Semi-Supervised Learning

The main models we have been studying (PAC, mistake-bound) are for supervised learning.
- Given labeled examples $S = \{(x_i, y_i)\}$, try to learn a good prediction rule.
- But often labeled data is rare 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?

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.

There’s also a book “Semi-supervised learning” on the topic.

Co-training

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

[Blum & Mitchell’98] motivated by [Yarowsky’95]
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: 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. $c_1(x_1) = c_2(x_2) = c(x)$

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 $Pr[h(x)=1|c(x)=1] > Pr[h(x)=1|c(x)=0] + \epsilon$.
  - 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)
A really simple learning algorithm

Claim: if data has a separator of margin $\gamma$, there’s a reasonable chance a random hyperplane will have error $\leq \frac{1}{2} - \frac{\gamma}{4}$. (all hyperplanes through origin)

Proof:
$\omega$ Pick a (positive) example $x$. Consider the 2-d plane defined by $x$ and target $w^*$.
$\omega$ $Pr_{\omega}(h \cdot x \leq 0 | h \cdot w^* \geq 0) \leq (\pi/2 - \gamma)/\pi = \frac{1}{2} - \gamma/\pi$.
$\omega$ So, $E_h[err(h) | h \cdot w^* \geq 0] \leq \frac{1}{2} - \gamma/\pi$.
$\omega$ Since $err(h)$ is bounded between 0 and 1, there must be a reasonable chance of success.

QED

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!
  - Repeat process multiple times.
  - Get 4 kinds of hyps: (close to $c$, close to $-c$, close to 1, close to 0)

• [BBY04] if don’t want to assume indep, and $C$ is learnable from positive data only, then suffices for $D^*$ to have expansion.

Transductive SVM

• 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)

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 algorithm.
- 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

Graph-based methods

- 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]

Proposed 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 $\chi$)
- 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 $C$).

Proposed 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.
Proposed 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 \( \chi \))
- Require \( \chi \) to be an expectation over individual examples:
  - \( \chi(h, D) = E_{x \in D} \chi(h, x) \) (compatibility of \( h \) with \( D \)), \( \chi(h, x) \in [0, 1] \)
  - \( \text{err}_{\text{unl}}(h) = 1 - \chi(h, D) \) (incompatibility of \( h \) with \( D \) (unlabeled error rate of \( h \))

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 \( \gamma \) of \( h \).
  - Can be written as an expectation over individual examples:
    \( \chi(h, D) = E_{x \in D} \chi(h, x) \)
    - \( \chi(h, x) = 1 \) if \( \text{dist}(x, h) \geq \gamma \)
    - \( \chi(h, x) = 0 \) if \( \text{dist}(x, h) < \gamma \)

Co-Training, Compatibility

- Co-training: examples come as pairs \( (x_1, x_2) \) and the goal is to learn a pair of functions \( (h_1, h_2) \).
- Hope is that the two parts of the example are consistent.
  - Legal (and natural) notion of compatibility:
    - the compatibility of \( (h_1, h_2) \) and \( D \):
      - can be written as an expectation over examples:
        \( \chi((h_1, h_2), (x_1, x_2)) = 1 \) if \( h_1(x_1) = h_2(x_2) \)
        \( \chi((h_1, h_2), (x_1, x_2)) = 0 \) if \( h_1(x_1) \neq h_2(x_2) \)

Sample Complexity - Uniform convergence bounds

Finite Hypothesis Spaces, Doubly Realizable Case

- Define \( C_{\gamma}(\epsilon) = \{ h \in C : \text{err}_{\text{unl}}(h) \leq \epsilon \} \).

Theorem

If we see \( m_u \geq \frac{1}{\epsilon^2} \left[ \ln \gamma + \frac{\ln \frac{1}{\delta}}{2} \right] \) unlabeled examples and \( m_l \geq \frac{1}{\epsilon^2} \left[ \ln |C_{\gamma}(\epsilon)| + \frac{\ln \frac{1}{\delta}}{4} \right] \) labeled examples, then with probability \( \geq 1 - 6 \), all \( h \in C \) with \( \text{err}(h) = 0 \) and \( \text{err}_{\text{unl}}(h) = 0 \) have \( \text{err}(h) \leq \epsilon \).

- Bound the # of labeled examples as a measure of the helpfulness of \( D \) with respect to \( \chi \)
  - a helpful distribution is one in which \( C_{\gamma}(\epsilon) \) is small

Semi-Supervised Learning

Natural Formalization (PAC_{\chi})

- We will say an algorithm "PAC_{\chi}-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 \( \ln |C_{\gamma}(\epsilon)| \) 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 \( t \in [0, 1] \), if we see

\[
\begin{align*}
\mu_0 & \geq \frac{2}{\varepsilon} \left[ \ln |C(t + 2\varepsilon)| + \ln \frac{4}{\delta} \right] \\
\mu & \geq \frac{1}{\varepsilon} \left[ \ln |C(t + 2\varepsilon)| + \ln \frac{2}{\delta} \right]
\end{align*}
\]

unlabeled examples and

\[
\begin{align*}
\mu_0 & \geq \frac{1}{\varepsilon} \left[ \ln |C(t + 2\varepsilon)| + \ln \frac{2}{\delta} \right] \\
\mu & \geq \frac{1}{\varepsilon} \left[ \ln |C(t + 2\varepsilon)| + \ln \frac{2}{\delta} \right]
\end{align*}
\]

labeled examples, then with prob. \( \geq 1 - \delta \), all \( h \in C \) with \( \tilde{\mu}(h) = 0 \) and \( \tilde{\mu}_\mu(h) \leq 1 + \varepsilon \) have \( \tilde{e}(h) \leq \varepsilon \), and furthermore all \( h \in C \) with \( \tilde{\mu}_\mu(h) \leq 1 + \varepsilon \) have \( \tilde{e}(h) \leq 1 + \varepsilon \).

**Implication** If \( \tilde{e}(\mu') \leq 1 + \varepsilon \) and \( \tilde{e}(\mu) = 0 \) then with probability \( \geq 1 - \delta \) the \( h \in C \) that optimizes \( \tilde{e}(h) \) and \( \tilde{e}_\mu(h) \) has \( \tilde{e}(h) \leq \varepsilon \).

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]

\[\text{\vspace{1cm}}\]