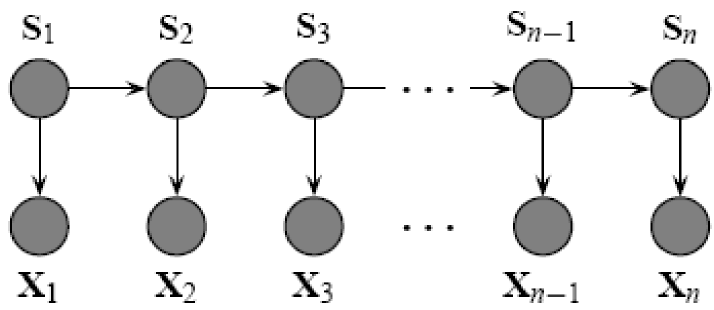
Conditional Random Field
CRF

Sudeshna Sarkar

2 Aug 2019

Hidden Markov Model



$$p(s,x) = p(s_1)\,p(x_1 \mid s_1)\prod_{i=2}^{n} p(s_i \mid s_{i-1})\,p(x_i \mid s_i)$$

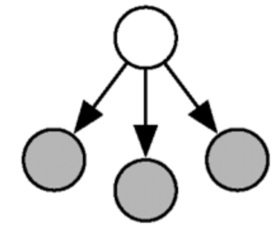
Cannot represent multiple interacting features or long range dependences between observed elements.

Discriminative Vs. Generative

$p(\mathbf{y}, \mathbf{x})$

- **Generative Model:** A model that generate observed data randomly
- **Naïve Bayes:** once the class label is known, all the features are independent

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

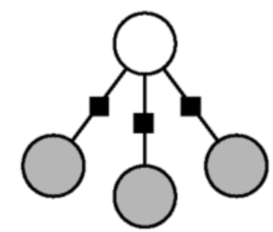


Naive Bayes

$p(\mathbf{y} \mid \mathbf{x})$

- **Discriminative:** Directly estimate the posterior probability; Aim at modeling the "discrimination" between different outputs
- **MaxEnt** classifier: linear combination of feature function in the exponent,

