Intelligent Systems (AI-2)

Computer Science cpsc422, Lecture 18

Oct, 20, 2017

Slide Sources
Raymond J. Mooney University of Texas at Austin

D. Koller, Stanford CS - Probabilistic Graphical Models

Lecture Overview

Probabilistic Graphical models

- Recap Markov Networks
- Recap one application
- Inference in Markov Networks (Exact and Approx.)
- Conditional Random Fields

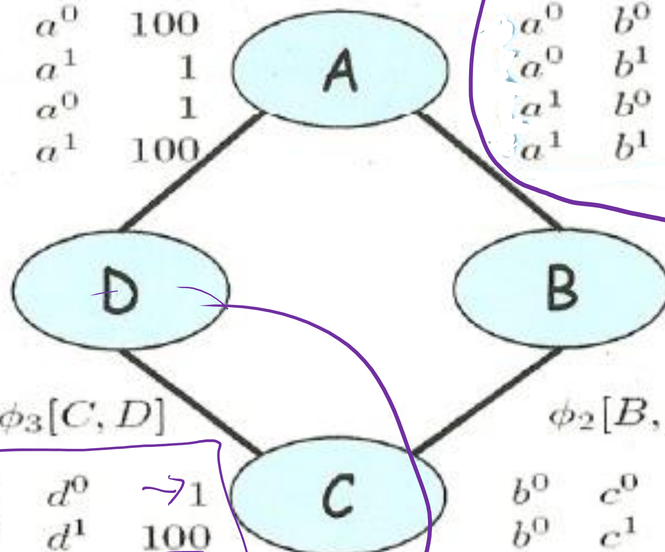
Parameterization of Markov Networks

$\phi_4[D, A]$

d^0	a^0	100
d^0	a^1	1
d^1	a^0	1
d^1	a^1	100

$\phi_1[A, B]$

a^0	b^0	30
a^0	b^1	5
a^1	b^0	1
a^1	b^1	10



$\phi_3[C, D]$

c^0	d^0	1
c^0	d^1	100
c^1	d^0	100
c^1	d^1	1

$\phi_2[B, C]$

b^0	c^0	100
b^0	c^1	1
b^1	c^0	1
b^1	c^1	100

X set of random vars: A factor is

$\Phi(Val(X)) \to \mathbb{R}$

Factors define the local interactions (like CPTs in Bnets)

What about the global model? What do you do with Bnets?

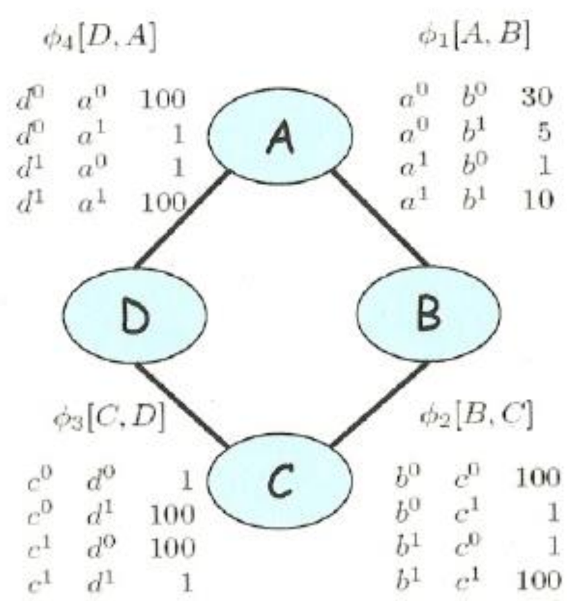
How do we combine local models?

As in BNets by multiplying them!

$$\tilde{P}(A,B,C,D) = \phi_1(A,B) \times \phi_2(B,C) \times \phi_3(C,D) \times \phi_4(A,D)$$

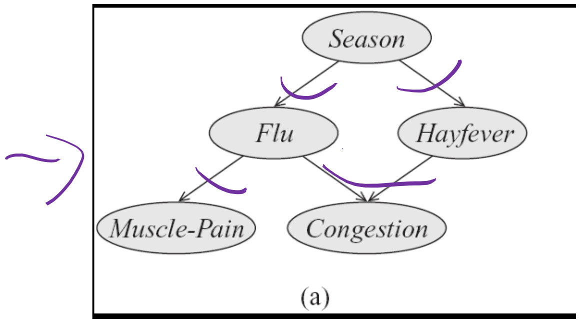
$$P(A,B,C,D) = \frac{1}{Z}\tilde{P}(A,B,C,D)$$

Assignment				Unnormalized	Normalized
a^0	b^0	c^0	d^0	300000	.04
a^0	b^0	c^0	d^1	300000	.04
a^0	b^0	c^1	d^0	300000	.04
a^0	b^0	c^1	d^1	30	4.1×10^{-6}
a^0	b^1	c^0	d^0	500	
a^0	b^1	c^0	d^1	500	
a^0	b^1	c^1	d^0	5000000	.69
a^0	b^1	c^1	d^1	500	
a^1	b^0	c^0	d^0	100	
a^1	b^0	c^0	d^1	1000000	
a^1	b^0	c^1	d^0	100	
a^1	b^0	c^1	d^1	100	
a^1	b^1	c^0	d^0	10	
a^1	b^1	c^0	d^1	100000	
a^1	b^1	c^1	d^0	100000	
a^1	b^1	c^1	d^1	100000	

