Manifold Regularization

Vikas Sindhwani
Department of Computer Science
University of Chicago
Joint Work with Mikhail Belkin and Partha Niyogi
The Problem of Learning

- $S = \{x_i, y_i\}_{i=1}^{l}$ is drawn from an unknown probability distribution $P_{X \times Y}$.

- A Learning Algorithm maps $S$ to an element $f_S$ of a hypothesis space $\mathcal{H}$ of functions mapping $X \rightarrow Y$.

- $f_S$ should provide good labels for future examples.

- Regularization: Choose a simple function that agrees with data.
The Problem of Learning

Notions of simplicity are the key to successful learning. Here’s a simple function that agrees with data.
Learning and Prior Knowledge

But Simplicity is a Relative Concept. Prior Knowledge of the Marginal can modify our notions of simplicity.
Motivation

- How can we exploit prior knowledge of the marginal distribution $P_X$?
- More practically, how can we use unlabeled examples drawn from $P_X$?
- Why is this important?
  - Natural Data has structure to exploit.
  - Natural Learning is largely semi-supervised.
  - Labels are Expensive, Unlabeled data is cheap and plenty.
Contributions

- A data-dependent, Geometric Regularization Framework for Learning from examples.
- Representer Theorems provide solutions.
- Extensions of SVM and RLS for Semi-supervised Learning.
- Regularized Spectral Clustering and Dimensionality Reduction.
- The problem of Out-of-sample extensions in graph methods is resolved.
- Good Empirical Performance.
Regularization with RKHS

- Learning in Reproducing Kernel Hilbert Spaces:

\[
    f^* = \arg\min_{f \in \mathcal{H}_K} \frac{1}{l} \sum_{i=1}^{l} V(f(x_i), y_i) + \gamma \|f\|_K^2
\]

- Regularized Least Squares (RLS):
  \[
  V(f(x_i), y_i) = (y_i - f(x_i))^2
  \]

- Support Vector Machine (SVM):
  \[
  V(f(x_i), y_i) = \max[0, 1 - y_i f(x_i)]
  \]
What are RKHS?

- Hilbert Spaces with a nice property:
  - If two functions \( f, g \in \mathcal{H} \) are close in the distance derived from the inner product, their values \( f(x), g(x) \) are close \( \forall x \in X \).

- Reproducing Property:
  - \( \mathcal{E}_x : f \mapsto f(x) \) is linear, continuous. By Reisz’s Representation theorem,
    \[ \exists K(x, .) \in \mathcal{H} : \mathcal{E}_x(f) = \langle f, K_x \rangle_{\mathcal{H}} = f(x). \]

- Kernel Function \( \leftrightarrow \) RKHS:
  - \( K(x, t) = K_x(t) = \langle K_x, K_t \rangle \)
Why RKHS?

- Rich Function Spaces with complexity control
  - e.g. Gaussian Kernel $K(x, y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$
  - $\|f^*\|^2_K = \int |\tilde{f}(\omega)|^2 \|\omega\|^2 \, d\omega$

- Representer Theorems show that the minimizer has the form:
  - $f^*(.) = \sum_{i=1}^{l} \alpha_i K(x_i, .)$ and therefore,
  - $\|f^*\|^2_K = \langle f^*, f^* \rangle_{\mathcal{H}_K} = \sum_{i,j=1}^{l} \alpha_i \alpha_j K(x_i, x_j)$

- Motivates kernelization (KPCA, KFD, etc).
- Good empirical performance.
Known Marginal

If $\mathcal{P}_X$ is known, solve:

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

- Extrinsic and Intrinsic Regularization
  - $\gamma_A$ controls complexity in ambient space.
  - $\gamma_I$ controls complexity in the intrinsic geometry of $\mathcal{P}_X$
Continuous Representer Theorem

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 exists and admits the following representation

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

where $\mathcal{M} = \text{supp}\{\mathcal{P}_X\}$ is the support of the marginal $\mathcal{P}_X$. 
A Manifold Regularizer