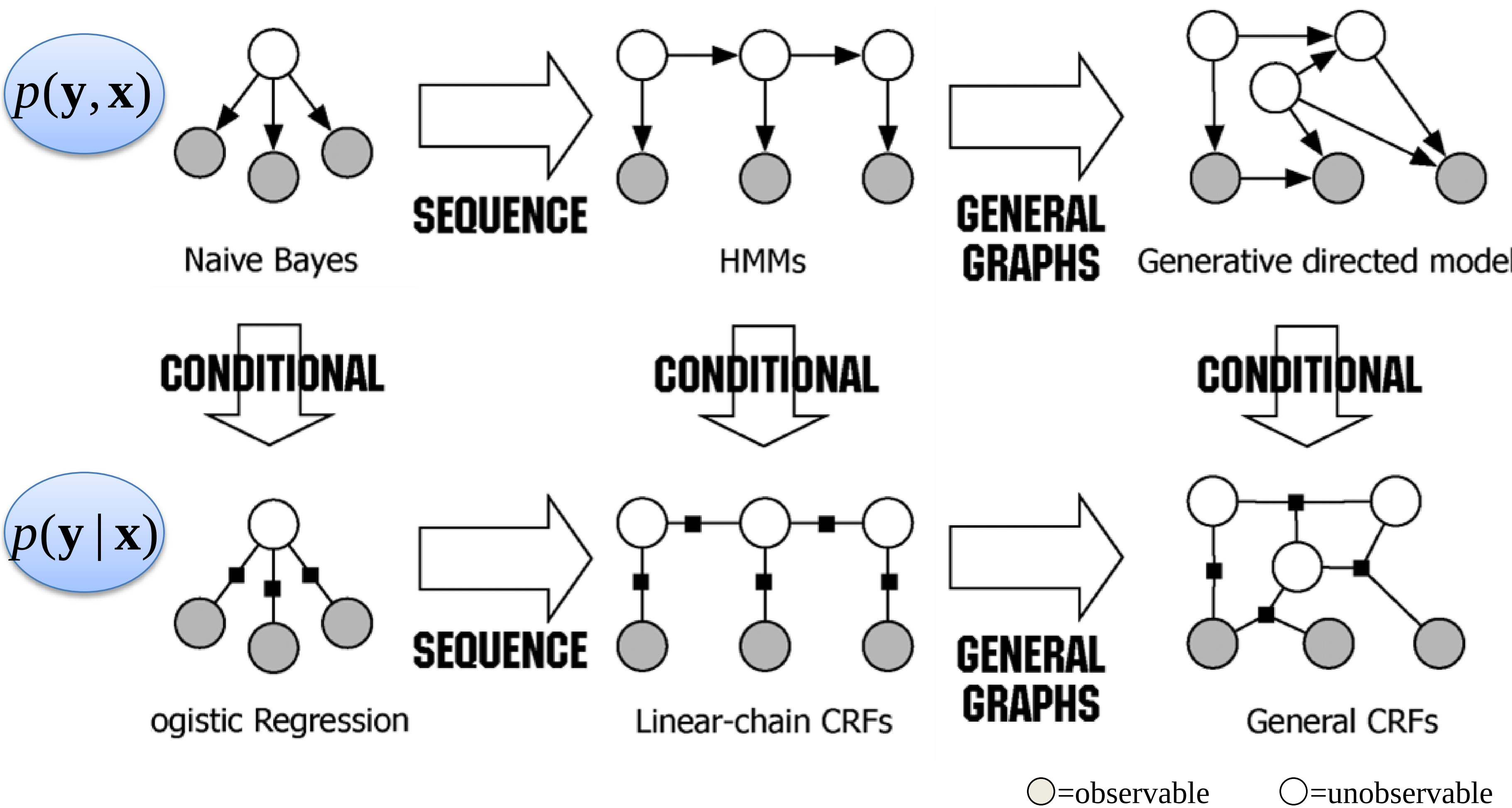
$$p(y|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left\{ \sum_{k=1}^{K} \theta_k f_k(y, \mathbf{x}) \right\}$$



Logistic Regression

Both generative models and discriminative models describe distributions over (y , x), but they work in different directions.

Discriminative Vs. Generative

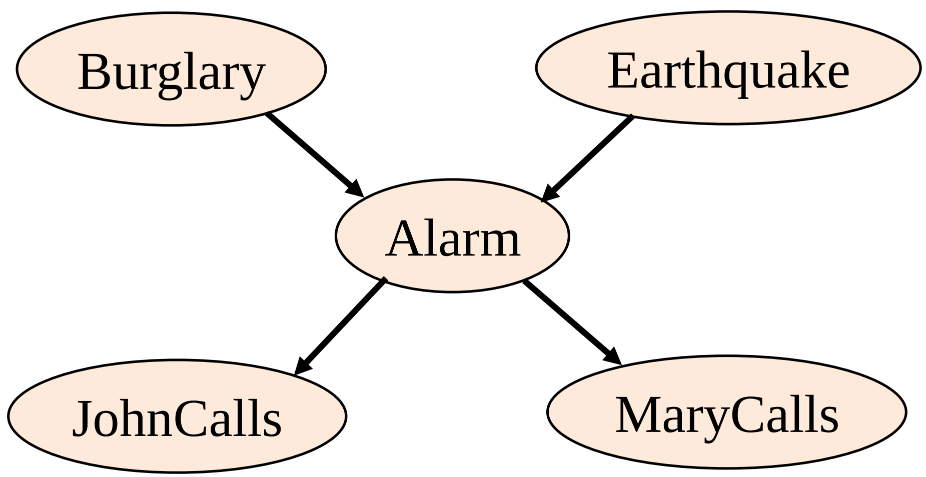


Graphical Models

- If no assumption of independence is made, then an exponential number of parameters must be estimated for sound probabilistic inference.
- If a blanket assumption of conditional independence is made, efficient training and inference is possible, but such a strong assumption is rarely warranted.
- **Graphical models** use directed or undirected graphs over a set of random variables to explicitly specify variable dependencies and allow for less restrictive independence assumptions while limiting the number of parameters that must be estimated.
  - **Bayesian Networks**: Directed acyclic graphs that indicate causal structure
  - **Markov Networks**: Undirected graphs that capture general dependencies

