February 16, 2004

John Lafferty
Zoubin Ghahramani
Joint work with:

Jerry Xiao Jin Zhu

Random Fields
Semi-supervised Learning with Gaussian
<table>
<thead>
<tr>
<th>Supervised (classification, regression, etc.)</th>
<th>Unsupervised (clustering etc.)</th>
<th>supervised learning</th>
<th>unsupervised learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>usage</td>
<td>labeled data</td>
<td>yes</td>
<td>no</td>
</tr>
</tbody>
</table>
|  | unlabeled data | yes | no | 1
Protein structure: months of X-ray crystallography

Text categorization: limited user patience

Speech recognition: slow to transcribe

Unlabeled data are often easy to obtain, a lot.

Labeled data are often hard, expensive, slow to obtain.

For Classification
<table>
<thead>
<tr>
<th>semi-supervised learning</th>
<th>supervised learning</th>
<th>unlabeled data</th>
<th>labeled data</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>( {x} )</td>
</tr>
<tr>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>((h', x))</td>
</tr>
</tbody>
</table>

Usage
How is it possible? Intuition
Some semi-supervised learning methods

- Graph methods → this talk
- co-training
- transductive SVM
- mixture models / EM
Mixture Models. Cluster then label
— harder to analyze
— same flavor

Cluster first, label next

• EM stuck in bad local minima
• Mixture assumption wrong

• Unlabeled data may hurt, when...

Mixture Models, Cluster then Label
Some natural text here...
Unlabeled data reduce version space size by

\[ E \text{inm } \& \text{ Mitchell 98} \]

Forcing \( h_1, h_2 \) to agree.
Graph Methods

Nodes: labeled and unlabeled data

Edges: local similarity

Labels: propagate
Energy

Energy: $\sum \frac{1}{2} \langle \psi | - | \psi \rangle \text{ with } \langle \psi | = (\mathcal{A}|\mathcal{H})$

Edges: Weight matrix $W$. (Very, very important)

Labels: Assume binary $\mathcal{A} \in \{0, 1\}$
Given labeled data:

Low energy $\rightarrow$ Label Propagation
Hard to compute: Boltzmann machines, MCMC.

Mode: equiv. graph mincut [Bham & Chawla 01], not unique.

\{ 0, 1 \} \not\exists \ni \{ f = f' \} (\lambda \not\exists \ni \exp (\lambda) d)

Markov Random Fields
Gaussian Random Fields

Why hard: \( y_i \in \{0, 1\} \) combinatorial optimization

Relax: \( y_i \in \mathbb{R} \)

\[
p(y) \propto \exp(-E(y)) \Bigg|_{y_L = f_L} = \exp \left( -\frac{1}{2} \sum_{i,j} w_{ij} (y_i - y_j)^2 \right) \Bigg|_{y_L = f_L} = \exp \left( -y^T \Delta y \right) \Bigg|_{y_L = f_L}
\]
\[
\begin{bmatrix}
\begin{array}{c|c}
\mu \nu \nabla & \tau \nu \nabla \\
\hline
\eta \tau \nabla & \tau \tau \nabla \\
\end{array}
\end{bmatrix} = \nabla
\]

The Laplacian

\[M - D = \nabla\]

\[
\begin{bmatrix}
\mu_m \nabla \\
\vdots & \ddots & 0 \\
0 & \ddots & \mu_1 \nabla \\
\end{bmatrix} = D
\begin{bmatrix}
\mu_1 m & \cdots & \mu_1 m \\
\mu_2 m & \cdots & \mu_2 m \\
\mu_m m & \cdots & \mu_m m \\
\end{bmatrix} = M
\]

The Laplacian
If $f_{\mathcal{H}} \mathcal{N}_{-\mathcal{H}}(\mathcal{H} \mathcal{N}) = f$

The mean on unlabeled data:

$\left( \begin{array}{c}
\mathcal{H} \mathcal{N} \\
\mathcal{H} \mathcal{N}
\end{array} \right) \mathcal{N}_{\mathcal{H} \mathcal{N}}$

Conditioned on labeled data:

$f_{\mathcal{H}} \mathcal{H} \mathcal{N}_{\mathcal{H} \mathcal{N}}(\mathcal{H} \mathcal{N} \mathcal{N}) = \mathcal{N} \mathcal{H}$

**Gaussian Random Fields**
\mathcal{f} \geq \forall \Omega \in \mathcal{f}, \mathcal{f} = \mathcal{f} \text{ or } 0 = \mathcal{f} \text{ is harmonic} \implies \text{ uniquely exists} \implies \text{ min energy state} \implies \text{ mode of the Gaussian Random Field} \implies \text{ The Mean } \mathcal{f}_\mu
(\text{reach label } I | \text{from } ?_d = ? f)

\text{Interpretation: Random Walks} f
\[ V = R \cdot I \]

Interpretation: Electric Networks
\[ 9.6 = \frac{|\Omega|}{|\Omega_f|} \]

subject to

\[ \lambda \cap_I \lambda = (\lambda)_I \]

minimize

edge: 90% class I

heuristic: incorporating class proportions. E.g. prior knowledge

naive: threshold at 0.5

Classification Using Gaussian Fields
OCR Digits (0-9)
20 Newsgroups (PC vs. MAC)
Learn the Graph
Learn the Graph