If $\mathcal{M}$, the support of the marginal is a compact submanifold $\mathcal{M} \subset X = \mathbb{R}^n$, it seems natural to choose:

$$\| f \|_I^2 = \int_{\mathcal{M}} \langle \nabla_{\mathcal{M}} f, \nabla_{\mathcal{M}} f \rangle$$

and to find $f^* \in \mathcal{H}_K$ that minimizes:

$$\frac{1}{l} \sum_{i=1}^{l} V(f(x_i), y_i) + \gamma_A \| f \|_K^2 + \gamma_I \int_{\mathcal{M}} \langle \nabla_{\mathcal{M}} f, \nabla_{\mathcal{M}} f \rangle$$
Laplace Beltrami Operator

The intrinsic regularizer is a quadratic form involving the Laplace-Beltrami operator on the manifold $\mathcal{L} f \overset{\text{def}}{=} -\text{div} \nabla \mathcal{M} f$:

$$\| f \|_I = \int_{\mathcal{M}} \langle \nabla \mathcal{M} f, \nabla \mathcal{M} f \rangle = \int_{\mathcal{M}} f \mathcal{L} f$$

because some calculus on manifolds establishes that for any vector field $V(= \nabla \mathcal{M} f)$,

$$\int_{\mathcal{M}} \langle V, \nabla \mathcal{M} f \rangle = -\int_{\mathcal{M}} \text{div}(V) f$$
Passage to the Discrete

- In reality, $\mathcal{M}$ is unknown and sampled only via examples $\{x_i\}_{i=1}^{l+u}$. Labels are not required for empirical estimates of $\|f\|_I^2$.

- Manifold $\mathcal{M} \leftrightarrow$ Graph $\mathcal{G}(V, E)$
  
  $V = \{x_i\}_{i=1}^{l+u} \quad E = \{(x_i, x_j) : x_i \sim_{W_{ij}} x_j\}$.

- Laplace Beltrami $\mathcal{L} \leftrightarrow$ Graph Laplacian $L$
  
  $L \overset{\text{def}}{=} D - W \quad D = \text{diag}\{D_{ii} = \sum_j W_{ij}\}$.

- $\|f\|_I^2 = \int_{\mathcal{M}} f \mathcal{L} f \leftrightarrow \|f\|_I^2 = f^T L f = \sum(f(x_i) - f(x_j))^2 W_{ij}$
Algorithms

- We have motivated the following optimization problem: Find a function \( f^* \in \mathcal{H}_K \) that minimizes:

\[
\frac{1}{l} \sum_{i=1}^{l} V(f(x_i), y_i) + \gamma_A \| f \|_K^2 + \frac{\gamma I}{(l + u)^2} f^T L f
\]

- Laplacian RLS
  \[ V(f(x_i), y_i) = (y_i - f(x_i))^2 \]

- Laplacian SVM
  \[ V(f(x_i), y_i) = \max [0, 1 - y_i f(x_i)] \]
Empirical Representor Theorem

The minimizer admits an expansion

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

Proof:
Write any \( f \in \mathcal{H}_K \) as \( \sum_{i=1}^{l+u} \alpha_i K(x_i, x) + f_\perp \)

- \( f(x_j) = \langle f, Kx_j \rangle = \sum_{i=1}^{l+u} \alpha_i K(x_i, x_j) \)
- \( f_\perp \) increases the norm. So \( f^*_\perp = 0 \).
Laplacian RLS

By the Representer Theorem, the problem becomes finite dimensional. For Laplacian RLS, we find $\alpha^* \in \mathcal{R}^{l+u}$ that minimizes:

$$\frac{1}{l} \| Y - JK\alpha \|^2 + \gamma_A \alpha^T K\alpha + \frac{\gamma I}{(u+l)^2} \alpha^T K L K \alpha$$

where $K : \text{Gram Matrix}; Y = [y_1, \ldots, y_l, 0 \ldots, 0]$ and $J = \text{diag}(1, \ldots, 1, 0, \ldots, 0)$. The solution is:

$$\alpha^* = (JK + \gamma_A l I + \frac{\gamma I l}{(u+l)^2} L K)^{-1} Y$$
Laplacian SVM

