Manifold Regularization

Mikhail Belkin, Partha Niyogi, Vikas Sindhwani
{misha,niyogi,vikass}@cs.uchicago.edu
Department of Computer Science
University of Chicago

Abstract

We propose a family of learning algorithms based on a new form of regularization that allows us to exploit the geometry of the marginal distribution. We focus on a semi-supervised framework that incorporates labeled and unlabeled data in a general-purpose learner. Some transductive graph learning algorithms and standard methods including Support Vector Machines and Regularized Least Squares can be obtained as special cases. We utilize properties of Reproducing Kernel Hilbert spaces to prove new Representer theorems that provide theoretical basis for the algorithms. As a result (in contrast to purely graph based approaches) we obtain a natural out-of-sample extension to novel examples and are thus able to handle both transductive and truly semi-supervised settings. We present experimental evidence suggesting that our semi-supervised algorithms are able to use unlabeled data effectively. In the absence of labeled examples, our framework gives rise to a regularized form of spectral clustering with an out-of-sample extension.

1 Introduction

The problem of learning from labeled and unlabeled data (*semi-supervised* and *transductive* learning) has attracted considerable attention in recent years (cf. [11, 7, 10, 15, 18, 17, 9]). In this paper, we consider this problem within a new framework for data-dependent regularization. Our framework exploits the geometry of the probability distribution that generates the data and incorporates it as an additional regularization term. We consider in some detail the special case where this probability distribution is supported on a submanifold of the ambient space.

Within this general framework, we propose two specific families of algorithms: the Laplacian Regularized Least Squares (hereafter LapRLS) and the Laplacian Support Vector Machines (hereafter LapSVM). These are natural extensions of RLS and SVM respectively. In addition, several recently proposed transductive methods (e.g., [18, 17, 1]) are also seen to be special cases of this general approach. Our solution for the semi-supervised case can be expressed as an expansion over labeled and unlabeled data points. Building on a solid theoretical foundation, we obtain a natural solution to the problem of out-of-sample extension (see also [6] for some recent work).When all examples are unlabeled, we obtain a new regularized version of spectral clustering.

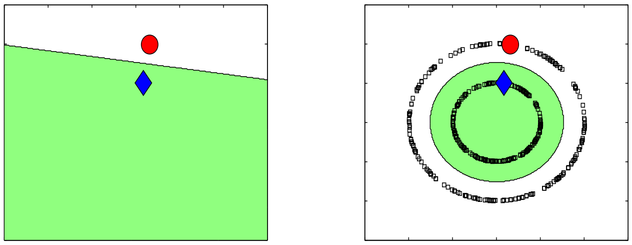
Our general framework brings together three distinct concepts that have received some independent recent attention in machine learning: Regularization in Reproducing Kernel Hilbert Spaces, the technology of Spectral Graph Theory and the geometric viewpoint of Manifold Learning algorithms.

2 The Semi-Supervised Learning Framework

First, we recall the standard statistical framework of learning from examples, where there is a probability distribution P on $X \times \mathbb{R}$ according to which training examples are generated. Labeled examples are (x, y) pairs drawn from P. Unlabeled examples are simply $x \in X$ drawn from the marginal distribution $\mathcal{P}_X$ of P.

One might hope that knowledge of the marginal $\mathcal{P}_X$ can be exploited for better function learning (e.g. in classification or regression tasks). Figure 1 shows how unlabeled data can radically alter our prior belief about the appropriate choice of classification functions. However, if there is no identifiable relation be-

Figure 1: Unlabeled data and prior beliefs



tween $\mathcal{P}_X$ and the conditional $\mathcal{P}(y|x)$, the knowledge of $\mathcal{P}_X$ is unlikely to be of much use. Therefore, we will make a specific assumption about the connection between the marginal and the conditional. We will assume that if two points $x_1, x_2 \in X$ are *close* in the *intrinsic* geometry of $\mathcal{P}_X$, then the conditional distributions $\mathcal{P}(y|x_1)$ and $\mathcal{P}(y|x_2)$ are similar. In other words, the conditional probability distribution $\mathcal{P}(y|x)$ varies smoothly along the geodesics in the intrinsic geometry of $\mathcal{P}_X$.

We utilize these geometric ideas to extend an established framework for function learning. A number of popular algorithms such as SVM, Ridge regression, splines, Radial Basis Functions may be broadly interpreted as regularization algorithms with different empirical cost functions and complexity measures in an appropriately chosen Reproducing Kernel Hilbert Space (RKHS).