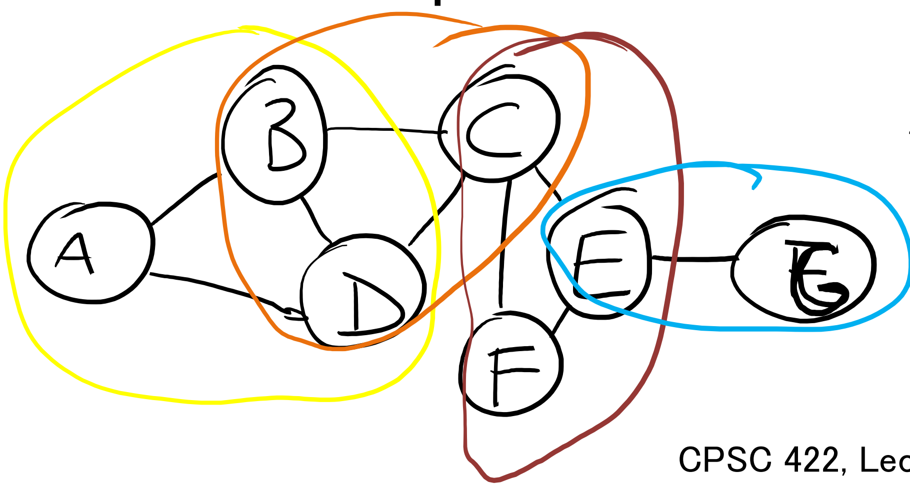


Step Back…. From structure to factors/potentials

In a Bnet the joint is factorized….



In a Markov Network you have one factor for each maximal clique



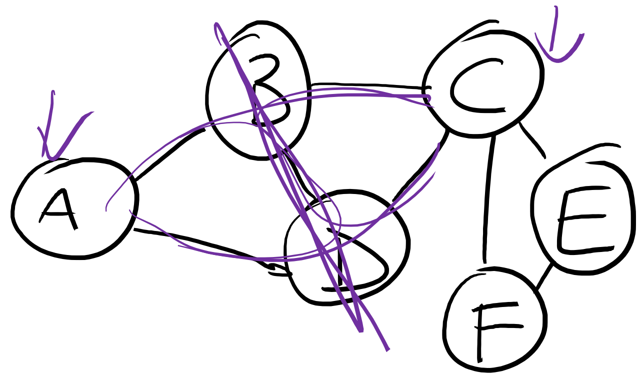
$\Phi_1(A\ B\ D)$

$\Phi_2(B\ D\ C)$ $\quad$ $\Phi_4(E\ G)$

$\Phi_3(C\ E\ F)$

General definitions

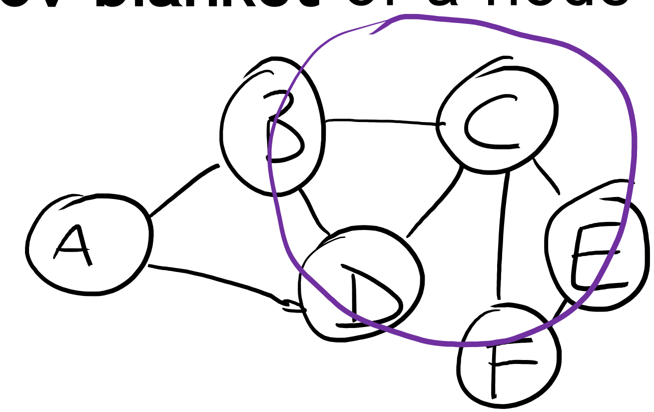
Two nodes in a Markov network are **independent** if and only if every path between them is cut off by evidence



eg for A C

So the **markov blanket** of a node is…?

eg for C



Lecture Overview

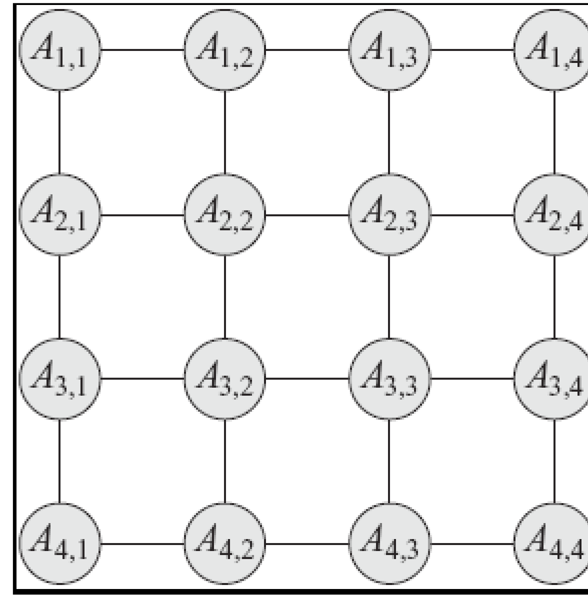
Probabilistic Graphical models

- Recap Markov Networks
- Applications of Markov Networks
- Inference in Markov Networks (Exact and Approx.)
- Conditional Random Fields