For Laplacian SVMs, we solve a QP:

\[ \beta^* = \arg\max_{\beta \in \mathcal{R}} \sum_{i=1}^{l} \beta_i - \frac{1}{2} \beta^T Q \beta \]

subject to:

\[ \sum_{i=1}^{l} \beta_i y_i = 0 \]
\[ 0 \leq \beta_i \leq \frac{1}{l} \]

where

\[ Q = Y J K (2\gamma_A I + 2 \frac{\gamma I}{(l+u)^2} L K)^{-1} J^T Y \]

and then invert a linear system:

\[ \alpha^* = (2\gamma_A I + 2 \frac{\gamma I}{(u+l)^2} L K)^{-1} J^T Y \beta^* \]
Manifold Regularization

- **Input**: $l$ labeled and $u$ unlabeled examples
- **Output**: $f : \mathcal{R}^n \mapsto \mathcal{R}$
- **Algorithm**:
  - Construct adjacency Graph. Compute Laplacian.
  - Choose Kernel $K(x, y)$. Compute Gram matrix $K$.
  - Choose $\gamma_A, \gamma_I$. (?)
  - Compute $\alpha^*$.
  - Output $f^*(x) = \sum_{i=1}^{l+u} \alpha_i^* K(x_i, x)$
### Unity of Learning

#### Supervised

- **SVM/RLS**
  \[
  \arg\min_{f \in \mathcal{H}_K} \frac{1}{l} \sum_{i=1}^{l} V(y_i, f(x_i)) + \gamma \|f\|_K^2
  \]

#### Partially Supervised

- **Graph Regularization**
  \[
  \arg\min_{f \in \mathcal{R}^{l+u}} \frac{1}{l} \sum_{i=1}^{l} V(y_i, f_i) + \gamma f^T L f
  \]
  - **Out-of-sample Extn.**
    \[
    \arg\min_{f \in \mathcal{H}_K} \frac{1}{l} \sum_{i=1}^{l} V(y_i, f_i) + \gamma f^T L f
    \]

#### Unsupervised

- **Graph Mincut**
  \[
  \arg\min_{f \in \{-1,+1\}^u} \frac{1}{4} \sum_{i,j=1}^{u} w_{ij} (f_i - f_j)^2
  \]
  - **Spectral Clustering**
    \[
    \arg\min_{f \in \mathcal{R}^u} \frac{1}{2} f^T L f
    \]
    - **Out-of-sample Extn.**
      \[
      \arg\min_{f \in \mathcal{H}_K} \frac{1}{2} f^T L f
      \]
    - **Reg. Spectral Clust.**
      \[
      \arg\min_{f \in \mathcal{H}_K} \frac{1}{2} f^T L f + \gamma \|f\|_K^2
      \]
Regularized Spectral Clustering

Unsupervised Manifold Regularization:

\[ f^* = \arg\min_{f \in \mathcal{H}_K} \gamma \|f\|^2_K + f^T L f \]

\[ 1^T f = 0; \quad \|f\|_2^2 = 1 \]

Representer Theorem: \( f^*(x) = \sum_{i=1}^{u} \alpha_i^* K(x_i, x) \)
leads to an eigenvalue problem:

\[ P(\gamma K + KLK)Pv = \lambda PK^2Pv \]

and \( \alpha^* = Pv^* \). \( v^* \) is the smallest-eigenvalue eigenvector; \( P \) projects orthogonal to \( K1 \).
Experiments: Synthetic

