Data, Structure and Geometry in Statistical Learning

John Lafferty

School of Computer Science
Carnegie Mellon University
“The coming century is surely the century of data. Our society is investing massively in the collection and processing of data of all kinds, on scales unimaginable until recently.”

—David Donoho (August 2000)

- Satellite Images
- DNA sequences
- Internet portals
- Sensor networks
- Text and hypertext
- Financial transactions
- Environment monitoring
- Astrostatistics...
Data Maven: IBM’s John Cocke

- Language models from data
- Machine translation data
- Optimizing compilers
A DISCOURSE

Concerning GRAVITY, and its Properties, wherein the Descent of Heavy Bodies, and the Motion of Project is briefly, but fully handled: Together with the Solution of a Problem of great Use in GUN- NERY. BY E. HALLEY.

NATURE amidst the great variety of Problems where-with she exercises the Wits of Philosophical men, scarce affords any one wherein the Effect is more visible, and the Cause more concealed than in those of the Phenomena of Gravity. Before we can go alone, we must learn to defend our selves from the violence of its Impulse, by not trusting the Center of Gravity of our Bodies beyond our reach; and yet the Acuteest Philosophers, and the Subtilest Enquirers into the Original of this Motion, have been so far from satisfying their Readers, that they themselves seem little to have understood the Consequences of their own Hypotheses.

Descartes his Notion, I must needs confess to be to me incomprehensible, while he will have the Particles of his Celestial matter, by being reflected on the Surface of the Earth, and so ascending therefrom, to drive down into their places those Territorial Bodies they find above them: This is as near as I can gather the Scope of the 20th, 21st, 22nd, and 23rd Sections of the last Book of his Principia Philosophia; yet neither he, nor any of his Followers can shew how a Body suspended in visuo athere, shall be carried downwards by a conti-
Relating Phenotype and Genotype

### Table: Gene Expression

<table>
<thead>
<tr>
<th>Gene/Protein</th>
<th>Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-cmyB (J22379)</td>
<td>0.12</td>
</tr>
<tr>
<td>Protoncombinator 1 (X59417)</td>
<td>0.05</td>
</tr>
<tr>
<td>MB f (E82350)</td>
<td>0.02</td>
</tr>
<tr>
<td>Cyclin D3 (M02297)</td>
<td>0.01</td>
</tr>
<tr>
<td>Myosin light chain (M31211)</td>
<td>0.00</td>
</tr>
<tr>
<td>RhoA (J24352)</td>
<td>-0.03</td>
</tr>
<tr>
<td>SNF2 (J29656)</td>
<td>-0.05</td>
</tr>
<tr>
<td>Brd1 (J50225)</td>
<td>-0.10</td>
</tr>
<tr>
<td>E2A (M15329)</td>
<td>-0.15</td>
</tr>
<tr>
<td>Inducible protein (J77738)</td>
<td>-0.20</td>
</tr>
<tr>
<td>Dynamin light chain (J39944)</td>
<td>-0.25</td>
</tr>
<tr>
<td>Topoisomerase II (JZ15115)</td>
<td>-0.30</td>
</tr>
<tr>
<td>THRI (M63469)</td>
<td>-0.35</td>
</tr>
<tr>
<td>Acyl-Coenzyme A dehydrogenase (M91432)</td>
<td>-0.40</td>
</tr>
<tr>
<td>SNF2 (J29775)</td>
<td>-0.45</td>
</tr>
<tr>
<td>Ca2+ATPase (J99801)</td>
<td>-0.50</td>
</tr>
<tr>
<td>SRRP (J29998)</td>
<td>-0.55</td>
</tr>
<tr>
<td>MCM4 (J58073)</td>
<td>-0.60</td>
</tr>
<tr>
<td>Deoxynucleosynthetic (U26306)</td>
<td>-0.65</td>
</tr>
<tr>
<td>Ov 18 (M31391)</td>
<td>-0.70</td>
</tr>
<tr>
<td>Rubidipin 5 (M06812)</td>
<td>-0.75</td>
</tr>
<tr>
<td>Heterochromatin protein (J35451)</td>
<td>-0.80</td>
</tr>
<tr>
<td>IL-7 (M99666)</td>
<td>-0.85</td>
</tr>
<tr>
<td>Adenosine deaminase (M17592)</td>
<td>-0.90</td>
</tr>
</tbody>
</table>

