Reading the Web: Advanced Statistical Language Processing

www.cs.cmu.edu/~tom/rtw09/

Machine Learning 10-709

October 8, 2009

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University
When can Unlabeled Data help supervised learning?

Problem setting (the PAC learning setting):
• Set $X$ of instances drawn from unknown distribution $P(X)$
• Wish to learn target function $f: X \rightarrow Y$ (or, $P(Y|X)$)
• Given a set $H$ of possible hypotheses for $f$

Given:
• i.i.d. labeled examples $L = \{\langle x_1, y_1 \rangle \ldots \langle x_m, y_m \rangle\}$
• i.i.d. unlabeled examples $U = \{x_{m+1}, \ldots x_{m+n}\}$

Wish to find hypothesis with lowest true error:
$$\hat{f} \leftarrow \arg \min_{h \in H} \Pr_{x \in P(X)} [h(x) \neq f(x)]$$
What if $x_i$ are fully observed?

Can use $U \rightarrow \hat{P}(X)$ to alter optimization problem

- Wish to find

$$\hat{f} \leftarrow \arg \min_{h \in H, x \in X} \sum_{x \in X} \delta(h(x) \neq f(x)) P(x)$$

- Often approximate as

$$\hat{f} \leftarrow \arg \min_{h \in H} \sum \frac{1}{|L|} \sum_{(x,y) \in L} \delta(h(x) \neq y)$$

and when $y = f(x)$, this is just

$$\hat{f} \leftarrow \arg \min_{h \in H} \sum_{x \in X} \delta(h(x) \neq f(x)) \frac{n(x, L)}{|L|}$$

- Can use $U$ for improved approximation:

$$\hat{f} \leftarrow \arg \min_{h \in H} \sum_{x \in X} \delta(h(x) \neq f(x)) \frac{n(x, L) + n(x, U)}{|L| + |U|}$$
What if CoTraining Assumption Not Perfectly Satisfied?

• Idea: Want classifiers that produce a maximally consistent labeling of the data
• If learning is an optimization problem, what function should we optimize?
Co-EM  [Nigam & Ghani, 2000; Jones 2005]

Idea:

• Like co-training, train two coupled functions
  – $P(\text{class } | X_1)$, $P(\text{class } | X_2)$

• Like EM, iterative probabilistic algorithm
  – Assign probabilistic values to unobserved class labels
  – Updating model parameters (= labels of other feature set)

Goal to learn $X_1 \rightarrow Y$, $X_2 \rightarrow Y$, $X_1 \times X_2 \rightarrow Y$

$$
P(Y|X_1 = k) = \sum_j P(Y|X_2 = j)P(X_2 = j|X_1 = k)
$$

$$
P(Y|X_2 = j) = \sum_k P(Y|X_1 = k)P(X_1 = k|X_2 = j)
$$
CoRegularization

Key idea:
• define explicit learning objective
• optimize it directly

What objective?
What Objective Function?

\[ E = E_1 + E_2 \]

\[ E_1 = \sum_{<x,y> \in L} (y - \hat{g}_1(x_1))^2 \]

\[ E_2 = \sum_{<x,y> \in L} (y - \hat{g}_2(x_2))^2 \]

Error on labeled examples
What Objective Function?

\[ E = E_1 + E_2 + c_3 E_3 \]

\[ E_1 = \sum_{<x,y> \in L} (y - \hat{g}_1(x_1))^2 \]

\[ E_2 = \sum_{<x,y> \in L} (y - \hat{g}_2(x_2))^2 \]

\[ E_3 = \sum_{x \in U} (\hat{g}_1(x_1) - \hat{g}_2(x_2))^2 \]

- Error on labeled examples
- Disagreement over unlabeled
What Objective Function?

\[ E = E_1 + E_2 + c_3 E_3 + c_4 E_4 \]

\[ E_1 = \sum_{<x,y> \in L} (y - \hat{g}_1(x_1))^2 \]

Error on labeled examples

\[ E_2 = \sum_{<x,y> \in L} (y - \hat{g}_2(x_2))^2 \]

Disagreement over unlabeled

\[ E_3 = \sum_{x \in U} (\hat{g}_1(x_1) - \hat{g}_2(x_2))^2 \]

Misfit to estimated class priors

\[ E_4 = \left( \frac{1}{|L|} \sum_{<x,y> \in L} y \right) - \left( \frac{1}{|L| + |U|} \sum_{x \in L \cup U} \frac{\hat{g}_1(x_1) + \hat{g}_2(x_2)}{2} \right)^2 \]
What Function Approximators?
What Function Approximators?

\[ \hat{g}_1(x) = \frac{1}{\sum w_{j,1} x_j} \frac{1}{1 + e^{j_x}} \]

\[ \hat{g}_2(x) = \frac{1}{\sum w_{k,2} x_k} \frac{1}{1 + e^{k_x}} \]