- SVM
  - $\gamma_A = 0.03125$, $\gamma_I = 0$

- Laplacian SVM
  - $\gamma_A = 0.03125$, $\gamma_I = 0.01$
  - $\gamma_A = 0.03125$, $\gamma_I = 1$
  - $\gamma_A = 1 \times 10^{-6}$, $\gamma_I = 1$
  - $\gamma_A = 0.0001$, $\gamma_I = 1$
  - $\gamma_A = 0.1$, $\gamma_I = 1$
Related Algorithms

- **Transductive SVMs** [Joachims, Vapnik]
  \[
  f^* = \arg\min_{f \in \mathcal{H}_K, y_{l+1}, \ldots y_{l+u}} \ C \sum_{i=1}^{l} (1 - y_i f(x_i))_+ + C^* \sum_{i=l+1}^{l+u} (1 - y_i f(x_i))_+ + \| f \|_K^2
  \]

- **Semi-supervised SVMs** [Bennet, Fung et al]
  \[
  f^* = \arg\min_{f \in \mathcal{H}_K, y_{l+1}, \ldots y_{l+u}} \ C \sum_{i=0}^{l} (1 - y_i f(x_i))_+ + \ C \sum_{i=l+1}^{l+u} \min \{ (1 - f(x_i))_+, (1 + f(x_i))_+ \} + \| f \|_K^2
  \]

- **Measure-based Reg.** [Bousquet et al]
  \[
  f^* = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{l} V(f(x_i), y_i) + \gamma \int_X \langle \nabla f(x), \nabla f(x) \rangle p(x) dx
  \]
Experiments : Synthetic Data

SVM

Transductive SVM

Laplacian SVM

TTI-C Talk September 14, 2004 – p.24
Experiments: Digits

RLS vs LapRLS

Error Rates

45 Classification Problems

SVM vs LapSVM

Error Rates

45 Classification Problems

TSVM vs LapSVM

Error Rates

45 Classification Problems

Out-of-Sample Extension

LapRLS (Unlabeled) vs LapRLS (Test)

LapSVM (Unlabeled) vs LapSVM (Test)

Performance Deviation

SVM (o), TSVM (x) Deviation

LapSVM Deviation
Experiments: Digits

**RLS vs LapRLS**

- **Average Error Rate**
- Number of Labeled Examples

**SVM vs LapSVM**

- **Average Error Rate**
- Number of Labeled Examples
Experiments : Speech

![Graphs showing error rates for different methods and datasets.](image)
Experiments : Speech

![Graphs showing error rates for different methods and experiments.](image)
## Experiments: Text

<table>
<thead>
<tr>
<th>Method</th>
<th>PRBEP</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>k-NN</td>
<td>73.2</td>
<td>13.3</td>
</tr>
<tr>
<td>SGT</td>
<td>86.2</td>
<td>6.2</td>
</tr>
<tr>
<td>Naive-Bayes</td>
<td>—</td>
<td>12.9</td>
</tr>
<tr>
<td>Cotraining</td>
<td>—</td>
<td>6.20</td>
</tr>
<tr>
<td>SVM</td>
<td>76.39 (5.6)</td>
<td>10.41 (2.5)</td>
</tr>
<tr>
<td>TSVM</td>
<td>88.15 (1.0)</td>
<td>5.22 (0.5)</td>
</tr>
<tr>
<td>LapSVM</td>
<td>87.73 (2.3)</td>
<td>5.41 (1.0)</td>
</tr>
<tr>
<td>RLS</td>
<td>73.49 (6.2)</td>
<td>11.68 (2.7)</td>
</tr>
<tr>
<td>LapRLS</td>
<td>86.37 (3.1)</td>
<td>5.99 (1.4)</td>
</tr>
</tbody>
</table>
Experiments: Text

Performance of RLS, LapRLS

Performance of SVM, LapSVM

LapSVM performance (Unlabeled)

LapSVM performance (Test)
Future Work

- Generalization as a function of labeled and unlabeled examples.
- Additional Structure: Structured Outputs, Invariances
- Active Learning, Feature Selection
- Efficient Algorithms: Linear Methods, Sparse Solutions
- Applications: Bioinformatics, Text, Speech, Vision, ...