Markov Networks Applications (1): Computer Vision

Called **Markov Random Fields**

- Stereo Reconstruction
- Image Segmentation
- Object recognition



Typically **pairwise MRF**

- Each *vars* correspond to a *pixel* (or *superpixel*)
- Edges (factors) correspond to interactions between adjacent pixels in the image
  - E.g., in segmentation: from generically penalize discontinuities, to road under car

Image segmentation

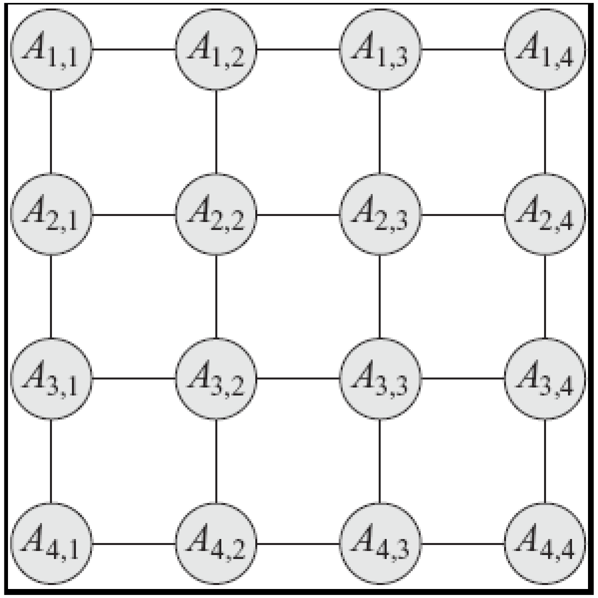
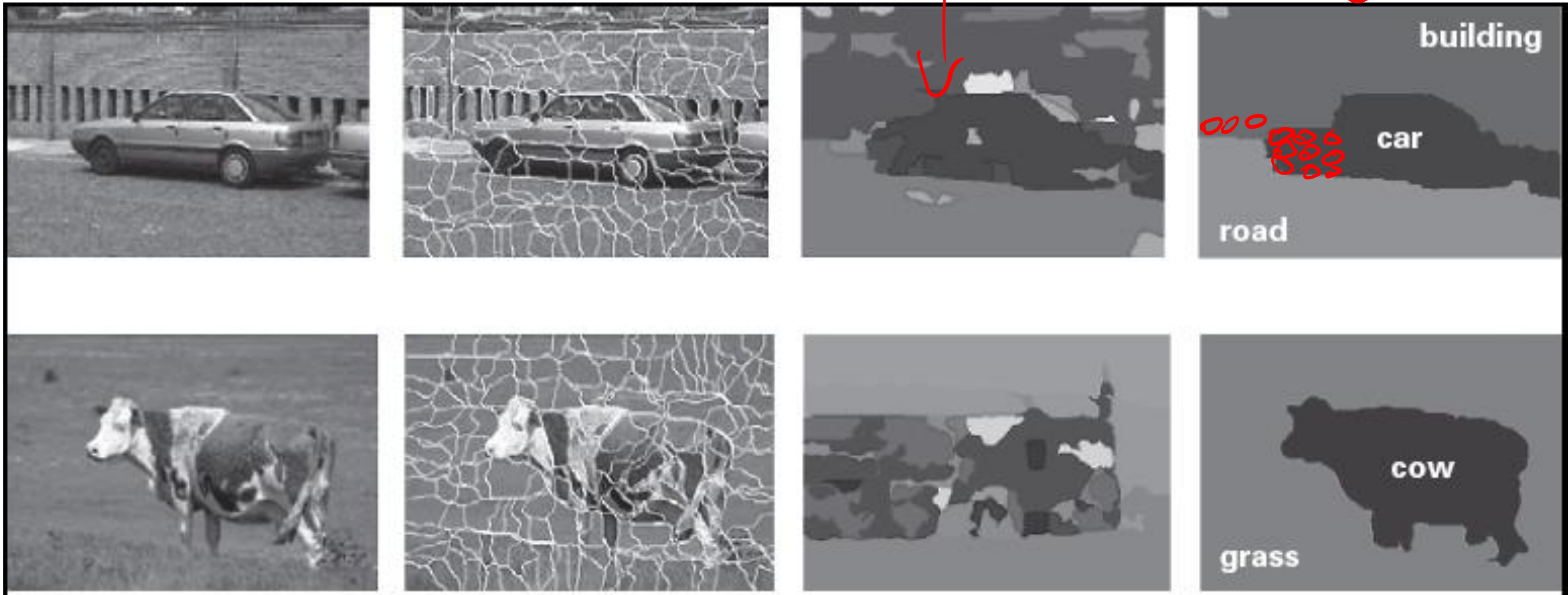


Image segmentation



(a) (b) (c) (d)

classifying each superpixel independently

with a Markov Random Field !

