A brief introduction to Conditional Random Fields

Mark Johnson

Macquarie University

April, 2005, updated October 2010

Talk outline

- Graphical models
- Maximum likelihood and maximum conditional likelihood estimation
- Naive Bayes and Maximum Entropy Models
- Hidden Markov Models
- Conditional Random Fields

Classification with structured labels

- Classification: predicting label y given features x

$$y^\star(x) \quad = \quad \arg\max_{y} P(y|x)$$

- Naive Bayes and Maxent models: label y is atomic, x can be structured (e.g., set of features)
- HMMs and CRFs are extensions of Naive Bayes and Maxent models where y is structured too
- HMMs and CRFs model dependencies between components y_i of label y
- Example: Part of speech tagging: x is a sequence of words, y is corresponding sequence of parts of speech (e.g., noun, verb, etc.)

Why graphical models?

- Graphical models depict *factorizations of probability distributions*
- Statistical and computational properties depend on the factorization
  - complexity of dynamic programming is size of a certain cut in the graphical model
- Two different (but related) graphical representations
  - *Bayes nets* (directed graphs; products of conditionals)
  - *Markov Random Fields* (undirected graphs; products of arbitrary terms)
- Each random variable X_i is represented by a node

Bayes nets (directed graph)

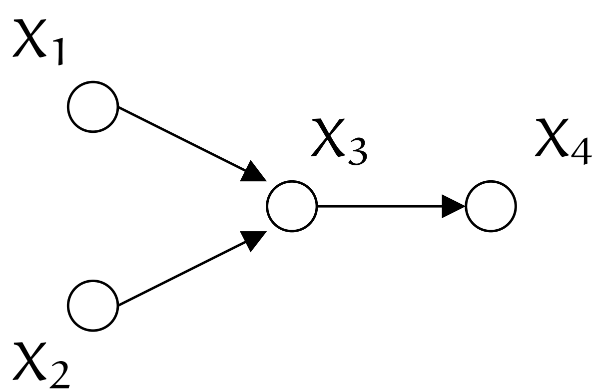
- Factorize *joint* $P(X_1, \ldots, X_n)$ into *product of conditionals*

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | X_{Pa(i)})$$

where $Pa(i) \subseteq (X_1, \ldots, X_{i-1})$

- The *Bayes net* contains an arc from each $j \in Pa(i)$ to i

$$P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2)P(X_3|X_1, X_2)P(X_4|X_3)$$



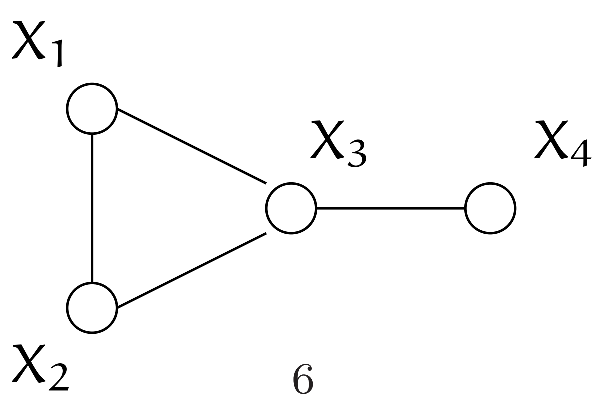
Markov Random Field (undirected)

- Factorize $P(X_1, \ldots, X_n)$ into product of *potentials* $g_c(X_c)$, where $c \subseteq (1, \ldots, n)$ and $c \in \mathcal{C}$ (a set of tuples of indices)

$$P(X_1, \ldots, X_n) = \frac{1}{Z} \prod_{c \in \mathcal{C}} g_c(X_c)$$

- If $i, j \in c \in \mathcal{C}$, then an edge connects i and j

$$\mathcal{C} = \{(1,2,3),(3,4)\}$$
$$P(X_1, X_2, X_3, X_4) = \frac{1}{Z} g_{123}(X_1, X_2, X_3)\, g_{34}(X_3, X_4)$$



A rose by any other name ...

