Semi-Supervised Learning of Sequence Models via Method of Moments

EMNLP - Empirical Methods for Natural Language Processing
November 1-6, 2016     Austin, Texas

Zita Marinho
IST, University of Lisbon
Robotics Institute, CMU
zmarinho@cmu.edu

André F. T. Martins
IT, IST, University of Lisbon
Unbabel
andre.martins@unbabel.com

Shay B. Cohen
School of Informatics
University of Edinburgh
scohen@inf.ed.ac.uk

Noah A. Smith
Computer Science & Eng.
University of Washington
nasmith@cs.washington.edu
observed data \{w_1, w_2, w_3, \ldots, w_6\}
labels \{y_1, y_2, y_3, \ldots, y_6\}
Semi-supervised sequence labeling with MoM

Introduction

Sequence Labeling

observed data \{w_1, w_2, w_3, \ldots, w_6\}

labels \{y_1, y_2, y_3, \ldots, y_6\}
Sequence Labeling

observed data \{w_1, w_2, w_3, \ldots, w_6\}
labels \{y_1, y_2, y_3, \ldots, y_6\}
K^6 possible assignments

\[ y_1, y_2, y_3, \ldots, y_6 \]

observed data \( \{w_1, w_2, w_3, \ldots, w_6\} \)

labels \( \{y_1, y_2, y_3, \ldots, y_6\} \)
Learn parameters?

\[ p(y_t | y_{t-1}) \]

\[ p(w_t | y_t) \]

- supervised learning
- unsupervised/semi-supervised (this talk)
Learn parameters?

Hidden Markov Model

- supervised learning
- unsupervised/semi-supervised (this talk)
- model can be extended to include features

Maximum Likelihood estimation (MLE)

- exact inference is hard
- EM sensitive to local optima (depends on initialization)
- EM expensive in large datasets (several inference passes)

Method of Moments estimation (MoM)

- computationally efficient
- no local optima
- one pass over data
### Hidden Markov Model

<table>
<thead>
<tr>
<th></th>
<th>via Maximum Likelihood Estimation</th>
<th>via Method of Moments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MLE</td>
<td>MLE</td>
</tr>
<tr>
<td>semi-supervised learning</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>unsupervised learning</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>


Arora et al., *A Practical Algorithm for Topic Modeling with Provable Guarantees*, ICML 2013
Learning sequence models via MoM

Outline

1. Learn HMM models via MoM
2. Solve a QP
3. Extend to feature-based model
4. Experiments
Key insight:

1. **Conditional Independence:**
   
   infer label by looking at context

2. **Anchor Trick:**
   
   learn a proxy for labels with anchors
1. Conditional Independence
1. Conditional Independence

context = \{ w_{-1}, w_{+1} \}
1. Conditional Independence

context

\[ y_{t-1} \rightarrow W_{t-1} \rightarrow \text{tasted} \]

\[ y_t \rightarrow W_t \rightarrow \text{like} \]

\[ y_{t+1} \rightarrow W_{t+1} \rightarrow \text{chimichangas} \]
1. Conditional Independence

![Diagram showing conditional independence in a sequence labeling problem]

- Context:
  - $y_{t-1}$
  - $y_t$
  - $y_{t+1}$

- Verb:
  - $W_{t-1}$
  - $W_t$
  - $W_{t+1}$

- Words:
  - fajitas
  - like
  - i

Problem Statement: Semi-supervised sequence labeling with MoM
“You shall know a word by the company it keeps.”

Firth, 1957
1. Conditional Independence

word \(\perp\) context | label

start \(\rightarrow\) \(y_1\) \(\rightarrow\) \(y_2\) \(\rightarrow\) \(y_3\) \(\rightarrow\) \(y_4\) \(\rightarrow\) \(y_5\) \(\rightarrow\) \(y_6\) \(\rightarrow\) \(y_7\) \(\rightarrow\) stop

\(W_1\) \(\rightarrow\) \(W_2\) \(\rightarrow\) \(W_3\) \(\rightarrow\) \(W_4\) \(\rightarrow\) \(W_5\) \(\rightarrow\) \(W_6\) \(\rightarrow\) \(W_7\)

\(W_1\) hehe \(W_2\) its \(W_3\) gonna \(W_4\) b \(W_5\) a \(W_6\) good \(W_7\) day
\[ p( \text{verb} \mid \text{be} ) = 1 \]
\[ p( \text{label} \neq \text{verb} \mid \text{be} ) = 0 \]

2. Anchor Trick

all instances of \( \text{be} = \text{verb} \)

Arora et al., *A Practical Algorithm for Topic Modeling with Provable Guarantees*, ICML 2013
More anchors per label

\[
\text{verb} = \text{b, be, are, is, am, have, going}
\]

more than 1 anchor word \(\longrightarrow\) less biased context estimates
How to find anchors?

