CS 6501 Natural Language Processing

Conditional Random Fields

Yangfeng Ji

September 26, 2019

Department of Computer Science
University of Virginia

Overview

1. Conditional Random Fields
2. Inference
3. Parameter Estimation
4. Applications of Sequence Labeling

Conditional Random Fields

Logistic Regression

A direct application of logistic regression:

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{\exp(\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}))}{\sum_{\boldsymbol{y}' \in \mathcal{Y}^T} \exp(\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}'))} \tag{1}$$

Huge $\mathcal{Y}^T$ causes the problems on

- decoding $\operatorname{argmax}_{\boldsymbol{y}' \in \mathcal{Y}^T} p(\boldsymbol{y}'|\boldsymbol{x})$

Logistic Regression

A direct application of logistic regression:

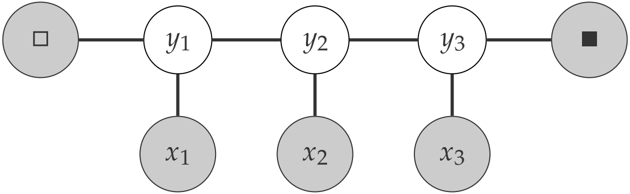
$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{\exp(\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}))}{\underbrace{\sum_{\boldsymbol{y}' \in \mathcal{Y}^T} \exp(\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}'))}_{\text{partition function } Z}} \tag{1}$$

Huge $\mathcal{Y}^T$ causes the problems on

- decoding $\text{argmax}_{\boldsymbol{y}' \in \mathcal{Y}^T} p(\boldsymbol{y}'|\boldsymbol{x})$
- computing the partition function with $|\mathcal{Y}^T| = K^T$ possible values

Conditional Random Fields

Graphical Model:



- Conditional independence
- Undirected graph
- Factorization over cliques

Decomposition of $f(x, y)$

Based on the dependency between x and y, the feature function can be factorized as

$$f(x, y) = \sum_{i=1}^{T} \underbrace{f_i(x_i, y_i, y_{i-1})}_{\text{local feature function}} \quad (2)$$

where

- i: the position to be tagged
- $y_i \in \mathcal{Y}$: POS tag at position i
- $y_{i-1} \in \mathcal{Y}$: POS tag at position $i-1$
- x_i: observation (word) at position i
- $f_i(x_i, y_i, y_{i-1})$ captures the dependency of (y_{i-1}, y_i) and (x_i, y_i)

Local Feature Function: Example

- standard features

[Lafferty et al., 2001]

Local Feature Function: Example

- standard features
- whether a spelling begins with upper case letter,
  - `IBM`, `Virginia`: PROPER NOUN

[Lafferty et al., 2001]

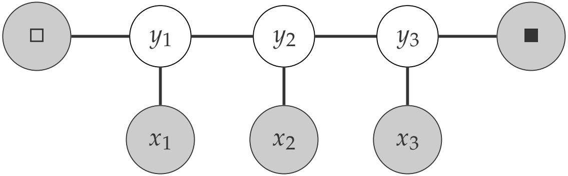
Local Feature Function: Example

- standard features
- whether a spelling begins with upper case letter,
  - `IBM`, `Virginia`: PROPER NOUN
- whether it ends in one of the following suffixes:
  - `-ies` e.g., `parties`: PROPER NOUN, PLURAL
  - `-ly` e.g., `extremely`, `loudly`: ADVERB
  - `-ing` e.g., : VERB, GERUND OR PRESENT PARTICIPLE
  - ...

[Lafferty et al., 2001]

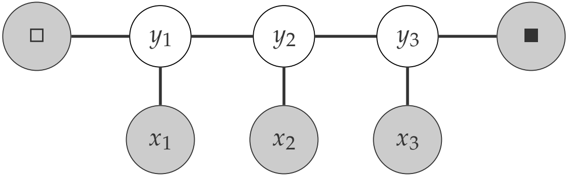
Graphical Model Representation

Conditional Random Fields:

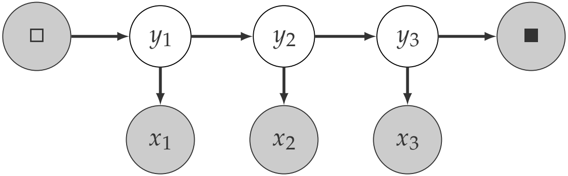


Graphical Model Representation

Conditional Random Fields:



