Semi-Supervised Learning and Learning via Similarity Functions: Two key settings for Data-Dependent Concept Spaces

Avrim Blum
Carnegie Mellon

[NIPS 2008 Workshop on Data Dependent Concept Spaces]
Theme of the workshop: Data-Dependent Concept Spaces

- Using the data to determine the set of functions you are viewing as most natural.
  - E.g., like large margin separators
  - Pre-bias + data ! Bias
  - Much is even without considering labels.
Theme of the workshop: Data-Dependent Concept Spaces

- This talk: two key settings where this is esp helpful.
  - Semi-supervised learning
  - Learning from similarity functions
Can we use unlabeled data to augment a small labeled sample to improve learning?

But unlabeled data is missing the most important info!!

But maybe can help us instantiate our biases.
Semi-Supervised Learning

1st half of this talk:

• Some examples of SSL algorithms using different kinds of biases.
• Unified framework for analyzing in terms of Distribution-dependent concept spaces
  [joint work with Nina Balcan]
  - Understand when unlabeled data can help
  - Help make sense of what’s going on.
Ex 1: Co-training

Many problems have two different sources of info you can use to determine label.

- Example $x$ is a pair $h x_1, x_2 i$

E.g., classifying webpages: can use words on page or words on links pointing to the page.
Ex 1: Co-training

Idea: Use small labeled sample to learn initial rules.

- E.g., “my advisor” pointing to a page is a good indicator it is a faculty home page.
- E.g., “I am teaching” on a page is a good indicator it is a faculty home page.
**Ex 1: Co-training**

Idea: Use small labeled sample to learn initial rules.
- E.g., “my advisor” pointing to a page is a good indicator it is a faculty home page.
- E.g., “I am teaching” on a page is a good indicator it is a faculty home page.

Then look for unlabeled examples where one rule is confident and the other is not. Have it label the example for the other.

Training 2 classifiers, one on each type of info. Using each to help train the other.
As a simple example, suppose $x_1, x_2 \in \mathbb{R}$. Target function is some interval $[a, b]$. 
**Ex 1: Co-training**

Or can directly optimize joint objective:

- Accuracy on labeled sample
- Agreement on unlabeled sample
Ex 1: Co-training

Turns out a number of problems can be set up this way.

E.g., [Levin-Viola-Freund03] identifying objects in images. Two different kinds of preprocessing.

E.g., [Collins&Singer99] named-entity extraction.
  - “I arrived in London yesterday”
  - ...

Ex 2: $S^3VM$ [Joachims98]

- Suppose we believe target separator goes through low density regions of the space/\textit{large margin}.
- Aim for separator with large margin \textit{wrt labeled and unlabeled data}. (L+U)
Suppose we believe target separator goes through low density regions of the space/large margin.

Aim for separator with large margin wrt labeled and unlabeled data. (L+U)

Unfortunately, optimization problem is now NP-hard.

- So do local optimization or branch-and-bound etc.
Ex 3: Graph-based methods

- Suppose we believe that very similar examples probably have the same label.
- If you have a lot of labeled data, this suggests a Nearest-Neighbor type of alg.
- If you have a lot of unlabeled data, perhaps can use them as “stepping stones”

E.g., handwritten digits [Zhu07]:

<table>
<thead>
<tr>
<th><code>0</code> <code>2</code></th>
<th><code>0</code> <code>1</code> <code>1</code> <code>2</code> <code>2</code> <code>2</code></th>
</tr>
</thead>
</table>
| not similar | ‘indirectly’ similar
with stepping stones  |
Ex 3: Graph-based methods

- Idea: construct a graph with edges between very similar examples.
- Unlabeled data can help “glue” the objects of the same class together.
Ex 3: Graph-based methods

- Idea: construct a graph with edges between very similar examples.
- Unlabeled data can help “glue” the objects of the same class together.
Ex 3: Graph-based methods

- Idea: construct a graph with edges between very similar examples.
- Unlabeled data can help “glue” the objects of the same class together.
- Solve for:
  - Minimum cut [BC, BLRR]
  - Minimum “soft-cut” [ZGL]
  \[ \sum_{e=(u,v)} (f(u)-f(v))^2 \]
  - Spectral partitioning [J]
  - ...
Several different approaches. Is there some unifying principle?

What should be true about the world for unlabeled data to help?
Semi-supervised model \cite{BB05}

High-level idea:

Unlabeled data useful if we have beliefs not only about the form of the target, but also about its relationship with the underlying distribution.
Semi-supervised model [BB05]

High-level idea:

In standard PAC, bias is a **set of or score over** possible targets.

**Concept class**

SSL: replace this with score over (f,D) pairs [fn from D to score over f’s]

**Distribution-dependent concept class**

➢ that can estimate from a finite sample
Semi-supervised model [BB05]

Augment notion of a concept class $C$ with a compatibility score $\chi(h, D)$ between a concept and the data distribution.