- Same fn form as Logistic regression, Max Entropy

- Use gradient descent to simultaneously learn g1 and g2, directly minimizing \( E = E1 + E2 + E3 + E4 \)
Gradient CoTraining

\[ \hat{g}_1(x) = \frac{1}{1 + e^j \sum w_{j,1} x_j} \]

\[ \hat{g}_2(x) = \frac{1}{1 + e^j \sum w_{j,2} x_j} \]

Gradient

\[ \nabla E[\vec{w}] \equiv \left[ \frac{\partial E}{\partial w_0}, \frac{\partial E}{\partial w_1}, \ldots, \frac{\partial E}{\partial w_n} \right] \]

Training rule:

\[ \Delta \vec{w} = -\eta \nabla E[\vec{w}] \]
Classifying Jobs for FlipDog

<table>
<thead>
<tr>
<th>Job Title</th>
<th>Company</th>
<th>Location</th>
<th>Date Posted</th>
</tr>
</thead>
<tbody>
<tr>
<td>C++/Java Consultants</td>
<td>Elite Placement Services</td>
<td>NA</td>
<td>November 01, 2000</td>
</tr>
<tr>
<td>Chief Software Architect</td>
<td>Elite Placement Services</td>
<td>Houston, TX</td>
<td>November 01, 2000</td>
</tr>
<tr>
<td>Web Application Developers</td>
<td>MI Systems, Inc.</td>
<td>Houston, TX</td>
<td>November 01, 2000</td>
</tr>
<tr>
<td>Sales Consulting Engineer</td>
<td>Visual Numerics, Inc.</td>
<td>Houston, TX</td>
<td>November 01, 2000</td>
</tr>
<tr>
<td>Peoplesoft Software Analyst (Systems Analyst III)</td>
<td>I.T. Staffing, Inc.</td>
<td>Houston, TX</td>
<td>October 27, 2000</td>
</tr>
</tbody>
</table>
CoRegularization
Classifying FlipDog job descriptions: SysAdmin vs. WebProgrammer

Final Accuracy
Labeled data alone: 86%
CoRegularization: 96%
CoRegularization
Classifying Upper Case sequences as Person Names

**Error Rates**

<table>
<thead>
<tr>
<th></th>
<th>25 labeled</th>
<th>2300 labeled</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5000 unlabeled</td>
<td>5000 unlabeled</td>
</tr>
<tr>
<td>Using labeled data only</td>
<td>.24</td>
<td>.13</td>
</tr>
<tr>
<td>CoRegularization</td>
<td>.15  *</td>
<td>.11  *</td>
</tr>
<tr>
<td>CoRegularization without fitting class priors (E4)</td>
<td>.27  *</td>
<td></td>
</tr>
</tbody>
</table>

* sensitive to weights of error terms E3 and E4
CoTraining/CoRegularization

- Unlabeled data improves supervised learning when example features are redundantly sufficient
  - Family of algorithms that train multiple classifiers

- Theoretical results
  - Expected error for rote learning
  - If \( X_1, X_2 \) conditionally independent given \( Y \), Then
    - PAC learnable from weak initial classifier plus unlabeled data
    - disagreement between \( g_1(x_1) \) and \( g_2(x_2) \) bounds final classifier error

- Many real-world problems of this type
  - Semantic lexicon generation [Riloff, Jones 99], [Collins, Singer 99]
  - Web page classification [Blum, Mitchell 98]
  - Word sense disambiguation [Yarowsky 95]
  - Speech recognition [de Sa, Ballard 98]
  - Visual classification of cars [Levin, Viola, Freund 03]
Coupled training type 2

Wish to learn $f_1: X \rightarrow Y_1$, $f_2: X \rightarrow Y_2$, such that: $(\forall x) \ g(f_1(x), f_2(x))$

e.g.

location: NounPhraseInSentence $\rightarrow \{0,1\}$

politician: NounPhraseInSentence $\rightarrow \{0,1\}$

g(y_1,y_2) = \text{not (and(y_1,y_2))}

Luke is mayor of Pittsburgh.
Coupling functions with different outputs

[Daume, 2008]

Wish to learn \( f_1: X \rightarrow Y_1 \), \( f_2: X \rightarrow Y_2 \), such that: \( (\forall x) \ g(f_1(x), f_2(x)) \)

Key theoretical question: what is sample complexity? How does it depend on the coupling constraint, \( g \)?

