Lecture Overview

• Recap: Naïve Markov – Logistic regression (simple CRF)
• CRFs: high-level definition
• CRFs Applied to sequence labeling
• NLP Examples: Name Entity Recognition, joint POS tagging and NP segmentation
• CFR + deep learning Example
Conditional Random Fields (CRFs)

- Model $P(Y_1 \ldots Y_k \mid X_1 \ldots X_n)$
- Special case of Markov Networks where all the $X_i$ are always observed

- Simple case $P(Y_1 \mid X_1 \ldots X_n)$
  - All vars are binary
  - $Y_i = \{0, 1\}$
  - $\forall i \ X_i = \{0, 1\}$
Some notation

\[ \prod \text{Predicate}(x) \text{ if } \text{Pred}(x) \text{ is true} \rightarrow 2 \]
\[ \text{if } \text{Pred}(x) \text{ is false} \rightarrow 0 \]

Examples:
\[ \sum \text{even}(x_i) \]
\[ \sum \text{prime}(x_i) \]
\[ \sum_{x_i \in X} x_i = 15 \]
\[ = 3 = 2 \]

CPSC 422, Lecture 18
What are the Parameters?

\[ \phi_i(x_i, y_1) = \exp \{ \omega_i \mathbb{1}[x_i = 1, y_1 = 1] \} \]

one such factor for each clique

also \[ \phi_0(y_1) = \exp \{ \omega_0 \mathbb{1}[y_1 = 1] \} \]

Example \[ \omega_2 = 1.5 \]

\[ \phi_2(x_2, y_1) \]

Example \[ \omega_0 = 0.4 \]

\[ \mathbb{1} \]

\[ \phi_0 \]

\[ \begin{array}{cccc}
X_2 & Y_1 & \phi_2 & e^{1.5} \\
0 & 1 & 1 & 1 \\
0 & 0 & 0 & 1 \\
1 & 1 & 1 & 1 \\
\end{array} \]
Let’s derive the probabilities we need

To compute

$$P(Y_1 | X_1 \ldots X_n) = \frac{P(Y_1, X_1 \ldots X_n)}{P(X_1 \ldots X_n)}$$

We compute

$$P(Y_1 = 1 | X_1 \ldots X_n) = \frac{P(Y_1 = 1, X_1 \ldots X_n)}{P(X_1 \ldots X_n)}$$
Let's derive the probabilities we need

$$
\phi_i(X_i, Y_1) = \exp \{ w_i \cdot 1 \{ X_i = 1, Y_1 = 1 \} \} \\
\text{how strongly } Y_2 = 1 \text{ given that } X_i = 1 \\\n\phi_0(Y_1) = \exp \{ w_0 \cdot 1 \{ Y_1 = 1 \} \}
$$

$$
\hat{P}(Y_1 = 1, X_1, X_2, \ldots, X_n) = \Phi_0(Y_1) \prod_{i=1}^{n} \Phi(X_i, Y_1)
$$

A. \ \ e^{\sum_{i=1}^{n} w_i} \\
B. \ \ e^{w_0 + \sum_{i=1}^{n} w_i} \\
C. \ \ e^{w_0 + \sum_{i=1}^{n} X_i} \\
D. \ \ e^{w_0 + \sum_{i=1}^{n} X_i} \\
E. \ \ e^{w_0 + \sum_{i=1}^{n} X_i}
$$
Let’s derive the probabilities we need

\[ \phi_i(X_i, Y_1) = \exp\{w_i \mid \{X_i = 1, Y_1 = 1\}\} \]

\[ \phi_0(Y_1) = \exp\{w_0 \mid \{Y_1 = 1\}\} \]

\[ \tilde{P}(Y_1 = 1, X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} \phi(X_i, Y) \]

\[ P(Y_1 = 1, X_1 = 0, X_2 = 1, X_3 = 1) \]

\[ e^{w_0 \times 1} \times e^{w_1 \times 0} \times e^{w_2 \times 1} \times e^{w_3 \times 1} = e^{w_0} \times e^{w_1 \times x_1} \times e^{w_2 \times x_2} \times e^{w_3 \times x_3} = e^{w_0 + \sum w_i x_i} \]
Let’s derive the probabilities we need

\[ \phi_i(X_i, Y_1) = \exp\{w_i \mathbb{1}\{X_i = 1, Y_1 = 1\}\} \]

\[ \phi_0(Y_1) = \exp\{w_0 \mathbb{1}\{Y_1 = 1\}\} \]

\[ \tilde{P}(Y_1 = 0, X_1, X_2, \ldots, X_n) = \Phi_0(Y_1) \prod_{i=1}^{n} \phi_i(X_i) \]