- MRFs have the same general form as *Maximum Entropy models*, *Exponential models*, *Log-linear models*, *Harmony models*, ...
- All of these have *the same generic form*

$$\begin{aligned} P(X) &= \frac{1}{Z} \prod_{c \in \mathcal{C}} g_c(X_c) \\ &= \frac{1}{Z} \exp \sum_{c \in \mathcal{C}} \log g_c(X_c) \end{aligned}$$

Potential functions as features

- If X is *discrete*, we can represent the potentials $g_c(X_c)$ as a *combination of indicator functions* $[\![X_c = x_c]\!]$, where $\mathcal{X}_c$ is the set of all possible values of X_c

$$\begin{aligned} g_c(X_c) &= \prod_{x_c \in \mathcal{X}_c} (\theta_{X_c = x_c})^{[\![X_c = x_c]\!]}, \text{ where } \theta_{X_c = x_c} = g_c(x_c) \\ \log g_c(X_c) &= \sum_{x_c \in \mathcal{X}_c} [\![X_c = x_c]\!] \phi_{X_c = x_c}, \text{ where } \phi_{X_c = x_c} = \log g_c(x_c) \end{aligned}$$

- View $[\![X_c = x_c]\!]$ as a *feature* which “fires” when the configuration x_c occurs

- $\phi_{X_c = x_c}$ is the *weight* associated with feature $[\![X_c = x_c]\!]$

A feature-based reformulation of MRFs

- Reformulating MRFs as features:

$$\begin{aligned} P(X) &= \frac{1}{Z}\prod_{c\in\mathcal{C}} g_c(X_c) \\ &= \frac{1}{Z}\prod_{c\in\mathcal{C},x_c\in\mathcal{X}_c} (\theta_{X_c=x_c})^{[\![X_c=x_c]\!]}, \text{where } \theta_{X_c=x_c} = g_c(x_c) \\ &= \frac{1}{Z}\exp\sum_{c\in\mathcal{C},X_c\in\mathcal{X}_c} [\![X_c = x_c]\!]\phi_{X_c=x_c}, \text{where } \phi_{X_c=x_c} = \log g_c(x_c) \end{aligned}$$

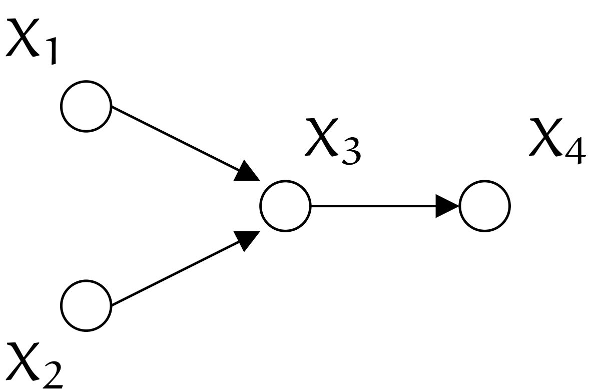
$$\begin{aligned} P(X) &= \frac{1}{Z} g_{123}(X_1, X_2, X_3)\ g_{34}(X_3, X_4) \\ &= \frac{1}{Z}\exp\left(\begin{array}{c} [\![X_{123} = 000]\!]\phi_{000} + [\![X_{123} = 001]\!]\phi_{001} + \ldots \\ [\![X_{34} = 00]\!]\phi_{00} + [\![X_{34} = 01]\!]\phi_{01} + \ldots \end{array} \right) \end{aligned}$$

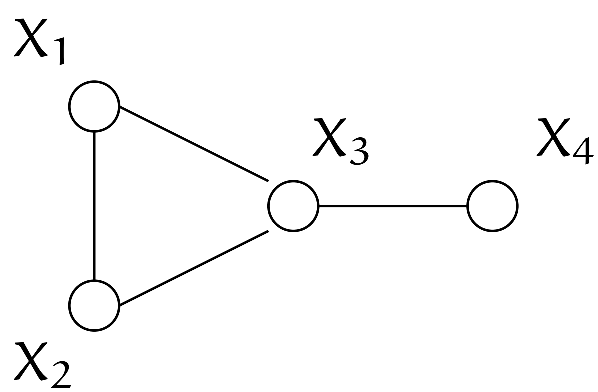
Bayes nets and MRFs

- MRFs are more general than Bayes nets
- Its easy to find the MRF representation of a Bayes net

$$P(X_1, X_2, X_3, X_4) = \underbrace{P(X_1)P(X_2)P(X_3|X_1, X_2)}_{g_{123}(X_1, X_2, X_3)} \underbrace{P(X_4|X_3)}_{g_{34}(X_3, X_4)}$$

