A Geometric Perspective on Data Analysis

Partha Niyogi

The University of Chicago

Collaborators: M. Belkin, X. He, A. Jansen, H. Narayanan, V. Sindhwani, S. Smale, S. Weinberger
Manifold Learning

Learning when data $\sim \mathcal{M} \subset \mathbb{R}^N$

- **Clustering:** $\mathcal{M} \rightarrow \{1, \ldots, k\}$
  connected components, min cut

- **Classification:** $\mathcal{M} \rightarrow \{-1, +1\}$
  $P$ on $\mathcal{M} \times \{-1, +1\}$

- **Dimensionality Reduction:** $f : \mathcal{M} \rightarrow \mathbb{R}^n \quad n << N$

- **$\mathcal{M}$ unknown:** what can you learn about $\mathcal{M}$ from data?
  e.g. dimensionality, connected components
  holes, handles, homology
  curvature, geodesics
An Acoustic Example

\[ u(t) \rightarrow l \leftarrow s(t) \]
An Acoustic Example

One Dimensional Air Flow

\[ (i) \quad \frac{\partial V}{\partial x} = -\frac{A}{\rho c^2} \frac{\partial P}{\partial t} \]

\[ (ii) \quad \frac{\partial P}{\partial x} = -\frac{\rho}{A} \frac{\partial V}{\partial t} \]

\[ V(x, t) = \text{volume velocity} \]

\[ P(x, t) = \text{pressure} \]
\[ u(t) = \sum_{n=1}^{\infty} \alpha_n \sin(n\omega_0 t) \in l_2 \]

\[ s(t) = \sum_{n=1}^{\infty} \beta_n \sin(n\omega_0 t) \in l_2 \]
Vocal Tract modeled as a sequence of tubes. (e.g. Stevens, 1998)
\( P \) on \( X \times Y \)

\[ X = \mathbb{R}^N; \quad Y = \{0, 1\}, \mathbb{R} \]

\((x_i, y_i)\) labeled examples

find \( f : X \rightarrow Y \) \hspace{1cm} \text{III Posed}
Regularization Principle

\[ f = \arg \min_{f \in H_K} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \gamma \| f \|_K^2 \]

Splines
Ridge Regression
SVM

- \( K : X \times X \rightarrow \mathbb{R} \) is a p.d. kernel
  - e.g. \( e^{-\|x-y\|^2 / \sigma^2} \), \((1 + x \cdot y)^d\), etc.

- \( H_K \) is a corresponding RKHS
  - e.g., certain Sobolev spaces, polynomial families, etc.
Simplicity is Relative
Simplicity is Relative
Intuitions

- $\text{supp } P_X$ has manifold structure

- *geodesic* distance v/s *ambient* distance

- geometric structure of data should be incorporated

- $f$ versus $f_M$
Manifold Regularization

\[
\min_{f \in H_K} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \gamma_A \|f\|^2_K + \gamma_I \|f\|^2_I
\]

\[\|f\|^2_I = \begin{cases} 
\text{Laplacian} & \int \langle \text{grad}_M f, \text{grad}_M f \rangle = \int f \Delta_M f \\
\text{Iterated Laplacian} & \int f \Delta_M^i f \\
\text{Heat kernel} & e^{-\Delta_M t} \\
\text{Differential Operator} & \int f(Df) 
\end{cases}\]

Representer Theorem: \( f = \sum_{i=1}^{n} \alpha_i K(x, x_i) + \int_M \alpha(y) K(x, y) \)

Belkin, Niyogi, Sindhwani (2004)
$\mathcal{M}$ is unknown but $x_1 \ldots x_M \in \mathcal{M}$

$$\|f\|_I^2 = \int_\mathcal{M} \langle \nabla_\mathcal{M} f, \nabla_\mathcal{M} f \rangle \approx \sum_{i \sim j} W_{ij} (f(x_i) - f(x_j))^2$$
\[ M \approx G = (V, E) \]

\[ e_{ij} \in E \text{ if } \|x_i - x_j\| < \epsilon \]

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

\[ \Delta_M \approx L = D - W \]

\[ \int \langle \text{grad} f, \text{grad} f \rangle \approx \sum_{i,j} W_{ij} (f(x_i) - f(x_j))^2 \]

\[ \int f(\Delta f) \approx f^T L f \]
\[
\frac{1}{n} \sum_{i=1}^{n} V(f(x_i), y_i) + \gamma_A \| f \|_K^2 + \gamma_I \sum_{i \sim j} W_{ij} (f(x_i) - f(x_j))^2
\]

**Representer Theorem:**
\[
f_{opt} = \sum_{i=1}^{n+m} \alpha_i K(x, x_i)
\]

\(V(f(x), y) = (f(x) - y)^2\): Least squares

\(V(f(x), y) = (1 - yf(x))_+\): Hinge loss (Support Vector Machines)
Ambient and Intrinsic Regularization