- Kernel alignment (research)
- Evidence maximization (local optima)
- Minimize label entropy (maximize label confidence heuristics)
<table>
<thead>
<tr>
<th>GF acc</th>
<th>0.6542</th>
</tr>
</thead>
<tbody>
<tr>
<td>98.02%</td>
<td>0.39%</td>
</tr>
<tr>
<td>94.70%</td>
<td>1.19%</td>
</tr>
</tbody>
</table>

OCR Digits "1" vs. "2", minimize entropy

Learn the Graph
modes. Low frequency = smoother.
Laplacian spectrum: vibration frequencies: \( \phi_i \perp \lambda_i \). The Kernel View
In general, a non-negative, monotonic decreasing function on the spectrum of the semi-supervised kernel [Smola & Kondor] emphasizes smooth components.

\[ \sum_{i} \phi_i \phi_i^\top \gamma_i \approx 1 - \nabla \approx K \]

\[ (\Lambda \Lambda_i - \Lambda_i - \exp) \hat{\alpha}_i \hat{d} = (\Lambda) \hat{d} \]

The Kernel View
Semi-supervised vs. standard kernels

<table>
<thead>
<tr>
<th>Ignore unlabeled data</th>
<th>Use unlabeled data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear, polynomial, RBF etc.</td>
<td>Semi-supervised kernels</td>
</tr>
</tbody>
</table>

28
Active + semi-supervised learning

Pool based

Given vs. query

\[\text{A LOT}\]

Unlabeled data

A few labeled data

Active Learning
Most ambiguous point not necessarily the best query:

Active Learning
\[
\sum_{\forall i \in \mathcal{I}} \min (\hat{f} - f, 0) = \text{err}
\]

estimated error

\[
\exists \hat{f} \rightarrow (I = \hat{\mathcal{I}}) \quad \text{approximation}
\]

\[
\sum_{\forall i \in \mathcal{I}} (\exists \hat{\mathcal{I}}(\exists \hat{\mathcal{I}} \neq (\hat{f})(\text{sgn}(0, 1))) = \text{err}
\]

generalization error

Active Learning
\[ \frac{\mathcal{H}_\mathcal{L}(\mathcal{L} \cup \nabla)}{\mathcal{L}(\mathcal{L} \cup \nabla)}(\mathcal{L}f - \mathcal{H}_\mathcal{L}) + \Omega f = (\mathcal{H}_\mathcal{L}, \mathcal{L}x) + \Omega f \]

re-train is fast

\((\mathcal{H}_\mathcal{L}, \mathcal{L}x) + \Omega f + (\mathcal{L}_\mathcal{L}(\mathcal{L}_\mathcal{L}(\mathcal{L}f - \mathcal{L}_\mathcal{L})x(0, \mathcal{L} \mathcal{L}(\mathcal{L}_\mathcal{L}(\mathcal{L}f - \mathcal{L}_\mathcal{L})x(0, \mathcal{L} \mathcal{L}(\mathcal{L}_\mathcal{L}(\mathcal{L}f - \mathcal{L}_\mathcal{L}

\text{select query } \gamma \text{ to minimize the estimated error}

\left( (\mathcal{H}_\mathcal{L}, \mathcal{L}x) + \Omega f - \mathcal{L}_\mathcal{L}(\mathcal{L}_\mathcal{L}(\mathcal{L}f - \mathcal{L}_\mathcal{L})x(0, \mathcal{L} \mathcal{L}(\mathcal{L}_\mathcal{L}(\mathcal{L}f - \mathcal{L}_\mathcal{L})x(0, \mathcal{L} \mathcal{L}(\mathcal{L}_\mathcal{L}(\mathcal{L}f - \mathcal{L}_\mathcal{L}

\text{estimated error after querying } \gamma

\textbf{Active Learning}
Active Learning: OCR Digits "1" vs. "2"
Active learning: 20 Newsroups PC vs. MAC
dense, sufficient

(sub-features, good, conditional independence)

\begin{align*}
x_2 & \\
\Rightarrow & \\
x_1 & \\
\Rightarrow & \\
x & \\
\{x_1, x_2\} = \text{split feature } x
\end{align*}
Co-training

Feature

train two classifiers, one on each sub-

Graph
Trans. SVM
Mixture
Method
Metric-based Model Selection

[Schuurmans & Southey 01]
Metric-based Model Selection

Consider 2 hypotheses $h_1, h_2$:

- both have zero training error
- disagree a lot on unlabeled data

$h_1, h_2$ can’t be both correct.
41

$[\text{szummer \& Jäakola 01}]$

$t$-step random walk: $(D_{1-W})^t$

$[\text{Kondor \& Lar fferty 02}]$

Diffusion kernel at time $t$: $e^{-t \Delta}$

Connection: Graph Kernels
$\mathcal{H} \wedge \nabla_{t-e} \int_0^\infty f = \tau_{-1} (\wedge \nabla) - \int f$

Connection: Graph Kernels