

Condition Random Fields for Rational Learning

Relational Data

- Dependencies exist between the entities that we wish model
- Each entity often has a rich set of features that aid classification.

Graphical Models

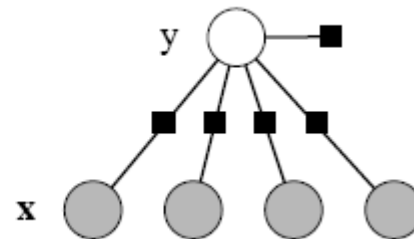
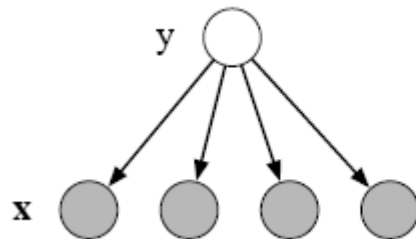
- To represent a distribution over a larger number of random variables by a product of local functions that each depend on only a small number of variables.

Graphical Models

- Finding the dependence structure among entities.
- They provide a simple way to view the probabilistic model
- Inference and learning are treated together
- Supervised and unsupervised learning are merged seamlessly
- Missing data handled nicely

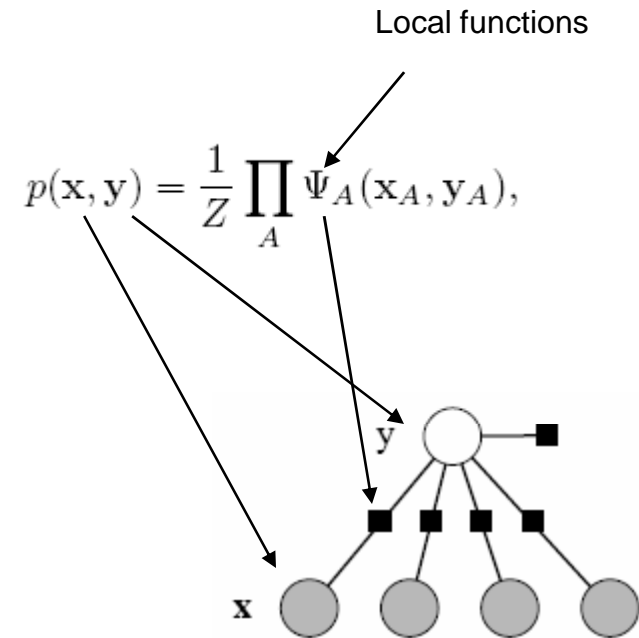
Graphical Models

- There are two kinds of graphical models
 - Directed graphs
 - Undirected graphs



Undirected Graphs

- The circles are variable nodes
- The shaded boxes are factor nodes.



Alternative Names

- Belief Networks
- Bayesian Networks
- Probabilistic Independence Networks
- Markov Random Fields
- Log Linear Models
- Influence Diagrams

Single Class Classification

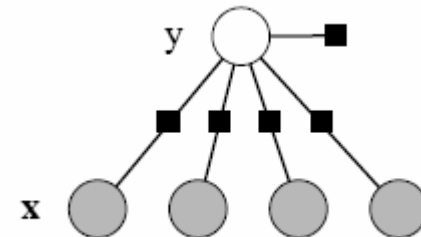
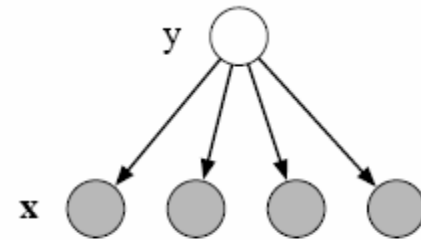
- Predicting a class y given a vector of features $\mathbf{x} = (x_1, x_2, \dots, x_k)$. Assume all the features are independent.
- Writing it as a factor graph.

$$\Psi(y) = p(y).$$

$$\Psi_k(y, x_k) = p(x_k|y)$$

- For each feature x_k

$$p(y, \mathbf{x}) = p(y) \prod_{k=1}^K p(x_k|y).$$



Sequence Models

- Given observed data sequences $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_t\}$
- A corresponding label sequence \mathbf{y}_t for each data sequence \mathbf{x}_t , and $Y = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_t\}$

Prediction Task

Given a sequence \mathbf{x} and model θ predict \mathbf{y}

Learning Task

Given training sets X and Y , learn the best model θ

Example label and observation sequence

label y <head> X-NNTP-Poster: NewsHound v1.33 **observation x**
<head>
<head> Archive-name: acorn/faq/part2
<head> Frequency: monthly
<head>
<question> 2.6) What configuration of serial cable should
<question> I use?
<answer> Here follows a diagram of the necessary
<answer> connections programs to work properly. They
<answer> are as far as know agreed upon by commercial
<answer> comms software developers fo
<answer>
<answer> Pins 1, 4, and 8 must be connected together
<answer> is to avoid the well known serial port chip bugs.

