Geometry of Semi-Supervised Learning

Mikhail Belkin
Partha Niyogi

The University of Chicago

Collaborators: V. Sindhwani, I. Matveeva, D. Surendran
Machine learning vs human learning

Human learning:

- Complex stimuli.
- Impoverished inputs.
- Robust.
- Extensive use of prior knowledge.
- Learning through mostly unlabeled data. Inference from few labeled examples.
Machine learning vs human learning

Human learning:

- Complex stimuli.
- Impoverished inputs.
- Robust.
- Extensive use of prior knowledge.
- Learning through mostly unlabeled data. Inference from few labeled examples.

Machine learning: (almost) none of the above!

Unlabeled data.
Machine learning vs human learning

Human learning:
- Complex stimuli.
- Impoverished inputs.
- Robust.
- Extensive use of prior knowledge.
- Learning through mostly unlabeled data. Inference from few labeled examples.

Machine learning: (almost) none of the above!

Unlabeled data.
Reasons to use unlabeled data in inference:

► Pragmatic:

Unlabeled data is everywhere. Need a way to use it.

► Philosophical:

The brain uses unlabeled data.
Intuition

Geometry of data changes our notion of similarity.
Intuition

Geometry of data changes our notion of similarity.
Intuition

Geometry of data changes our notion of similarity.
Intuition

Geometry of data changes our notion of similarity.
Manifold assumption

Geometry is important.
Manifold assumption

Geometry is important.
Geometry is important.
Cluster assumption
Cluster assumption
Geometry is important.
Geometry is important.
Unlabeled data to estimate geometry.
Manifold/geometric assumption: functions of interest are smooth with respect to the underlying geometry.
Manifold assumption

**Manifold/geometric assumption:** functions of interest are smooth with respect to the underlying geometry.

Probabilistic setting: Map $X \rightarrow Y$. Probability distribution $P$ on $X \times Y$.

Regression/(two class)classification: $X \rightarrow \mathbb{R}$. 
Manifold/geometric assumption: functions of interest are smooth with respect to the underlying geometry.

Probabilistic setting: Map $X \rightarrow Y$. Probability distribution $P$ on $X \times Y$.

Regression/(two class)classification: $X \rightarrow \mathbb{R}$.

Probabilistic version: conditional distributions $P(y|x)$ are smooth with respect to the marginal $P(x)$. 

Manifold assumption
What is smooth?

Function $f : X \rightarrow \mathbb{R}$. Penalty at $x \in X$:

$$\frac{1}{\delta k} \int_{\text{small } \delta} (f(x) - f(x + \delta))^2 p(x) d\delta \approx \|\nabla f\|^2 p(x)$$

Total penalty – Laplace operator:

$$\int_X \|\nabla f\|^2 p(x) = \langle f, \mathcal{L}_p f \rangle_X$$
What is smooth?

Function $f : X \rightarrow \mathbb{R}$. Penalty at $x \in X$:

$$\frac{1}{\delta^k} \int_{\text{small } \delta} (f(x) - f(x + \delta))^2 p(x) d\delta \approx \|\nabla f\|^2 p(x)$$

Total penalty – Laplace operator:

$$\int_X \|\nabla f\|^2 p(x) = \langle f, \mathcal{L}_p f \rangle_X$$

Two-class classification – conditional $P(1|x)$.

Manifold assumption: $\langle P(1|x), \mathcal{L}_p P(1|x) \rangle_X$ is small.
Algorithmic framework: Laplacian

Natural smoothness functional (analogue of $\text{grad}$):

$$S(f) = (f_1 - f_2)^2 + (f_1 - f_3)^2 + (f_2 - f_3)^2 + (f_3 - f_4)^2 + (f_4 - f_5)^2 + (f_4 - f_5)^2 + (f_5 - f_6)^2$$

