Intelligent Systems (AI–2)

Computer Science cpsc422, Lecture 19

Oct, 23, 2017

Slide Sources
Raymond J. Mooney University of Texas at Austin

D. Koller, Stanford CS – Probabilistic Graphical Models
D. Page, Whitehead Institute, MIT

Several Figures from
“Probabilistic Graphical Models: Principles and Techniques” D. Koller, N. Friedman 2009
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
Let’s derive the probabilities we need

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

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

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

\[
\begin{align*}
P(Y_1 \mid x_1, \ldots, x_n) &= \frac{\widetilde{P}(Y_1, x_1, \ldots, x_n)}{\widetilde{P}(x_1, \ldots, x_n)} \\
&\approx \phi_0(Y_1) \prod_{i=1}^{n} \phi_i(x_i, y_i) \\
&\approx P(Y_1 = 0, x_1, \ldots, x_n) \\
&\approx \Phi_0 + \sum \phi_i \chi_i \end{align*}
\]
Continue…..

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

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

\[ P(Y_i = 1 \mid x_1, \ldots, x_n) = \left\{ \begin{array}{ll}
\frac{1}{e^{-z} + 1} & e^{-z} \\
\frac{e^{-z}}{e^{-z} + 1} & e^{-z} + 1
\end{array} \right\} \]
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

CPSC 422, Lecture
Logistic Regression as a Markov Net (CRF)

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 $P(X, Y)$
Let’s generalize …. 

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

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

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

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
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”??
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)

The University of British Columbia

is in Vancouver B.C.
Linear chain CRF parameters

With two factors “types” for each word

\[ \phi_t^1(Y_t, Y_{t-1}) \quad \phi_t^1(Y_t, Y_{t+1}) \]

Dependency between neighboring target vars

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

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)

\[ \phi_t(Y_t, X_{tk}) = \exp\{w_{tk} \times \bigwedge \{Y_t = \text{I-LOC} , X_{tk} = 1 \} \} \]

appears in atlas of location names
Features can also be

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

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

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

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. Only one is correct out of 2.

![Diagram with labeled entities and options for answer selection]
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!
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:

\[ \Psi_t(j, i, x) \overset{\text{def}}{=} p(y_t = j|y_{t-1} = i)p(x_t = x|y_t = j). \] (4.1)

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 21)
- Can perform multiple tasks simultaneously (see slide 23)
- *Standard Inference algorithm* for HMM can be applied
- *Practical Learning algorithms* exist
- State-of-the-art 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
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)
Input → Deep Convolutional Neural Network → Aeroplane Coarse Score map → Bi-linear Interpolation → Final Output

Fully Connected CRF
Applications of AI
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
Midterm, Wed, Oct 25, we will start at noon sharp

How to prepare…

• Go to Office Hours (extra hours offered)
• 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 Fri

• Start Logics
• Revise Logics from 322!