

Hidden Markov Models and Conditional Random Fields

Lecture 13: October 23, 2013

CS886-2 Natural Language Understanding
University of Waterloo

CS886 Lecture Slides (c) 2013 P. Poupart

1

Sequence Data

- So far, we assumed that the data instances are classified independently
 - More precisely, we assumed that the data is iid (identically and independently distributed)
 - E.g., entity recognition, part-of-speech tagging, speech recognition
- Language data naturally forms a sequence where adjacent labels are correlated
 - Text: sequence of characters/words
 - Speech: sequence of audio frames/phonemes

CS886 Lecture Slides (c) 2013 P. Poupart

2

Named Entity Recognition

Kofi Atta Annan is a Ghanaian diplomat who served as the seventh Secretary General of the United Nations from January 1, 1997, to January 1, 2007, serving two five-year terms. Annan was the co-recipient of the Nobel Peace Prize in October 2001.

Kofi Annan was born on April 8, 1938, to Victoria and Henry Reginald Annan in Kumasi, Ghana. He is a twin, an occurrence that is regarded as special in Ghanaian culture. Efua Atta, his twin sister, shares the same middle name, which means 'twin'. As with most Akan names, his first name indicates the day of the week he was born: 'Kofi' denotes a boy born on a Friday. The name Annan can indicate that a child was the fourth in the family, but in his family it was simply a name which Annan inherited from his parents.

In 1962, Annan started working as a Budget Officer for the World Health Organization, an agency of the United Nations. From 1974 to 1976, he was the Director of Tourism in Ghana. Annan then returned to work for the United Nations as an Assistant Secretary General in three consecutive positions.

| |
|--------------|
| Person |
| Location |
| Organization |
| Date |
| Nationality |
| Title |

CS886 Lecture Slides (c) 2013 P. Poupart

3

Part-of-speech Tagging

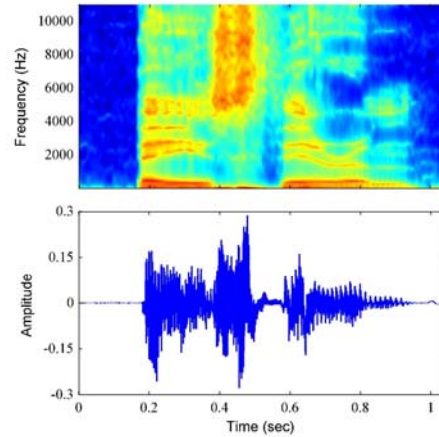
The suburb of Saffron Park lay on the sunset side of London, as red and ragged as a cloud of sunset. It was built of a bright brick throughout; its skyline was fantastic, and even its ground plan was wild. It had been the outburst of a speculative builder, faintly tinged with art, who called its architecture sometimes Elizabethan and sometimes Queen Anne, apparently under the impression that the two sovereigns were identical. It was described with some justice as an artistic colony, though it never in any definable way produced any art.

The/DT suburb/NN of/IN Saffron/NNP Park/NNP lay/VBD on/IN the/DT sunset/JJ side/NN of/IN London/NNP
 It/PRP was/VBD built/VBN of/IN a/DT bright/JJ brick/NN throughout/IN ;/: its/PRP\$ skyline/NN was/VBD f
 It/PRP had/VBD been/VBN the/DT outburst/NN of/IN a/DT speculative/JJ builder/NN ,, faintly/RB tingec
 It/PRP was/VBD described/VBN with/IN some/DT justice/NN as/IN an/DT artistic/JJ colony/NN ,,/, though

CS886 Lecture Slides (c) 2013 P. Poupart

4

Speech Recognition



| b | ey | z | th | ih | er | em |
 | Bayes' | Theorem |

CS886 Lecture Slides (c) 2013 P. Poupart

5

Classification

- Extension of some classification models for sequence data

| | Independent classification | Correlated classification |
|-----------------------|----------------------------|---------------------------------|
| Generative models | Naïve Bayes model | Hidden Markov Model |
| Discriminative models | Logistic Regression | Conditional Random Field |

CS886 Lecture Slides (c) 2013 P. Poupart

6

Hidden Markov Model

- Graphical Model

- Parameterization
 - Transition distribution:
 - Emission distribution:
- Joint distribution:

CS886 Lecture Slides (c) 2013 P. Poupart

7

Assumptions

- **Stationary Process:** transition and emission distributions are identical at each time step

$$\Pr(X_t|Y_t) = \Pr(X_{t+1}|Y_{t+1}) \quad \forall t$$

$$\Pr(Y_t|Y_{t-1}) = \Pr(Y_{t+1}|Y_t) \quad \forall t$$

- **Markovian Process:** next state is independent of previous states given the current state

$$\Pr(Y_{t+1}|Y_t, Y_{t-1}, \dots, Y_0) = \Pr(Y_{t+1}|Y_t) \quad \forall t$$

CS886 Lecture Slides (c) 2013 P. Poupart

8

Conditional Random Field

- Graphical Model

- Parameterization
 - Transition potential:
 - Emission potential:
- Conditional distribution:

CS886 Lecture Slides (c) 2013 P. Poupart

9

Assumptions

- **Stationary Process:** transition and emission distributions are identical at each time step

$$f(X_t, Y_t) = f(X_{t+1}, Y_{t+1}) \quad \forall t$$

$$f(Y_t, Y_{t-1}) = f(Y_{t+1}, Y_t) \quad \forall t$$

- **Markovian Process:** next state is independent of previous states given the current state

$$\Pr(Y_{t+1} | Y_t, Y_{t-1}, \dots, Y_0) = \Pr(Y_{t+1} | Y_t) \quad \forall t$$

CS886 Lecture Slides (c) 2013 P. Poupart

10

Text Tagging

- Hidden Markov Model:
 - Y : term tag (i.e., entity class, part-of-speech tag)
 - X : features related to a term
 - $f(y_{t-1}, y_t)$: correlation between adjacent tags
 - $f(x_t, y_t)$: correlation between the features and the tag of a term
- **Joint classification:** $\operatorname{argmax}_{y_{1:t}} \Pr(y_{1:t} | x_{1:t})?$

Inference in sequence models

- Four common tasks:
 - **Monitoring:** $\Pr(y_t | x_{1..t})$
 - **Prediction:** $\Pr(y_{t+k} | x_{1..t})$
 - **Hindsight:** $\Pr(y_k | x_{1..t})$ where $k < t$
 - **Most likely explanation:**

$$\operatorname{argmax}_{y_0, \dots, y_t} \Pr(y_{0..t} | x_{1..t})$$
- What algorithms should we use?

Hindsight

- $\Pr(y_k | x_{1..t})$ for $k < t$: distribution over a past state given observations
- Example: determine entity type of a word in a sentence
- computation:

$$\Pr(y_k | x_{1..t}) \propto \Pr(y_k, x_{k+1..t} | x_{1..k}) \text{ by conditioning}$$

$$= \Pr(y_k | x_{1..k}) \Pr(x_{k+1..t} | y_k) \text{ by chain rule}$$

Forward-backward algorithm

1. Compute $\Pr(y_k | x_{1..k})$ by forward computation

For $i = 1$ to k do

$$\Pr(y_i | x_{1..i}) \propto f(x_i, y_i) \sum_{y_{i-1}} f(y_{i-1}, y_i) \Pr(y_{i-1} | x_{1..i-1})$$

End
2. Compute $\Pr(x_{k+1..t} | y_k)$ by backward computation

For $j = t$ downto k do

$$\Pr(x_{j+1..t} | y_j) = \sum_{y_{j+1}} f(y_j, y_{j+1}) f(x_{j+1}, y_{j+1}) \Pr(x_{j+2..t} | y_{j+1})$$

End
3. $\Pr(y_k | x_{k+1..t}) \propto \Pr(y_k | x_{1..k}) \Pr(x_{k+1..t} | y_k)$
 - Linear complexity in t

Most likely explanation

- $\operatorname{argmax}_{y_{0..t}} \Pr(y_{0..t}|x_{1..t})$: most likely state sequence given observations

- Example: speech recognition

- Computation:

$$\max_{y_{0..t}} \Pr(y_{0..t}|x_{1..t}) \propto \max_{y_t} f(x_t, y_t) \max_{y_{0..t-1}} \Pr(y_{0..t}|x_{0..t-1})$$

- Recursive computation:

$$\begin{aligned} \max_{y_{0..i-1}} \Pr(y_{0..i}|x_{0..i-1}) &\propto \\ \max_{y_{i-1}} f(y_{i-1}, y_i) f(x_{i-1}, y_{i-1}) &\max_{y_{0..i-2}} \Pr(y_{0..i-1}|x_{0..i-2}) \end{aligned}$$

Viterbi Algorithm

1. Compute $\max_{y_{0..t}} \Pr(y_{0..t}|x_{1..t})$ by dynamic programming

$$\max_{y_0} \Pr(y_{0..1}) \propto \max_{y_0} f(y_0, y_1)$$

For $i = 1$ to $t - 1$ do

$$\begin{aligned} \max_{y_{0..i}} \Pr(y_{0..i+1}|x_{0..i}) &\propto \\ \max_{y_i} f(y_i, y_{i+1}) f(x_i, y_i) &\max_{y_{0..i-1}} \Pr(y_{0..i}|x_{0..i-1}) \end{aligned}$$

End

$$\max_{y_{0..t}} \Pr(y_{0..t}|x_{1..t}) \propto \max_{y_t} f(x_t, y_t) \max_{y_{0..t-1}} \Pr(y_{0..t}|x_{0..t-1})$$

- Linear complexity in t