- *Moralization*, i.e, "marry the parents"

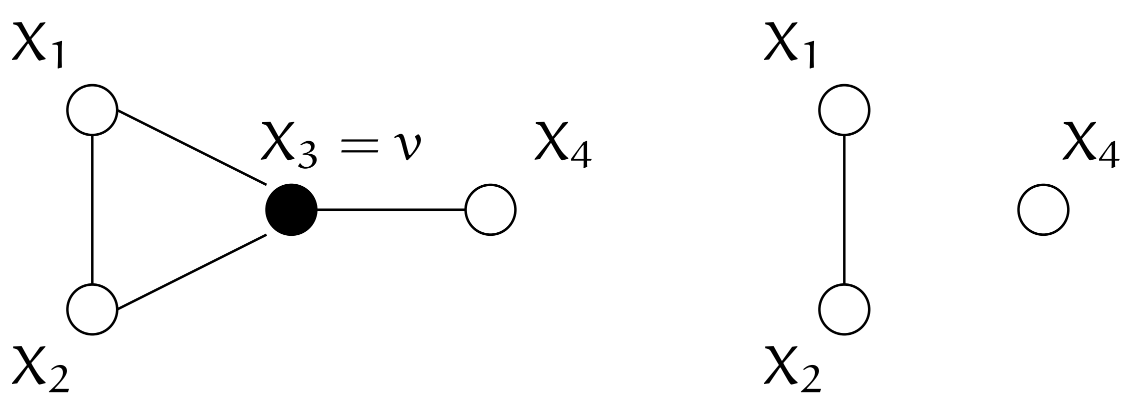




Conditionalization in MRFs

- Conditionalization is *fixing the value of some variables*
- To get a MRF representation of the conditional distribution, *delete nodes whose values are fixed and arcs connected to them*

$$
\begin{aligned}
P(X_1, X_2, X_4 | X_3 = v) &= \frac{1}{Z \;\; P(X_3 = v)} \; g_{123}(X_1, X_2, v) \; g_{34}(v, X_4) \\
&= \frac{1}{Z'(v)} \quad g'_{12}(X_1, X_2) \quad g'_4(X_4)
\end{aligned}
$$



Classification

- Given value of X, predict value of Y
- Given a probabilistic model $P(Y|X)$, predict:

$$y^\star(x) = \arg\max_y P(y|x)$$

- In general we must learn $P(Y|X)$ from data $D = ((x_1, y_1), \ldots, (x_n, y_n))$
- Restrict attention to a *parametric model class* P_θ parameterized by parameter vector θ
  - learning is estimating θ from D

ML and CML Estimation

- Maximum likelihood estimation (MLE) picks the θ that makes the data $D = (x, y)$ as *likely as possible*

$$\hat{\theta} = \arg\max_{\theta} P_{\theta}(x, y)$$

- Conditional maximum likelihood estimation (CMLE) picks the θ that maximizes *conditional likelihood* of the data $D = (x, y)$

$$\hat{\hat{\theta}} = \arg\max_{\theta} P_{\theta}(y|x)$$

- $P(X, Y) = P(X)P(Y|X)$, so CMLE *ignores* $P(X)$

MLE and CMLE example

- $X, Y \in \{0, 1\}$, $\theta \in [0, 1]$, $P_\theta(X = 1) = \theta$, $P_\theta(Y = X|X) = \theta$

  Choose X by flipping a coin with weight θ, then set Y to same value as X if flipping same coin again comes out 1.