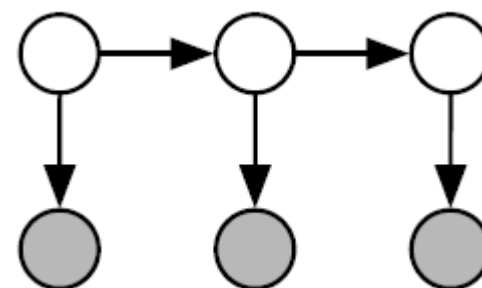
label sequence

v

observation sequence x

Sequence Models HMM

- (Jointed Distribution Model)
 - Two Assumptions
 - Each state depends only on its predecessor.
 - Each observation variable x_t depends only on the current state y_t .
 - Three probability distributions
 - $P(y_1)$ over initial state
 - The transition distribution $P(y_t|y_{t-1})$
 - Observation distribution $p(x_t|y_t)$.



HMMs

$$p(\mathbf{y}, \mathbf{x}) = \prod_{t=1}^T p(y_t|y_{t-1})p(x_t|y_t),$$

Generative Modeling (HMM)

- Joint distribution $p(y,x) = p(x)*p(x|y)$
- Need to model $p(x)$ (difficult)
- Given training set X with label sequences Y :
 - Train a model θ that maximizes $P(X, Y | \theta)$
 - Maximizes the conditional likelihood.
 - For a new data sequence \mathbf{x} , the predicted label \mathbf{y} maximizes $P(\mathbf{y} | \mathbf{x}) = P(\mathbf{y} | \mathbf{x}, \theta)P(\mathbf{x} | \theta)$

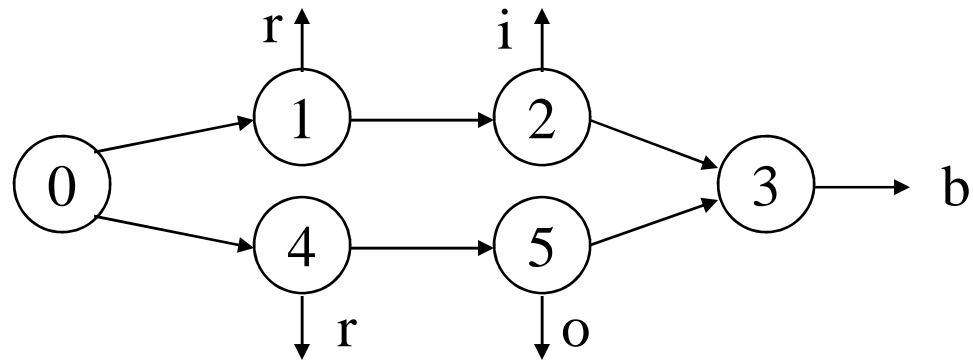
Discriminative Modeling (**MEMM**)

- Conditional distribution $p(y|x)$
- Does not need $p(x)$
- Given training set X with label sequences Y :
 - Train a model θ that maximizes $P(Y | X, \theta)$
 - Maximizes the joint likelihood.
 - For a new data sequence \mathbf{x} , the predicted label \mathbf{y} maximizes $P(\mathbf{y} | \mathbf{x}, \theta)$