For a Mercer kernel $K : X \times X \to \mathbb{R}$, there is an associated RKHS $\mathcal{H}_K$ of functions $X \to \mathbb{R}$ with the corresponding norm $\| \, \|_K$. Given a set of labeled examples (x_i, y_i), $i = 1, \ldots, l$ the standard framework estimates an unknown function by minimizing

$$f^* = \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \frac{1}{l} \sum_{i=1}^{l} V(x_i, y_i, f) + \gamma \|f\|_K^2 \qquad (1)$$

where V is some loss function, such as squared loss $(y_i - f(x_i))^2$ for RLS or the soft margin loss function for SVM. Penalizing the RKHS norm imposes smoothness conditions on possible solutions. The classical Representer Theorem states that the solution to this minimization problem exists in $\mathcal{H}_K$ and can be written as

$$f^*(x) = \sum_{i=1}^{l} \alpha_i K(x_i, x) \qquad (2)$$

Therefore, the problem is reduced to optimizing over the finite dimensional space of coefficients α_i, which is the algorithmic basis for SVM, Regularized Least Squares and other regression and classification schemes.

Our goal is to extend this framework by incorporating additional information about the geometric structure of the marginal $\mathcal{P}_X$. We would like to ensure that the solution is smooth with respect to both the ambient space and the marginal distribution $\mathcal{P}_X$. To achieve that, we introduce an additional regularizer:

$$f^* = \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \frac{1}{l} \sum_{i=1}^{l} V(x_i, y_i, f) + \gamma_A \|f\|_K^2 + \gamma_I \|f\|_I^2 \quad (3)$$

where $\|f\|_I^2$ is an appropriate penalty term that should reflect the intrinsic structure of $\mathcal{P}_X$. Here γ_A controls the complexity of the function in the *ambient* space while γ_I controls the complexity of the function in the *intrinsic* geometry of $\mathcal{P}_X$. Given this setup one can prove the following representer theorem:

Theorem 2.1. *Assume that the penalty term $\|f\|_I$ is sufficiently smooth with respect to the RKHS norm $\|f\|_K$. Then the solution f^* to the optimization problem in Eqn 3 above exists and admits the following representation*

$$f^*(x) = \int_{\mathcal{M}} \alpha(y) K(x, y) \, d\mathcal{P}_X(y) + \sum_{i=1}^{l} \alpha_i K(x_i, x) \quad (4)$$

where $\mathcal{M} = \operatorname{supp}\{\mathcal{P}_X\}$ is the support of the marginal $\mathcal{P}_X$.

The proof of this theorem runs over several pages and is omitted for lack of space. See [4] for details including the exact statement of the smoothness conditions.

In most applications, however, we do not know $\mathcal{P}_X$. Therefore we must attempt to get empirical estimates of $\|f\|_I$. Note that in order to get such empirical estimates it is sufficient to have *unlabeled* examples.

A case of particular recent interest is when the support of $\mathcal{P}_X$ is a compact submanifold $\mathcal{M} \subset X = \mathbb{R}^n$. In that case, a natural choice for $\|f\|_I$ is $\int_{\mathcal{M}} \langle \nabla_{\mathcal{M}} f, \nabla_{\mathcal{M}} f \rangle$. The optimization problem becomes

$$f^* = \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \frac{1}{l} \sum_{i=1}^{l} V(x_i, y_i, f) + \gamma_A \|f\|_K^2 + \gamma_I \int_{\mathcal{M}} \langle \nabla_{\mathcal{M}} f, \nabla_{\mathcal{M}} f \rangle \qquad (5)$$

The term $\int_{\mathcal{M}} \langle \nabla_{\mathcal{M}} f, \nabla_{\mathcal{M}} f \rangle$ may be approximated on the basis of labeled and unlabeled data using the graph Laplacian ([1]). Thus, given a set of l labeled examples $\{(x_i, y_i)\}_{i=1}^{l}$ and a set of u unlabeled examples $\{x_j\}_{j=l+1}^{j=l+u}$, we consider the following optimization problem :

$$f^* = \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \frac{1}{l} \sum_{i=1}^{l} V(x_i, y_i, f) + \gamma_A \|f\|_K^2 + \frac{\gamma_I}{(u+l)^2} \hat{f}^T L \hat{f} \qquad (6)$$

where $\hat{f} = [f(x_1), \ldots, f(x_{l+u})]^T$, and L is the graph Laplacian given by $L = D - W$ where W_{ij} are the edge weights in the data adjacency graph. Here, the diagonal matrix D is given by $D_{ii} = \sum_{j=1}^{l+u} W_{ij}$. The normalizing coefficient $\frac{1}{(u+l)^2}$ is the natural scale factor for the empirical estimate of the Laplace operator. On a sparse adjacency graph it may be replaced by $\sum_{i,j=1}^{l+u} W_{ij}$.

The following simple version of the representer theorem shows that the minimizer has an expansion in terms of both labeled and unlabeled examples and is a key to our algorithms.

