8. Sequence Labeling with Conditional Random Fields

Dietrich Klakow

Overview

- Sequence Labeling
- Bayesian Networks
- Markov Random Fields
- Conditional Random Fields
- Software example

Background Reading

Hanna M. Wallach

Conditional Random Fields: An Introduction.

Technical Report MS-CIS-04-21. Department of Computer and Information Science, University of Pennsylvania, 2004.

http://www.inference.phy.cam.ac.uk/hmw26/papers/crf_intro.pdf

Sequence Labeling Tasks

Sequence: a sentence

Pierre
Vinken
,
61
years
old
,
will
join
the
board
as
a
nonexecutive
director
Nov.
29
.

POS Labels

Pierre	NNP
Vinken	NNP
,	,
61	CD
years	NNS
old	JJ
,	,
will	MD
join	VB
the	DT
board	NN
as	IN
a	DT
nonexecutive	JJ
director	NN
Nov.	NNP
29	CD
.	.

Chunking

Task: find phrase boundaries:

$[_{NP}$ He $]$ $[_{VP}$ reckons $]$ $[_{NP}$ the current account deficit $]$ $[_{VP}$ will narrow $]$ $[_{PP}$ to $]$ $[_{NP}$ only £ 1.8 billion $]$ $[_{PP}$ in $]$ $[_{NP}$ September $]$.

Chunking

Pierre	———	B-NP
Vinken	———	I-NP
,	———	O
61	———	B-NP
years	———	I-NP
old	———	B-ADJP
,	———	O
will	———	B-VP
join	———	I-VP
the	———	B-NP
board	———	I-NP
as	———	B-PP
a	———	B-NP
nonexecutive	———	I-NP
director	———	I-NP
Nov.	———	B-NP
29	———	I-NP
.	———	O

Named Entity Tagging

Pierre	———	B-PERSON
Vinken	———	I-PERSON
,	———	O
61	———	B-DATE:AGE
years	———	I-DATE:AGE
old	———	I-DATE:AGE
,	———	O
will	———	O
join	———	O
the	———	O
board	———	B-ORG_DESC:OTHER
as	———	O
a	———	O
nonexecutive	———	O
director	———	B-PER_DESC
Nov.	———	B-DATE:DATE
29	———	I-DATE:DATE
.	———	O

Supertagging

Pierre	N/N
Vinken	N
,	,
61	N/N
years	N
old	(S[adj]\NP)\NP
,	,
will	(S[dcl]\NP)/(S[b]\NP)
join	((S[b]\NP)/PP)/NP
the	NP[nb]/N
board	N
as	PP/NP
a	NP[nb]/N
nonexecutive	N/N
director	N
Nov.	((S\NP)\(S\NP))/N[num]
29	N[num]
.	.

Hidden Markov Model

HMM: just an Application of a Bayes Classifier

$$(\hat{\pi}_1, \hat{\pi}_2 ... \hat{\pi}_N) = \underset{\pi_1, \pi_2 .. \pi_N}{\arg\max} \left[P(x_1, x_2 ... x_N, \pi_1, \pi_2 ... \pi_N) \right]$$