A. 1  
B. \(e^{w_0}\)  
C. 0  
D. \(e^{\sum_{i=1}^{n} w_i}\)
Let’s derive the probabilities we need

\[ P(Y_1 = 1, x_1, \ldots, x_n) = \exp(w_0 + \sum_{i=1}^{n} w_i x_i) \]
\[ P(Y_1 = 0, x_1, \ldots, x_n) = 1 \]

\[ P(Y_1 = 1 | x_1, \ldots, x_n) = \frac{P(Y_1 = 1) P(X_1, \ldots, X_n)}{\sum_{Y_1} P(Y_1 = 1) P(X_1, \ldots, X_n)} \]
\[ = \frac{\exp(w_0 + \sum w_i x_i)}{1 + \exp(w_0 + \sum w_i x_i)} \]

The sigmoid function \( \frac{e^x}{1 + e^x} \) or \( \frac{1}{1 + e^{-x}} \) can be used.
Let’s derive the probabilities we need

\[ P(Y_1 = 1, x_1, \ldots, x_n) = \exp(w_0 + \sum_{i=1}^{n} w_i x_i) \]

\[ P(Y_1 = 0, x_1, \ldots, x_n) = 1 \]

\[ P(Y_1 = 1 | x_1, \ldots, x_n) = \frac{\hat{P}(Y_1 = 1) \cdot x_1 \cdot \ldots \cdot x_n}{P(x_1, \ldots, x_n)} \approx \frac{e^z}{1 + e^z} \]

\[ \approx \frac{e^{-z}}{e^{-z} + 1} \]
Sigmoid Function used in Logistic Regression

- Great practical interest
- Number of parameters $w_i$ is linear instead of exponential in the number of parents
- Natural model for many real-world applications
- Naturally aggregates the influence of different parents
Logistic regression is a simple Markov Net (a CRF) *aka naïve markov model*

- But only models the **conditional distribution**, $P(Y \mid X)$ and not the full joint $P(X, Y)$
Let’s generalize ….

Assume that you always observe a set of variables $X = \{X_1 \ldots X_n\}$ and you want to predict one or more variables $Y = \{Y_1 \ldots Y_k\}$.

A CRF is an undirected graphical model whose nodes corresponds to $X \cup Y$.

$\phi_1(D_1) \ldots \phi_m(D_m)$ represent the factors which annotate the network (but we disallow factors involving only vars in $X$ – why?)

They would be

A. too large
B. constant
C. difficult to acquire

$$P(Y \mid X) = \frac{1}{Z(X)} \left( \prod_{i=1}^{m} \phi_i(D_i) \right)$$

$$Z(X) = \sum_{Y} \left( \prod_{i=1}^{m} \phi_i(D_i) \right)$$
Lecture Overview

- Recap: Naïve Markov – Logistic regression (simple CRF)
- CRFs: high-level definition
- CRFs Applied to sequence labeling
- NLP Examples: Name Entity Recognition, joint POS tagging and NP segmentation
- CFR + deep learning Example
Sequence Labeling

Linear-chain CRF
Increase representational Complexity: Adding Features to a CRF

- Instead of a single observed variable $X_i$ we can model multiple features $X_{ij}$ of that observation.
CRFs in Natural Language Processing

- One target variable $Y$ for each word $X$, encoding the possible labels for $X$
- Each target variable is connected to a set of feature variables that capture properties relevant to the target distinction.

Is the word capitalized? Does the word end in “ing”?