Lecture Overview

<u>**Probabilistic Graphical models**</u>

- Recap Markov Networks
- Applications of Markov Networks
- **Inference in Markov Networks (Exact and Approx.)**
- **Conditional Random Fields**

Variable elimination algorithm for Bnets

Given a network for $P(Z, Y_1, \cdots, Y_j\ Z_1, \cdots, Z_i)$, :

To compute $P(Z| Y_1=v_1, \cdots, Y_j=v_j)$:

1. Construct a factor for each conditional probability.
2. Set the observed variables to their observed values.
3. Given an elimination ordering, simplify/decompose sum of products
4. Perform products and sum out Z_i
5. Multiply the remaining factors Z
6. Normalize: divide the resulting factor $f(Z)$ by $\sum_Z f(Z)$.

Variable elimination algorithm for Markov Networks…..

same!

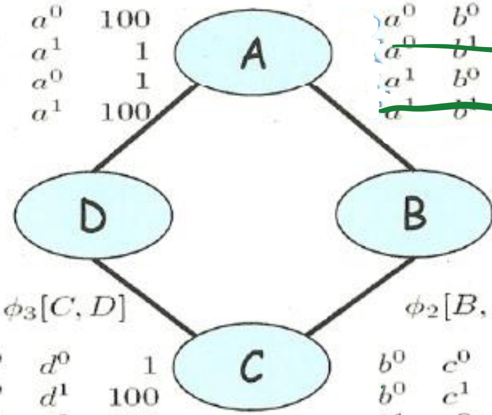
Variable Elimination on MN: Example

$\phi_4[D, A]$

d^0	a^0	100
d^0	a^1	1
d^1	a^0	1
d^1	a^1	100

$\phi_1[A, B]$

a^0	b^0	30
~~a^0~~	~~b^1~~	~~5~~
a^1	b^0	1
~~a^1~~	~~b^1~~	~~10~~



$\phi_3[C, D]$

c^0	d^0	1
c^0	d^1	100
c^1	d^0	100
c^1	d^1	1

$\phi_2[B, C]$

b^0	c^0	100
b^0	c^1	1
~~b^1~~	~~c^0~~	~~1~~
~~b^1~~	~~c^1~~	~~100~~

Example compute

$P(D \mid b^0)$ — Z — $B = Y_1$

Set observed vars

Elimination ordering: A C

$$\alpha \sum_C \Phi_3 \Phi \sum_A \Phi_1 \Phi_4$$

Now it is just a matter of multiplying factors and summing out vars

Normalize at the end!

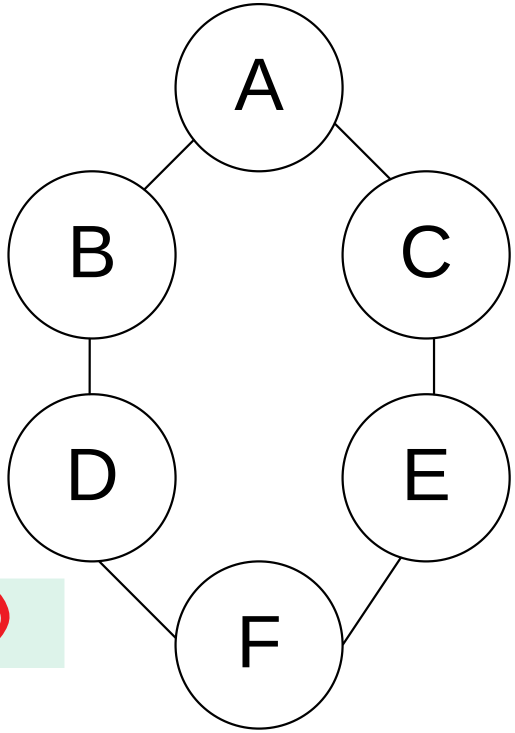
Gibbs sampling for Markov Networks

Example: P(D | C=0)

Note: never change evidence!

Resample non-evidence variables in a pre-defined order or a random order

Suppose we begin with A



What do we need to sample?

A. P(A | B=0)

B. P(A | B=0, C=0)

C. P(B=0, C=0**| A)**

A	B	C	D	E	F
1	0	0	1	1	0

Gibbs sampling MN: what to sample

For Bnets $P(x_i'|mb(X_i)) = \alpha P(x_i'|parents(X_i))\Pi_{Z_j \in Children(X_i)} P(z_j|parents(Z_j))$

For Markov Networks just the product of the factors (normalized)

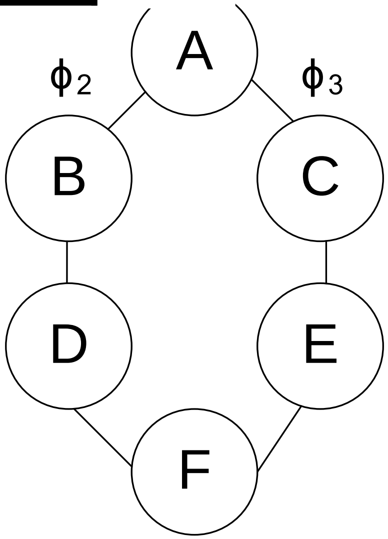
Resample probability distribution of P(A|BC)