$x_1, x_2 ... x_N$: observation/input sequence

$\pi_1, \pi_2 ... \pi_N$: label sequence

Decomposition of Probabilities

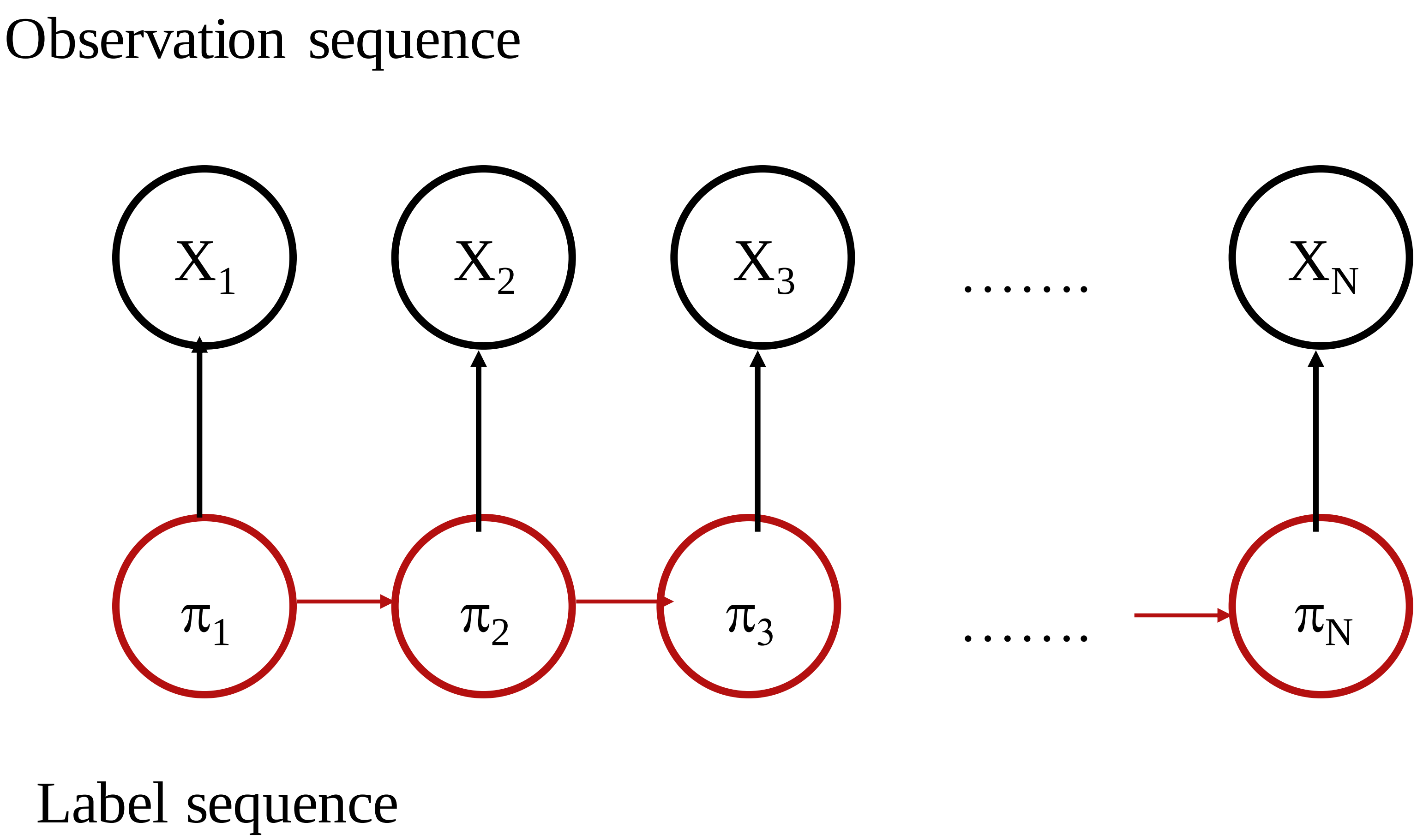
$$P(x_1, x_2..x_N, \pi_1, \pi_2..\pi_N)$$

$$= \prod_{i=1}^{N} P(x_i \mid \pi_i) P(\pi_i \mid \pi_{i-1})$$

$P(\pi_i \mid \pi_{i-1})$: transition probability

$P(x_i \mid \pi_i)$: emission probability

Graphical view HMM



Criticism

- HMMs model only limited dependencies

$\mapsto$ come up with more flexible models

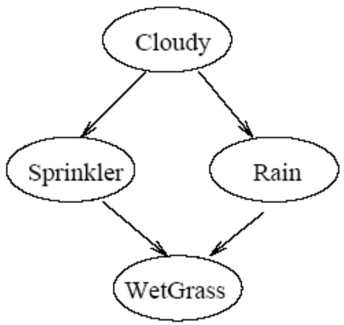
$\mapsto$ come up with graphical description