### Baby Chart

<table>
<thead>
<tr>
<th>Month</th>
<th>Name</th>
<th>Sex</th>
<th>Birth Weight</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>June</td>
<td>James</td>
<td>M</td>
<td>3.2 kg</td>
<td>56 cm</td>
</tr>
<tr>
<td>July</td>
<td>Sarah</td>
<td>F</td>
<td>3.5 kg</td>
<td>58 cm</td>
</tr>
<tr>
<td>August</td>
<td>Thomas</td>
<td>M</td>
<td>3.8 kg</td>
<td>60 cm</td>
</tr>
<tr>
<td>September</td>
<td>David</td>
<td>M</td>
<td>4.0 kg</td>
<td>62 cm</td>
</tr>
</tbody>
</table>

**Remarks:**
- James was born with a minor birthmark on his left arm.
- Sarah was delivered by cesarean section due to breech presentation.
- Thomas had a few transient irritability episodes that resolved by the end of the month.
- David was exclusively breastfed for the first week and gradually transitioned to formula.

**Adm. No.:** 23 8 15
“Sitting inside Akamai’s Network Operations Command Center, the command room for 15,000 high-speed servers stationed around the globe, (gives) a God’s-eye view of the Internet, monitoring its health in real time.”

*Wired*, July 2003
New York Times, October 12, 2004:

“At St. Helens, Little Lava but Plenty of Data”
NCAR Mass Storage System

- Central facility for storing data used and generated by climate models, field experiments, and other earth-science models
- Current size (October 2003): 1.5 petabytes
- Growth rate: 50 terabytes/month
- Coupled to, but exceeds Moore’s law growth for cycles
Garbage Out: Municipal Waste

Municipal Waste (millions of tons) vs. Population (millions)

<table>
<thead>
<tr>
<th>Waste Importers</th>
<th>Million Tons</th>
<th>Waste Exporters</th>
<th>Million Tons</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Pennsylvania</td>
<td>9.76</td>
<td>1. New York</td>
<td>5.60</td>
</tr>
<tr>
<td>2. Virginia</td>
<td>3.89</td>
<td>2. New Jersey</td>
<td>1.80</td>
</tr>
<tr>
<td>4. Illinois</td>
<td>1.54</td>
<td>4. Maryland</td>
<td>1.54</td>
</tr>
<tr>
<td>5. Indiana</td>
<td>1.53</td>
<td>5. Massachusetts</td>
<td>1.21</td>
</tr>
</tbody>
</table>
Structured Prediction

Given: Structured input—sequence, or graph-structured

Objective: Predict a classification label for each node

- Fundamental importance in many areas
  - Speech, natural language processing, text analysis, web search, biosequence analysis, etc.

- Potential for significant theoretical and practical advances
  - Conditional graphical models
  - Relax traditional i.i.d. assumptions in statistical inference/learning
Conditional Random Fields

- Conditional model $p(y \mid x)$
  - Incorporate domain knowledge without increasing state space, using “features”
  - Do not make strong independence assumptions
  - Comparable to HMMs in computationally efficiency

- Promising results in
  - Tagging, parsing (Collins, 2002), (Sha and Pereira, 2003)
  - Information extraction (Pinto et al., 2003)
  - Image processing (Kumar and Hebert, 2004)
  - Gene finding (Pereira et al. 2004)

- Recent developments
Using Labeled and Unlabeled Data

Standard paradigm in machine learning is supervised learning

- Examples given by “teacher” to “student”
- Well-understood theoretically, “startling success”

Labeled data is limited

- Labeling examples often very expensive
- Example: protein shape classification, crystallography

How can *unlabeled* data be most effectively leveraged?

