Semi-Supervised Learning

Given labeled examples $S = \{(x_i, y_i)\}$, try to learn a good prediction rule.

Often labeled data is limited or expensive.

On the other hand, often unlabeled data is plentiful and cheap.

- Documents, images, OCR, web-pages, protein sequences, ...

Can we use unlabeled data to help?

Unlabeled data is missing the most important info! But maybe still has useful regularities that we can use. E.g., OCR.

This is a question a lot of people in ML have been interested in. A number of interesting methods have been developed.

Today:

- Discuss several methods for trying to use unlabeled data to help.
- Extension of PAC model to make sense of what's going on.

Plan for today

Methods:

- Co-training
- Transductive SVM
- Graph-based methods

Model:

- Augmented PAC model for SSL.

Co-training

[Blum&Mitchell'98] motivated by [Yarowsky'95]

Yarowsky's Problem & Idea:

Some words have multiple meanings (e.g., "plant"). Want to identify which meaning was intended in any given instance.

Standard approach: learn function from local context to desired meaning, using labeled data. 

"...nuclear power plant generated..."

Idea: use fact that in most documents, multiple uses have same meaning. Use to transfer confident predictions over.

There's also a book "Semi-supervised learning" on the topic.
Co-training

Actually, many problems have a similar characteristic.

- Examples $x$ can be written in two parts $(x_1, x_2)$.
- Either part alone is in principle sufficient to produce a good classifier.
- E.g., speech+video, image and context, web page contents and links.
- So if confident about label for $x_1$, can use to impute label for $x_2$, and vice versa. Use each classifier to help train the other.

Example: intervals

Suppose $x_1 \in \mathbb{R}, x_2 \in \mathbb{R}, c_1 = [a_1, b_1], c_2 = [a_2, b_2]$.

Co-Training Theorems

- [BM98] if $x_1, x_2$ are independent given the label: $D = p(D_1^+ \times D_2^+) + (1-p)(D_1^- \times D_2^-)$, and if $C$ is SQ-learnable, then can learn from an initial "weakly-useful" $h_1$ plus unlabeled data.
- Def: $h$ is weakly-useful if $\Pr[h(x)=1|c(x)=1] > \Pr[h(x)=1|c(x)=0] + \epsilon$. (same as weak hyp if target $c$ is balanced)
- E.g., say "syllabus" appears on 1/3 of course pages but only 1/6 of non-course pages.
- Use as noisy label. Like classification noise with potentially asymmetric noise rates $\alpha, \beta$.
- Can learn so long as $\alpha \beta < 1 - \epsilon$.
  (helpful trick: balance data so observed labels are 50/50)

Example: classifying webpages

- Co-training: Agreement between two parts
  - examples contain two sets of features, i.e. an example is $x = (x_1, x_2)$ and the belief is that the two parts of the example are sufficient and consistent, i.e. if $c_i, c_j$ such that $c_i(x_1), c_j(x_2) = c(x)$

Co-Training Theorems

- [BM98] if $x_1, x_2$ are independent given the label: $D = p(D_1^+ \times D_2^+) + (1-p)(D_1^- \times D_2^-)$, and if $C$ is SQ-learnable, then can learn from an initial "weakly-useful" $h_1$ plus unlabeled data.
- [BB05] in some cases (e.g., LTFs), you can use this to learn from a single labeled example.
**Co-Training Theorems**

- [BM98] if $x_1, x_2$ are independent given the label: $D = p(D_1^+ \times D_2^-) + (1-p)(D_1^- \times D_2^+)$, and $C$ is SQ-learnable, then can learn from an initial "weakly-useful" $h_1$ plus unlabeled data.
- [BB05] in some cases (e.g., LTFs), you can use this to learn from a single labeled example!
  - Pick random hyperplane and boost.
  - Repeat process multiple times.
  - Get 4 kinds of hyps: (close to $c$, close to $\neg c$, close to 1, close to 0)

**Co-Training and expansion**

Want initial sample to expand to full set of positives after limited number of iterations.

**Transductive SVM [Joachims98]**

- 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. Algorithm instead does local optimization.
  - Start with large margin over labeled data. Induces labels on U.
  - Then try flipping labels in greedy fashion.

**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, suggests a graph-based method.
**Graph-based methods**

- Transductive approach. (Given $L + U$, output predictions on $U$).
- Construct a graph with edges between very similar examples.
- Solve for:
  - Minimum cut
  - Minimum “soft-cut” [ZGL]
  - Spectral partitioning

**PAC-SSL Model [BB05]**

- Augment the notion of a concept class $C$ with a notion of compatibility $\chi$ between a concept and the data distribution.
  - “learn $C$” becomes “learn $(C, \chi)$” (i.e. learn class $C$ under compatibility notion $\chi$)
- To do this, need unlabeled data to allow us to uniformly estimate compatibilities well.
- Require that the degree of compatibility be something that can be estimated from a finite sample.

**PAC-SSL Model [BB05]**

- Suppose just two labels: 0 & 1.
- Solve for labels $f(x)$ for unlabeled examples $x$ to minimize:
  - $\sum_{e=(u,v)}|f(u)-f(v)|$ [soln = minimum cut]
  - $\sum_{e=(u,v)}(f(u)-f(v))^2$ [soln = electric potentials]

**How can we think about these approaches to using unlabeled data in a PAC-style model?**

- “learn $C$” becomes “learn $(C, \chi)$” (i.e. learn class $C$ under compatibility notion $\chi$)
- Express relationships that one hopes the target function and underlying distribution will possess.
- Idea: use unlabeled data & the belief that the target is compatible to reduce $C$ down to just (the highly compatible functions in $C$).
Margins, Compatibility

• Margins: belief is that should exist a large margin separator.