CPSC 422, Lecture 19
Slide 18
Named Entity Recognition Task

- Entity often span multiple words "British Columbia"
- Type of an entity may not be apparent for individual words "University of British Columbia"
- Let’s assume three categories: **Person, Location, Organization**
- BIO notation (for sequence labeling)

possible

```
B-PER  I-PER  B-LOC  I-LOC
```

labels

```
B-ORG  I-ORG  OTHER
```

The University of British Columbia

```
0  B-ORG  I-ORG  I-ORG  I-ORG
```

is in Vancouver B.C.

```
0  0  B-LOC  I-LOC
```

Linear chain CRF parameters

With two factors “types” for each word

\[ \phi^1_t(Y_t, Y_{t-1}), \phi^1_t(Y_t, Y_{t+1}) \]

Dependency between neighboring target vars

\[ \phi^2_t(Y_t, X_1, \ldots, X_T) \]

Dependency between target variable and its context in the word sequence, which can include also features of the words (capitalized, appear in an atlas of location names, etc.)

Factors are similar to the ones for the Naïve Markov (logistic regression)

\[ \phi_t(Y_t, X_{tk}) = \exp\{w_{tk} \times \prod \{Y_t = \text{I-LOC}, X_{tk} = 1\}\} \]
Features can also be

- The word
- Following word
- Previous word
More on features

Including features that are **conjunctions of simple features** increases accuracy

\[
\prod \{ Y_t = 1 - \text{PER}_t, X_{t+1}, k = "spoke" \} \\
\prod \{ Y_t = 1 - \text{PER}_t, X_{t-1}, k = "Mrs." \}
\]

Total number of features can be \(10^5 - 10^6\)

However features are sparse i.e. most features are 0 for most words
Linear-Chain Performance

Per-token/word accuracy in the high 90% range for many natural datasets. Label is wrong for 2 words out of 9.

Per-field precision and recall are more often around 80-95%, depending on the dataset. Entire Named Entity Phrase must be correct.

The University of British Columbia is in Vancouver, B.C.

- Only one is correct out of 2

Per-word accuracy?
Per-field precision?

A. 1/2
B. 7/9
C. 7/9
Skip-Chain CRFs

Include additional factors that connect non-adjacent target variables

E.g., When a word occur multiple times in the same documents

Graphical structure over Y can depend on the values of the Xs!
Lecture Overview

- Recap: Naïve Markov – Logistic regression (simple CRF)
- CRFs: high-level definition
- CRFs Applied to sequence labeling
- **NLP Examples:** Name Entity Recognition, joint POS tagging and NP segmentation
- CFR + deep learning Example
Coupled linear-chain CRFs

- Linear-chain CRFs can be combined to perform multiple tasks simultaneously

- Performs part-of-speech labeling and noun-phrase segmentation
Coupled linear-chain CRFs

- Linear-chain CRFs can be combined to perform multiple tasks simultaneously

- Performs part-of-speech labeling and noun-phrase segmentation
Inference in CRFs (just intuition)

An HMM can be viewed as a factor graph

\[ p(y, x) = \prod_t \Psi_t(y_t, y_{t-1}, x_t) \]

where \( Z = 1 \), and the factors are defined as:

\[ \Psi_t(j, i, x) \triangleq p(y_t = j | y_{t-1} = i)p(x_t = x | y_t = j). \]

Forward / Backward / Smoothing and Viterbi can be rewritten (not trivial!) using these factors

Then you plug in the factors of the CRFs and all the algorithms work fine with CRFs! 😊
CRFs Summary

- Ability to incorporate arbitrary overlapping local and global features
- Graphical structure over Y can depend on the values of the Xs (see slide 24)
- Can perform multiple tasks simultaneously (see slide 26)
- *Standard Inference algorithm* for HMM can be applied
- *Practical Learning algorithms* exist
- Strong baseline on many labeling tasks (*deep learning recently shown to be often better when large training data are available... current research on ensembling them!*)

See MALLET package for CRF implementation
Probabilistic Graphical Models

From “Probabilistic Graphical Models: Principles and Techniques” D. Koller, N. Friedman 2009
Combining CRFs and Neural Models

SEMANTIC IMAGE SEGMENTATION WITH DEEP CONVOLUTIONAL NETS AND FULLY CONNECTED CRFS

International Conference on Learning Representations (ICLR), San Diego, California, USA, May 2015.