Basic fact:

$$S(f) = \sum_{i \sim j} (f_i - f_j)^2 = \frac{1}{2} f^t L f$$
Algorithmic framework
Algorithmic framework

\[ W_{ij} = e^{-\frac{||x_i - x_j||^2}{t}} \]

\[ Lf(x_i) = f(x_i) \sum_j e^{-\frac{||x_i - x_j||^2}{t}} - \sum_j f(x_j) e^{-\frac{||x_i - x_j||^2}{t}} \]

\[ f^t Lf = 2 \sum_{i \sim j} e^{-\frac{||x_i - x_j||^2}{t}} (f_i - f_j)^2 \]
Algorithmic framework

\[ W_{ij} = e^{-\frac{||x_i-x_j||^2}{t}} \]

\[ L f(x_i) = f(x_i) \sum_j e^{-\frac{||x_i-x_j||^2}{t}} - \sum_j f(x_j) e^{-\frac{||x_i-x_j||^2}{t}} \]

\[ f^t L f = 2 \sum_{i \sim j} e^{-\frac{||x_i-x_j||^2}{t}} (f_i - f_j)^2 \]
Semi-supervised learning

Learning from labeled and unlabeled data.

- Unlabeled data is everywhere. Need to use it.
- Natural learning is semi-supervised.

Labeled data: \((x_1, y_1), \ldots, (x_l, y_l) \in \mathbb{R}^N \times \mathbb{R}\)
Unlabeled data: \(x_{l+1}, \ldots, x_{l+u} \in \mathbb{R}^N\)

Need to reconstruct

\[ f_{L,U} : \mathbb{R}^N \rightarrow \mathbb{R} \]
Regularization

Estimate $f : \mathbb{R}^N \rightarrow \mathbb{R}$

Data: $(x_1, y_1), \ldots, (x_l, y_l)$

Regularized least squares (hinge loss for SVM):

$$f^* = \arg\min_{f \in \mathcal{H}} \frac{1}{l} \sum (f(x_i) - y_i)^2 + \lambda \|f\|_K^2$$

fit to data + smoothness penalty

$\|f\|_K$ incorporates our smoothness assumptions. Choice of $\|f\|_K$ is important.
Algorithm: RLS/SVM

Solve: \[ f^* = \arg\min_{f \in \mathcal{H}} \frac{1}{l} \sum_{i=1}^{l} (f(x_i) - y_i)^2 + \lambda \| f \|_K^2 \]

\( \| f \|_K \) is a Reproducing Kernel Hilbert Space norm with kernel \( K(x, y) \).

Can solve explicitly (via Representer theorem):

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

\[ [\alpha_1, \ldots, \alpha_l]^t = (K + \lambda I)^{-1} [y_1, \ldots, y_l]^t \]

\( (K)_{ij} = K(x_i, x_j) \)
$\gamma_A = 0.03125 \quad \gamma_I = 0$

Toy example
Toy example

SVM

Laplacian SVM

Laplacian SVM

$\gamma_A = 0.03125 \quad \gamma_i = 0$

$\gamma_A = 0.03125 \quad \gamma_i = 0.01$

$\gamma_A = 0.03125 \quad \gamma_i = 1$
Manifold regularization

Data space $X$.

$$
\begin{align*}
    f^* &= \arg\min_{f \in \mathcal{H}} \frac{1}{l} \sum_{i=1}^{l} (f(x_i) - y_i)^2 + \lambda_A \|f\|_K^2 + \lambda_I \|f\|_I^2 \\
    &= \text{fit to data + extrinsic smoothness + intrinsic smoothness}
\end{align*}
$$
Data space $X$.

$$f^* = \arg\min_{f \in \mathcal{H}} \frac{1}{l} \sum_{i=1}^{l} (f(x_i) - y_i)^2 + \lambda_A \| f \|_K^2 + \lambda_I \| f \|_I^2$$

fit to data + extrinsic smoothness + intrinsic smoothness

$$\| f \|_I^2 = \langle f, Df \rangle \quad D : \text{RKHS} \rightarrow L^2 \text{ is bounded}.$$ 

**Theorem** [Intrinsic Representer theorem]

$$f^*(\cdot) = \sum_{i=1}^{l} \alpha_i K(x_i, \cdot) + \int_X \alpha(x) K(x, \cdot) \, d\mu_x$$
What is the nature of $\|f\|_I$?