Key insight:
- \( g \) will be most useful if the probability that it is satisfied by a high error \( f \) applied to a random \( x \), is low
Coupling functions with different outputs

[Daume, 2008]

Consider simpler one-sided learning of $f_2$, given we know $f_1$

1: Learn $h_2$ directly on $D$
2: For each example $(x, y_1) \in D^{unlab}$
3: Compute $y_2 = h_2(x)$
4: If $\chi(y_1, y_2)$, add $(x, y_2)$ to $D$
5: Relearn $h_2$ on the (augmented) $D$
6: Go to (2) if desired

**Definition 4.** We say the discrimination of $\chi$ for $h^0$ is $Pr_D[\chi(f_1(x), h^0(x))]^{-1}$. 
Coupling functions with different outputs

[Daume, 2008]

Theorem 1. Suppose $C_2$ is PAC-learnable with noise in the structured setting, $h_2^0$ is a weakly useful predictor of $f_2$, and $\chi$ is correct with respect to $D, f_1, f_2, h_2^0$, and has discrimination $\geq 2(|Y| - 1)$. Then $C_2$ is also PAC-learnable with one-sided hints.

(here $|Y| = |Y_1| \times |Y_2|$ is the number of values the two functions can take on)
Structured Output Learning

Suppose we wish to learn $f: X \rightarrow Y$ where $Y$ is a vector, tree, or graph?

Want to learn simultaneously the dependencies among components of $Y$, and their dependence on $X$.

Conditional Random Fields

see Sutton & McCallum, “An Introduction to Conditional Random Fields for Relational Learning”
Conditional Random Fields

Figure 1.2 Diagram of the relationship between naive Bayes, logistic regression, HMMs, linear-chain CRFs, generative models, and general CRFs.
an undirected graphical model as the set of all distributions that can be written in the form

\[ p(x, y) = \frac{1}{Z} \prod_A \Psi_A(x_A, y_A), \quad \text{(1.1)} \]

for any choice of factors \( F = \{ \Psi_A \} \), where \( \Psi_A : \mathcal{V}^n \to \mathbb{R}^+ \). (These functions are also called local functions or compatibility functions.) We will occasionally use the term random field to refer to a particular distribution among those defined by an undirected model. To reiterate, we will consistently use the term model to refer to a family of distributions, and random field (or more commonly, distribution) to refer to a single one.

The constant \( Z \) is a normalization factor defined as

\[ Z = \sum_{x, y} \prod_A \Psi_A(x_A, y_A), \quad \text{(1.2)} \]
**Definition 1.1**

Let $Y, X$ be random vectors, $\Lambda = \{\lambda_k\} \in \mathbb{R}^K$ be a parameter vector, and $\{f_k(y, y', x_t)\}_{k=1}^K$ be a set of real-valued feature functions. Then a **linear-chain conditional random field** is a distribution $p(y|x)$ that takes the form

$$p(y|x) = \frac{1}{Z(x)} \exp \left\{ \sum_{k=1}^K \lambda_k f_k(y_t, y_{t-1}, x_t) \right\},$$

(1.16)

where $Z(x)$ is an instance-specific normalization function

$$Z(x) = \sum_y \exp \left\{ \sum_{k=1}^K \lambda_k f_k(y_t, y_{t-1}, x_t) \right\}.$$

(1.17)

We have just seen that if the joint $p(y, x)$ factorizes as an HMM, then the associated conditional distribution $p(y|x)$ is a linear-chain CRF. This HMM-like CRF is
Figure 1.3  Graphical model of an HMM-like linear-chain CRF.

Figure 1.4  Graphical model of a linear-chain CRF in which the transition score depends on the current observation.
Figure 1.5 Graphical representation of a skip-chain CRF. Identical words are connected because they are likely to have the same label.

“skip” links added only for identical, capitalized tokens

| $w_t = w$ |
| $w_t$ matches [A-Z] [a-z]+ |
| $w_t$ matches [A-Z] [A-Z]+ |
| $w_t$ matches [A-Z] |
| $w_t$ matches [A-Z]+ |
| $w_t$ matches [A-Z]+[a-z]+[A-Z]+[a-z] |
| $w_t$ appears in list of first names, last names, honorifics, etc. |
| $w_t$ appears to be part of a time followed by a dash |
| $w_t$ appears to be part of a time preceded by a dash |
| $w_t$ appears to be part of a date |
| $T_t = T$ for all $k$ and $\delta \in [-4, 4]$ |

Table 1.1 Input features $q_k(x, t)$ for the seminars data. In the above $w_t$ is the word at position $t$, $T_t$ is the POS tag at position $t$, $w$ ranges over all words in the training data, and $T$ ranges over all part-of-speech tags returned by the Brill tagger. The “appears to be” features are based on hand-designed regular expressions that can span several tokens.
<table>
<thead>
<tr>
<th>System</th>
<th>stime</th>
<th>etime</th>
<th>location</th>
<th>speaker</th>
<th>overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIEN Peshkin and Pfeffer [2003]</td>
<td>96.0</td>
<td>98.8</td>
<td>87.1</td>
<td>76.9</td>
<td>89.7</td>
</tr>
<tr>
<td>Linear-chain CRF</td>
<td>97.5</td>
<td>97.5</td>
<td>88.3</td>
<td>77.3</td>
<td>90.2</td>
</tr>
<tr>
<td>Skip-chain CRF</td>
<td>96.7</td>
<td>97.2</td>
<td>88.1</td>
<td>80.4</td>
<td>90.6</td>
</tr>
</tbody>
</table>

