Semi-Supervised Learning

Maria-Florina Balcan

10/24/2018

Readings:


Fully Supervised Learning

Data Source

Distribution D on X

Labeled Examples

(x_1, c^*(x_1)), …, (x_m, c^*(x_m))

h : X → Y

Alg. outputs

Learning Algorithm

Expert / Oracle

c^* : X → Y
Fully Supervised Learning

Data Source: Distribution $D$ on $X$

Labeled Examples: $(x_1, c^*(x_1)), \ldots, (x_m, c^*(x_m))$

Alg. outputs: $h : X \rightarrow Y$

Expert / Oracle: $c^* : X \rightarrow Y$

$S_l = \{(x_1, y_1), \ldots, (x_m, y_m)\}$

$x_i$ drawn i.i.d from $D$, $y_i = c^*(x_i)$

Goal: $h$ has small error over $D$.

$\text{err}_D(h) = \Pr_{x \sim D} (h(x) \neq c^*(x))$
Classic Paradigm Insufficient Nowadays

Modern applications: **massive amounts** of raw data.

Only a tiny fraction can be annotated by human experts.

Protein sequences  Billions of webpages  Images
Modern applications: **massive amounts** of raw data.

Techniques that best utilize data, **minimizing need for expert/human intervention**.

Paradigms where there has been great progress.

- Semi-supervised Learning, (Inter)active Learning.
Active Learning

Learning Algorithm

Unlabeled data

raw data

Classifier

face

not face

Expert Labeler
Semi-Supervised Learning

Learning Algorithm

Classifier

Raw data

Unlabeled data

Labeled data

face

not face

Expert Labeler
Semi-Supervised Learning

Data Source

Learning Algorithm

Unlabeled examples

Unlabeled examples

Expert / Oracle

Labeled Examples

Algorithm outputs a classifier

\[ S_l = \{(x_1, y_1), ..., (x_m, y_m)\} \]

\[ x_i \text{ drawn i.i.d from } D, \quad y_i = c^*(x_i) \]

\[ S_u = \{x_1, ..., x_m\} \text{ drawn i.i.d from } D \]

Goal: \ h \ has small error over \ D. \n
\[ \text{err}_D(h) = \Pr_{x \sim D}(h(x) \neq c^*(x)) \]
Semi-supervised Learning

• Major topic of research in ML.
• Several methods have been developed to try to use unlabeled data to improve performance, e.g.:
  - Transductive SVM [Joachims ‘99]
  - Co-training [Blum & Mitchell ‘98]
  - Graph-based methods [B&C01], [ZGL03]

Workshops [ICML ‘03, ICML’ 05, ...]

Books:
• Semi-Supervised Learning, MIT 2006
  O. Chapelle, B. Scholkopf and A. Zien (eds)
• Introduction to Semi-Supervised Learning, Morgan & Claypool, 2009 Zhu & Goldberg
Semi-supervised Learning

- Major topic of research in ML.
- Several methods have been developed to try to use unlabeled data to improve performance, e.g.:
  - Transductive SVM [Joachims ‘99]
  - Co-training [Blum & Mitchell ‘98]
  - Graph-based methods [B&CO1], [ZGL03]

Test of time awards at ICML!

Both wide spread applications and solid foundational understanding!!!
Semi-supervised Learning

- Major topic of research in ML.
- Several methods have been developed to try to use unlabeled data to improve performance, e.g.:
  - Transductive SVM [Joachims ‘99]
  - Co-training [Blum & Mitchell ‘98]
  - Graph-based methods [B&C01], [ZGL03]

Today: discuss these methods.

Very interesting, they all exploit unlabeled data in different, very interesting and creative ways.
Unlabeled data useful if we have beliefs not only about the form of the target, but also about its relationship with the underlying distribution.

Key Insight

Semi-supervised learning: no querying. Just have lots of additional unlabeled data.

A bit puzzling; unclear what unlabeled data can do for us.... It is missing the most important info. How can it help us in substantial ways?
Semi-supervised SVM
[Joachims ‘99]
Margins based regularity

Target goes through low density regions (large margin).