*Making theoretical and practical progress on the labeled/unlabeled data question is currently one of the most important and interesting problems in machine learning*
Labeled and Unlabeled Data as a Graph
Detour: Kernels


- A powerful & elegant tool—dramatic improvements in accuracy (and sometimes computational efficiency)
- Non-linear classifiers, implicit representation feature space
- Popular indoor sport—“Kernel $X$”, where $X$ is PCA, ICA, Clustering, . . .
- *Primarily a tool for Euclidean space*
Linear Classifiers vs. Kernel Classifiers

\[
\hat{f}(x) = \sum_{i=1}^{N} \alpha_i y_i \langle x, x_i \rangle
\]

\[
\hat{f}(x) = \sum_{i=1}^{N} \alpha_i y_i K(x, x_i)
\]
Popular Kernels

Gaussian: \(K_\sigma(x, x') = \exp \left(-\frac{\|x - x'\|^2}{2\sigma^2}\right)\)

Polynomial: \(K_d(x, x') = (1 + \langle x, x' \rangle)^d\)

- Processing data into vectors in \(\mathbb{R}^n\) is the “dirty laundry” of machine learning (Dietterich); proper set of features is crucial.
- Often data is “shoehorned” into Euclidean space (e.g., vector space representation for text processing).
- *Are there methods that take advantage of the “natural” structure of the data and problem?*
What if data lies on a graph or other discrete structure?

Late autumn turns into early winter in the backyard. The change of the guard is complete. Long gone are the birds of summer; settled in are the birds of winter. Bird feeding gets busier with each passing colder day. Especially in the northerly, winter residents have adapted their routines to these changing times. Cornus and juniper seeds are consumed during longer, colder nights, again in late afternoon, they feed heavily to fuel their furnaces for the long night ahead.

Throughout the country, chickadees, in bands of three to eight birds, fill into feeders, snap up sunflower seeds, and then spring off into the nearest shrubs to set them. With seeds placed between their minute feet, they hammer the seeds with tiny, sharp, black bills, breaking them open and consuming the kernels. Then back to the feeder for another seed.

Often cavorting with the chickadees are tufted titmice, white-breasted nuthatches, brown creepers, and hairy woodpeckers. Banding together means more eyes to find food, and to spot danger from predators.

Cardinals are year-round residents in their range and are regular to the bird feeders that offer good feeding and abundant sunflower seeds for quick cracking. The cardinals are more “popular” than the others...first in the morning and last in the evening...at the feeders.

Goldfinches, dressed in their dull olive-brown winter gowns, are here and there, and everywhere. The flock in the morning may return in the evening, but it is more likely a different band of birds, as goldfinches are constantly on the move in winter.

Now is the time to watch for visitors from the north when food shortages in boreal forest habitats may force the birds south. Get ready for pileated, purple finches, indigos, evening and pine grosbeaks, crossbills, three-toed and black-backed woodpeckers and snowy owls to invade the backyard in search of food, as far south as Georgia.

This is also the time when resident only begin tuning up for the winter courting season. On cold, clear and quiet nights, great horned and barn owls howl, and owlets waltz quietly.

In the North, bird balls with heathers are popular attractants for birds in late autumn. When natural sources of water may be scarce or frozen, if the water is freezing, birds will pass on bathing, but still drink.

Natural growing evergreen trees, decorated for the birds with string of reds, popcorn, nuts, and berries, are a great seasonal enrichment.

---

Structured Data

Late autumn turns into early winter in the backyard. The change of the guard is complete. Long gone are the birds of summer; settled in are the birds of winter. Bird feeding gets busier with each passing colder day. Especially in the northerly, winter residents have adapted their routines to these changing times. Cornus and juniper seeds are consumed during longer, colder nights, again in late afternoon, they feed heavily to fuel their furnaces for the long night ahead.