Hidden Markov Models:



Inference

Decode $p(\boldsymbol{y}|\boldsymbol{x})$

With this local feature function:

$$\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{T} \boldsymbol{f}_i(x_i, y_i, y_{i-1}) \tag{3}$$

Decode $p(\boldsymbol{y}|\boldsymbol{x})$

With this local feature function:

$$\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{T} \boldsymbol{f}_i(x_i, y_i, y_{i-1}) \tag{3}$$

$$\begin{aligned}
\operatorname*{argmax}_{\boldsymbol{y} \in \mathcal{Y}^T} p(\boldsymbol{y}|\boldsymbol{x}) &= \operatorname*{argmax}_{\boldsymbol{y} \in \mathcal{Y}^T} \frac{\exp(\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}))}{\sum_{\boldsymbol{y}' \in \mathcal{Y}^T} \exp(\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}'))} \\
&= \operatorname*{argmax}_{\boldsymbol{y} \in \mathcal{Y}^T} \exp(\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}))
\end{aligned}$$

Decode $p(\boldsymbol{y}|\boldsymbol{x})$

With this local feature function:

$$\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{T} \boldsymbol{f}_i(x_i, y_i, y_{i-1}) \tag{3}$$

$$\begin{aligned}
\operatorname*{argmax}_{\boldsymbol{y}\in\mathcal{Y}^T} p(\boldsymbol{y}|\boldsymbol{x}) &= \operatorname*{argmax}_{\boldsymbol{y}\in\mathcal{Y}^T} \frac{\exp(\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x},\boldsymbol{y}))}{\sum_{\boldsymbol{y}'\in\mathcal{Y}^T}\exp(\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x},\boldsymbol{y}'))} \\
&= \operatorname*{argmax}_{\boldsymbol{y}\in\mathcal{Y}^T} \exp(\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})) \\
&= \operatorname*{argmax}_{\boldsymbol{y}\in\mathcal{Y}^T} \boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})
\end{aligned}$$

Decode $p(\boldsymbol{y}|\boldsymbol{x})$

With this local feature function:

$$\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{T} f_i(x_i, y_i, y_{i-1}) \tag{3}$$

$$\begin{aligned}
\operatorname*{argmax}_{\boldsymbol{y}\in\mathcal{Y}^T} p(\boldsymbol{y}|\boldsymbol{x}) &= \operatorname*{argmax}_{\boldsymbol{y}\in\mathcal{Y}^T} \frac{\exp(\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x},\boldsymbol{y}))}{\sum_{\boldsymbol{y}'\in\mathcal{Y}^T}\exp(\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x},\boldsymbol{y}'))} \\
&= \operatorname*{argmax}_{\boldsymbol{y}\in\mathcal{Y}^T} \exp(\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})) \\
&= \operatorname*{argmax}_{\boldsymbol{y}\in\mathcal{Y}^T} \boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x},\boldsymbol{y}) \\
&= \operatorname*{argmax}_{\boldsymbol{y}\in\mathcal{Y}^T} \boldsymbol{\theta}^\top \sum_{i=1}^{T} f_i(x_i, y_i, y_{i-1})
\end{aligned}$$

Decode $p(\boldsymbol{y}|\boldsymbol{x})$

With this local feature function:

$$\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{T} \boldsymbol{f}_i(x_i, y_i, y_{i-1}) \tag{3}$$

$$\begin{aligned}
\operatorname*{argmax}_{\boldsymbol{y} \in \mathcal{Y}^T} p(\boldsymbol{y}|\boldsymbol{x}) &= \operatorname*{argmax}_{\boldsymbol{y} \in \mathcal{Y}^T} \frac{\exp(\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}))}{\sum_{\boldsymbol{y}' \in \mathcal{Y}^T} \exp(\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}'))} \\
&= \operatorname*{argmax}_{\boldsymbol{y} \in \mathcal{Y}^T} \exp(\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y})) \\
&= \operatorname*{argmax}_{\boldsymbol{y} \in \mathcal{Y}^T} \boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}) \\
&= \operatorname*{argmax}_{\boldsymbol{y} \in \mathcal{Y}^T} \boldsymbol{\theta}^\top \sum_{i=1}^{T} \boldsymbol{f}_i(x_i, y_i, y_{i-1}) \\
&= \operatorname*{argmax}_{\boldsymbol{y} \in \mathcal{Y}^T} \sum_{i=1}^{T} \boldsymbol{\theta}^\top \boldsymbol{f}_i(x_i, y_i, y_{i-1})
\end{aligned}$$