• assume we are looking for linear separator
• belief: should exist one with large separation

Labeled data only

SVM

Transductive SVM
Transductive Support Vector Machines

Optimize for the separator with large margin wrt labeled and unlabeled data. [Joachims ‘99]

Input: $S_l = \{(x_1, y_1), \ldots, (x_{m_l}, y_{m_l})\}$

$S_u = \{x_1, \ldots, x_{m_u}\}$

$$\text{argmin}_w \|w\|^2 \text{ s.t.}:
\begin{align*}
\cdot \quad & y_i w \cdot x_i \geq 1, \text{ for all } i \in \{1, \ldots, m_l\} \\
\cdot \quad & \tilde{y}_u w \cdot x_u \geq 1, \text{ for all } u \in \{1, \ldots, m_u\} \\
\cdot \quad & \tilde{y}_u \in \{-1, 1\} \text{ for all } u \in \{1, \ldots, m_u\}
\end{align*}$$

Find a labeling of the unlabeled sample and $w$ s.t. $w$ separates both labeled and unlabeled data with maximum margin.
Transductive Support Vector Machines

Optimize for the separator with large margin wrt labeled and unlabeled data. [Joachims ‘99]

Input: $S_l = \{(x_1, y_1), \ldots, (x_{m_l}, y_{m_l})\}$
$S_u = \{x_1, \ldots, x_{m_u}\}$

$$\arg\min_w \|w\|^2 + C \sum_i \xi_i + C \sum_u \xi_u$$

- $y_i w \cdot x_i \geq 1 - \xi_i$, for all $i \in \{1, \ldots, m_l\}$
- $\hat{y_u} w \cdot x_u \geq 1 - \xi_u$, for all $u \in \{1, \ldots, m_u\}$
- $\hat{y_u} \in \{-1, 1\}$ for all $u \in \{1, \ldots, m_u\}$

Find a labeling of the unlabeled sample and $w$ s.t. $w$ separates both labeled and unlabeled data with maximum margin.
Transductive Support Vector Machines

Optimize for the separator with large margin wrt labeled and unlabeled data.

Input: $S_l = \{(x_1, y_1), \ldots, (x_{m_l}, y_{m_l})\}$
$S_u = \{x_1, \ldots, x_{m_u}\}$

$$\text{argmin}_w \|w\|^2 + C \sum_i \xi_i + C \sum_u \xi_u$$

- $y_i w \cdot x_i \geq 1 - \xi_i$, for all $i \in \{1, \ldots, m_l\}$
- $\widetilde{y}_u w \cdot x_u \geq 1 - \xi_u$, for all $u \in \{1, \ldots, m_u\}$
- $\widetilde{y}_u \in \{-1, 1\}$ for all $u \in \{1, \ldots, m_u\}$

NP-hard..... Convex only after you guessed the labels... too many possible guesses...
Transductive Support Vector Machines

Optimize for the separator with large margin wrt labeled and unlabeled data.

Heuristics (Joachims) high level idea:

- First maximize margin over the labeled points
- Use this to give initial labels to unlabeled points based on this separator.
- Try flipping labels of unlabeled points to see if doing so can increase margin

Keep going until no more improvements. Finds a locally-optimal solution.
Experiments [Joachims99]

Figure 6: Average P/R-breakeven point on the Reuters dataset for different training set sizes and a test set size of 3,299.
Transductive Support Vector Machines

Helpful distribution

Non-helpful distributions

Margin not satisfied

Margin satisfied

1/γ² clusters, all partitions separable by large margin
Co-training

[Blum & Mitchell ’98]

Different type of underlying regularity assumption:
Consistency or Agreement Between Parts
Co-training: Self-consistency

Agreement between two parts: co-training \cite{Blum-Mitchell98}.

- examples contain two sufficient sets of features, \( x = \langle x_1, x_2 \rangle \)
- belief: the parts are consistent, i.e. \( \exists c_1, c_2 \text{ s.t. } c_1(x_1) = c_2(x_2) = c^*(x) \)

For example, if we want to classify web pages: \( x = \langle x_1, x_2 \rangle \) as faculty member homepage or not
Iterative 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.

Idea: Use unlabeled data to propagate learned information.
Iterative 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.