Theorem 2.2. *The minimizer of optimization problem 6 admits an expansion*

$$f^*(x) = \sum_{i=1}^{l+u} \alpha_i K(x_i, x) \qquad (7)$$

in terms of the labeled and unlabeled examples.

The proof is a variation of the standard orthogonality argument, which we omit for lack of space.

Remarks : (a) Other natural choices of $\|\ \|_I$ exist. Examples are (i) heat kernel (ii) iterated Laplacian (iii) kernels in geodesic coordinates. The above kernels are geodesic analogs of similar kernels in Euclidean space. (b) Note that K restricted to $\mathcal{M}$ (denoted by $K_{\mathcal{M}}$) is also a kernel defined on $\mathcal{M}$ with an associated RKHS $\mathcal{H}_{\mathcal{M}}$ of functions $\mathcal{M} \to \mathbb{R}$. While this might suggest $\|f\|_I = \|f|_{\mathcal{M}}\|_{K_M}$ ($f|_{\mathcal{M}}$ is f restricted to $\mathcal{M}$) as a reasonable choice for $\|f\|_I$, it turns out, that for the minimizer f^* of the corresponding optimization problem we get $\|f^*\|_I = \|f^*\|_K$, yielding the same solution as standard regularization, although with a different γ.

3 Algorithms

We now present solutions to the optimization problem posed in Eqn (6). To fix notation, we assume we have l labeled examples $\{(x_i, y_i)\}_{i=1}^{l}$ and u unlabeled examples $\{x_j\}_{j=l+1}^{j=l+u}$. We use K interchangeably to denote the kernel function or the Gram matrix.

3.1 Laplacian Regularized Least Squares (LapRLS)

The Laplacian Regularized Least Squares algorithm solves Eqn (6) with the squared loss function: $V(x_i, y_i, f) = (y_i - f(x_i))^2$. Since the solution is of the form given by (7), the objective function can be reduced to a convex differentiable function of the $(l+u)$-dimensional expansion coefficient vector $\alpha = [\alpha_1, \ldots, \alpha_{l+u}]^T$ whose minimizer is given by :

$$\alpha^* = (JK + \gamma_A l I + \frac{\gamma_I l}{(u+l)^2} LK)^{-1} Y \qquad (8)$$

Here, K is the $(l+u) \times (l+u)$ Gram matrix over labeled and unlabeled points; Y is an $(l+u)$ dimensional label vector given by - $Y = [y_1, \ldots, y_l, 0, \ldots, 0]$ and J is an $(l+u) \times (l+u)$ diagonal matrix given by - $J = diag(1, \ldots, 1, 0, \ldots, 0)$ with the first l diagonal entries as 1 and the rest 0.

Note that when $\gamma_I = 0$, Eqn (8) gives zero coefficients over unlabeled data. The coefficients over labeled data are exactly those for standard RLS.

3.2 Laplacian Support Vector Machines (LapSVM)

Laplacian SVMs solve the optimization problem in Eqn. 6 with the soft margin loss function defined as $V(x_i, y_i, f) = \max(0, 1 - y_i f(x_i)), y_i \in \{-1, +1\}$. Introducing slack variables, using standard Lagrange Multiplier techniques used for deriving SVMs [16], we first arrive at the following quadratic program in l dual variables β :

$$\beta^\star = \max_{\beta \in \mathbb{R}^l} \sum_{i=1}^{l} \beta_i - \frac{1}{2}\beta^T Q \beta \qquad (9)$$

subject to the contraints : $\sum_{i=1}^{l} y_i \beta_i = 0, \quad 0 \leq \beta_i \leq \frac{1}{l} \quad , i = 1, ...l$, where

$$Q = YJK(2\gamma_A I + 2\frac{\gamma_I}{(u+l)^2} LK)^{-1} J^T Y \qquad (10)$$

Here, Y is the diagonal matrix $Y_{ii} = y_i$, K is the Gram matrix over both the labeled and the unlabeled data; L is the data adjacency graph Laplacian; and J is an $l \times (l+u)$ matrix given by - $J_{ij} = 1$ if $i = j$ and x_i is a labeled example, and $J_{ij} = 0$ otherwise. To obtain the optimal expansion coefficient vector $\alpha^* \in \mathbb{R}^{(l+u)}$, one has to solve the following linear system after solving the quadratic program above :

$$\alpha^* = (2\gamma_A I + 2\frac{\gamma_I}{(u+l)^2} LK)^{-1} J^T Y \beta^\star \qquad (11)$$

One can note that when $\gamma_I = 0$, the SVM QP and Eqns (10,11), give zero expansion coefficients over the unlabeled data. The expansion coefficients over the labeled data and the Q matrix are as in standard SVM, in this case. Laplacian SVMs can be easily implemented using standard SVM software and packages for solving linear systems.