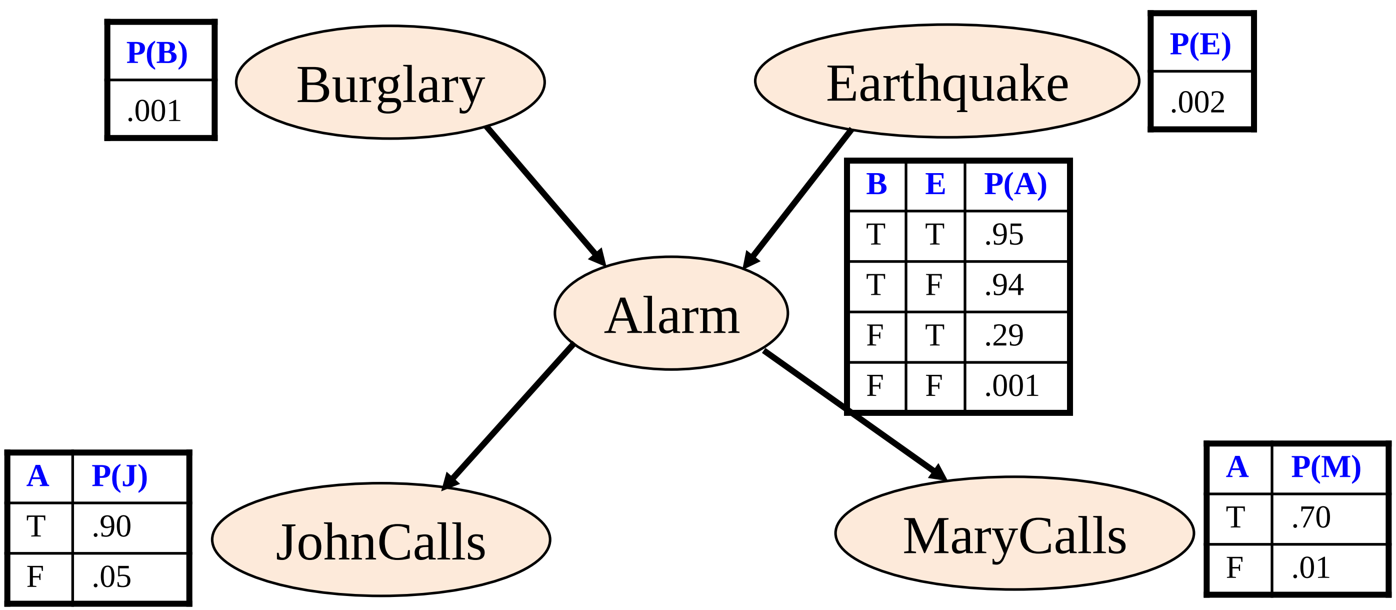
Bayesian Networks

- Directed Acyclic Graph (DAG)
  - Nodes are random variables
  - Edges indicate causal influences



Conditional Probability Tables

- Each node has a **conditional probability table** (**CPT**) that gives the probability of each of its values given every possible combination of values for its parents (conditioning case).
  - Roots (sources) of the DAG that have no parents are given prior probabilities.



P(B)
.001

P(E)
.002

B	E	P(A)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(J)
T	.90
F	.05

A	P(M)
T	.70
F	.01

Joint Distributions for Bayes Nets

- A Bayesian Network implicitly defines a joint distribution.

$$P(x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} P(x_i \mid \mathrm{Parents}(X_i))$$