Throughout the country, chickadees, in bands of three to eight birds, fill into feeders, snap up sunflower seeds, and then spring off into the nearest shrubs to set them. With seeds placed between their minute feet, they hammer the seeds with tiny, sharp, black bills, breaking them open and consuming the kernels. Then back to the feeder for another seed.

Often cavorting with the chickadees are tufted titmice, white-breasted nuthatches, brown creepers, and hairy woodpeckers. Banding together means more eyes to find food, and to spot danger from predators.

Cardinals are year-round residents in their range and are regular to the bird feeders that offer good feeding and abundant sunflower seeds for quick cracking. The cardinals are more “popular” than the others...first in the morning and last in the evening...at the feeders.

Goldfinches, dressed in their dull olive-brown winter gowns, are here and there, and everywhere. The flock in the morning may return in the evening, but it is more likely a different band of birds, as goldfinches are constantly on the move in winter.

Now is the time to watch for visitors from the north when food shortages in boreal forest habitats may force the birds south. Get ready for pileated, purple finches, indigos, evening and pine grosbeaks, crossbills, three-toed and black-backed woodpeckers and snowy owls to invade the backyard in search of food, as far south as Georgia.

This is also the time when resident only begin tuning up for the winter courting season. On cold, clear and quiet nights, great horned and barn owls howl, and owlets waltz quietly.

In the North, bird balls with heathers are popular attractants for birds in late autumn. When natural sources of water may be scarce or frozen, if the water is freezing, birds will pass on bathing, but still drink.

Natural growing evergreen trees, decorated for the birds with string of reds, popcorn, nuts, and berries, are a great seasonal enrichment.
Laplacian on a Riemannian Manifold

On a Riemannian manifold with metric $g$, the geometric Laplacian is $\text{div} \circ \nabla$ or

$$\Delta f = \frac{1}{\sqrt{\det g}} \partial_i \left( g^{ij} \sqrt{\det g} \partial_i f \right)$$

$$= \sum_{ij} g^{ij} \partial_j \partial_i f + \text{(lower order terms)}$$

More generally, $\Delta = d^* d + d d^*$. 
Laplacian: Invariant Definition

\[ \Delta f(p) = - \lim_{r \to 0} \frac{2n}{r^2} \left( f(p) - \frac{1}{\text{vol}(S_r)} \int_{S_r(p)} f \right) \]

\( f \) is harmonic if \( \Delta f = 0 \)
Think of edge $e$ as “tangent vector” at $e_-$. 

For $f : V \to \mathbb{R}$, $df : E \to \mathbb{R}$ is the 1-form

$$df(e) = f(e_+) - f(e_-)$$

Then $\Delta = d^* d$ is discrete analogue of $\text{div} \circ \nabla$. 
Combinatorial Laplacian

\[ \Delta = D - W \] is, self-adjoint, positive, and an \textit{averaging operator}

\[ \Delta f(x) = \sum_{y \sim x} w_{xy}(f(x) - f(y)) \]

\[ = d(x) f(x) - \sum_{x \sim y} w_{xy} f(y) \]

We say \( f \) is \textit{harmonic} if \( \Delta f = 0 \).
Diffusion Kernels on Graphs

Let $\Delta$ be the graph Laplacian. In analogy with the continuous setting,

$$\frac{d}{dt}K_t = \Delta K_t$$

is the *heat equation* on a graph. Solution

$$K_t = e^{t\Delta}$$

is called the *heat kernel* or *diffusion kernel*.
Building Up Kernels

If $K_t^{(i)}$ are kernels on $\mathcal{X}_i$ then $K_t = \bigotimes_{i=1}^{n} K_t^{(i)}$ is a kernel on $\mathcal{X}_1 \times \ldots \times \mathcal{X}_n$.