For example:

$$\|f\|_I^2 = \int_X \|\nabla_X f\|_X^2 \, d\mu_X$$

Any differential operator on the space $X$, e.g. $\mathcal{L}^n$. Diffusions and other kernels on the manifold.

**Problem:** $X$ is usually not known!
Data-dependent regularization

Estimate $f : \mathbb{R}^N \rightarrow \mathbb{R}$

Labeled data: $(x_1, y_1), \ldots, (x_l, y_l)$
Unlabeled data: $x_{l+1}, \ldots, x_{l+u}$

$$f^* = \arg \min_{f \in \mathcal{H}} \frac{1}{l} \sum_{i=1}^{l} (f(x_i) - y_i)^2 + \lambda_A \|f\|_K^2 + \lambda_I \|f\|_I^2$$

Empirical estimate:

$$\|f\|_I^2 = \frac{1}{(l + u)^2} [f(x_1), \ldots, f(x_{l+u})] L [f(x_1), \ldots, f(x_{l+u})]^t$$
Representer theorem (discrete case):

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

Explicit solution for quadratic loss:

\[ \tilde{\alpha} = (JK + \lambda_A l I + \frac{\lambda I l}{(u + l)^2} L K)^{-1} [y_1, \ldots, y_l, 0, \ldots, 0]^t \]

\[ (K)_{ij} = K(x_i, x_j), \quad J = \text{diag}(1, \ldots, 1, 0, \ldots, 0) \]
Laplacian Regularized Least Squares demo [link]

Available at
http://people.cs.uchicago.edu/~mrainey/jlapvis/JLapVis.html
Experimental results: USPS

- RLS vs LapRLS
- SVM vs LapSVM
- TSVM vs LapSVM

Error Rates vs Classification Problems

Out-of-Sample Extension

Std Deviation of Error Rates

SVM (o), TSVM (x), LapSVM Std Dev
## Experimental comparisons

<table>
<thead>
<tr>
<th>Dataset</th>
<th>g50c</th>
<th>Coil20</th>
<th>Uspst</th>
<th>mac-win</th>
<th>WebKB (link)</th>
<th>WebKB (page)</th>
<th>WebKB (page+link)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM (full labels)</td>
<td>3.82</td>
<td>0.0</td>
<td>3.35</td>
<td>2.32</td>
<td>6.3</td>
<td>6.5</td>
<td>1.0</td>
</tr>
<tr>
<td>RLS (full labels)</td>
<td>3.82</td>
<td>0.0</td>
<td>2.49</td>
<td>2.21</td>
<td>5.6</td>
<td>6.0</td>
<td>2.2</td>
</tr>
<tr>
<td>SVM (l labels)</td>
<td>8.32</td>
<td>24.64</td>
<td>23.18</td>
<td>18.87</td>
<td>25.6</td>
<td>22.2</td>
<td>15.6</td>
</tr>
<tr>
<td>RLS (l labels)</td>
<td>8.28</td>
<td>25.39</td>
<td>22.90</td>
<td>18.81</td>
<td>28.0</td>
<td>28.4</td>
<td>21.7</td>
</tr>
<tr>
<td>Graph-Reg</td>
<td>17.30</td>
<td>6.20</td>
<td>21.30</td>
<td>11.71</td>
<td>22.0</td>
<td>10.7</td>
<td>6.6</td>
</tr>
<tr>
<td>TSVM</td>
<td>6.87</td>
<td>26.26</td>
<td>26.46</td>
<td>7.44</td>
<td>14.5</td>
<td>8.6</td>
<td>7.8</td>
</tr>
<tr>
<td>Graph-density</td>
<td>8.32</td>
<td>6.43</td>
<td>16.92</td>
<td>10.48</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>▽TSVM</td>
<td>5.80</td>
<td>17.56</td>
<td>17.61</td>
<td>5.71</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>LDS</td>
<td>5.62</td>
<td>4.86</td>
<td>15.79</td>
<td>5.13</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>LapSVM</td>
<td>5.44</td>
<td>3.66</td>
<td>12.67</td>
<td>10.41</td>
<td>18.1</td>
<td>10.5</td>
<td>6.4</td>
</tr>
<tr>
<td>LapRLS</td>
<td>5.18</td>
<td>3.36</td>
<td>12.69</td>
<td>10.01</td>
<td>19.2</td>
<td>11.0</td>
<td>6.9</td>
</tr>
<tr>
<td>LapSVM_{joint}</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>5.7</td>
<td>6.7</td>
<td>6.4</td>
</tr>
<tr>
<td>LapRLS_{joint}</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>5.6</td>
<td>8.0</td>
<td>5.8</td>
</tr>
</tbody>
</table>
Continuous spectral clustering