Bayesian Networks

Example for Bayesian Network

From Russel
and Norvig 95
AI: A Modern Approach

P(C=F)	P(C=T)
0.5	0.5



C	P(S=F)	P(S=T)
F	0.5	0.5
T	0.9	0.1

C	P(R=F)	P(R=T)
F	0.8	0.2
T	0.2	0.8

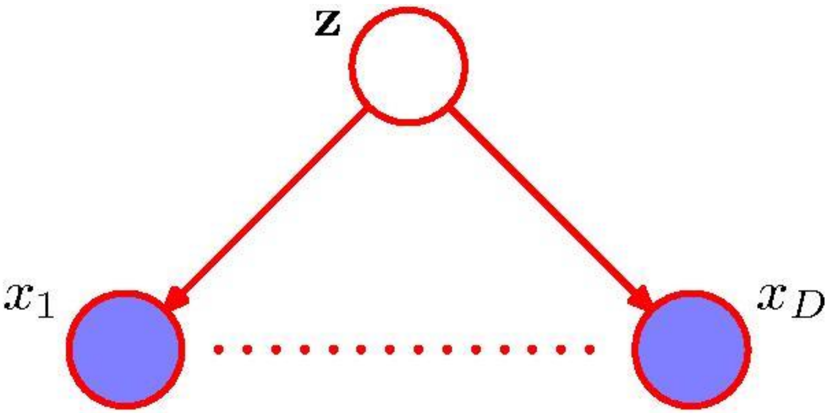
S	R	P(W=F)	P(W=T)
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

Corresponding joint distribution

$$P(C,S,R,W) =$$
$$P(W \mid S,R)P(S \mid C)P(R \mid C)P(C)$$

Naïve Bayes

Observations x_1, …. x_D are assumed to be independent

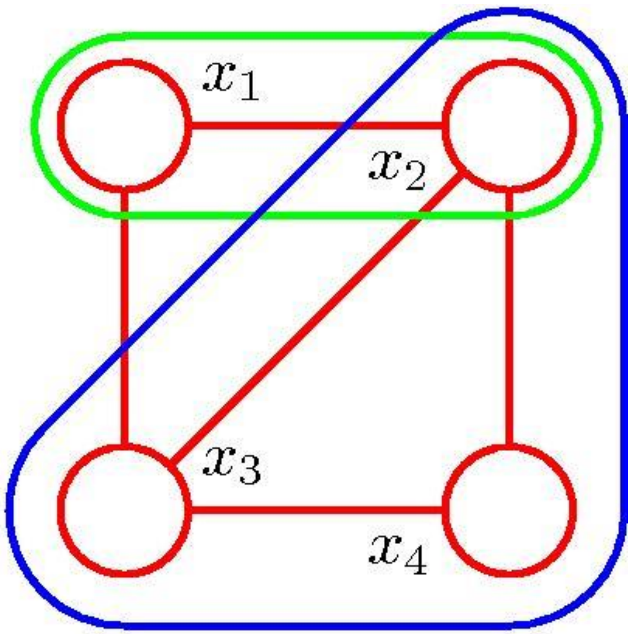


$$\prod_{i=1}^{D} P(x_i \mid z)$$

Markov Random Fields

- Undirected graphical model
- New term:
- *clique* in an undirected graph:
  - Set of nodes such that every node is connected to every other node
- *maximal clique*: there is no node that can be added without add without destroying the clique property