- Example

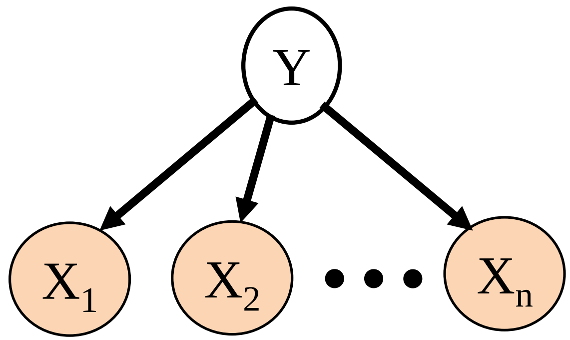
$$P(J \wedge M \wedge A \wedge \neg B \wedge \neg E)$$

$$= P(J \mid A) P(M \mid A) P(A \mid \neg B \wedge \neg E) P(\neg B) P(\neg E)$$

$$= 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998 = 0.00062$$

Naïve Bayes as a Bayes Net

- Naïve Bayes is a simple Bayes Net



- Priors P(Y) and conditionals P(X_i|Y) for Naïve Bayes provide CPTs for the network.

Markov Networks

- Undirected graph over a set of random variables, where an edge represents a dependency.
- The **Markov blanket** of a node, *X*, in a Markov Net is the set of its neighbors in the graph (nodes that have an edge connecting to *X*).
- Every node in a Markov Net is conditionally independent of every other node given its Markov blanket.

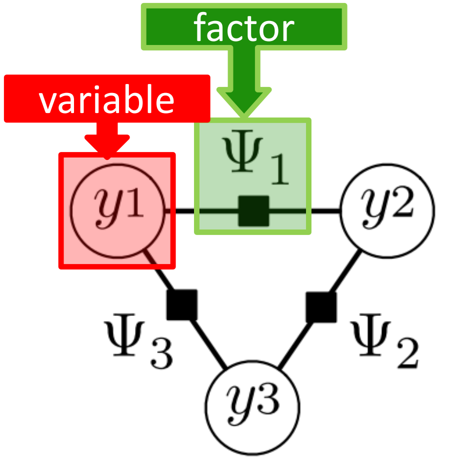
Distribution for a Markov Network

- The distribution of a Markov net is most compactly described in terms of a set of **potential functions** (a.k.a. factors, compatibility functions), ψ_k, for each clique, k, in the graph.
- For each joint assignment of values to the variables in clique k, ψ_k assigns a non-negative real value that represents the compatibility of these values.
- The joint distribution of variables $\boldsymbol{y}$:

$$p(\mathbf{y}) = \frac{1}{Z}\prod_C \psi_C(\mathbf{y}_C), \quad Z = \sum_{\mathbf{y}}\prod_C \psi_C(\mathbf{y}_C)$$