**Idea:** Use unlabeled data to **propagate** learned information.

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.
Iterative Co-Training

Works by using unlabeled data to propagate learned information.

- Have learning algos $A_1$, $A_2$ on each of the two views.
- Use labeled data to learn two initial hyp. $h_1$, $h_2$.

Repeat

- Look through unlabeled data to find examples where one of $h_i$ is confident but other is not.
- Have the confident $h_i$ label it for algorithm $A_{3-i}$. 
Original Application: Webpage classification

12 labeled examples, 1000 unlabeled

<table>
<thead>
<tr>
<th>Method</th>
<th>Page-based</th>
<th>Hyperlink-based</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>Std. Supervised</td>
<td>12.9</td>
<td>12.4</td>
<td>11.1</td>
</tr>
<tr>
<td>Co-training</td>
<td>6.2</td>
<td>11.6</td>
<td>5.0</td>
</tr>
<tr>
<td>Just say neg</td>
<td>22</td>
<td>22</td>
<td>22</td>
</tr>
</tbody>
</table>

(sample run)
Iterative Co-Training
A Simple Example: Learning Intervals

Use labeled data to learn $h_1^1$ and $h_2^1$

Use unlabeled data to bootstrap

Use labeled data to learn $h_1^2$ and $h_2^2$
Expansion, Examples: Learning Intervals

Consistency: zero probability mass in the regions

Non-expanding (non-helpful) distribution

Expanding distribution
Co-training: Theoretical Guarantees

What properties do we need for co-training to work well?
We need assumptions about:
1. the underlying data distribution
2. the learning algos on the two sides

[Blum & Mitchell, COLT '98]
1. Independence given the label
2. Alg. for learning from random noise.

[Balcan, Blum, Yang, NIPS 2004]
1. Distributional expansion.
2. Alg. for learning from positive data only.
Co-training/Multi-view SSL: Direct Optimization of Agreement

Input: \( S_l = \{(x_1, y_1), \ldots, (x_{m_l}, y_{m_l})\} \)
\( S_u = \{x_1, \ldots, x_{m_u}\} \)

\[
\arg\min_{h_1, h_2} \sum_{l=1}^{2} \sum_{i=1}^{m_l} l(h_1(x_i), y_i) + C \sum_{i=1}^{m_u} \text{agreement}(h_1(x_i), h_2(x_i))
\]

Each of them has small labeled error
Regularizer to encourage agreement over unlabeled data

E.g.,
Co-training/Multi-view SSL: Direct Optimization of Agreement

Input: $S_l = \{(x_1, y_1), \ldots, (x_{m_l}, y_{m_l})\}$
$S_u = \{x_1, \ldots, x_{m_u}\}$

$$\arg\min_{h_1, h_2} \sum_{l=1}^{2} \sum_{i=1}^{m_l} l(h_l(x_i), y_i) + c \sum_{i=1}^{m_u} \text{agreement}(h_1(x_i), h_2(x_i))$$

- $l(h(x_i), y_i)$ loss function
  - E.g., square loss $l(h(x_i), y_i) = (y_i - h(x_i))^2$
  - E.g., 0/1 loss $l(h(x_i), y_i) = 1_{y_i \neq h(x_i)}$

E.g.,

Original Application: Webpage classification

12 labeled examples, 1000 unlabeled

<table>
<thead>
<tr>
<th></th>
<th>Page-based</th>
<th>Hyperlink-based</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>Std. Supervised</td>
<td>12.9</td>
<td>12.4</td>
<td>11.1</td>
</tr>
<tr>
<td>Co-training</td>
<td>6.2</td>
<td>11.6</td>
<td>5.0</td>
</tr>
<tr>
<td>Just say neg</td>
<td>22</td>
<td>22</td>
<td>22</td>
</tr>
</tbody>
</table>

(sample run)
Many Other Applications

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”
  ...

Central to NELL (Never Ending Learning)!
See http://rtw.ml.cmu.edu/rtw/
Similarity Based Regularity
[Blum&Chwala01], [ZhuGhahramaniLafferty03]
Graph-based Methods

• Assume we are given a pairwise similarity fnc and that very similar examples probably have the same label.