- small labeled corpus
- small lexicon

<table>
<thead>
<tr>
<th>Austin</th>
<th>noun</th>
</tr>
</thead>
<tbody>
<tr>
<td>airport</td>
<td></td>
</tr>
<tr>
<td>playground</td>
<td></td>
</tr>
<tr>
<td>am, be, is, are</td>
<td>verb</td>
</tr>
<tr>
<td>go, make, made</td>
<td></td>
</tr>
<tr>
<td>become</td>
<td></td>
</tr>
<tr>
<td>he, it, she</td>
<td>pron</td>
</tr>
<tr>
<td>so, on, of</td>
<td>adp</td>
</tr>
</tbody>
</table>
co-occurrences in data

\[ W_{t-1} \quad W_t \quad W_{t+1} \quad W_{t+2} \]

context

Andrew fights like Jet Li.
Ann sings like me.

eat Fruit like cherry.
Children like ice-cream.
Method of moments

$\begin{align*}
  W_{t-1} & \quad W_t & \quad W_{t+1} & \quad W_{t+2} \\
  \text{context} & \\
\end{align*}$

$Q$

$p(\text{context} | \text{word})$

Andrew fights **like** Jet Li.  
Ann sings **like** me.

eat Fruit **like** cherry. 
Children **like** ice-cream.
Let there be love.

Bill will be a ninja.

<table>
<thead>
<tr>
<th>word</th>
<th>context</th>
</tr>
</thead>
<tbody>
<tr>
<td>fights</td>
<td>a</td>
</tr>
<tr>
<td>cherry</td>
<td>me</td>
</tr>
<tr>
<td>there</td>
<td>will</td>
</tr>
<tr>
<td>love</td>
<td>Q</td>
</tr>
</tbody>
</table>

Method of moments
1. Conditional Independence

\[ p(\text{context} | \text{word}) = \sum_{\text{labels}} p(\text{label} | \text{word}) \cdot p(\text{context} | \text{label}) \]
Method of moments

1. Conditional Independence

\[ p(\text{context} \mid \text{word}) = \sum_{\text{labels}} p(\text{label} \mid \text{word}) \quad p(\text{context} \mid \text{label}) \]

2. Anchor Trick

\[ p(\text{context} \mid \text{word}) = \sum_{\text{labels}} p(\text{label} \mid \text{word}) \quad p(\text{context} \mid \text{anchors}) \]
Outline

1. Learn HMM models via MoM

2. Solve a QP

3. Extend to feature-based model

4. Experiments
Method of Moments

\[ p(\text{context} \mid \text{word}) = p(\text{label} \mid \text{word}) \times p(\text{context} \mid \text{label}) \]

\[ Q = \gamma \Gamma R \]

\[ \gamma = \text{argmin} \| q - R \gamma \|^2 \]

\[ 0 \leq \gamma \leq 1 \]

\[ \sum_{\text{labels}} \gamma = 1 \]
Method of Moments

\[ p(\text{context} | \text{word}) = p(\text{label} | \text{word}) \times p(\text{context} | \text{label}) \]

\[ \gamma = \text{argmin} \| q - R\gamma \|^2 + \lambda \| \gamma_{\text{sup}} - \gamma \|^2 \]

\[ 0 \leq \gamma \leq 1 \]

\[ \sum_{\text{labels}} \gamma = 1 \]
Method of Moments

\[ p(\text{context} \mid \text{word}) \cdot p(\text{label} \mid \text{word}) \cdot p(\text{context} \mid \text{label}) = q \cdot \Gamma \cdot R \]

\[ y = \arg\min_{\gamma} \| q - R \gamma \|_2^2 + \lambda \| \gamma_{\text{sup}} - \gamma \|_2^2 \]

\[ 0 \leq \gamma \leq 1 \]

\[ \sum_{\text{labels}} \gamma = 1 \]

estimated from labeled data

estimated from unlabeled data
Learn parameters?

\[ p(\text{label} \mid \text{word}) \]

\[ \gamma \]

coefficients

\[ \overset{\downarrow}{\text{Observation Matrix}} \]

Bayes’ Rule

\[ p(\text{word} \mid \text{label}) = \frac{p(\text{word})}{p(\text{label})} \]

\[ p(\text{label}) = \sum_{\text{words}} \gamma p(\text{word}) \]
Learn parameters?

**Observation Matrix**

Bayes’ Rule

\[ p(\text{word} \mid \text{label}) = \gamma \frac{p(\text{word})}{p(\text{label})} \]

**Transition Matrix**

- estimate from labeled data only
Learning sequence models via MoM

Outline

1. Learn HMM models via MoM

2. Relax the notion of anchors

3. Solve a QP

4. Experiments
Semi-supervised Twitter POS tagging

Twitter dataset

2.7 M unlabeled tweets
1000-100 labeled tweets

12 Universal POS

≈ 200k words

hehe its gonna b a good day

x  prt  verb  verb  det  adj  noun

Slav Petrov et al., A Universal Part-of-Speech Tagset, 2011
Owoputi et al., Improved part-of-speech tagging for online conversational text with word clusters. 2013
Twitter POS tagging