Liang-Chieh Chen Univ. of California, Los Angeles; George Papandreou Google Inc.; Iasonas Kokkinos INRIA; Kevin Murphy Google Inc.; Alan L. Yuille Univ. of California, Los Angeles

1. Use CNN to generate a rough prediction of segmentation (smooth, blurry heat map)
2. Refine this prediction with a conditional random field (CRF)
422 big picture

Deterministic

- First Order Logics
- Full Resolution
- SAT

Stochastic

- Belief Nets
  - Approx. : Gibbs
- Markov Chains and HMMs
  - Forward, Viterbi...
  - Approx. : Particle Filtering
- Undirected Graphical Models
  - Markov Networks
  - Conditional Random Fields
- Markov Decision Processes and Partially Observable MDP
  - Value Iteration
  - Approx. Inference
  - Reinforcement Learning

Applications of AI

StarAI (statistical relational AI)
Hybrid: Det + Sto
- Prob CFG
- Prob Relational Models
- Markov Logics

Representation
Reasoning Technique
Learning Goals for today’s class

You can:

• Provide general definition for CRF
• Apply CRFs to sequence labeling
• Describe and justify features for CRFs applied to Natural Language processing tasks
• Explain benefits of CRFs
How to prepare…

- Go to Office Hours
- Learning Goals (look at the end of the slides for each lecture – complete list has been posted)
- Revise all the clicker questions and practice exercises
- More practice material has been posted
- Check questions and answers on Piazza

Next class Wed

- Start Logics
- Revise Logics from 322!
Conditional Random Fields (CRFs)

- Model $P(Y_1 \ldots Y_k | X_1 \ldots X_n)$
- Special case of Markov Networks where all the $X_i$ are always observed

- Simple case $P(Y_1 | X_1 \ldots X_n)$
  
  *All vars are binary*

  $Y_i = \{0, 1\}$

  $\forall i \ X_i = \{0, 1\}$
What are the Parameters?

\[ \phi_i(x_i, y_i) = \exp \{ w_i \mathbb{1}[x_i = 1, y_i = 1] \} \]

One such factor for each clique

Also \[ \phi_0(y_i) = \exp \{ w_0 \mathbb{1}[y_i = 1] \} \]

Example \( w_2 = 1.5 \) \[ \phi_2(x_2, y_i) \]

Example \( w_0 = 0.4 \) \[ \phi_0 \]
Let’s derive the probabilities we need

\[ \phi_i(X_i, Y_1) = \exp\{ w_i \, \mathbb{1}\{X_i = 1, Y_1 = 1\} \} \]

how strongly \( Y_2 = 1 \) given that \( X_i = 1 \)

\[ \phi_0(Y_1) = \exp\{ w_0 \, \mathbb{1}\{Y_1 = 1\} \} \]

\[ P(Y_1 \mid x_1, \ldots, x_n) = \frac{\phi_0(Y_1) \prod_{i=1}^n \phi(X_i, Y_i)}{P(Y_1 = 1, x_1, \ldots, x_n) + \prod_{i=1}^n \phi(X_i, Y_i)} \]

\[ P(Y_1, x_1, \ldots, x_n) \approx \frac{\phi_0(Y_1) \prod_{i=1}^n \phi(X_i, Y_i)}{P(Y_1 = 1, x_1, \ldots, x_n)} \]

\[ P(Y_1 = 1, x_1, \ldots, x_n) = \]

\[ P(Y_1 = 0, x_1, \ldots, x_n) = \]
Let’s derive the probabilities we need

\[ \phi_i(X_i, Y_1) = \exp\{w_i \mathbf{1}\{X_i = 1, Y_1 = 1\}\} \]

\[ \phi_0(Y_1) = \exp\{w_0 \mathbf{1}\{Y_1 = 1\}\} \]

\[ \hat{P}(Y_1 = 1 | X_1, X_2, \ldots, X_n) = \overline{\phi_0(Y_1)} \prod_{i=1}^{n} \phi_i(X_i, Y_1) \]

\[ \hat{P}(Y_1 = 1 | X_1 = 0, X_2 = 1, X_3 = 1) = e^{w_0 + \sum w_i X_i} \]