- Given data $D = ((x_1, y_1), \ldots, (x_n, y_n))$,

$$\begin{aligned} \hat{\theta} &= \frac{\sum_i^n [\![x_i = 1]\!] + [\![x_i = y_i]\!]}{2n} \\ \hat{\hat{\theta}} &= \frac{\sum_i^n [\![x_i = y_i]\!]}{n} \end{aligned}$$

- CMLE *ignores* $P(X)$, so *less efficient* if model *correctly* relates $P(Y|X)$ and $P(X)$

- But if model *incorrectly relates* $P(Y|X)$ and $P(X)$, MLE converges to wrong θ
  - e.g., if x_i are chosen by some different process entirely

Complexity of decoding and estimation

- Finding $y^\star(x) = \arg\max_y P(y|x)$ is *equally hard* for Bayes nets and MRFs with similar architectures

- A Bayes net is a product of independent conditional probabilities

  $\Rightarrow$ MLE is *relative frequency* (easy to compute)

  – *no closed form for CMLE* if conditioning variables have parents

- A MRF is a product of arbitrary potential functions g

  – estimation involves learning values of each g takes

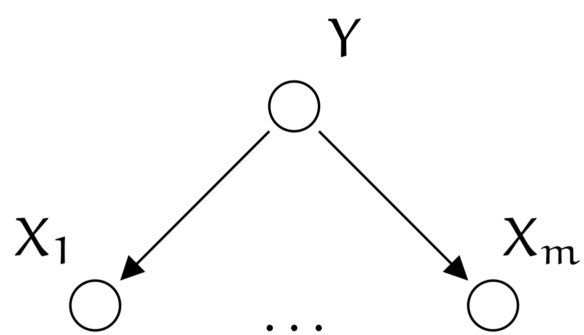
  – partition function Z changes as we adjust g

  $\Rightarrow$ usually *no closed form for MLE and CMLE*

Multiple features and Naive Bayes

- Predict label Y from features $X_1, \ldots, X_m$

$$
\begin{aligned}
P(Y|X_1, \ldots, X_m) &\propto P(Y) \prod_{j=1}^{m} P(X_j | Y, X_1, \ldots, X_{j-1}) \\
&\approx P(Y) \prod_{j=1}^{m} P(X_j | Y)
\end{aligned}
$$

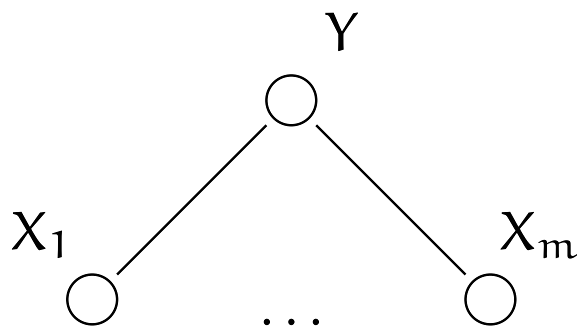


- Naive Bayes estimate is MLE $\widehat{\theta} = \arg\max_{\theta} P(x_1, \ldots, x_n, y)$
  - Trivial to compute (relative frequency)
  - May be poor if X_j aren't really conditionally independent

Multiple features and MaxEnt

- Predict label Y from features $X_1, \ldots, X_m$

$$P(Y|X_1, \ldots, X_m) \;\propto\; \prod_{j=1}^{m} g_j(X_j, Y)$$



- MaxEnt estimate is CMLE $\hat{\hat{\theta}} = \arg\max_\theta P(y|x_1, \ldots, x_m)$
  - Makes *no assumptions* about $P(X)$
  - Difficult to compute (iterative numerical optimization)

Sequence labeling

- Predict labels $\mathsf{Y}_1, \ldots, \mathsf{Y}_m$ given features $\mathsf{X}_1, \ldots, \mathsf{X}_m$
- Example: Parts of speech

$$\begin{array}{lclllll} \mathsf{Y} & = & \text{DT} & \text{JJ} & \text{NN} & \text{VBS} & \text{JJR} \\ \mathsf{X} & = & \text{the} & \text{big} & \text{dog} & \text{barks} & \text{loudly} \end{array}$$