“learn $C$” becomes “learn $(C, \chi)$” (learn class $C$ under $\chi$)

$(C, \chi)$ express beliefs/hopes about the world.

Idea I: use unlabeled data & belief that the target is compatible to reduce $C$ down to just $\{\text{the highly compatible functions in } C\}$.

Idea II: require that the degree of compatibility can be estimated from a finite sample.
Formally

- Require compatibility $\chi(h,D)$ to be expectation over individual examples. (don’t need to be so strict but this is cleanest)
  - View incompatibility as unlabeled loss fn $l_{unl}(h,x) \in [0,1]$.
- Unlabeled error rate $\text{Err}_{unl}(h) = E_{x \sim D}[l_{unl}(h,x)]$
- $\chi(h,D) = 1 - \text{Err}_{unl}(h)$

Co-training: $\text{Err}_{unl}(h) =$ probability mass on pts $hx_1,x_2$ where $h_1(x_1) \neq h_2(x_2)$
Formally

- Require compatibility $\chi(h,D)$ to be expectation over individual examples. (don’t need to be so strict but this is cleanest)
  - View incompatibility as unlabeled loss fn $l_{\text{unl}}(h,x)\in [0,1]$.
- Unlabeled error rate $\text{Err}_{\text{unl}}(h) = E_{x \sim D}[l_{\text{unl}}(h,x)]$
- $\chi(h,D) = 1 - \text{Err}_{\text{unl}}(h)$

**$S^3\text{VM}$**: $\text{Err}_{\text{unl}}(h) =$ probability mass near to separator $h$. 

\[ l_{\text{unl}}(h,x) \]
Can use to prove sample bounds

Simple example theorem: (believe target is fully compatible)

Define $C_D(\varepsilon) = \{h \in C : \text{err}_{\text{unl}}(h) \cdot \varepsilon\}$.

If we see

$$m_u \geq \frac{1}{\varepsilon} \left[ \ln |C| + \ln \frac{2}{\delta} \right]$$

unlabeled examples and

$$m_l \geq \frac{1}{\varepsilon} \left[ \ln |C_D(\varepsilon)| + \ln \frac{2}{\delta} \right]$$

labeled examples, then with probability $\geq 1 - \delta$, all $h \in C$ with $\tilde{\text{err}}(h) = 0$ and $\tilde{\text{err}}_{\text{unl}}(h) = 0$ have $\text{err}(h) \leq \varepsilon$.

Bound the # of labeled examples as a measure of the helpfulness of $D$ wrt our incompatibility score.
- a helpful distribution is one in which $C(\varepsilon)$ is small
Can use to prove sample bounds

Simple example theorem: (believe target is fully compatible)

Define $C_D(\varepsilon) = \{h \in C : \text{err}_{unl}(h) \cdot \varepsilon\}$.

If we see

$$m_u \geq \frac{1}{\varepsilon} \left[ \ln |C| + \ln \frac{2}{\delta} \right]$$

unlabeled examples and

$$m_l \geq \frac{1}{\varepsilon} \left[ \ln |C_D(\varepsilon)| + \ln \frac{2}{\delta} \right]$$

labeled examples, then with probability $\geq 1 - \delta$, all $h \in C$ with $\hat{\text{err}}(h) = 0$ and $\hat{\text{err}}_{unl}(h) = 0$ have $\text{err}(h) \leq \varepsilon$.

E.g., Co-training: if all $x_1 = x_2$ then target compatible but so is every other $h \in C$. Not so helpful...
Can use to prove sample bounds

Simple example theorem: (believe target is fully compatible)

Define $C_D(\varepsilon) = \{h \in C : \text{err}_{\text{unl}}(h) \cdot \varepsilon\}$.

If we see

$$m_u \geq \frac{1}{\varepsilon} \left[ \ln |C| + \ln \frac{2}{\delta} \right]$$

unlabeled examples and

$$m_l \geq \frac{1}{\varepsilon} \left[ \ln |C_D(\varepsilon)| + \ln \frac{2}{\delta} \right]$$

labeled examples, then with probability $\geq 1 - \delta$, all $h \in C$ with $\hat{\text{err}}(h) = 0$ and $\hat{\text{err}}_{\text{unl}}(h) = 0$ have $\text{err}(h) \leq \varepsilon$.

Extend to case where target not fully compatible.

Then care about $\{h \in C : \text{err}_{\text{unl}}(h) \cdot \varepsilon + \text{err}_{\text{unl}}(f^*)\}$.