**Table 1.2**  Comparison of $F_1$ performance on the seminars data. The top line gives a dynamic Bayes net that has been previously used on this data set. The skip-chain CRF beats the previous systems in overall $F_1$ and on the speaker field, which has proved to be the hardest field of the four. Overall $F_1$ is the average of the $F_1$ scores for the four fields.
Homework 3

- [http://www.cs.cmu.edu/~tom/10709_fall09/hw3.html](http://www.cs.cmu.edu/~tom/10709_fall09/hw3.html)
Further Reading


- S. Dasgupta, et al., “PAC Generalization Bounds for Co-training”, *NIPS 2001*

Some nodes are more important than others  [Jones, 2005]

Can use this for active learning...

<table>
<thead>
<tr>
<th>Noun-phrase</th>
<th>Outdegree</th>
</tr>
</thead>
<tbody>
<tr>
<td>you</td>
<td>1656</td>
</tr>
<tr>
<td>we</td>
<td>1479</td>
</tr>
<tr>
<td>it</td>
<td>1173</td>
</tr>
<tr>
<td>company</td>
<td>1043</td>
</tr>
<tr>
<td>this</td>
<td>635</td>
</tr>
<tr>
<td>all</td>
<td>520</td>
</tr>
<tr>
<td>they</td>
<td>500</td>
</tr>
<tr>
<td>information</td>
<td>448</td>
</tr>
<tr>
<td>us</td>
<td>367</td>
</tr>
<tr>
<td>any</td>
<td>339</td>
</tr>
<tr>
<td>products</td>
<td>332</td>
</tr>
<tr>
<td>i</td>
<td>319</td>
</tr>
<tr>
<td>site</td>
<td>314</td>
</tr>
<tr>
<td>one</td>
<td>311</td>
</tr>
<tr>
<td>1996</td>
<td>282</td>
</tr>
<tr>
<td>he</td>
<td>269</td>
</tr>
<tr>
<td>customers</td>
<td>269</td>
</tr>
<tr>
<td>these</td>
<td>263</td>
</tr>
<tr>
<td>them</td>
<td>263</td>
</tr>
<tr>
<td>time</td>
<td>234</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Context</th>
<th>Outdegree</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;&lt;x&gt;&gt; including</td>
<td>683</td>
</tr>
<tr>
<td>including &lt;&lt;x&gt;&gt;</td>
<td>612</td>
</tr>
<tr>
<td>&lt;&lt;x&gt;&gt; provides</td>
<td>565</td>
</tr>
<tr>
<td>provides &lt;&lt;x&gt;&gt;</td>
<td>565</td>
</tr>
<tr>
<td>provide &lt;&lt;x&gt;&gt;</td>
<td>390</td>
</tr>
<tr>
<td>&lt;&lt;x&gt;&gt; include</td>
<td>389</td>
</tr>
<tr>
<td>include &lt;&lt;x&gt;&gt;</td>
<td>375</td>
</tr>
<tr>
<td>&lt;&lt;x&gt;&gt; provide</td>
<td>364</td>
</tr>
<tr>
<td>one of &lt;&lt;x&gt;&gt;</td>
<td>354</td>
</tr>
<tr>
<td>&lt;&lt;x&gt;&gt; made</td>
<td>345</td>
</tr>
<tr>
<td>&lt;&lt;x&gt;&gt; offers</td>
<td>338</td>
</tr>
<tr>
<td>offers &lt;&lt;x&gt;&gt;</td>
<td>320</td>
</tr>
<tr>
<td>&lt;&lt;x&gt;&gt; said</td>
<td>287</td>
</tr>
<tr>
<td>&lt;&lt;x&gt;&gt; used</td>
<td>283</td>
</tr>
<tr>
<td>includes &lt;&lt;x&gt;&gt;</td>
<td>279</td>
</tr>
<tr>
<td>to provide &lt;&lt;x&gt;&gt;</td>
<td>266</td>
</tr>
<tr>
<td>use &lt;&lt;x&gt;&gt;</td>
<td>263</td>
</tr>
<tr>
<td>like &lt;&lt;x&gt;&gt;</td>
<td>260</td>
</tr>
<tr>
<td>variety of &lt;&lt;x&gt;&gt;</td>
<td>252</td>
</tr>
<tr>
<td>&lt;&lt;x&gt;&gt; includes</td>
<td>250</td>
</tr>
</tbody>
</table>