Factorization

Factorize $\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y})$ with respect to timestep i

$$\begin{aligned}\sum_{i=1}^{T} \boldsymbol{\theta}^\top \boldsymbol{f}_i(x_i, y_i, y_{i-1}) = & \underbrace{\sum_{j \le i-1} \boldsymbol{\theta}^\top \boldsymbol{f}_j(x_j, y_j, y_{j-1})}_{\text{past}} \\ & + \underbrace{\boldsymbol{\theta}^\top \boldsymbol{f}_i(x_i, y_i, y_{i-1})}_{\text{present}} \\ & + \underbrace{\sum_{k \ge i+1} \boldsymbol{\theta}^\top \boldsymbol{f}_k(x_k, y_k, y_{k-1})}_{\text{future}}\end{aligned} \tag{4}$$

Viterbi Algorithm

$$s_i(k, k') = \boldsymbol{\theta}^\top \boldsymbol{f}_i(x_i, y_{i-1} = k', y_i = k)$$

Algorithm 11 The Viterbi algorithm. Each $s_m(k, k')$ is a local score for tag $y_m = k$ and $y_{m-1} = k'$.

for $k \in \{0, \ldots K\}$ **do**
  $v_1(k) = s_1(k, \Diamond)$
for $m \in \{2, \ldots, M\}$ **do**
  for $k \in \{0, \ldots, K\}$ **do**
    $v_m(k) = \max_{k'} s_m(k, k') + v_{m-1}(k')$
    $b_m(k) = \operatorname{argmax}_{k'} s_m(k, k') + v_{m-1}(k')$
$y_M = \operatorname{argmax}_k s_{M+1}(\blacklozenge, k) + v_M(k)$
for $m \in \{M - 1, \ldots 1\}$ **do**
  $y_m = b_m(y_{m+1})$
return $\boldsymbol{y}_{1:M}$

[Eisenstein, 2018]

Parameter Estimation

Parameter Estimation: Logistic regression

When label y is still a random variable

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \frac{\exp(\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x},y))}{\sum_{y'\in\mathcal{Y}} \exp(\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x},y'))} \tag{5}$$

the derivative with respect $\boldsymbol{\theta}$

$$\frac{\partial \log p(y|\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \boldsymbol{f}\left(\boldsymbol{x},y\right) - \mathbb{E}_{Y|X}[\boldsymbol{f}\left(\boldsymbol{x},y\right)] \tag{6}$$

where

$$\mathbb{E}_{Y|X}[\boldsymbol{f}\left(\boldsymbol{x},y\right)] = \sum_{y\in\mathcal{Y}} \left\{p(y|\boldsymbol{x})\boldsymbol{f}(\boldsymbol{x},y)\right\} \tag{7}$$

Parameter Estimation: CRFs

When label $\boldsymbol{y}$ is a sequence

$$\frac{\partial \log p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \boldsymbol{f}(\boldsymbol{x},\boldsymbol{y}) - \mathbb{E}_{Y|X}[\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})] \tag{8}$$

where

$$\mathbb{E}_{Y|X}[\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})] = \sum_{\boldsymbol{y}\in\mathcal{Y}^T} \left\{ p(\boldsymbol{y}|\boldsymbol{x})\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y}) \right\} \tag{9}$$

and

$$\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{T} \boldsymbol{f}_i(\boldsymbol{x}, y_{i-1}, y_i) \tag{10}$$