150 training labeled sequences

<table>
<thead>
<tr>
<th>Method</th>
<th>Tagging Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM</td>
<td>71.7</td>
</tr>
<tr>
<td>EM</td>
<td>77.2</td>
</tr>
<tr>
<td>Self-training</td>
<td>78.2</td>
</tr>
<tr>
<td>AHMM</td>
<td>84.3</td>
</tr>
</tbody>
</table>
Twitter POS tagging

1000 training labeled sequences

HMM tagging accuracy

- HMM: 81.1
- EM: 83.1
- self-training: 86.1
- AHMM: 88.0

Bar chart showing the tagging accuracy for different models with 1000 training labeled sequences.
Learning sequence models via MoM

Outline

1. Learn HMM models via MoM
2. Relax the notion of anchors
3. Extend to feature HMM
4. Experiments
Log-linear model

$\phi$ (word) →
- is upper
- is title
- is digit
- is url
- starts #
- is emoticon

\[
\begin{align*}
&\text{start} & y_1 & y_2 & y_3 & y_4 & y_5 & y_6 & y_7 & \text{stop} \\
&W_1 & W_2 & W_3 & \text{w} & W_5 & W_6 & W_7 \\
&:\) & \text{wait} & \text{now} & \text{I} & \text{am} & \text{goin} & 2 \\
\end{align*}
\]

1. Conditional Independence

word \perp context \mid label

ϕ (word)
ψ (context)
Log-linear model

\[
\begin{align*}
\psi(\text{context}) & = \phi(\text{word}) \times \psi(\text{context}) \\
Q & = \Gamma \\
\Phi(\text{word}) | \text{label} & = \frac{E[\psi(\text{context}) | \text{label}]}{E[\Phi(\text{word})]} \\
\Gamma & = \frac{E[\Phi(\text{word}) | \text{label} \] p(\text{label})}{E[\Phi(\text{word})]} \\
\end{align*}
\]
\[ \gamma = \arg\min_{\gamma} \left\| q - R\gamma \right\|^2 + \lambda \left\| Y_{\text{sup}} - Y \right\|^2 \]

\[ \sum_{\text{labels}} Y = 1 \]

- solve per feature dimension \( \Phi_j \)
Learn parameters?

\[ \gamma = \frac{\mathbb{E}[\Phi(\text{word}) \mid \text{label}]}{\mathbb{E}[\Phi(\text{word})]} \frac{\mathbb{E}[\Phi(\text{word})]}{p(\text{label})} \]

mean parameters

\[ \mu = \mathbb{E}[\Phi(\text{word}) \mid \text{label}] = \gamma \frac{\mathbb{E}[\Phi(\text{word})]}{p(\text{label})} \]
Learn parameters?

mean parameters \[\mu\] → canonical parameters \[\theta_y\]

Fenchel-Legendre Duality

\[\theta^*_y = \arg\max_{\theta_y} \theta_y^\top \mu_y - \log Z_y\]

partition function

\[Z_y = \sum_w \exp (\theta_y^\top t_w)\]
Algorithm

- Compute moments
- Find anchors
- Solve QP

\[ Q \]

\[ R \]

\[ \Gamma \]

**Mean parameters** \( \mu \)

**Canonical parameters** \( \theta_y \)

- Solve maxent problem

Approximate times:
- \(~10s\) min
- \(~10s\) sec
- \(~10s\) min
- \(~10s\) min
- \(~2-3h\)
- \(~secs\)
Algorithm

compute moments

find anchors

solve QP

\[ \mu \]

mean parameters

\[ \theta_y \]

canonical parameters

solve maxent problem

\[ \Gamma \]

\[ Q \]

find anchors

supervision

solve QP

\[ R \]
Learning sequence models via MoM

Outline

1. Learn HMM models via MoM
2. Relax the notion of anchors
3. Solve a QP
4. Experiments
Twitter POS tagging

150 training labeled sequences

feature HMM

- HMM
- EM
- self-training
- AHMM

Tagging accuracy:
- 81.8
- 81.8
- 83.4
- 85.3
Twitter POS tagging

1000 training labeled sequences

feature HMM

- HMM
- EM
- self-training
- AHMM

Tagging accuracy:
- 89.1
- 89.1
- 89.4
- 89.1
Twitter POS tagging

Tagging accuracy vs. labeled training size

- feature HMM
- HMM
- anchor FHMM
Twitter POS tagging

1000 training sequences

![Bar chart showing training time (in hours) for different methods: Brown Clusters, EM, self-training, and AHMM. The chart indicates that Brown Clusters have the highest training time of 42.0 hours, followed by EM (14.9 hours), self-training (10.3 hours), and AHMM (3.8 hours).]
Conclusions

- MoM algorithm for semi-supervised learning
- Flexible method
  (easy to add supervision)
- Fast to train
  (only one pass over the data)
- Particularly good with little supervision

Thank you!

zmarinho@cmu.edu