D is helpful if this is small
Can use to prove sample bounds

**Agnostic version:**  (not sure how compatible target is)

If we choose $h = \arg\min_{h'}[\hat{\text{err}}(h') + \hat{\text{pen}}(h')]$ then whp

$$\text{err}(h) \leq \min_t [\text{err}(f_t^* + \hat{\text{pen}}(f_t^*)) + 5\sqrt{\frac{\log(8/\delta)}{m_l}}}$$

where $f_t^*$ is best fn of unlabeled error $\leq t$ and $\hat{\text{pen}}(h) = \sqrt{\frac{24}{m_l}}\log(8\Pi[C_S(\text{err}_{unl}(h) + 2\varepsilon), S, m_l])$.

Furthermore if unlabeled set is large, get $\hat{\text{pen}} \approx \text{pen}$.

**Notes:** unlabeled set needs to be large compared to $\max[\dim(C), \dim(\chi_C)]$

where $\chi_C$ = class induced by $\chi$ & $C$: $\chi_h(x) = \chi(h,x)$
Examples

- Target is large-margin separator, distribution is uniform on small number $k$ of well-shaped blobs.

- If # of unlabeled is suff lg, need only $O(k/\varepsilon)$ labeled examples.
Interesting feature

\(\varepsilon\)-Cover bounds often give substantial gain compared to uniform convergence bounds.

- **Nice example:**
  - Examples are *pairs* of points in \(\{0,1\}^n\).
  - \(C\) = linear separators
  - Example is compatible if both parts positive or both parts negative (like simple form of co-training).
  - \(D\) = independent given the label.

- If \(\text{poly}(n)\) unlabeled, \(o(\log n)\) labeled, then still many bad compatible, consistent hypotheses.

- But \(\varepsilon\)-cover size is constant. All compatible fns are close to (a) target, (b) \(\pm\)target, (c) 1, or (d) 0.
Conclusions part 1

• Data-dependent concept spaces can give a unifying view of semi-supervised learning.
• Beliefs over how target should relate to distribution.
  - View as notion of unlabeled error rate.
  - Unlabeled data used to estimate your bias.
  - Can do DDSRM to trade off.
Topic 2:
Learning with similarity functions

[joint work with Nina Balcan and Nati Srebro]
A kernel \( K \) is a legal def of dot-product: fn s.t. there exists an implicit mapping \( \Phi_K \) such that \( K(x, y) = \Phi_K(x) \cdot \Phi_K(y) \).

- E.g., \( K(x, y) = (x \cdot y + 1)^d \).
- \( \Phi_K: \text{(n-diml space)} \to \text{(n^d-diml space)} \).

Point is: many learning algos can be written so only interact with data via dot-products.
- If replace \( x \cdot y \) with \( K(x, y) \), it acts implicitly as if data was in higher-dimensional \( \Phi \)-space.
Moreover, generalize well if good margin

- If data is lin. separable by margin $\gamma$ in $\Phi$-space, then need sample size only $\tilde{O}(1/\gamma^2)$ to get confidence in generalization.

Assume $|\Phi(x)| \cdot 1$.

- Kernels found to be useful in practice for dealing with many, many different kinds of data.
Moreover, generalize well if good margin

...but there’s a bit of a disconnect:

- Intuitively we think of a good kernel for some problem as one that makes sense as a good notion of similarity.
- But **Theory** talks about margins in implicit high-dimensional $\Phi$-space.
Can we define a notion of a good measure of similarity that...

1. Talks in terms of more intuitive quantities (no implicit high-diml spaces, no PSD requirement, etc)

2. Is broad: includes usual notion of “good kernel” (one that induces a large margin separator in $\Phi$-space).

3. Maybe even allows one to learn classes that have no large-margin kernels?

Use notion of data-dependent concept spaces
Warmup property

• Say have a learning problem $P$ (distribution $D$ over examples labeled by unknown target $f$).
• Sim fn $K:(x,y)![-1,1]$.
• Consider asking for property that most examples $x$ satisfy:
  "on average, more similar to points $y$ of its own type than to points $y$ of the other type, by some gap $\gamma$"

$$E_{y \sim D}[K(x,y)|l(y)=l(x)] - E_{y \sim D}[K(x,y)|l(y) \neq l(x)] + \gamma$$

Then this would be great for learning
Just do average-nearest-nbr

At least a $1-\epsilon'$ prob mass of $x$ satisfy:

$$E_{y \sim D}[K(x,y) | l(y) = l(x)], E_{y \sim D}[K(x,y) | l(y) \neq l(x)] + \gamma$$

- Draw $S^+$ of $O((1/\gamma^2) \ln 1/\epsilon\delta)$ positive examples.
- Draw $S^-$ of $O((1/\gamma^2) \ln 1/\epsilon\delta)$ negative examples.
- Classify new examples based on which gives better score.
- With prob $\geq 1-\delta$, $S^+$ and $S^-$ give you error rate $\cdot \epsilon + \epsilon'$. 
But not broad enough

- $K(x,y) = x \& y$ has good separator but doesn't satisfy condition. (half of + are more similar to - that to typical +)

Avg simil to negs is 0.5, but to pos is only 0.25
But not broad enough

- Idea: would work if we didn’t pick y’s from top-left.
- Broaden to say: OK if 9 region $R$ s.t. most $x$ are on average more similar to $y_2R$ of same label than to $y_2R$ of other label. (even if don’t know $R$ in advance)
Broader defn...