Example



cliques: green and blue

maximal clique: blue

Factorization

x : all nodes $x_1 ... x_N$

x_C : nodes in clique C

C_M : set of all maximal cliques

$\Psi_C(x_C)$: potential function $(\Psi_C(x_C) \geq 0)$

Joint distribution described by graph

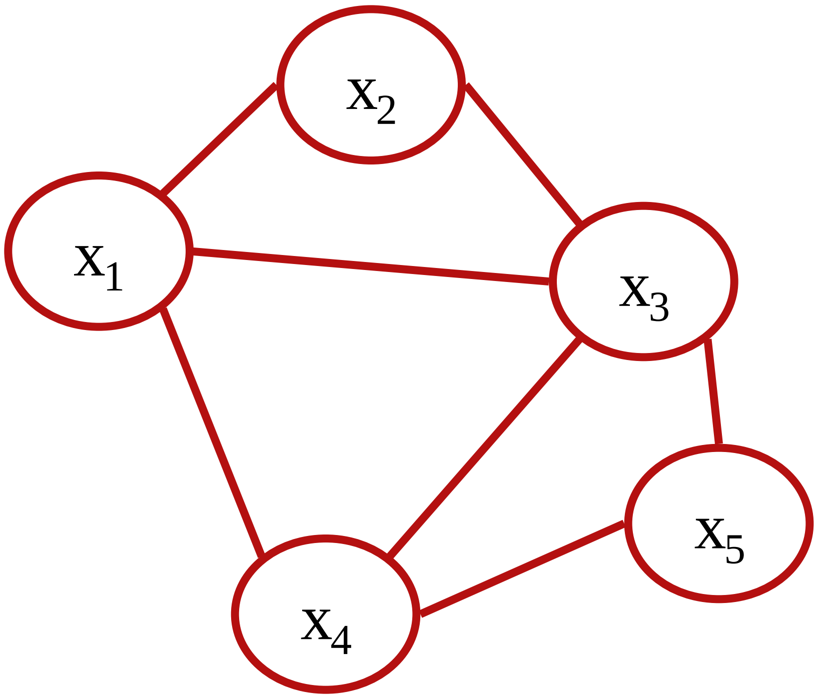
$$p(x) = \frac{1}{Z} \prod_{C \in C_M} \Psi_C(x_C)$$

Normalization

$$Z = \sum_x \prod_{C \in C_M} \Psi_C(x_C)$$

Z is sometimes call the *partition function*

Example



What are the maximum cliques?
Write down joint probability described by this graph