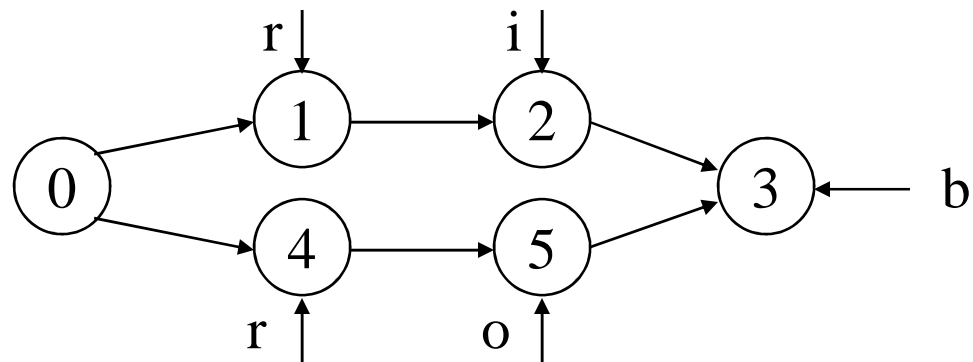
rib/rob models

Training data {<rib, 123>, <rob, 453>}

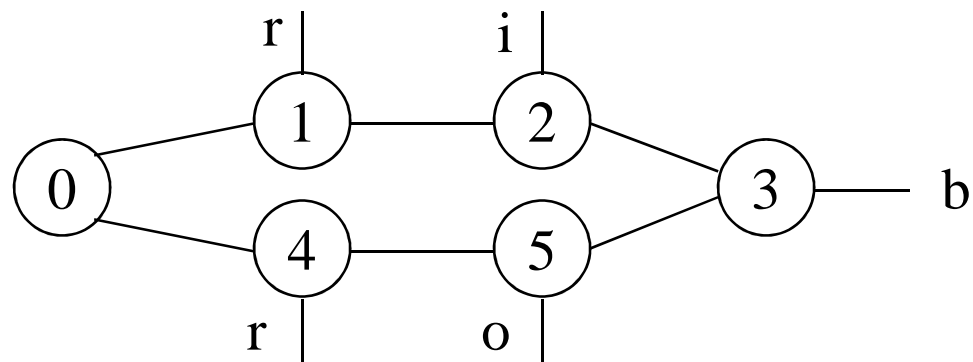
HMM



MEMM



CRF



Conditional Random Fields

- Combining discriminative and generative model

$$p(\mathbf{y}, \mathbf{x}) = \prod_{t=1}^T p(y_t | y_{t-1}) p(x_t | y_t).$$

$$\begin{aligned} \theta &= \{\lambda_{ij}, \mu_{oi}\} \\ \lambda_{ij} &= \log p(y' = i | y = j) \end{aligned}$$

$$p(\mathbf{y}, \mathbf{x}) = \frac{1}{Z} \exp \left\{ \sum_t \sum_{i,j \in S} \lambda_{ij} \mathbf{1}_{\{y_t=i\}} \mathbf{1}_{\{y_{t-1}=j\}} + \sum_t \sum_{i \in S} \sum_{o \in O} \mu_{oi} \mathbf{1}_{\{y_t=i\}} \mathbf{1}_{\{x_t=o\}} \right\}$$

Added parameterization but not distribution

Each feature function has the form

$$p(\mathbf{y}, \mathbf{x}) = \frac{1}{Z} \exp \left\{ \sum_{k=1}^K \lambda_k f_k(y_t, y_{t-1}, x_t) \right\}$$

$$f_{ij}(y, y', x) = \mathbf{1}_{\{y=i\}} \mathbf{1}_{\{y'=j\}} \text{ for each transition } (i, j)$$

$$f_{io}(y, y', x) = \mathbf{1}_{\{y=i\}} \mathbf{1}_{\{x=o\}} \text{ for each state-observation pair } (i, o)$$

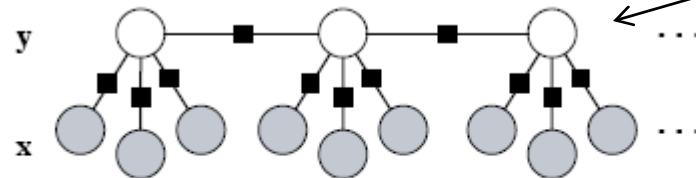
Conditional Random Fields

- The last step, writing the conditional distribution $p(\mathbf{y}|\mathbf{x})$

$$p(\mathbf{y}|\mathbf{x}) = \frac{p(\mathbf{y}, \mathbf{x})}{\sum_{\mathbf{y}'} p(\mathbf{y}', \mathbf{x})} = \frac{\exp \left\{ \sum_{k=1}^K \lambda_k f_k(y_t, y_{t-1}, x_t) \right\}}{\sum_{\mathbf{y}'} \exp \left\{ \sum_{k=1}^K \lambda_k f_k(y'_t, y'_{t-1}, x_t) \right\}} \longleftarrow Z(\mathbf{x})$$



$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp \left\{ \sum_{k=1}^K \lambda_k f_k(y_t, y_{t-1}, x_t) \right\}$$



Transition score is the same for each transition between two states.

Graphical model of an HMM-like linear-chain CRF.