The Manifold Regularization algorithms and some connections are presented in the table below. For Graph Regularization and Label Propagation see [12, 3, 18].

	Manifold Regularization Algorithms
Input:	l labeled examples $\{(x_i, y_i)\}_{i=1}^{l}$, u unlabeled examples $\{x_j\}_{j=l+1}^{l+u}$
Output:	Estimated function $f : \mathbb{R}^n \to \mathbb{R}$
Step 1	► Construct data adjacency graph with $(l+u)$ nodes using, e.g, k nearest neighbors. Choose edge weights W_{ij}, e.g. binary weights or heat kernel weights $W_{ij} = e^{-\|x_i - x_j\|^2/4t}$.
Step 2	► Choose a kernel function $K(x, y)$. Compute the Gram matrix $K_{ij} = K(x_i, x_j)$.
Step 3	► Compute graph Laplacian matrix : $L = D - W$ where D is a diagonal matrix given by $D_{ii} = \sum_{j=1}^{l+u} W_{ij}$.
Step 4	► Choose γ_A and γ_I.
Step 5	► Compute α^* using Eqn (8) for squared loss (Laplacian RLS) or using Eqns (10,11) together with the SVM QP solver for soft margin loss (Laplacian SVM).
Step 6	► Output function $f^*(x) = \sum_{i=1}^{l+u} \alpha_i^* K(x_i, x)$.
	Connections to other algorithms
$\gamma_A \geq 0 \;\; \gamma_I \geq 0$	*Manifold Regularization*
$\gamma_A \geq 0 \;\; \gamma_I = 0$	*Standard Regularization (RLS or SVM)*
$\gamma_A = 0 \;\; \gamma_I > 0$	*Out-of-sample extension for Graph Regularization (RLS or SVM)*
$\gamma_A = 0 \;\; \gamma_I \to 0$	*Out-of-sample extension for Label Propagation (RLS or SVM)*
$\gamma_A \to 0 \;\; \gamma_I = 0$	*Hard margin (RLS or SVM)*

4 Related Work

In this section we survey various approaches to semi-supervised and transductive learning and highlight connections of Manifold Regularization to other algorithms.

Transductive SVM (TSVM) [16], [11]: TSVMs are based on the following optimization principle :

$$f^* = \underset{f\in\mathcal{H}_K, y_{l+1},\cdots y_{l+u}}{\operatorname{argmin}} \; C\sum_{i=0}^{l}(1 - y_i f(x_i))_+ \\ + C^* \sum_{i=l+1}^{l+u} (1 - y_i f(x_i))_+ + \|f\|_K^2$$

which proposes a joint optimization of the SVM objective function over binary-valued labels on the unlabeled data and functions in the RKHS. Here, C, C^* are parameters that control the relative hinge-loss over labeled and unlabeled sets. The joint optimization is implemented in [11] by first using an inductive SVM to label the unlabeled data and then iteratively solving SVM quadratic programs, at each step switching labels to improve the objective function. However this procedure is susceptible to local minima and requires an unknown, possibly large number of label switches before converging. Note that even though TSVM were inspired by transductive inference, they do provide an out-of-sample extension.

Semi-Supervised SVMs (S^{3}VM) [5] : S^{3}VM incorporate unlabeled data by including the minimum hinge-loss for the two choices of labels for each unlabeled example. This is formulated as a mixed-integer program for linear SVMs in [5] and is found to be intractable for large amounts of unlabeled data. The presentation of the algorithm is restricted to the linear case.

Measure-Based Regularization [9]: The conceptual framework of this work is closest to our approach. The authors consider a gradient based regularizer that penalizes variations of the function more in high density regions and less in low density regions leading to the following optimization principle:

$$f^* = \underset{f\in\mathcal{F}}{\operatorname{argmin}} \sum_{i=0}^{l} V(f(x_i), y_i) + \\ \gamma \int_X \langle \nabla f(x), \nabla f(x)\rangle p(x)dx$$

where p is the density of the marginal distribution $\mathcal{P}_X$. The authors observe that it is not straightforward to find a kernel for arbitrary densities p, whose associated RKHS norm is $\int \langle \nabla f(x), \nabla f(x)\rangle p(x)dx$. Thus, in the absence of a representer theorem, the authors propose to perform minimization over the linear space $\mathcal{F}$ generated by the span of a fixed set of basis functions chosen apriori. It is also worth noting that while [9] uses the gradient $\nabla f(x)$ in the ambient space, we use the penalty functional associated with the gradient $\nabla_{\mathcal{M}} f$ over a submanifold. In a situation where the data truly lies on or near a submanifold $\mathcal{M}$, the difference between these two penalizers can be significant since smoothness in the normal direction to the data manifold is irrelevant to classification or regression. The algorithm in [9] does not demonstrate performance improvements in real world experiments.