B=0 ; C=0

	A=1	A=0
C=1	~~1~~	~~2~~
C=0	3	4

	A=1	A=0
~~B=1~~	~~1~~	~~5~~
B=0	4.3	0.2

A	B	C	D	E	F
1	0	0	1	1	0
?	0	0	1	1	0



$$\Phi_2 \times \Phi_3 =$$

A=1	A=0
12.9	0.8

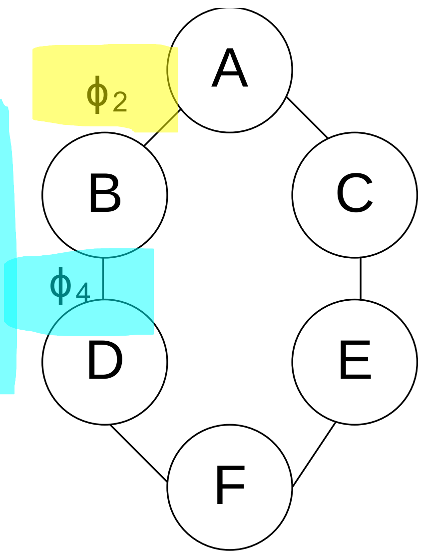
A=1	A=0
0.95	0.05

Example: Gibbs sampling

Resample probability distribution of B given A D

	A=1	A=0
B=1	1	5
B=0	4.3	0.2

A	B	C	D	E	F
1	0	0	1	1	0
1	0	0	1	1	0
1	?	0	1	1	0



	D=1	D=0
B=1	1	2
B=0	2	1

$\phi_2 \times \phi_4 =$

B=1	B=0
1	??

B=1	B=0
0.11	0.89

A. 10

B. 0.4

C. 8.6

iclicker

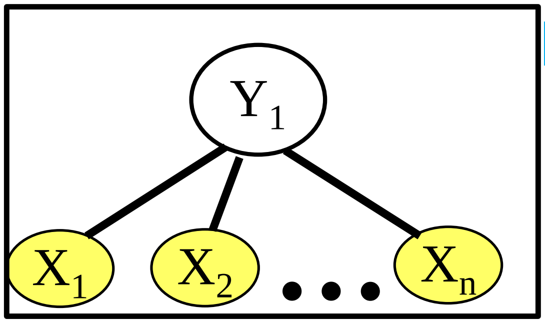
Lecture Overview

Probabilistic Graphical models

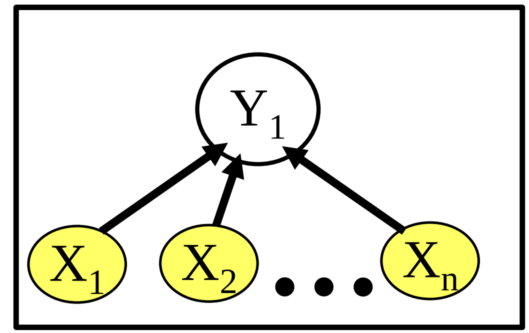
- Recap Markov Networks
- Applications of Markov Networks
- Inference in Markov Networks (Exact and Approx.)
- **Conditional Random Fields**

We want to model $P(Y_1 | X_1 .. X_n)$

… where all the X_i are always observed



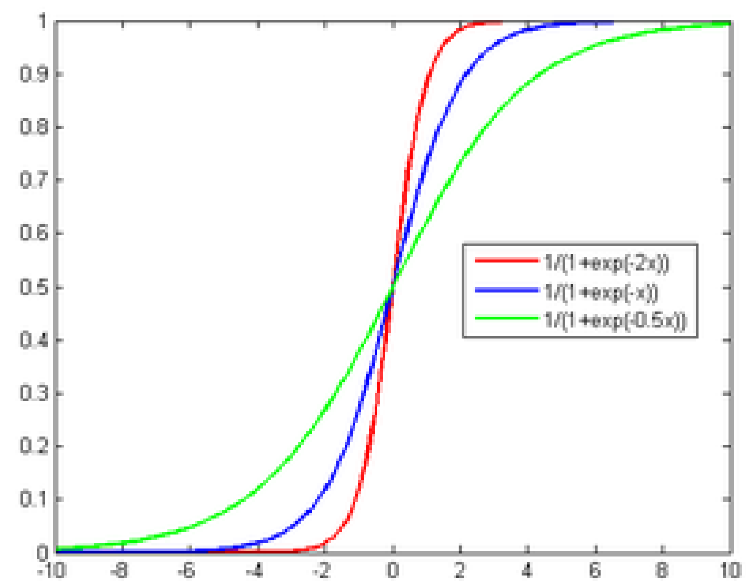
MN



BN

- Which model is simpler, MN or BN?