For the hypercube, we get

$$K_t(x, x') \propto (\tanh t)^{d(x, x')}$$

Hamming distance

Other graphs with explicit diffusion kernels:

- Infinite trees (Chung & Yau, 1999)
- Cycles
- Rooted trees
- Strings with wildcards
Diffusion on Statistical Models

- “Similarity” between $x$ and $x'$ may be difficult to quantify—even for domain experts.

- Generative models are fairly easy to derive. A good generative model will use domain knowledge.

- Fisher information on the family “codifies” this into a distance. Gives statistical family structure of a Riemannian manifold.

- Associate each data point with a model: $x \mapsto \theta(x)$, and consider diffusion kernel on information manifold.
Information Geometry

\[ \mathcal{F} = \{ p(\cdot | \theta), \ \theta \in \Theta \subset \mathbb{R}^d \} \]  
\( d \)-dimensional statistical family.

**Fisher information matrix** \([g_{ij}(\theta)]\)

\[
g_{ij}(\theta) = \int_X \partial_i \log p(x | \theta) \partial_j \log p(x | \theta) p(x | \theta) \, dx
\]

Defines a Riemannian metric on \( \Theta \), which is invariant under reparameterization.

See (Amari, ’00; Kass & Voss, ’97)
Special Case: Multinomial

\[ d(\theta, \theta') = 2 \arccos \left( \sum_{i=1}^{d+1} \sqrt{\theta_i \theta'_i} \right) \]

\[ \sqrt{} \text{ maps simplex to the sphere} \]
Approximation of the Heat Kernel

Much of modern differential geometry is based on Laplacians, heat kernels and related constructions. Differential geometry has developed the *parametrix expansion* for the heat kernel.

For the multinomial, leads to approximation

\[
K_t(x, x') \approx \frac{1}{(4\pi t)^{d/2}} \exp \left( -\frac{1}{t} \arccos^2 \left( \sum_{i=1}^{d+1} \sqrt{\theta_i(x)\theta_i(x')} \right) \right)
\]

Proven to be very effective for text classification.
SVM Decision Boundaries: Trinomial

Gaussian kernel

Information diffusion kernel
Special Case: Spherical Normal

\[ p(x \mid \theta) = \mathcal{N}(\mu, \sigma I). \] Information metric: hyperbolic space.

Decision boundary for the diffusion kernel makes intuitive sense, since as \( \sigma \downarrow 0 \), mean is known with increasing certainty.
Bounds on Covering Numbers and Generalization Error

SVM hypothesis class covering numbers \( \mathcal{N}(\epsilon, \mathcal{F}_R(x)) \) for the diffusion kernel \( K_t \) satisfy

\[
\log \mathcal{N}(\epsilon, \mathcal{F}_R(x)) = O \left( \left( \frac{V}{t^2} \right) \log \frac{d+2}{2} \left( \frac{1}{\epsilon} \right) \right)
\]

Based on eigenvalue bounds in differential geometry:

\[
c_1(d) \left( \frac{j}{V} \right)^{\frac{2}{d}} \leq \mu_j \leq c_2(d) \left( \frac{j + 1}{V} \right)^{\frac{2}{d}}
\]

Better error bounds available based on Rademacher averages.
Labeled and Unlabeled Data as a Graph
Labeled and Unlabeled Data as a Graph

- Idea: Construct a random field on graph
- Intuition: Similar examples have similar labels
- Information “propagates” from labeled examples
- Graph encodes prior information—may be inaccurate or misleading!
Random Fields

- Energy function \( E(y) = \sum_{ij} w_{ij} (y_i - y_j)^2 \) with (conditional) random field \( p(y | x) \propto \exp (-\beta H(y)) \)

- Discrete case, \( y_i \in \{+1, -1\} \)
  - Most probable = graph mincuts (Blum and Chawla, 2001). Not unique.
  - Intractable to compute marginals

- Our framework: “relaxation” to continuous case, \( y_i \in \mathbb{R} \)
  - Convex, unique mode
  - Gaussian field, efficient algorithms
Gaussian Fields and Harmonic Functions