$\mapsto$ white board

Energy Function

Define

$$\Psi_C(x_C) = e^{-E_C(x_C)}$$

Insert into joint distribution

$$p(x) = \frac{1}{Z} e^{-\sum_{C \in C_M} E_C(x_C)}$$

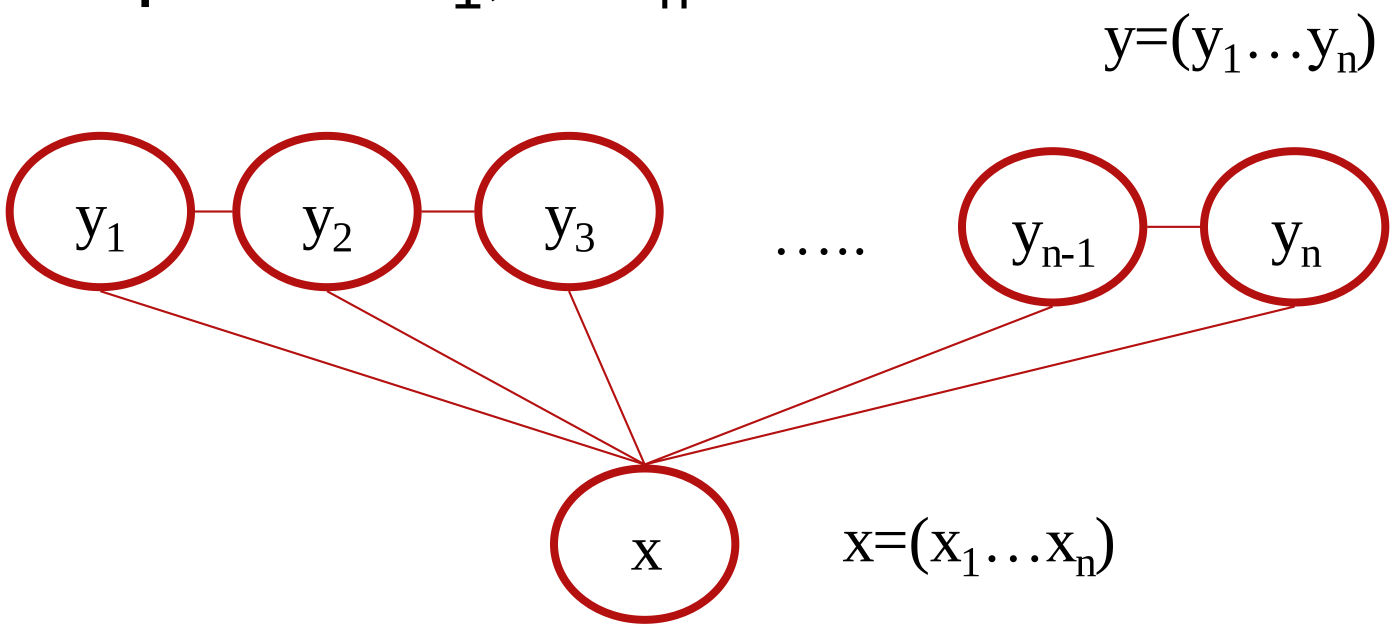
Conditional Random Fields

Definition

Markov random field

were each random variable y_i

is conditioned on the complete input sequence $x_1, \ldots x_n$



Distribution

Distribution

$$p(y \mid x) = \frac{1}{Z(x)} e^{-\sum_{i=1}^{n}\sum_{j=1}^{N} \lambda_j f_j(y_{i-1}, y_i, x, i)}$$

λ_j : parameters to be trained

$f_j(y_{i-1}, y_i, x, i)$: feature function

Example feature functions

Modeling transitions

$$f_1(y_{i-1}, y_i, x, i) = \begin{cases} 1 \text{ if } \mathrm{y_{i\text{-}1}} = IN \text{ and } \mathrm{y_i} = NNP \\ 0 \text{ else} \end{cases}$$

Modeling emissions

$$f_2(y_{i-1}, y_i, x, i) = \begin{cases} 1 \text{ if } \mathrm{y_i} = NNP \text{ and } \mathrm{x_i} = September \\ 0 \text{ else} \end{cases}$$

Training

- Like in maximum entropy models

Generalized iterative scaling

- Convergence:

  $p(y|x)$ is a convex function

  $\mapsto$ unique maximum

  Convergence is slow

  Improved algorithms exist

Decoding: Auxiliary Matrix

Define additional start symbol

y_0=START

and stop symbol

y_{n+1}=STOP

Define matrix $M^i(x)$

such that

$$\left[M^i(x)\right]_{y_{i-1}y_i} = M^i_{y_{i-1}y_i}(x) = e^{-\sum_{j=1}^{N}\lambda_j f_j(y_{i-1}, y_i, x, i)}$$

Reformulate Probability

With that definition we have

$$p(y \mid x) = \frac{1}{Z(x)} \prod_{i=1}^{n+1} M^{i}_{y_{i-1} y_i}(x)$$

with

$$Z(x) = \sum_{y_1} \sum_{y_2} \sum_{y_3} \ldots \sum_{y_n} M^{1}_{y_0 y_1}(x) M^{2}_{y_1 y_2}(x) \ldots M^{n+1}_{y_n y_{n+1}}(x)$$

Use Matrix Properties

Use matrix product

$$\left[M^{1}(x)M^{2}(x)\right]_{y_0 y_2} = \sum_{y_1} M^{1}_{y_0 y_1}(x) M^{2}_{y_1 y_2}(x)$$

with

$$Z(x) = \left[M^{1}(x)M^{2}(x)\ldots M^{n+1}(x)\right]_{y_0 = START,\, y_{n+1} = STOP}$$

Viterbi Decoding

- Matrix M replaces the product of transition and emission probability
- Decoding can be done in Viterbi style
- Effort:
  - linear in length of sequence
  - quadratic in the number of labels

Software

CRF++

- See http://crfpp.sourceforge.net/

Summary

- Sequence labeling problems
- CRFs are
  - flexible
  - Expensive to train
  - Fast to decode