- Naturally aggregates the influence of different parents



Conditional Random Fields (CRFs)

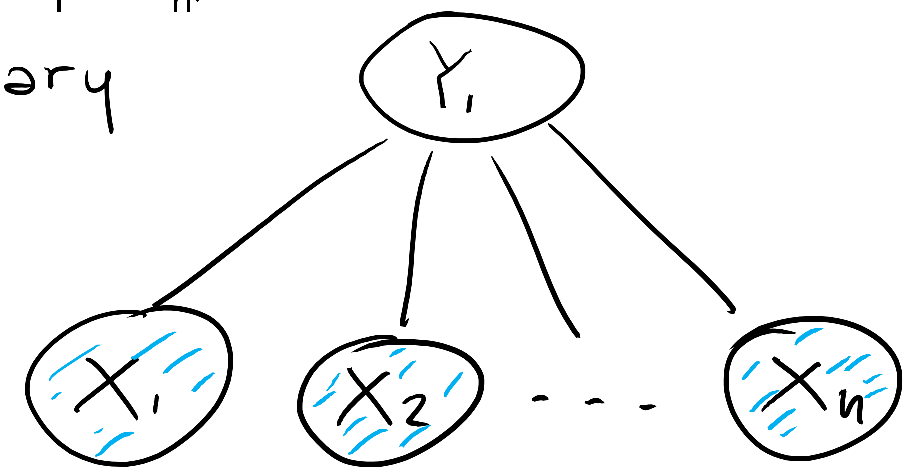
- Model $P(Y_1 .. Y_k \mid X_1 .. X_n)$
- Special case of Markov Networks where all the X_i are always observed

- Simple case $P(Y_1 \mid X_1 \cdots X_n)$

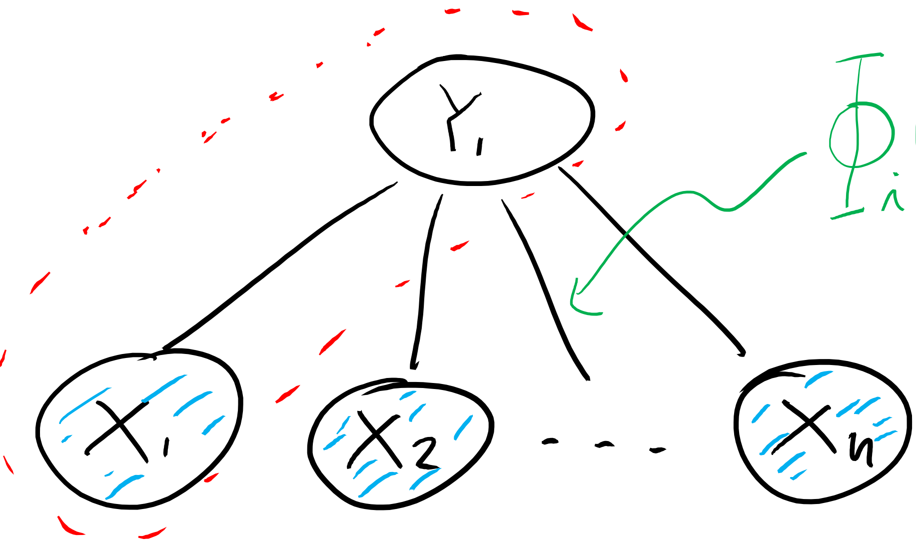
all vars are binary

$Y_1 = \{0, 1\}$

$\forall i \; X_i = \{0, 1\}$



What are the Parameters?



$\Phi_i(X_i, Y_1) = \exp\{w_i * \mathbb{1}\{X_i = 1, Y_1 = 1\}\}$

one such factor for each clique

also $\Phi_0(Y_1) = \exp\{w_0 \mathbb{1}\{Y = 1\}\}$

Example $w_2 = 1.5$ $\Phi_2(X_2, Y_1)$

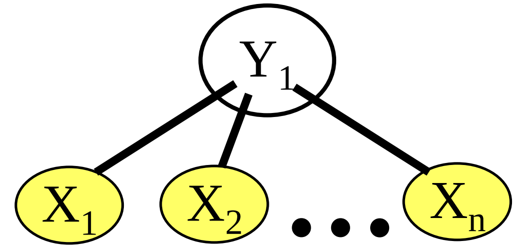
X_2	Y_1	Φ_2
1	1	$e^{1.5}$
0	1	1
1	0	1
0	0	1

Example $w_0 = .4$

Y_1	Φ_0
0	1
1	$e^{.4}$

Let's derive the probabilities we need

$$\phi_i(X_i, Y_1) = \exp\{w_i * \mathbb{1}\{X_i = 1, Y_1 = 1\}\}$$



$$\phi_0(Y_1) = \exp\{w_0 * \mathbb{1}\{Y_1 = 1\}\}$$

$$\tilde{P}(Y_1 = 1, x_1, x_2, \ldots, x_n) = \Phi_0(Y_1) * \prod_{i=1}^{n} \Phi_i(x_i, Y_1)$$