Isoperimetric inequalities. Cheeger constant.

\[ h = \inf \frac{\text{vol}^{n-1}(\delta M_1)}{\min(\text{vol}^n(M_1), \text{vol}^n(M - M_1))} \]

\[ h \leq \frac{\sqrt{\lambda_1}}{2} \]  

[Cheeger]
Spectral clustering

\[ L = \begin{pmatrix} 
2 & -1 & -1 & 0 & 0 & 0 \\
-1 & 2 & -1 & 0 & 0 & 0 \\
-1 & -1 & 3 & -1 & 0 & 0 \\
0 & 0 & -1 & 3 & -1 & -1 \\
0 & 0 & 0 & -1 & 2 & -1 \\
0 & 0 & 0 & -1 & -1 & 2 
\end{pmatrix} \]
Spectral clustering

Unnormalized clustering:

\[ L \mathbf{e}_1 = \lambda_1 \mathbf{e}_1 \quad \mathbf{e}_1 = [-0.46, -0.46, -0.26, 0.26, 0.46, 0.46] \]
Spectral clustering

Unnormalized clustering:

$$Le_1 = \lambda_1 e_1 \quad e_1 = [-0.46, -0.46, -0.26, 0.26, 0.46, 0.46]$$

$$L = \begin{pmatrix} 2 & -1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & -1 \\ 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & -1 & 2 \end{pmatrix}$$
Spectral clustering

Unnormalized clustering:

\[ L e_1 = \lambda_1 e_1 \quad e_1 = [-0.46, -0.46, -0.26, 0.26, 0.46, 0.46] \]

Normalized clustering:

\[ L e_1 = \lambda_1 D e_1 \quad e_1 = [-0.31, -0.31, -0.18, 0.18, 0.31, 0.31] \]

\[ L = \begin{pmatrix}
2 & -1 & -1 & 0 & 0 & 0 \\
-1 & 2 & -1 & 0 & 0 & 0 \\
-1 & -1 & 3 & -1 & 0 & 0 \\
0 & 0 & -1 & 3 & -1 & -1 \\
0 & 0 & 0 & -1 & 2 & -1 \\
0 & 0 & 0 & -1 & -1 & 2
\end{pmatrix} \]
Regularized spectral clustering

\[ f^* = \arg\min_{f \in \mathcal{H}_K} \lambda \|f\|_K^2 + \sum_{i \sim j} (f(x_i) - f(x_j))^2 \]

Representer theorem:

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

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

\[ (\alpha_1, \ldots, \alpha_u) = Pv \]

Out-of-sample extension for spectral clustering.

Belkin Niyogi Sindhwani 04 Related work: Bengio, et al 04, Vert, Yamanishi 04
Regularized spectral clustering

\[ \gamma_A = 1e^{-06} \quad \gamma_I = 1 \]

\[ \gamma_A = 0.0001 \quad \gamma_I = 1 \]

\[ \gamma_A = 0.1 \quad \gamma_I = 1 \]