Conditional Random Fields

- CRFs allow transition (i,j) score to depend on the current observation vector by adding $\mathbf{1}_{\{y_t=j\}} \mathbf{1}_{\{y_{t-1}=1\}} \mathbf{1}_{\{x_t=o\}}$

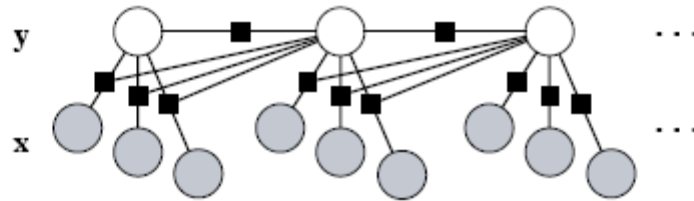


Figure 1.4 Graphical model of a linear-chain CRF in which the transition score depends on the current observation.

Parameter Estimation

Input: training data $D = \{(x^{(i)}, y^{(i)})\}$, where $i = 1 \dots N$ with empirical $\underline{p}(\mathbf{x}, \lambda)$ distribution. Where X is a sequence of inputs and y is a sequence of predictions.

Output: parameters $\Theta = (\lambda_1, \lambda_2, \dots)$

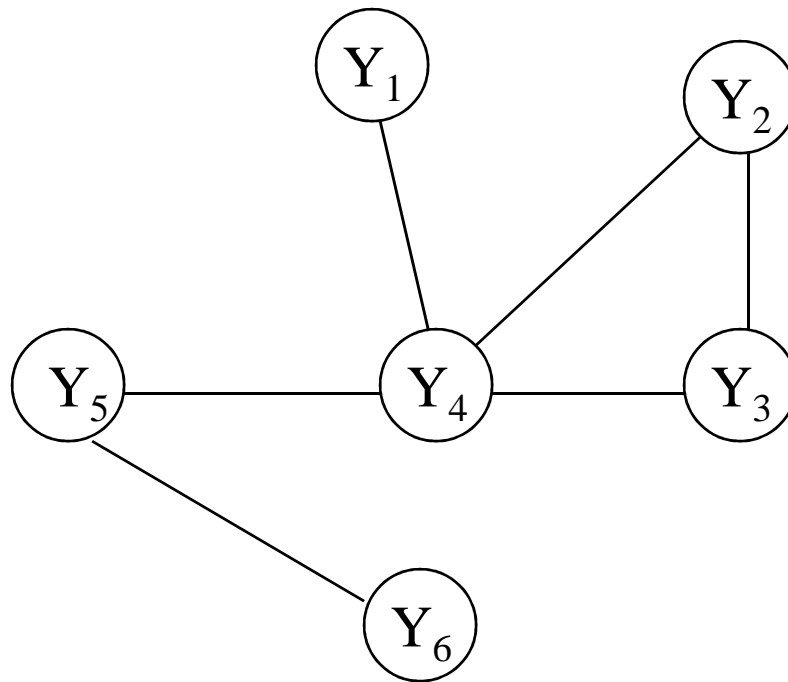
Maximize: the log-likelihood objective function:

$$O(\Theta) = \sum_{i=1}^N \log p_{\Theta}(\mathbf{y}^{(i)} | \mathbf{x}^{(i)})$$

Random Field Example

Let $G = (Y, E)$ be a graph where each vertex Y_v is a random variable
Suppose $P(Y_v | \text{all other } Y) = P(Y_v | \text{neighbors}(Y_v))$ then Y is a
random field

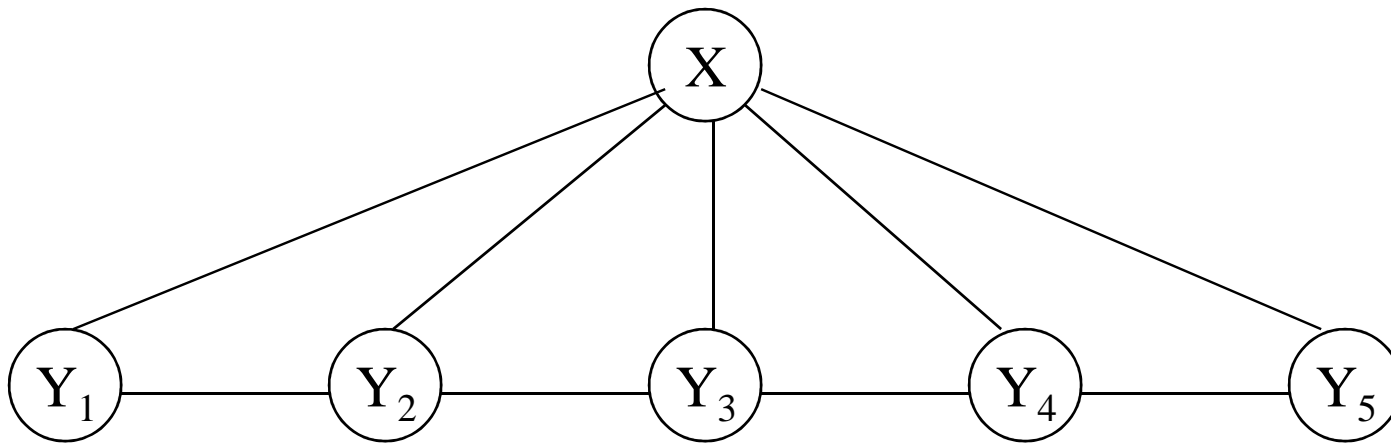
Example:



- $P(Y_5 | \text{all other } Y) = P(Y_5 | Y_4, Y_6)$

Conditional Random Field Example

Suppose $P(Y_v | X, \text{all other } Y) = P(Y_v | X, \text{neighbors}(Y_v))$
then X with Y is a **conditional** random field

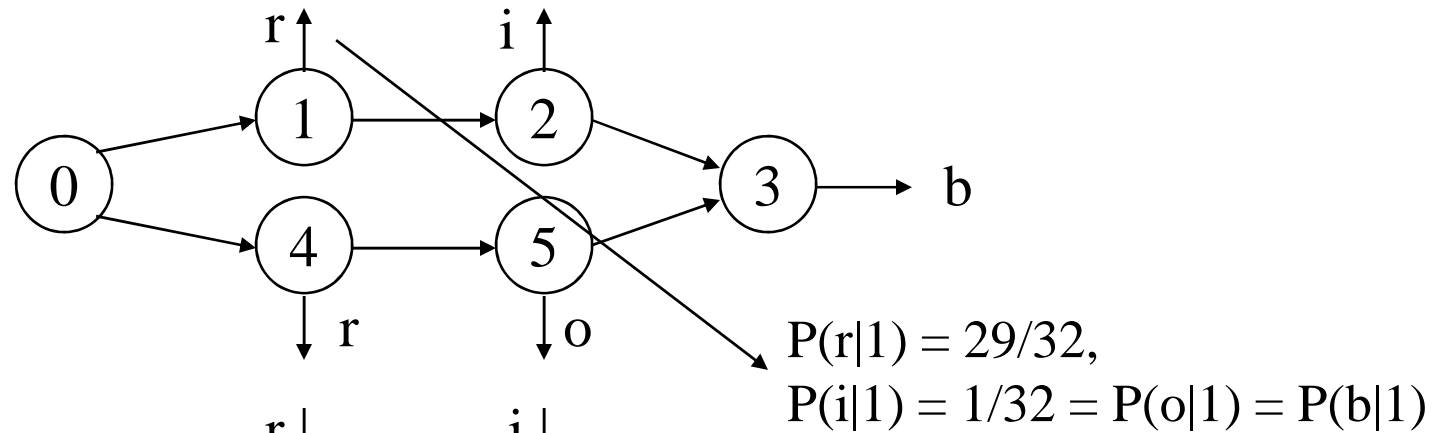


- $P(Y_3 | X, \text{all other } Y) = P(Y_v | X, Y_2, Y_4)$
- Think of X as observations and Y as labels

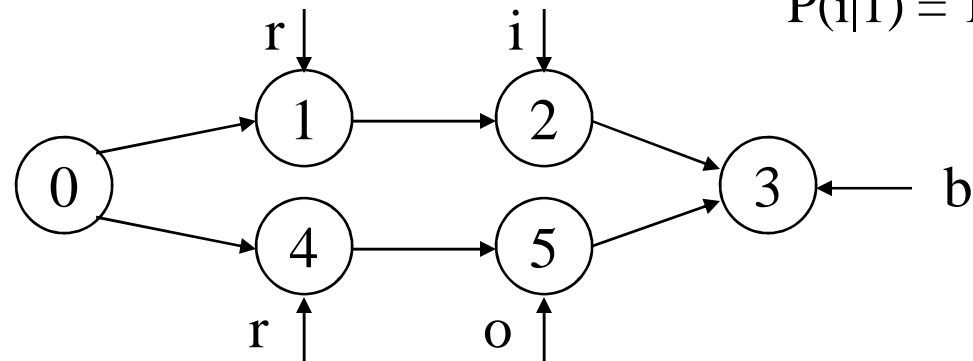
rib/rob models

Training data {<rib, 123>, <rob, 453>}

HMM



MEMM



CRF