A. $e^{\sum_{1}^{n} w_i}$

B. $e^{w_0 + \sum_{1}^{n} w_i * x_i}$

D $e^{w_0 + \sum_{1}^{n} w_i}$

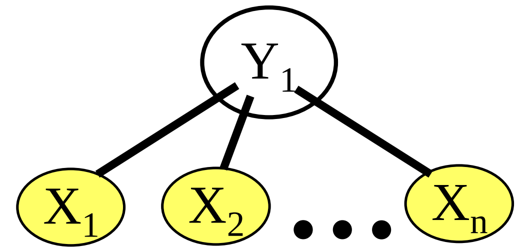
C. $e^{w_0 + \sum_{1}^{n} x_i}$

iclicker

Let's derive the probabilities we need

$$\phi_i(X_i, Y_1) = \exp\{w_i \mathbb{1}\{X_i = 1, Y_1 = 1\}\}$$

$$\phi_0(Y_1) = \exp\{w_0 \mathbb{1}\{Y_1 = 1\}\}$$



$$\tilde{P}(Y_1 = 1, x_1, x_2, \ldots, x_n) = \Phi_0(Y_1) * \prod_{i=1}^{n} \Phi_i(x_i, Y_1)$$

example

$$P(Y_1 = 1, X_1 = 0, X_2 = 1, X_3 = 1)$$

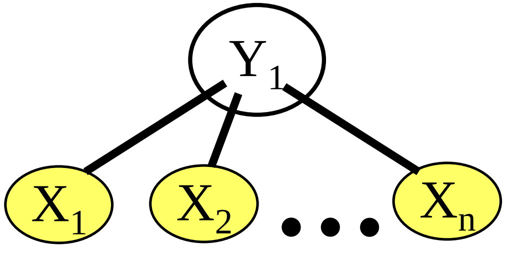
$$e^{w_0 * 1} * e^{w_1 * 0} * e^{w_2 * 1} * e^{w_3 * 1}$$

$$e^{w_0} * e^{w_1 * x_1} * e^{w_2 * x_2} * e^{w_3 * x_3} =$$

$$= \boxed{e^{w_0 + \sum w_i x_i}}$$

Let's derive the probabilities we need

$$\phi_i(X_i, Y_1) = \exp\{w_i \mathbb{1}\{X_i = 1, Y_1 = 1\}\}$$



$$\phi_0(Y_1) = \exp\{w_0 \mathbb{1}\{Y_1 = 1\}\}$$

$$\tilde{P}(Y_1 = 0, x_1, x_2, \ldots, x_n) = \Phi_0(Y_1) * \prod_{i=1}^{n} \Phi_i(x_i, Y_1)$$

A. 1

B. e^{w_0}

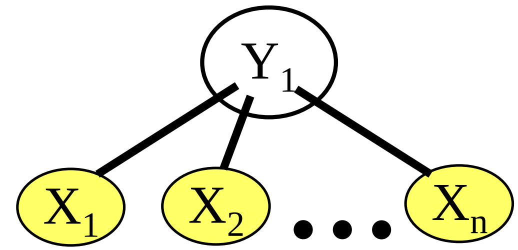
C. 0

D. $e^{\sum_{i}^{n} w_i}$

Let's derive the probabilities we need

a $\tilde{P}(Y_1 = 1, x_1, \ldots, x_n) = \exp(w_0 + \sum_{i=1}^{n} w_i x_i)$

b $\tilde{P}(Y_1 = 0, x_1, \ldots, x_n) = 1$



$$P(Y_1 = 1 \mid x_1, \ldots, x_n) = \frac{\tilde{P}(Y_1 = 1, x_1, \ldots, x_n)}{P(x_1, \ldots, x_n)}$$

sum of a and b

$$= \frac{\exp(w_0 + \sum w_i x_i)}{1 + \exp(w_0 + \sum w_i x_i)}$$

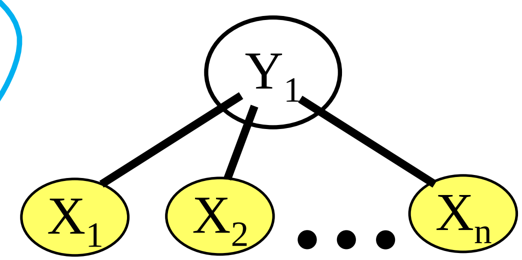
z

sigmoid function $\frac{e^{z}}{1 + e^{z}}$ or $\frac{1}{e^{-z} + 1}$

Let's derive the probabilities we need

a $\tilde{P}(Y_1 = 1, x_1, \ldots, x_n) = \exp\Big(\underbrace{w_0 + \sum_{i=1}^{n} w_i x_i}_{z}\Big)$