Expectation

$$\begin{aligned}
\mathbb{E}_{Y|X}[\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})] &= \sum_{\boldsymbol{y}\in\mathcal{Y}^T} \left\{ p(\boldsymbol{y} \mid \boldsymbol{x}) \boldsymbol{f}(\boldsymbol{x},\boldsymbol{y}) \right\} \\
&= \sum_{\boldsymbol{y}\in\mathcal{Y}^T} \left\{ p(\boldsymbol{y} \mid \boldsymbol{x}) \sum_{i=1}^{T} \boldsymbol{f}_i(x_i, y_{i-1}, y_i) \right\} \\
&= \sum_{\boldsymbol{y}\in\mathcal{Y}^T} \sum_{i=1}^{T} \left\{ p(\boldsymbol{y} \mid \boldsymbol{x}) \boldsymbol{f}_i(x_i, y_{i-1}, y_i) \right\} \\
&= \sum_{i=1}^{T} \sum_{\boldsymbol{y}\in\mathcal{Y}^T} \left\{ p(\boldsymbol{y} \mid \boldsymbol{x}) \boldsymbol{f}_i(x_i, y_{i-1}, y_i) \right\} \\
&= \sum_{i=1}^{T} \sum_{y_{i-1}\in\mathcal{Y}; y_i\in\mathcal{Y}} \left\{ p(y_{i-1}, y_i \mid \boldsymbol{x}) \boldsymbol{f}_i(\boldsymbol{x}, y_{i-1}, y_i) \right\}
\end{aligned}$$

Applications of Sequence Labeling

Applications

- Part-of-Speech taggins [Eisenstein, 2018, section 8.1]
- Named entity recognition (NER) [Eisenstein, 2018, section 8.3]
- Dialogue act identification [Eisenstein, 2018, section 8.6]

Parts of Speech

- *"Open classes"*
  - Nouns
  - Verbs
  - Adjectives
  - Adverbs
  - Numbers
- *"Closed classes"*
  - Modal verbs (e.g., `can, should`)
  - Prepositions (e.g., `on, to`)
  - Particles (e.g., `off, up`)
  - Determiners (e.g., `the, some`)
  - Pronouns (e.g., `she, they`)
  - Conjunctions (e.g., `and, or`)

[Smith, 2017]

Penn Treebank Tagset

Tag	Description	Example	Tag	Description	Example
CC	Coordin. Conjunction	*and, but, or*	SYM	Symbol	+,%, &
CD	Cardinal number	*one, two, three*	TO	"to"	*to*
DT	Determiner	*a, the*	UH	Interjection	*ah, oops*
EX	Existential 'there'	*there*	VB	Verb, base form	*eat*
FW	Foreign word	*mea culpa*	VBD	Verb, past tense	*ate*
IN	Preposition/sub-conj	*of, in, by*	VBG	Verb, gerund	*eating*
JJ	Adjective	*yellow*	VBN	Verb, past participle	*eaten*
JJR	Adj., comparative	*bigger*	VBP	Verb, non-3sg pres	*eat*
JJS	Adj., superlative	*wildest*	VBZ	Verb, 3sg pres	*eats*
LS	List item marker	*1, 2, One*	WDT	Wh-determiner	*which, that*
MD	Modal	*can, should*	WP	Wh-pronoun	*what, who*
NN	Noun, sing. or mass	*llama*	WP$	Possessive wh-	*whose*
NNS	Noun, plural	*llamas*	WRB	Wh-adverb	*how, where*
NNP	Proper noun, singular	*IBM*	$	Dollar sign	$
NNPS	Proper noun, plural	*Carolinas*	#	Pound sign	#
PDT	Predeterminer	*all, both*	"	Left quote	(' or ")
POS	Possessive ending	*'s*	"	Right quote	(' or ")
PRP	Personal pronoun	*I, you, he*	(	Left parenthesis	([, (, {, <)
PRP$	Possessive pronoun	*your, one's*	)	Right parenthesis	( ],), }, >)
RB	Adverb	*quickly, never*	,	Comma	,
RBR	Adverb, comparative	*faster*	.	Sentence-final punc	(. ! ?)
RBS	Adverb, superlative	*fastest*	:	Mid-sentence punc	(: ; ... – -)
RP	Particle	*up, off*			

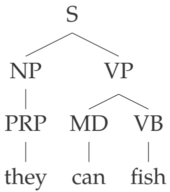
45 taggs, about 40 pages of guidelines [Marcus et al., 1993]

Why We Need POS?

- Disambiguation
  - $\texttt{they}_{\text{PRP}}$ $\texttt{can}_{\text{MD}}$ $\texttt{fish}_{\text{VB}}$

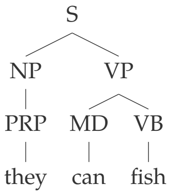
Why We Need POS?