Graph Based Approaches See e.g., [7, 10, 15, 17, 18, 1]: A variety of graph based methods have been proposed for transductive inference. However, these methods do not provide an out-of-sample extension. In [18], nearest neighbor labeling for test examples is proposed once unlabeled examples have been labeled by transductive learning. In [10], test points are approximately represented as a linear combination of training and unlabeled points in the feature space induced by the kernel. We also note the very recent work [6] on out-of-sample extensions for semi-supervised learning. For Graph Regularization and Label Propagation see [12, 3, 18]. Manifold regularization provides natural out-of-sample extensions to several graph based approaches. These connections are summarized in the Table on page 5.

Other methods with different paradigms for using unlabeled data include Cotraining [8] and Bayesian Techniques, e.g., [14].

5 Experiments

We performed experiments on a synthetic dataset and two real world classification problems arising in visual and speech recognition. Comparisons are made with inductive methods (SVM, RLS). We also compare with Transductive SVM (e.g., [11]) based on our survey of related algorithms in Section 4. For all experiments, we constructed adjacency graphs with 6 nearest neighbors. Software and Datasets for these experiments are available at **http://manifold.cs.uchicago.edu/manifold_regularization**. More detailed experimental results are presented in [4].

5.1 Synthetic Data : Two Moons Dataset

The two moons dataset is shown in Figure 2. The best decision surfaces across a wide range of parameter settings are also shown for SVM, Transductive SVM and Laplacian SVM. The dataset contains 200 examples with only 1 labeled example for each class. Figure 2 demonstrates how TSVM fails to find the optimal solution. The Laplacian SVM decision boundary seems to be intuitively most satisfying. Figure 3 shows how increasing the intrinsic regularization allows effective use of unlabeled data for constructing classifiers.

5.2 Handwritten Digit Recognition

In this set of experiments we applied Laplacian SVM and Laplacian RLSC algorithms to 45 binary classification problems that arise in pairwise classification of handwritten digits. The first 400 images for each digit in the USPS training set (preprocessed using PCA to 100 dimensions) were taken to form the training set and 2 of these were randomly labeled. The remaining images formed the test set. Polynomial Kernels of degree 3 were used, and $\gamma l = 0.05 (C = 10)$ was set for inductive methods following experiments reported in [13]. For manifold regularization, we chose to split the same weight in the ratio $1 : 9$ so that $\gamma_A l = 0.005, \frac{\gamma_I l}{(u+l)^2} = 0.045$. The observations reported in this section hold consistently across a wide choice of parameters. In Figure 4, we compare the error rates of Laplacian algorithms, SVM and TSVM, at the precision-recall breakeven points in the ROC curves (averaged over 10 random choices of labeled examples) for the 45 binary classification problems. The following comments can be made: (a) Manifold regularization results in significant improvements over inductive classification, for both RLS and SVM, and either compares well or significantly outperforms TSVM across the 45 classification problems. Note that TSVM solves multiple quadratic programs in the size of the labeled and unlabeled sets whereas LapSVM solves a single QP in the size of the labeled set, followed by a linear system. This resulted in substantially faster training times for LapSVM in this experiment. (b) Scatter plots of performance on test and unlabeled data sets confirm that the out-of-sample extension is good for both LapRLS and LapSVM. (c) Finally, we found Laplacian algorithms to be significantly more stable with respect to choice of the labeled data than the inductive methods and TSVM, as shown in the scatter plot in Figure 4 on standard deviation of error rates. In Figure 5, we plot performance as a function of number of labeled examples.

5.3 Spoken Letter Recognition

This experiment was performed on the Isolet database of letters of the English alphabet spoken in isolation (available from the UCI machine learning repository). The data set contains utterances of 150 subjects who spoke the name of each letter of the English alphabet twice. The speakers are grouped into 5 sets of 30 speakers each, referred to as isolet1 through isolet5. For the purposes of this experiment, we chose to train on the first 30 speakers (isolet1) forming a training set of 1560 examples, and test on isolet5 containing 1559 examples (1 utterance is missing in the database due to poor recording). We considered the task of classifying the first 13 letters of the English alphabet from the last 13. The experimental set-up is meant to simulate a real-world situation: we considered 30 binary classification problems corresponding to 30 splits of the training data where all 52 utterances

Figure 2: Two Moons Dataset: Best decision surfaces using RBF kernels for SVM, TSVM and Laplacian SVM. Labeled points are shown in color, other points are unlabeled.