b $\tilde{P}(Y_1 = 0, x_1, \ldots, x_n) = 1$

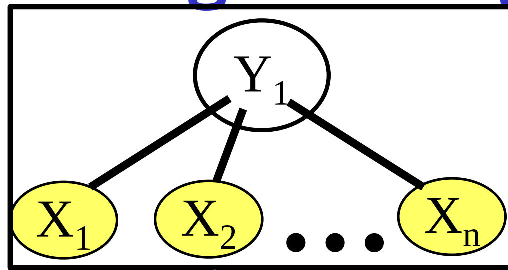


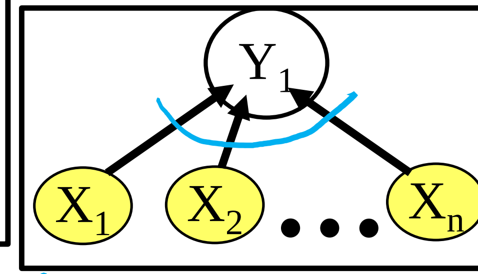
$$P(Y_1 = 1 \mid x_1, \ldots, x_n) = \frac{\tilde{P}(Y_1 = 1, x_1, \ldots, x_n)}{P(x_1, \ldots, x_n)}$$

$\leftarrow$ sum of a and b

$$= \boxed{\frac{e^{z}}{1 + e^{z}}} \cdot \frac{e^{-z}}{e^{-z}} \qquad \boxed{\frac{1}{e^{-z} + 1}}$$

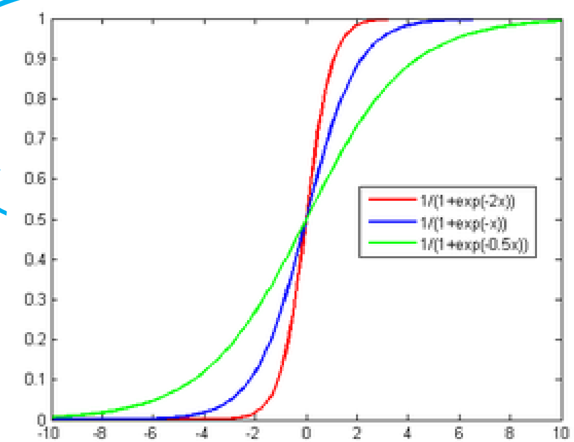
Sigmoid Function used in Logistic Regression

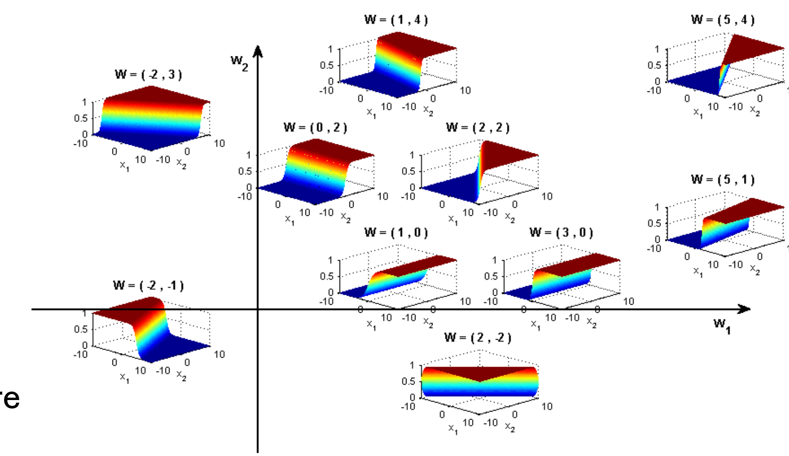




- Great practical interest
- Number of param w_i is linear instead of exponential in the number of parents
- Natural model for many real-world applications
- Naturally aggregates the influence of different parents

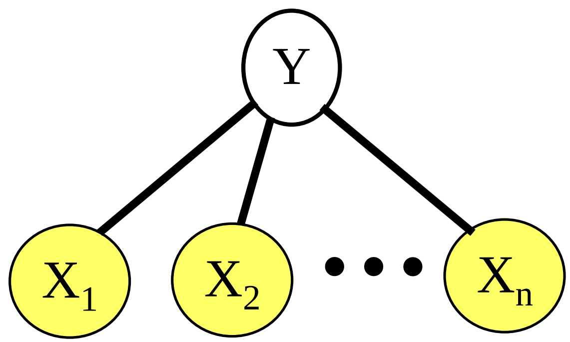
$\frac{1}{1+e^{-x}}$





Logistic Regression as a Markov Net (CRF)

Logistic regression is a simple Markov Net (a CRF)
aka **naïve markov model**



- But only models the **conditional distribution**, P(*Y* | ***X***) and not the full joint P(*X*, *Y*)

Learning Goals for today's class

You can:

- Perform Exact and Approx. Inference in Markov Networks
- Describe a few applications of Markov Networks
- Describe a natural parameterization for a Naïve Markov model (which is a simple CRF)
- Derive how P(Y|X) can be computed for a Naïve Markov model
- Explain the discriminative vs. generative distinction and its implications

Next class Mon Linear-chain CRFs

To Do Revise generative temporal models (HMM)

Midterm, Wed, Oct 25,
we will start at noon sharp

How to prepare….

- Go to **Office Hours** (extra one Mon 3-4 my office)
- **Learning Goals** (look at the end of the slides for each lecture – complete list has been posted)
- Revise all the **clicker questions** and **practice exercises**
- **More practice material** has been posted
- Check questions and answers on Piazza