• If we have a lot of labeled data, this suggests a Nearest-Neighbor type of algorithm.

• If you have a lot of unlabeled data, perhaps can use them as “stepping stones”.

<table>
<thead>
<tr>
<th>0 2</th>
<th>0 2 2 2 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>not similar</td>
<td>‘indirectly’ similar with stepping stones</td>
</tr>
</tbody>
</table>

E.g., handwritten digits [Zhu07]:
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.
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.

Graph-based Methods

Often, transductive approach. (Given L + U, output predictions on U). Are allowed to output any labeling of $L \cup U$.

Main Idea:

• Construct graph $G$ with edges between very similar examples.

• Might have also glued together in $G$ examples of different classes.

• Run a graph partitioning algorithm to separate the graph into pieces.

Several methods:
- Minimum/Multiway cut [Blum&Chawla01]
- Minimum “soft-cut” [ZhuGhahramaniLafferty'03]
- Spectral partitioning
- ...
**Objective:** Solve for labels on unlabeled points that minimize total weight of edges whose endpoints have different labels. (i.e., the total weight of bad edges)

- If just two labels, can be solved efficiently using max-flow min-cut algorithms
  - Create super-source $s$ connected by edges of weight $\infty$ to all $+$ labeled pts.
  - Create super-sink $t$ connected by edges of weight $\infty$ to all $-$ labeled pts.
  - Find minimum-weight $s$-$t$ cut
**Minimum “soft cut”**

[ZhuGhahramaniLafferty'03]

**Objective** Solve for probability vector over labels \( f_i \) on each unlabeled point \( i \).

(labeled points get coordinate vectors in direction of their known label)

- Minimize \( \sum_{e=(i,j)} w_e \| f_i - f_j \|^2 \)
  - where \( \| f_i - f_j \| \) is Euclidean distance.

- Can be done efficiently by solving a set of linear equations.
How to Create the Graph

- Empirically, the following works well:
  1. Compute distance between $i, j$
  2. For each $i$, connect to its kNN. k very small but still connects the graph
  3. Optionally put weights on (only) those edges
     \[
     \exp \left( -\frac{\|x_i - x_j\|^2}{2\sigma^2} \right)
     \]
  4. Tune $\sigma$
What You Should Know

• Unlabeled data useful if we have beliefs not only about the form of the target, but also about its relationship with the underlying distribution.

• Different types of algorithms (based on different beliefs).
  
  - **Transductive SVM** [Joachims ’99]
  - **Co-training** [Blum & Mitchell ’98]
  - **Graph-based methods** [B&C01], [ZGL03]
Additional Material on Co-training
Co-training [BlumMitchell’98]

Say that \( h_1 \) is a \textbf{weakly-useful predictor} if
\[
\Pr[h_1(x) = 1|c_1(x) = 1] > \Pr[h_1(x) = 1|c_1(x) = 0] + \gamma.
\]

Has higher probability of saying positive on a true positive than it does on a true negative, by at least some gap \( \gamma \)

Say we have enough labeled data to produce such a starting point.

\textbf{Theorem:} if \( C \) is learnable from random classification noise, we can use a weakly-useful \( h_1 \) plus \textbf{unlabeled} data to create a strong learner under independence given the label.
Co-training: Benefits in Principle

[BalcanBlum05]: Under independence given the label, any pair \( \langle h_1, h_2 \rangle \) with high agreement over unlabeled data must be close to:

- \( \langle c_1^*, c_2^* \rangle \), \( \langle \neg c_1^*, \neg c_2^* \rangle \), \( \langle \text{true}, \text{true} \rangle \), or \( \langle \text{false}, \text{false} \rangle \)

![Diagram](image-url)
Co-training: Benefits in Principle

[BalcanBlum05]: Under independence given the label, any pair $\langle h_1, h_2 \rangle$ with high agreement over unlabeled data must be close to:

- $\langle c_1^*, c_2^* \rangle$, $\langle \neg c_1^*, \neg c_2^* \rangle$, $\langle \text{true, true} \rangle$, or $\langle \text{false, false} \rangle$

E.g.,

Because of independence, we will see a lot of disagreement....