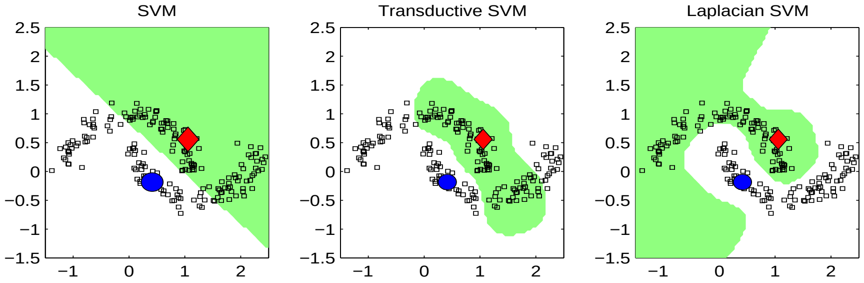


Figure 3: Two Moons Dataset: Laplacian SVM with increasing intrinsic regularization.

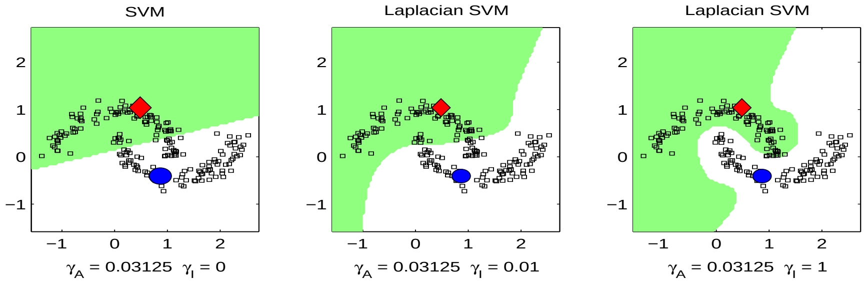


Figure 4: USPS Experiment - Error Rates at Precision-Recall Breakeven points for 45 binary classification problems

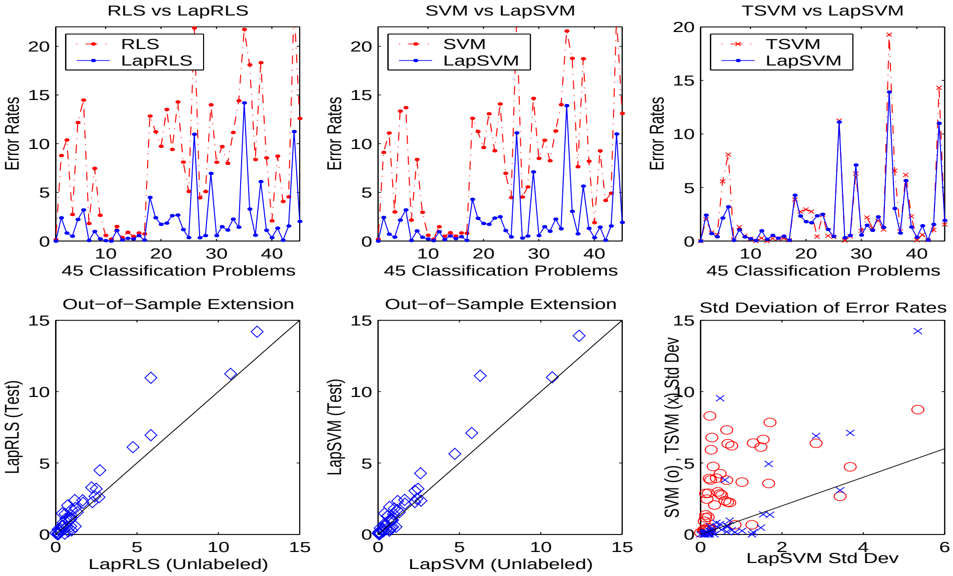


Figure 5: USPS Experiment - Mean Error Rate at Precision-Recall Breakeven points as a function of number of labeled points (T: Test Set, U: Unlabeled Set)

