15-859(B) Machine Learning Theory

Semi-Supervised Learning

Avrim Blum
04/22/09

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?

Semi-Supervised Learning

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.

Semi-Supervised Learning

Can we use unlabeled data to help?
• 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

[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.
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. $c_1(x_1)\leq c_2(x_2)\leq c(x)$

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: $D = p(D_1^+ \times D_2^+ \times (1-p)(D_1^- \times D_2^-)$, and if $C$ is SQ-learnable, then can learn from an initial "weakly-useful" $h_i$ 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.

Co-Training Theorems

- [BM98] if $x_1$, $x_2$ are independent given the label: $D = p(D_1^+ \times D_2^+ \times (1-p)(D_1^- \times D_2^-)$, and if $C$ is SQ-learnable, then can learn from an initial "weakly-useful" $h_i$ plus unlabeled data.
- 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)
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!
  - 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

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]

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

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 $\chi$)
- 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 $C$).

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]
- 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 $D$,
  - $\chi(h,x) \in [0,1]$
  - $err_{unl}(h)=1-\chi(h,D)$ 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

\[ \chi(h,D) = \mathbb{E}_{x \in D}[\chi(h,x)] \]

- \( \chi(h,x) = 0 \) if \( \text{dist}(x,h) \leq \gamma \)
- \( \chi(h,x) = 1 \) if \( \text{dist}(x,h) \geq \gamma \)

**Sample Complexity - Uniform convergence bounds**

**Finite Hypothesis Spaces, Doubly Realizable Case**

- Define \( C_{0,\gamma}(\epsilon) = \{ h \in C : \text{err}_{unl}(h) \leq \epsilon \} \).

**Theorem**

If we see

\[ m_u \geq \frac{1}{\epsilon} \left[ \ln \frac{1}{\epsilon} + \ln \frac{2}{\delta} \right] \]

unlabeled examples and

\[ m_l \geq \frac{1}{\epsilon^2} \left[ \ln |C_{0,\gamma}(\epsilon)| + \ln \frac{2}{\delta} \right] \]

labeled examples, then with probability \( \geq 1 - \delta \), all \( h \in C' \) with \( \text{err}_{\gamma}(h) = 0 \) and \( \text{err}_{unl}(h) = 0 \) 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 \( C_{0,\gamma}(\epsilon) \) is small

**Semi-Supervised Learning**

**Natural Formalization (PAC)**

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

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

unlabeled examples and

\[ m_l \geq \frac{2}{\epsilon^2} \left[ \ln |C_{0,\gamma}(\epsilon)| + \ln \frac{2}{\delta} \right] \]

labeled examples, then with probability \( \geq 1 - \delta \), all \( h \in C' \) with \( \text{err}_{\gamma}(h) = 0 \) and \( \text{err}_{unl}(h) \leq t + \epsilon \) have \( \text{err}(h) \leq \epsilon \), and furthermore all \( h \in C' \) with \( \text{err}_{unl}(h) = t + \epsilon \) have \( \text{err}_{\gamma}(h) \leq t + \epsilon \).

**Implication** if \( \text{err}_{\gamma}(c') \leq t + \epsilon \) and \( \text{err}_{\gamma}(c) = 0 \) then with probability \( \geq 1 - \delta \) the \( h \in C' \) that optimizes \( \text{err}_{\gamma}(h) \) and \( \text{err}_{unl}(h) \) has \( \text{err}(h) \leq \epsilon \).
Infinite hypothesis spaces / VC-dimension

Infinite Hypothesis Spaces
Assume $\chi(h,x) \in \{0,1\}$ and $\chi(C) = \{\chi_{h} : h \in C\}$ where $\chi_{h}(x) = \chi(h,x)$. $C(m,D)$ - expected # of splits of m points from D with concepts in C.

Theorem

- For algorithms that behave in a specific way:
  - first use the unlabeled data to choose a representative set of compatible hypotheses
  - then use the labeled sample to choose among these

E.g., in case of co-training linear separators with independence assumption:
- $\chi$-cover of compatible set $= \{0, 1, c^*, \neg c^*\}$

E.g., Transductive SVM when data is in two blobs.

- Ways unlabeled data can help in this model
  - If the target is highly compatible with D and have enough unlabeled data to estimate $\chi$ over all $h \in C$, then can reduce the search space (from C down to just those $h \in C$ whose estimated unlabeled error rate is low).
  - By providing an estimate of D, unlabeled data can allow a more refined distribution-specific notion of hypothesis space size (such as Annealed VC-entropy or the size of the smallest $\epsilon$-cover).
  - If D is nice so that the set of compatible $h \in C$ has a small $\epsilon$-cover and the elements of the cover are far apart, then can learn from even fewer labeled examples than the $1/\epsilon$ needed just to verify a good hypothesis.

- $\epsilon$-Cover-based bounds
  - For algorithms that behave in a specific way:
    - first use the unlabeled data to choose a representative set of compatible hypotheses
    - then use the labeled sample to choose among these

E.g., in case of co-training linear separators with independence assumption:
- $\epsilon$-cover of compatible set $= \{0, 1, c^*, \neg c^*\}$

E.g., Transductive SVM when data is in two blobs.

- Ways unlabeled data can help in this model
  - If the target is highly compatible with D and have enough unlabeled data to estimate $\chi$ over all $h \in C$, then can reduce the search space (from C down to just those $h \in C$ whose estimated unlabeled error rate is low).
  - By providing an estimate of D, unlabeled data can allow a more refined distribution-specific notion of hypothesis space size (such as Annealed VC-entropy or the size of the smallest $\epsilon$-cover).
  - If D is nice so that the set of compatible $h \in C$ has a small $\epsilon$-cover and the elements of the cover are far apart, then can learn from even fewer labeled examples than the $1/\epsilon$ needed just to verify a good hypothesis.