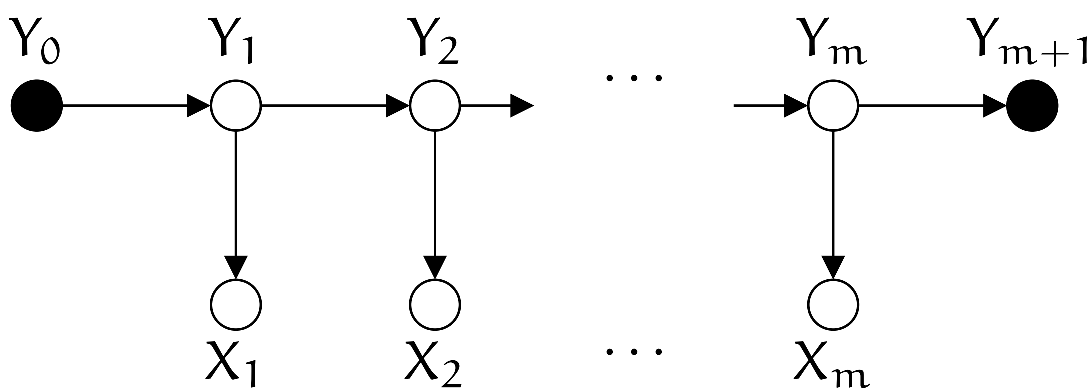
- Example: Named entities

$$\begin{array}{lclllll} \mathsf{Y} & = & \text{[NP} & \text{NP} & \text{NP]} & - & - \\ \mathsf{X} & = & \text{the} & \text{big} & \text{dog} & \text{barks} & \text{loudly} \end{array}$$

- Example: $\mathsf{X}_1, \ldots, \mathsf{X}_m$ are image regions, each X_j is labeled Y_j

Hidden Markov Models

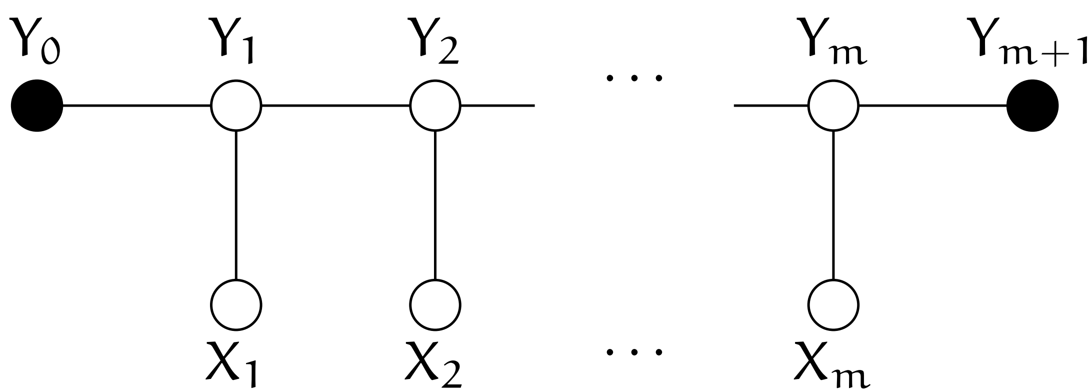
$$P(X,Y) = \left(\prod_{j=1}^{m} P(Y_j|Y_{j-1})P(X_j|Y_j)\right) P(Y_m, \mathrm{stop})$$



- Usually assume *time invariance* or *stationarity* i.e., $P(Y_j|Y_{j-1})$ and $P(X_j|Y_j)$ do not depend on j
- HMMs are Naive Bayes models with compound labels Y
- Estimator is MLE $\widehat{\theta} = \arg\max_\theta P_\theta(x, y)$

Conditional Random Fields

$$P(Y|X) = \frac{1}{Z(x)} \left(\prod_{j=1}^{m} f(Y_j, Y_{j-1}) g(X_j, Y_j) \right) f(Y_m, \text{stop})$$



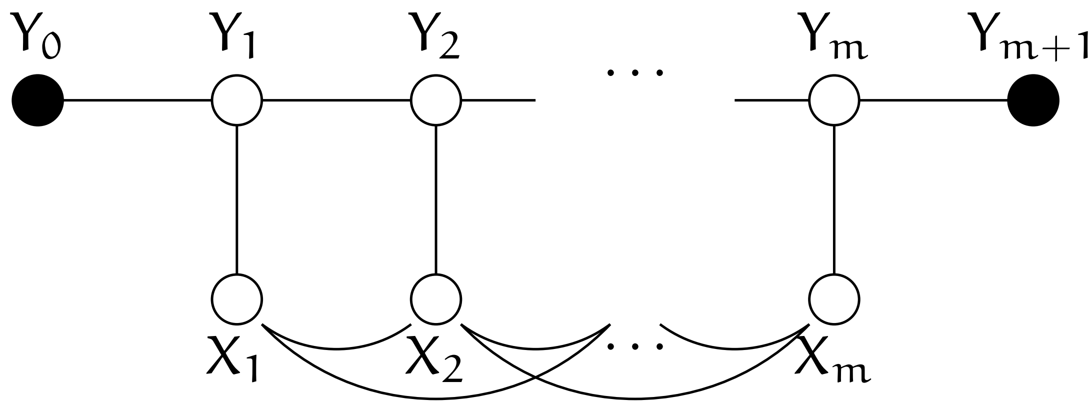
- *time invariance* or *stationarity*, i.e., f and g don't depend on j
- CRFs are MaxEnt models with compound labels Y
- Estimator is CMLE $\hat{\hat{\theta}} = \arg\max_\theta P_\theta(y|x)$