• Incompatibility of h and D (unlabeled error rate of h) - the probability mass within distance γ of h.

• Can be written as an expectation over individual examples

\[ \chi(h,D) = \mathbb{E}_{x \in D} [\chi(h,x)] = \frac{1 - e^{-\gamma}}{2\gamma} \]

• Illegal notion of compatibility: the largest γ s.t. D has probability mass exactly zero within distance γ of h.

Co-Training, Compatibility

• Co-training: examples come as pairs (x₁, x₂) and the goal is to learn a pair of functions (h₁, h₂).

• Hope is that the two parts of the example are consistent.

• Legal (and natural) notion of compatibility:
  - the compatibility of (h₁, h₂) and D:

\[ \mathbb{P}(x₁, x₂) | h₁(x₁) = h₂(x₂) \]

• Can be written as an expectation over examples:

\[ \chi((h₁, h₂), (x₁, x₂)) = 1 \text{ if } h₁(x₁) = h₂(x₂) \]

\[ \chi((h₁, h₂), (x₁, x₂)) = 0 \text{ if } h₁(x₁) \neq h₂(x₂) \]

Sample Complexity - Uniform convergence bounds

Finite Hypothesis Spaces, Doubly Realizable Case

• Define \( C_0(\epsilon) = \{ h \in C : \text{err}_D(h) \leq \epsilon \} \).

Theorem

If we use

\[ m_u \geq \frac{1}{\epsilon} \left[ \ln |C| + \ln \frac{4}{\delta} \right] \]

unlabeled examples and

\[ m_l \geq \frac{1}{\epsilon} \left[ \ln |C_0(D, \epsilon)| + \ln \frac{2}{\delta} \right] \]

labeled examples, then with probability ≥ 1 - δ, all \( h \in C \) with \( \text{err}(h) = 0 \) and \( \epsilon \text{err}_{D_{\text{unl}}}(h) \leq \epsilon \) have \( \epsilon \text{err}_{D_{\text{l}}}(h) \leq \epsilon \).

• Bound the # of labeled examples as a measure of the helpfulness of D with respect to \( \chi \),

  a helpful distribution is one in which \( C_0(\epsilon) \) is small.

Semi-Supervised Learning

Natural Formalization (PACₐₐ)

• We will say an algorithm "PACₐₐ-learns" if it runs in poly time using samples poly in respective bounds.

• E.g., can think of ln|C| as # bits to describe target without knowing D, and ln|C₀(\epsilon)| as number of bits to describe target knowing a good approximation to D, given the assumption that the target has low unlabeled error rate.

Target in C, but not fully compatible

Finite Hypothesis Spaces - \( c' \) not fully compatible

Theorem

Given \( \epsilon \in [0, 1] \), if we see

\[ m_u \geq \frac{2}{\epsilon^2} \left[ \ln |C| + \ln \frac{4}{\delta} \right] \]

unlabeled examples and

\[ m_l \geq \frac{1}{\epsilon} \left[ \ln |C_{D_\text{unl}}(\epsilon + 2\epsilon)| + \ln \frac{2}{\delta} \right] \]

labeled examples, then with probability ≥ 1 - δ, all \( h \in C \) with \( \epsilon \text{err}_{D_{\text{unl}}}(h) = 0 \) and \( \epsilon \text{err}_{\text{D}_{\text{l}}}(h) \leq \epsilon \) have \( \epsilon \text{err}_{\text{D}_{\text{l}}}(h) \leq \epsilon + \epsilon \).

Implication: if \( \epsilon \text{err}_{\text{D}_{\text{l}}}(\epsilon') \leq \epsilon \) and \( \epsilon \text{err}_{\text{D}_{\text{l}}}(\epsilon) = 0 \) then with probability ≥ 1 - δ the \( h \in C \) that optimizes \( \epsilon \text{err}_{\text{D}_{\text{l}}}(h) \) and \( \epsilon \text{err}_{\text{D}_{\text{unl}}}(h) \) has \( \epsilon \text{err}(h) \leq \epsilon \).
Infinite Hypothesis Spaces / VC-dimension

Infinite Hypothesis Spaces
Assume \( \chi(h, x) \in \{0, 1\} \) and \( \chi(C) = \{\chi(h, x) \mid h \in C\} \) where \( \chi(C) = \chi(h, x) \).

\( C(m, D) \) - expected \# of splits of \( m \) points from \( D \) with concepts in \( C \).

Theorem

\[

m_u = O \left( \frac{\log(\text{VCdim}(C))}{\epsilon^2} + \frac{1}{\epsilon^2} \log \frac{2}{\delta} \right)
\]

unlabeled examples and

\[

m_l \geq \frac{2}{\epsilon} \log \left( \frac{2^m + \log \frac{2}{\delta}}{\delta} \right)
\]

labeled examples, where

\[

s = C_m(\epsilon + 2\delta/2m_u, D)
\]

are sufficient so that with probability at least \( 1 - \delta \), all \( h \in C \) with \( \hat{\epsilon}(h) = 0 \) and \( \epsilon_{unl}(h) \leq \epsilon \) have \( \epsilon(h) \leq \epsilon \), and furthermore all \( h \in C \) have

\[

\epsilon(h) - \epsilon_{unl}(h) \leq \epsilon
\]

Implication: If \( \epsilon_{unl}(h) \leq \epsilon \), then with prob. \( \geq 1 - \delta \), the \( h \in C \) that optimizes both \( \hat{\epsilon}(h) \) and \( \epsilon_{unl}(h) \) has \( \epsilon(h) \leq \epsilon \).

\[\text{No class on 4/21. Instead, go to Andy Carlson's thesis defense at 2:00pm in 8102 if you can. Slip homework under my office door 8111}\]