$$\psi_C(\mathbf{y}_C) \geq 0$$

Typically $\psi_C(\mathbf{y}_C) = \exp\{-E(\mathbf{y}_C)\}$

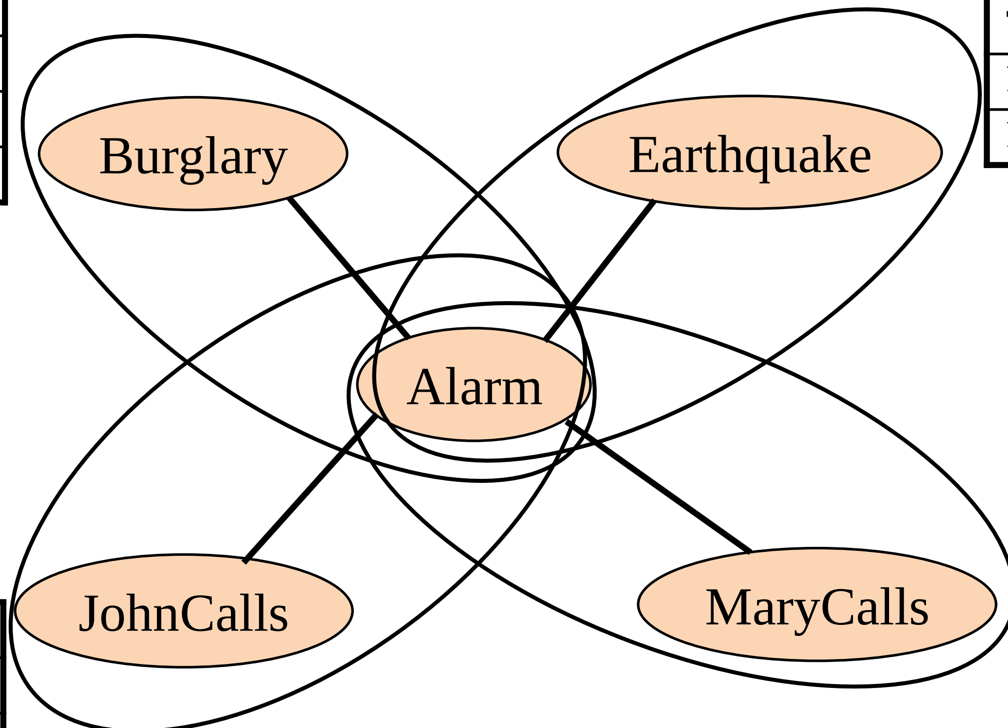


$$\boxed{p(y_1, y_2, y_3) \propto \Psi_1(y_1, y_2)\Psi_2(y_2, y_3)\Psi_3(y_1, y_3)}$$