Example: 

\[ e^{w_0 x_1} \times e^{w_1 x_1} \times e^{w_2 x_2} \times e^{w_3 x_3} = e^{w_0 + \sum w_i x_i} \]
Let’s derive the probabilities we need

\[ \phi_i(X_i, Y_1) = \exp\{w_i \perp \{X_i = 1, Y_1 = 1\}\} \]

\[ \phi_0(Y_1) = \exp\{w_0 \perp \{Y_1 = 1\}\} \]

\[ \tilde{P}(Y_1 = 0, X_1, X_2, \ldots, X_n) = \frac{\phi_0(Y_1) \prod_{i=1}^{n} \phi(X_i, Y_1)}{1} \]
Let's derive the probabilities we need

\( P(Y_1 = 1, x_1, \ldots, x_n) = \exp(w_0 + \sum_{i=1}^{n} w_i x_i) \)

\( P(Y_1 = 0, x_1, \ldots, x_n) = 1 \)

\[ P(Y_1 = 1 \mid x_1, \ldots, x_n) = \frac{\tilde{P}(Y_1 = 1 \mid x_1, \ldots, x_n)}{\tilde{P}(x_1, \ldots, x_n)} \]

\[ = \frac{\exp(w_0 + \sum w_i x_i)}{1 + \exp(w_0 + \sum w_i x_i)} \]
Continue.....

\( P(Y_i = 1 \mid x_1, \ldots, x_n) = \frac{e^{w_0 + \sum w_i x_i}}{1 + e^{w_0 + \sum w_i x_i}} \)

\[
1 - \frac{1}{e^{-2} + 1} = \frac{e^2}{2 + e^2} \cdot \frac{e^{-2}}{e^{-2}} = \frac{1}{e^{-2} + 1}
\]

\[
P(Y_i = 1 \mid x_1, \ldots, x_n) = \left\{ \frac{1}{e^{-2} + 1}, \frac{e^{-2}}{e^{-2} + 1} \right\}
\]
END in class 2017
Let’s derive the probabilities we need

\[ \phi_i(X_i, Y_1) = \exp\{w_i \mid \{X_i = 1, Y_1 = 1\}\} \]

\[ \phi_0(Y_1) = \exp\{w_0 \mid \{Y_1 = 1\}\} \]

How strongly \( Y_2 = 1 \) given that \( X_1 = 1 \)

\[
P(Y_1 \mid x_1, \ldots, x_n) = \frac{\hat{P}(Y_1, x_1, \ldots, x_n)}{\hat{P}(x_1, \ldots, x_n)} \approx \frac{\hat{P}(Y_1) \prod_{i=1}^{n} \phi_i(x_i, y_i)}{\hat{P}(x_1, \ldots, x_n)}
\]

\[
\approx \hat{P}(Y_1 = 0, x_1, \ldots, x_n) = 1 \]

\[
\approx \hat{P}(Y_1 = 1, x_1, \ldots, x_n) = e^{w_0 + \sum w_i x_i}
\]
$P(Y_i = 1 | x_1, \ldots, x_n) = \frac{\exp(w_0 + \sum w_i x_i)}{1 + \exp(w_0 + \sum w_i x_i)}$

$= \frac{\exp(2)}{1 + \exp(2)} \frac{\exp(-2)}{1 + \exp(-2)} = \frac{1}{e^{-2} + 1}$

$P(Y_i = 1 | x_1, \ldots, x_n) = \left\{ \frac{1}{e^{-2} + 1}, \frac{e^{-2}}{e^{-2} + 1} \right\}$
Sigmoid Function used in Logistic Regression

- Great practical interest
- Number of param $w_i$ is linear instead of exponential in the number of parents
- Natural model for many real-world applications
- Naturally aggregates the influence of different "parents"
Logistic regression is a simple Markov Net (a CRF) *aka* naïve markov model

- But only models the **conditional distribution**, $P(Y | X)$ and not the full joint $P(X, Y)$