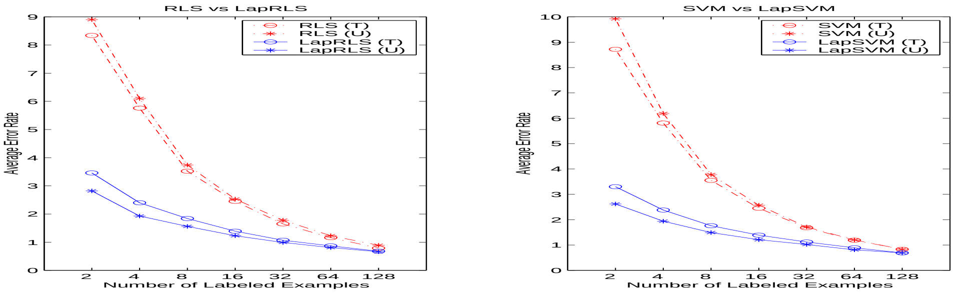
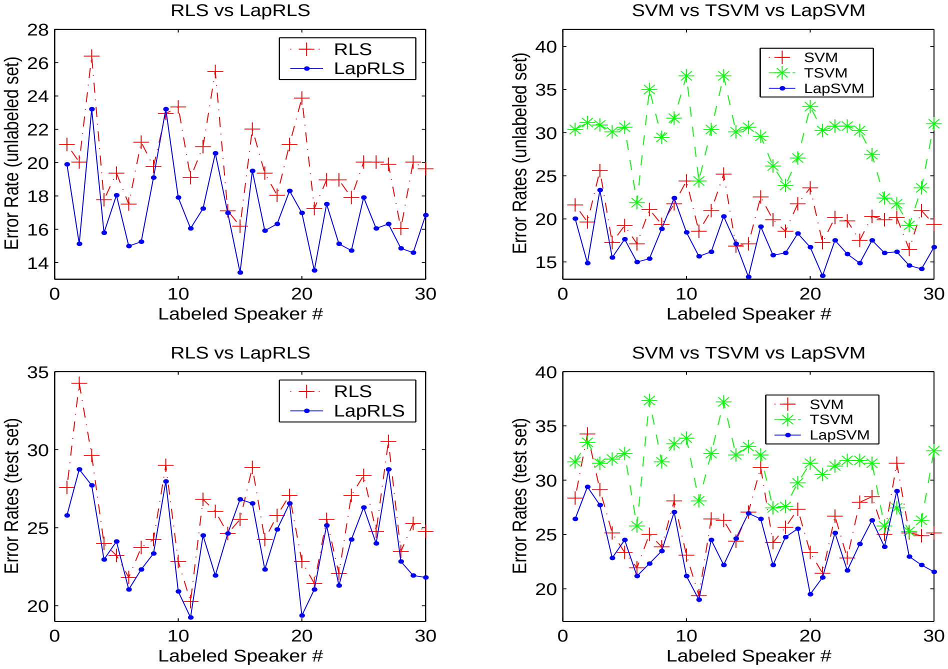


Figure 6: Isolet Experiment - Error Rates at precision-recall breakeven points of 30 binary classification problems



of one speaker were labeled and all the rest were left unlabeled. The test set is composed of entirely new speakers, forming the separate group isolet5.

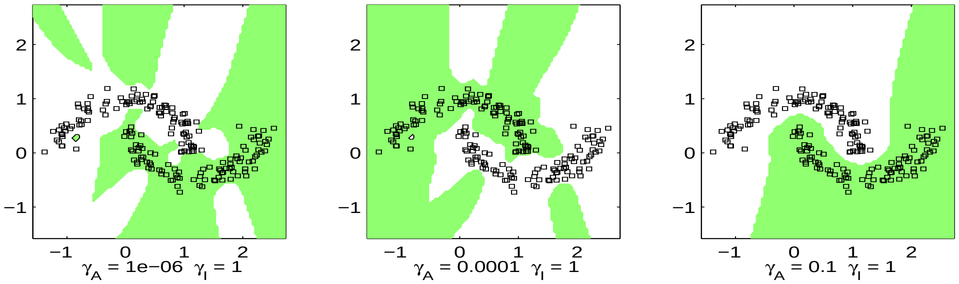
We chose to train with RBF kernels of width $\sigma = 10$ (this was the best value among several settings with respect to 5-fold cross-validation error rates for the fully supervised problem using standard SVM). For SVM and RLSC we set $\gamma l = 0.05$ ($C = 10$) (this was the best value among several settings with respect to mean error rates over the 30 splits). For Laplacian RLS and Laplacian SVM we set $\gamma_A l = \frac{\gamma_I l}{(u+l)^2} = 0.005$. In Figure 6, we compare these algorithms. The following comments can be made: (a) LapSVM and LapRLS make significant performance improvements over inductive methods and TSVM, for predictions on unlabeled speakers that come from the same group as the labeled speaker, over all choices of the labeled speaker. (b) On Isolet5 which comprises of a separate group of speakers, performance improvements are smaller but consistent over the choice of the labeled speaker. This can be expected since there appears to be a systematic bias that affects all algorithms, in favor of same-group speakers. For further details, see [4].