Sample Markov Network

B	A	ϕ_1
T	T	100
T	F	1
F	T	1
F	F	200

E	A	ϕ_2
T	T	50
T	F	10
F	T	1
F	F	200



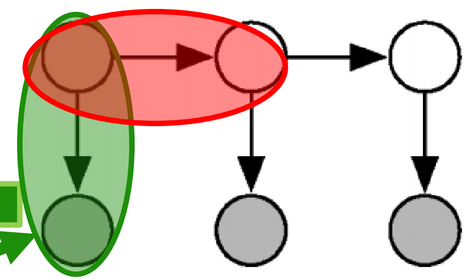
J	A	ϕ_3
T	T	75
T	F	10
F	T	1
F	F	200

M	A	ϕ_4
T	T	50
T	F	1
F	T	10
F	F	200

Sequence prediction

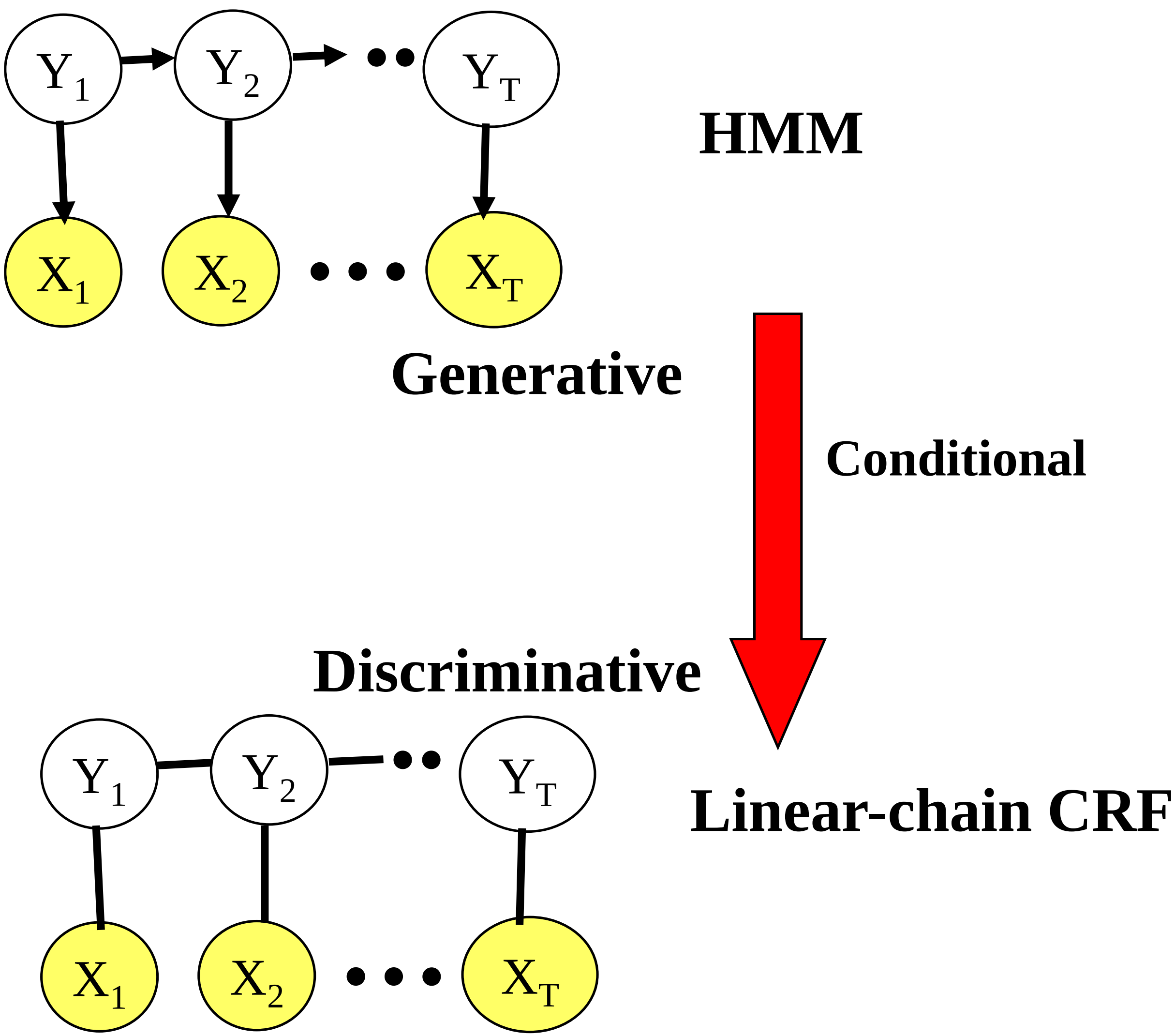
- NER: identifying and classifying proper names in text,
  - Set of observation, $X = \{x_t\}_{t=1}^{\mathrm{T}}$
  - Set of underlying sequence of states, $Y = \{y_t\}_{t=1}^{\mathrm{T}}$
- HMM is generative:

$$p(\mathbf{y}, \mathbf{x}) = \prod_{t=1} \overbrace{p(y_t | y_{t-1})}^{\text{Transition probability}} \underbrace{p(x_t | y_t)}_{\text{Observation probability}}$$



- Doesn’t model long-range dependencies
- Not practical to represent multiple interacting features (hard to model p(x))
- The primary advantage of CRFs over hidden Markov models is their conditional nature, resulting in the relaxation of the independence assumptions
- And it can handle overlapping features

Sequence Labeling



Simple Linear Chain CRF Features

- Modeling the conditional distribution is similar to that used in multinomial logistic regression.
- Create feature functions $f_k(Y_t, Y_{t-1}, X_t)$
  - Feature for each state transition pair i, j
    - $f_{i,j}(Y_t, Y_{t-1}, X_t) = 1$ if $Y_t = i$ and $Y_{t-1} = j$ and 0 otherwise
  - Feature for each state observation pair i, o
    - $f_{i,o}(Y_t, Y_{t-1}, X_t) = 1$ if $Y_t = i$ and $X_t = o$ and 0 otherwise
- **Note**: number of features grows quadratically in the number of states (i.e. tags).