Decoding and Estimation

- HMMs and CRFs have *same complexity of decoding* i.e., computing $y^\star(x) = \arg\max_y P(y|x)$
  - dynamic programming algorithm (Viterbi algorithm)
- Estimating a HMM from labeled data (x, y) is *trivial*
  - HMMs are Bayes nets $\Rightarrow$ MLE is relative frequency
- Estimating a CRF from labeled data (x, y) is *difficult*
  - Usually *no closed form* for partition function $Z(x)$
  - Use *iterative numerical optimization procedures* (e.g., Conjugate Gradient, Limited Memory Variable Metric) to maximize $P_\theta(y|x)$

When are CRFs better than HMMs?

- When HMM independence assumptions are wrong, i.e., there are dependences between X_j not described in model

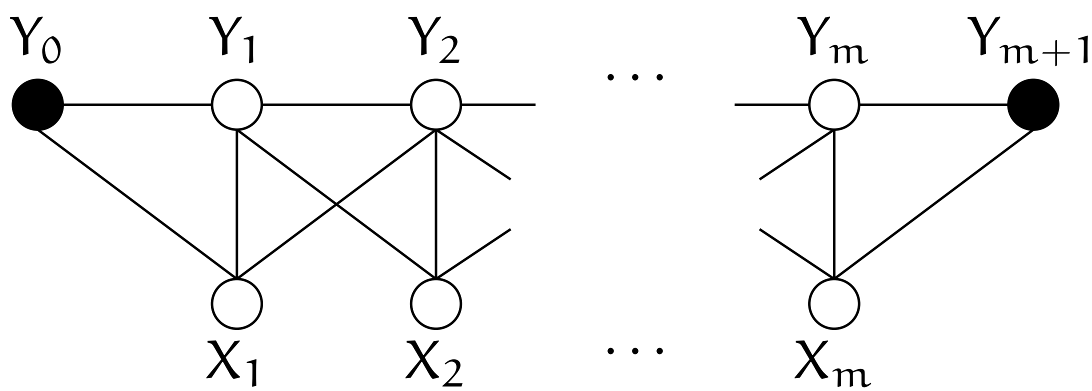


- HMM uses MLE $\Rightarrow$ models joint $P(X, Y) = P(X)P(Y|X)$
- CRF uses CMLE $\Rightarrow$ models conditional distribution $P(Y|X)$
- Because CRF uses CMLE, it makes no assumptions about $P(X)$
- *If* $P(X)$ *isn't modeled well by HMM, don't use HMM!*

Overlapping features

- Sometimes label Y_j depends on X_{j-1} and X_{j+1} as well as X_j

$$P(Y|X) = \frac{1}{Z(x)} \left(\prod_{j=1}^{m} f(X_j, Y_j, Y_{j-1}) g(X_j, Y_j, Y_{j+1}) \right)$$



- Most people think this would be difficult to do in a HMM

Summary

- HMMs and CRFs both associate a *sequence* of labels $(Y_1, \ldots, Y_m)$ to items $(X_1, \ldots, X_m)$
- HMMs are Bayes nets and estimated by MLE
- CRFs are MRFs and estimated by CMLE
- HMMs assume that X_j are *conditionally independent*
- CRFs do not assume that the X_j are conditionally independent
- The Viterbi algorithm computes $y^\star(x)$ for both HMMs and CRFs
- HMMs are trivial to estimate
- CRFs are difficult to estimate
- It is easier to add new features to a CRF
- There is no EM version of CRF

HMM with longer range dependencies