5.4 Regularized Spectral Clustering and Data Representation

When all training examples are unlabeled, the optimization problem of our framework, expressed in Eqn 6 reduces to the following clustering objective function :

$$\min_{f \in \mathcal{H}_K} \gamma \|f\|_K^2 + \hat{f}^T L \hat{f} \qquad (12)$$

Figure 7: Two Moons Dataset: Regularized Clustering



where $\gamma = \frac{\gamma_A u^2}{\gamma_I}$ is a regularization parameter that controls the complexity of the clustering function. To avoid degenerate solutions we need to impose some additional conditions (cf. [2]). It can be easily seen that a version of Representer theorem holds so that the minimizer has the form $f^* = \sum_{i=1}^{u} \alpha_i K(x_i, \cdot)$ By substituting back in Eqn. 12, we come up with the following optimization problem:

$$\alpha = \underset{\substack{\mathbf{1}^T K\alpha=0\\ \alpha^T K^2\alpha=1}}{\text{argmin}} \; \gamma\|f\|_K^2 + \hat{f}^T L \hat{f}$$

where $\mathbf{1}$ is the vector of all ones and $\alpha = (\alpha_1, \ldots, \alpha_u)$ and K is the corresponding Gram matrix.

Letting P be the projection onto the subspace of $\mathbb{R}^u$ orthogonal to $K\mathbf{1}$, one obtains the solution for the constrained quadratic problem, which is given by the generalized eigenvalue problem

$$P(\gamma K + KLK)P\mathbf{v} = \lambda P K^2 P\mathbf{v} \qquad (13)$$

The final solution is given by $\alpha = P\mathbf{v}$, where $\mathbf{v}$ is the eigenvector corresponding to the smallest eigenvalue.

The framework for clustering sketched above provides a regularized form spectral clustering, where γ controls the smoothness of the resulting function in the ambient space. We also obtain a natural out-of-sample extension for clustering points not in the original data set. Figure 7 shows the results of this method on a two-dimensional clustering problem.

By taking multiple eigenvectors of the system in Eqn. 13 we obtain a natural regularized out-of-sample extension of Laplacian eigenmaps [1]. This leads to new method for dimensionality reduction and data representation. Further study of this approach is a direction of future research.

Acknowledgments. We are grateful to Marc Coram, Steve Smale and Peter Bickel for intellectual support and to NSF funding for financial support. We would like to acknowledge the Toyota Technological Institute for its support for this work.

References

[1] M. Belkin, P. Niyogi, *Using Manifold Structure for Partially Labeled Classification*, NIPS 2002.

[2] M. Belkin, P. Niyogi, *Laplacian Eigenmaps for Dimensionality Reduction and Data Representation*, Neural Computation, June 2003

[3] M. Belkin, I. Matveeva, P. Niyogi, *Regression and Regularization on Large Graphs*, COLT 2004.

[4] M. Belkin, P. Niyogi, V. Sindhwani, *Manifold Regularization : A Geometric Framework for Learning From Examples*, Technical Report, Univ. of Chicago, Department of Computer Science, TR-2004-06. Available at : http://www.cs.uchicago.edu/research/publications/techreports/TR-2004-06

[5] K. Bennett and A. Demirez, *Semi-Supervised Support Vector Machines*, NIPS 1998

[6] Y. Bengio, O. Delalleau and N.Le Roux, *Efficient Non-Parametric Function Induction in Semi-Supervised Learning*, Technical Report 1247, DIRO, University of Montreal, 2004.

[7] A. Blum, S. Chawla, *Learning from Labeled and Unlabeled Data using Graph Mincuts*, ICML 2001.

[8] A. Blum, T. Mitchell, *Combining Labeled and Unlabeled Data with Co-Training*, COLT 1998

[9] O. Bousquet, O. Chapelle, M. Hein, *Measure Based Regularization*, NIPS 2003

[10] Chapelle, O., J. Weston and B. Schoelkopf, *Cluster Kernels for Semi-Supervised Learning*, NIPS 2002.

[11] T. Joachims, *Transductive Inference for Text Classification using Support Vector Machines*, ICML 1999.

[12] A. Smola and R. Kondor, *Kernels and Regularization on Graphs*, COLT/KW 2003.

[13] B. Schoelkopf, C.J.C. Burges, V. Vapnik, *Extracting Support Data for a Given Task*, KDD95.

[14] M. Seeger *Learning with Labeled and Unlabeled Data*, Tech Report. Edinburgh University (2000)

[15] Martin Szummer, Tommi Jaakkola, *Partially labeled classification with Markov random walks*, NIPS 2001,.

[16] V. Vapnik, *Statistical Learning Theory*, Wiley-Interscience, 1998.

[17] D. Zhou, O. Bousquet, T.N. Lal, J. Weston and B. Schoelkopf, *Learning with Local and Global Consistency*, NIPS 2003.

[18] X. Zhu, J. Lafferty and Z. Ghahramani, *Semi-supervised learning using Gaussian fields and harmonic functions*, ICML 2003.