Conditional Distribution for Linear Chain CRF

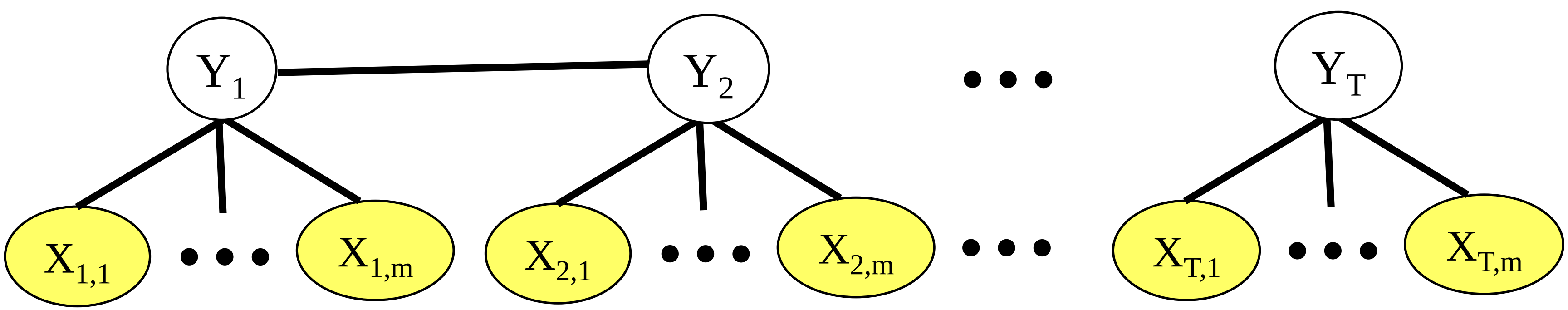
- Using these feature functions for a simple linear chain CRF, we can define:

$$P(Y \mid X) = \frac{1}{Z(X)} \exp\left(\sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k f_k(Y_t, Y_{t-1}, X_t)\right)$$

$$Z(X) = \sum_{Y} \exp\left(\sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k f_k(Y_t, Y_{t-1}, X_t)\right)$$

Adding Token Features to a CRF

- Can add token features $X_{i,j}$



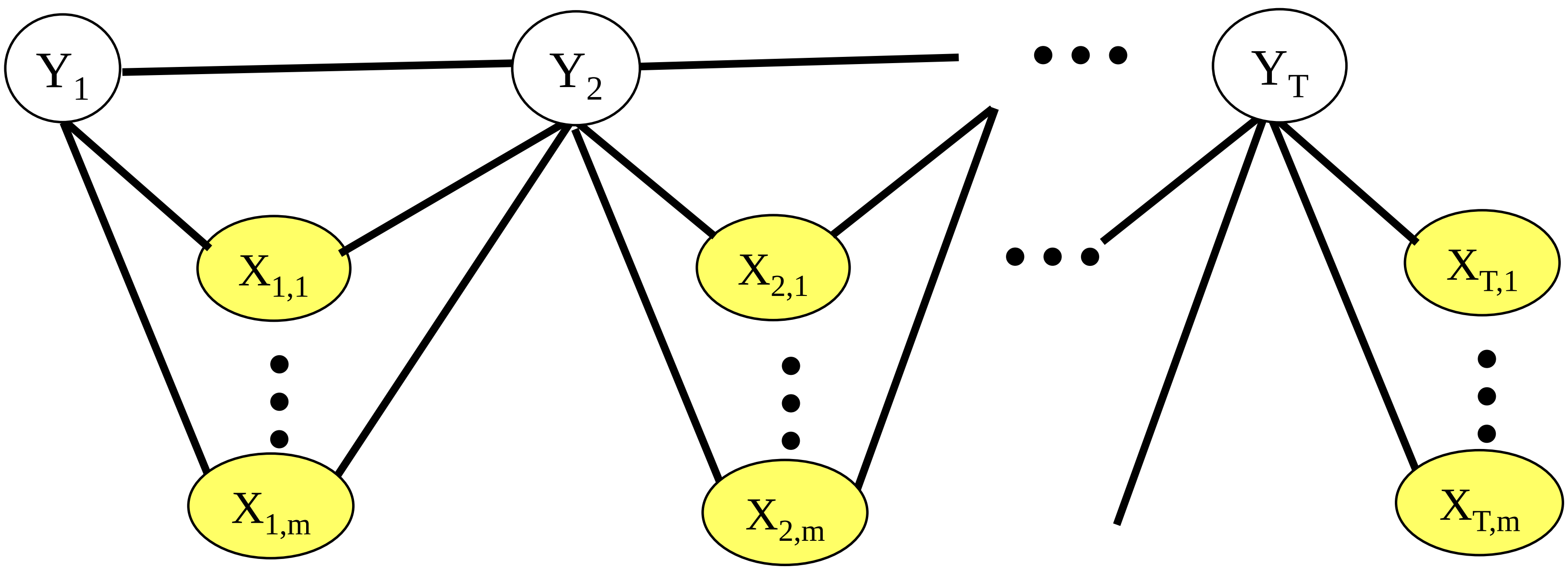
- Can add additional feature functions for each token feature to model conditional distribution.