- Ask that exists a set $R$ of “reasonable” $\gamma$ (allow probabilistic) s.t. almost all $x$ satisfy

$$E_{\gamma}[l(x)l(y)K(x,y) | y \in R], \gamma$$

- Formally, say $K$ is $(\varepsilon', \gamma, \tau)$-good if have hinge-loss $\varepsilon'$, and $\Pr(R) \geq \tau$.

- **Thm 1:** this is a legitimate way to think about good kernels:
  - If kernel has margin $\gamma$ in implicit space, then for any $\tau$ is $(\tau, \gamma^2, \tau)$-good in this sense.
Broader defn...

• Ask that exists a set $R$ of “reasonable” $\gamma$ (allow probabilistic) s.t. almost all $x$ satisfy

$$E_y[I(x)I(y)K(x,y) \mid y \in R], \gamma$$

• Formally, say $K$ is $(\varepsilon', \gamma, \tau)$-good if have hinge-loss $\varepsilon'$, and $\Pr(R) \geq \tau$.

• Thm 2: this is sufficient for learning.
  - If $K$ is $(\varepsilon', \gamma, \tau)$-good, $\varepsilon' \leq \varepsilon$, can learn to error $\varepsilon$ with
  $$O((1/\varepsilon \gamma^2) \log(1/\varepsilon \gamma \tau))$$ labeled examples.

  [and $\tilde{O}(1/(\gamma^2 \tau))$ unlabeled examples]
How to use such a sim fn?

- Assume exists $R$ s.t. $Pr_y[R]$, $\tau$ and almost all $x$ satisfy

  $$E_y[l(x)l(y)K(x,y) | y \in R], \gamma$$

- Draw $S = \{y_1, \ldots, y_n\}$, $n^{1/4}1/(\gamma^2 \tau)$. [could be unlabeled]

- View as “landmarks”, use to map new data:

  $$F(x) = [K(x,y_1), \ldots, K(x,y_n)].$$

- Whp, exists separator of good $L_1$ margin in this space: $w = [0,0,1/n_R,1/n_R,0,0,0,-1/n_R,0]$  
  ($n_R = \# y_i \in R$)

---

![Diagram](image_url)
How to use such a sim fn?

• Assume exists \( R \) s.t. \( \Pr_y[R] \geq \tau \) and almost all \( x \) satisfy

\[
E_y[\mathbb{1}(x)\mathbb{1}(y)K(x,y) | y \in R], \gamma
\]

- Draw \( S = \{y_1, \ldots, y_n\}, \ n^{\frac{1}{4}}1/(\gamma^2 \tau) \).
- View as “landmarks”, use to map new data:

\[
F(x) = [K(x,y_1), \ldots, K(x,y_n)].
\]

- Whp, \textbf{exists} separator of good \( L_1 \) margin in this space:

\[
w=[0,0,1/n_R, 1/n_R, 0, 0, 0, -1/n_R, 0]
\]

\( (n_R = \# y_i \in R) \)

- So, take new set of examples, project to this space, and run good \( L_1 \) alg (Winnow).
Interesting property

• Notice that using landmarks to define an explicit Data-dependent concept space:
  - \( F(x) = [K(x, y_1), \ldots, K(x, y_n)] \).
  - Considering linear separators of large \( L_1 \) margin in this space. (which again is data-dependent)
Relation between “large margin kernel” and “good similarity fn”

- Any large margin kernel is also good similarity function (though can potentially incur quadratic penalty in sample cplx).
- But, can also get exponential improvement.
- Can define C, D with
  - Similarity fn \((0,1,1/|C|)\)-good for all \(c_i \in 2^C\).

\[\textnormal{hinge loss} \quad \textnormal{margin} \quad \textnormal{Pr}(R_i)\]

\[\text{Gives bound } O(\varepsilon^{-1} \log(|C|))\]

- But any kernel has margin \(O(1/|C|^{1/2})\) for some \(c \in 2^C\).
Summary

- Defined property of similarity fn that is sufficient to be able to learn well.
  - Can apply to similarity fns that aren’t positive-semidefinite (or even symmetric).
  - Includes usual notion of “good kernels” modulo some loss in parameters, but also potential gain.
  - Uses data-dependent concept spaces.
- Open questions: properties suggesting other algs? E.g., [WangYangFeng07]