SVM

γ_A = 0.03125  γ_I = 0

Laplacian SVM

γ_A = 0.03125  γ_I = 0.01

Laplacian SVM

γ_A = 0.03125  γ_I = 1
Experimental Results: USPS

1. RLS vs LapRLS
2. SVM vs LapSVM
3. TSVM vs LapSVM

Out-of-Sample Extension

Std Deviation of Error Rates

SVM (○), TSVM (×) Std Dev

LapSVM (Unlabeled) vs LapSVM (Test)
Experimental Results: Isolet

![Graphs showing error rates for RLS vs LapRLS and SVM vs TSVM vs LapSVM for unlabeled and test sets.](image)
## Experimental comparisons

<table>
<thead>
<tr>
<th>Dataset → Algorithm</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>( \text{LapSVM}_{\text{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>( \text{LapRLS}_{\text{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>
Graph and Manifold Laplacian

Fix $f : X \to \mathbb{R}$.
Fix $x \in \mathcal{M}$

$$(L_nf) = \sum_j (f(x) - f(x_j)) e^{-\frac{\|x-x_j\|^2}{4t_n}}$$

Put $t_n = n^{-k-2-\alpha}$, where $\alpha > 0$

with prob. 1, $\lim_{n \to \infty} \frac{(4\pi t_n)^{-\frac{k+1}{2}}}{n} (L_nf)|x = \Delta_M f |x$

Belkin (2003), Belkin and Niyogi (2004,2005)
also Lafon (2004), Coifman et al
$x_1, \ldots, x_n \in \mathcal{M} \subset \mathbb{R}^N$

Can you learn **qualitative** features of $\mathcal{M}$?

- Can you tell a torus from a sphere?
- Can you tell how many connected components?
- Can you tell the dimension of $\mathcal{M}$?

(e.g. Carlsson, Zamorodian, Edelsbrunner, de Silva et al)
$x_1, \ldots, x_n \in \mathcal{M} \subset \mathbb{R}^N$

$U = \bigcup_{i=1}^{n} B_\epsilon(x_i)$

If $\epsilon$ well chosen, then $U$ deformation retracts to $\mathcal{M}$.

Homology of $U$ is constructed using the nerve of $U$ and agrees with the homology of $\mathcal{M}$.
\( \mathcal{M} \subset \mathbb{R}^d \) with cond. no. \( \tau \\
\bar{x} = \{x_1, \ldots, x_n\} \sim \) uniformly sampled i.i.d. 
\( 0 < \epsilon < \frac{\tau}{2} \)

\[ \beta = \frac{\text{vol}(\mathcal{M})}{(\sin^{-1}(\epsilon/2\tau))^k \text{vol}(B_\epsilon/8)} \]

Let \( U = \bigcup_{x \in \bar{x}} B_\epsilon(x) \)

\[ n > \beta (\log(\beta) + \log(\frac{1}{\delta})) \]

with prob. \( > 1 - \delta \), 
homology of \( U \) equals the homology of \( \mathcal{M} \)

(Niyogi, Smale, Weinberger, 2004)
Spectral Clustering

Isoperimetric inequalities. Cheeger constant.

\[ \delta M_1 = \min \left( \frac{\text{vol}^{n-1}(\delta M_1)}{\min(\text{vol}^n(M_1), \text{vol}^n(M - M_1))}, \sqrt{\lambda_1} \right) \leq \frac{\sqrt{\lambda_1}}{2} \]

[Cheeger]
Clustering Digits

Laplacian Eigenmap

PCA
Clustering Speech
Volume of a convex body is deterministically hard to compute. (Bárány and Füredi, 1987).


Current best $O^*(d^4)$ by Lovász and Vempala.
Volume of a convex body is deterministically hard to compute. (Bárány and Füredi, 1987).


Current best $O^*(d^4)$ by Lovász and Vempala.

Surface volume at least as hard as volume. Grötschel, Lovász and Schrijver (1987) list as open problem.

Heat Equation $\Delta u = u_t$ with $u(x, 0) = f(x)$

Solution $u(x, t) = f(x) \ast K_t(x, y)$

$$F_t(M) = \sqrt{\frac{\pi}{t}} \int_{\mathbb{R}^d \setminus M} u(x, t)dx$$
[Theorem 1] \( \lim_{t \to 0} F_t(M) = |\partial M| \)

[Theorem 2] Let \( M \) be a convex body in \( \mathbb{R}^d \) such that

(i) \( B_r \subset M \subset B_R \)

(ii) \( \partial M \) has condition \( 1/\tau \)

Then, it is possible to find the surface area of \( M \) in time

\[
O^* \left( \frac{d^4}{\epsilon^2} + \frac{d^{3.5} R^3}{r^2 \tau \epsilon^3} \right),
\]

with error of \( \epsilon \) with prob. \( > 3/4 \).

Belkin, Narayanan, Niyogi (2005)
Important Issues

- How to handle noise theoretically and practically?
- How to choose the graph neighborhood correctly?
- How often do manifolds arise in natural data? What is the right metric on these manifolds?
- What are other ways in which one might utilize the geometry of natural distributions?
- Identify real problems where this approach can make a difference.
- Complexity estimates and provably correct algorithms rather than heuristics.