Features in POS Tagging

- For POS Tagging, use lexicographic features of tokens.
  - Capitalized?
  - Start with numeral?
  - Ends in given suffix (e.g. “s”, “ed”, “ly”)?

Enhanced Linear Chain CRF (standard approach)

- Can also condition transition on the current token features.



- Add feature functions:
  - $f_{i,j,k}(Y_t, Y_{t-1}, X)$ 1 if $Y_t = i$ and $Y_{t-1} = j$ and $X_{t-1,k} = 1$ and 0 otherwise

Supervised Learning (Parameter Estimation)

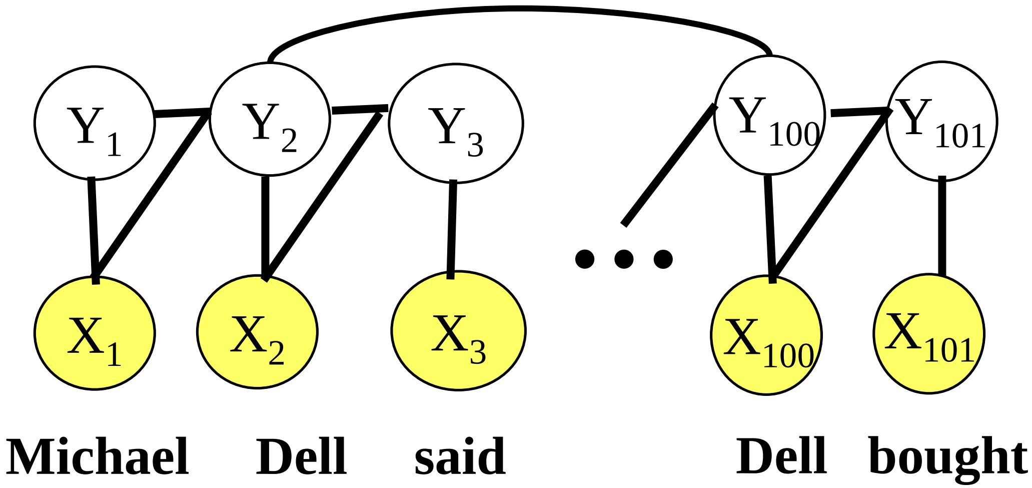
- As in logistic regression, use L-BFGS optimization procedure, to set λ weights to maximize CLL of the supervised training data.
- See paper for details.

Sequence Tagging (Inference)

- Variant of Viterbi algorithm can be used to efficiently, $O(TN^2)$, determine the globally most probable label sequence for a given token sequence using a given log-linear model of the conditional probability $P(Y \mid X)$.
- See paper for details.

Skip-Chain CRFs

- Can model some long-distance dependencies (i.e. the same word appearing in different parts of the text) by including long-distance edges in the Markov model.



- Additional links make exact inference intractable, so must resort to approximate inference to try to find the most probable labeling.

CRF Results

- Experimental results verify that they have superior accuracy on various sequence labeling tasks.
  - Part of Speech tagging
  - Noun phrase chunking
  - Named entity recognition
  - Semantic role labeling
- However, CRFs are much slower to train and do not scale as well to large amounts of training data.
  - Training for POS on full Penn Treebank (~1M words) currently takes "over a week."
- Skip-chain CRFs improve results on IE.

CRF Summary

- CRFs are a discriminative approach to sequence labeling whereas HMMs are generative.
- Discriminative methods are usually more accurate since they are trained for a specific performance task.
- CRFs also easily allow adding additional token features without making additional independence assumptions.
- Training time is increased since a complex optimization procedure is needed to fit supervised training data.
- CRFs are a state-of-the-art method for sequence labeling.