Combinatorial Laplacian \( \Delta = \begin{bmatrix} \Delta_{LL} & \Delta_{LU} \\ \Delta_{UL} & \Delta_{UU} \end{bmatrix} = D - W \), where

\[
W = \begin{bmatrix} w_{11} & \ldots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{n1} & \ldots & w_{nn} \end{bmatrix} \quad D = \begin{bmatrix} \sum w_1. & \cdots & 0 \\ 0 & \ddots & \sum w_n. \end{bmatrix}
\]

Field is clamped on the labeled data. On unlabeled data, Gaussian:

\[
y_U \sim \mathcal{N} \left( f_U, \frac{1}{2} \Delta_{UU}^{-1} \right) \\
f_U = -\Delta_{UU}^{-1} \Delta_{UL} f_L \quad \text{harmonic}
\]
Random Walks and Electric Networks

\[ R_{ij} = \frac{1}{w_{ij}} \]
Text Classification: Newsgroups

Red and blue curves use $\approx 2000$ unlabeled documents with 10-nearest neighbor graph from tf.idf and cosine similarity. Black line is standard “state-of-the-art” support vector machine.
Digit Recognition (10 classes)

![Graph showing accuracy vs. labeled set size for different models. The graph plots accuracy on the y-axis and labeled set size on the x-axis. Key lines represent different models: GF, thresh, and SVM. The GF model shows the highest accuracy across various set sizes.]
Active Learning (Experimental Design)

- Idea: Choose “query” point to be labeled
- Common heuristic: most uncertain example
- Our approach: minimize (estimated) expected risk, combining semi-supervised and active learning
- Can be implemented efficiently in Gaussian field/process view
Active Learning: Digit Classification

![Graph showing accuracy versus labeled set size for different methods. The graph includes lines for Active Learning, Random Query, and SVM.]
Why Does This Work So Well?

- Data matches our assumptions: clusters in the graph have like labels.
- In newsgroups data, like email, one posting will quote part of a previous posting, creating “threads of discussion” that are well-captured by the graph.
- In active learning, a single label suffices to disambiguate the entire thread.

If your computer could ask you just one question, should it ask the one it’s most unsure about, or the one from which it can learn the most? Graph structure allows the program a way to estimate the impact of each such question.
Recent Work: “Free Food Cam” Data

- Person identification in low-quality video
- Goal: Develop machine learning methods that will complement image processing techniques
Graph Structure for “Free Food Cam” Data

- Images are minimally processed:
  - Background subtraction, identifies “person-blobs”
  - Face detection (from Henry Schneiderman)

- Graph as three types of edges:
  - Images close in time – 2 seconds
  - Color histogram of foreground “blobs” – 3 nearest neighbors
  - Faces close in Euclidean distance – 1 nearest neighbor

- Baseline: SVM (3 kernels): 30% accuracy.

- Gaussian field: 85% accuracy for recognizing 10 people over several days, with only 10 labeled images/person (first day).
Closing Comments (1/2)

- We’ve begun “The Data Century”
  - Data, not computing, will drive computer science in the future
  - The distinction between statistics and computer science (and even AI) will continue to blur

- Structured prediction
  - Remarkable progress has been made on the “iid” case
  - Theory and methods for structured data are now emerging
  - Should be well developed within five–ten years

- Geometric perspective in statistical learning
  - Geometry of apparently non-geometric objects such as text documents, graphs, statistical models
  - Machinery of Riemannian structures
• Fundamental limits for using unlabeled data
  ▶ Several promising statistical learning approaches have been developed. Solid theoretical grounding is still lacking.
  ▶ Capacity and risk
  ▶ Consistency and error bounds
  ▶ Graph/manifold construction
Acknowledgements

I’ve had the great pleasure of collaborating with several fantastic people on some of the ideas & directions discussed here:

Avrim Blum, Zoubin Ghahramani, Risi Kondor, Guy Lebanon, Andrew McCallum, Fernando Pereira, Jerry Zhu