- Disambiguation
  - they$_{\mathrm{PRP}}$ can$_{\mathrm{MD}}$ fish$_{\mathrm{VB}}$
- Basic component for syntactic parsing



Why We Need POS?

- Disambiguation
  - `they`$_{\text{PRP}}$ `can`$_{\text{MD}}$ `fish`$_{\text{VB}}$
- Basic component for syntactic parsing



- Word prediction in speech recognition
  - Personal pronouns (`I`, `you`, `he`) are likely to be followed by verbs

Applications

- ✓ Part-of-Speech taggins [Eisenstein, 2018, section 8.1]
- Named entity recognition (NER) [Eisenstein, 2018, section 8.3]
- Dialogue act identification [Eisenstein, 2018, section 8.6]

Named Entity Recognition

Example

```
Atlantis touched down at Kennedy Space Center
```

Named Entity Recognition

Example

[Atlantis]$_{\text{MSIC}}$ touched down at [Kennedy Space Center]$_{\text{LOC}}$

Named Entity Recognition

Example

`[Atlantis]`$_{\text{MSIC}}$ `touched down at [Kennedy Space Center]`$_{\text{LOC}}$

Tag set

- B: beginning
- I: inside
- O: outside

Category

- Person
- Location
- Organization
- Msic

Named Entity Recognition

Example

`[Atlantis]`$_{\text{MSIC}}$ `touched down at [Kennedy Space Center]`$_{\text{LOC}}$

Tag set

- B: beginning
- I: inside
- O: outside

Category

- Person
- Location
- Organization
- Msic

BIO Annotation

Atlantis	touched	down	at	Kennedy	Space	Center	.
B_{MSIC}	O	O	O	B_{LOC}	I_{LOC}	I_{LOC}	O

Another Type of NER

For understanding scientific articles and academic papers

Computer Science:
This paper addresses the task of **[named entity recognition]**$_{\text{Task}}$, using **[conditional random fields]**$_{\text{Process}}$. Our method is evlauated on the **[ConLL NER Corpus]**$_{\text{Material}}$.

Physics:
[Local field effects] $_{\text{Process}}$ on spontaneous emission rates within **[nanostructure photonics material]**$_{\text{Material}}$ for example are familiar, and have been well used.

Material Science:
The **[Kelvin probe force microscopy technique]** $_{\text{Process}}$ allows **[detection of local EWF]**$_{\text{Task}}$ between an **[atomic force micorscopy]**$_{\text{Material}}$ and **[metal surface]**$_{\text{Material}}$.

[Luan et al., 2017]

Applications

- ✓ Part-of-Speech taggins [Eisenstein, 2018, section 8.1]
- ✓ Named entity recognition (NER) [Eisenstein, 2018, section 8.3]
- ► Dialogue act identification [Eisenstein, 2018, section 8.6]

Dialog Act Identification

Dialogue acts are labels over utterances in a dialogue, corresponding roughly to the speaker's intention.

Speaker	Dialogue Act	Utterance
A	YES-NO-QUESTION	*So do you go college right now?*
A	ABANDONED	*Are yo-*
B	YES-ANSWER	*Yeah,*
B	STATEMENT	*It's my last year [laughter].*
A	DECLARATIVE-QUESTION	*You're a, so you're a senior now.*
B	YES-ANSWER	*Yeah,*
B	STATEMENT	*I'm working on my projects trying to graduate [laughter]*
A	APPRECIATION	*Oh, good for you.*
B	BACKCHANNEL	*Yeah.*

- Sequence labeling over utterances
- For better understanding a conversation

Reference

Eisenstein, J. (2018).
Natural Language Processing.
MIT Press.

Lafferty, J., McCallum, A., and Pereira, F. (2001).
Conditional random fields: Probabilistic models for segmenting and labeling sequence data.
In *ICML*.

Luan, Y., Ostendorf, M., and Hajishirzi, H. (2017).
Scientific information extraction with semi-supervised neural tagging.
arXiv preprint arXiv:1708.06075.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993).
Building a large annotated corpus of english: The penn treebank.
Computational linguistics, 19(2):313–330.

Smith, N. A. (2017).
Natural language processing: Lecture notes.