Natural Language Processing (CSE 517): Graphical Models

Noah Smith

University of Washington
nasmith@cs.washington.edu

February 8–10, 2016

Notation

Let $\boldsymbol{V} = \langle V_1, V_2, \ldots, V_\ell \rangle$ be a collection of random variables (not necessarily a sequence).

$\mathrm{Val}(V)$ will denote the values of a r.v. V.

$\boldsymbol{V}_I$ denotes a subset of the r.v.s $\boldsymbol{V}$ with indices $i \in I$.

$\boldsymbol{V}_{\neg I} = \boldsymbol{V} \setminus \boldsymbol{V}_I$

Recall:

- $p(\boldsymbol{V}) = \prod_{i=1}^{\ell} p(V_i \mid V_1, \ldots, V_{i-1})$ (always true, for any ordering)
- $p(\boldsymbol{V}_I, \boldsymbol{V}_J \mid \boldsymbol{V}_K) = p(\boldsymbol{V}_I \mid \boldsymbol{V}_K) \cdot p(\boldsymbol{V}_J \mid \boldsymbol{V}_K)$ if and only if $\boldsymbol{V}_I \perp \boldsymbol{V}_J \mid \boldsymbol{V}_K$ (conditional independence)
- $p(\boldsymbol{V}_I = \boldsymbol{v}_I) = \sum_{\boldsymbol{v}_{\neg I} \in \mathrm{Val}(\boldsymbol{V}_{\neg I})} p(\boldsymbol{V}_I = \boldsymbol{v}_I, \boldsymbol{V}_{\neg I} = \boldsymbol{v}_{\neg I})$ (marginalization)

Factor Graphs

Two kinds of vertices:

- Random variables (denoted by circles, "V_i")
- Factors (denoted by squares, "f_j")

The graph is *bipartite*; every edge connects some variable to some factor. Let $I_j \subseteq \{1, \ldots, \ell\}$ be the set of variables f_j is connected to.

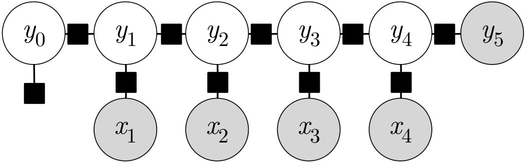
Factor f_j defines a map $\mathrm{Val}(\boldsymbol{V}_{I_j}) \to \mathbb{R}_{\geq 0}$.

The graph and factors define a probability distribution:

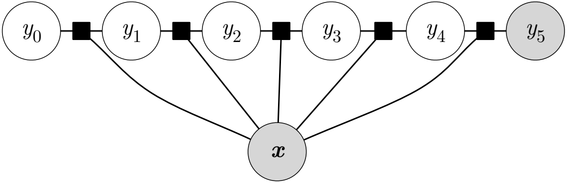
$$p(\boldsymbol{V} = \boldsymbol{v}) \propto \prod_j f_j(\boldsymbol{v}_{I_j})$$

Factor Graphs We've Seen Before

Hidden Markov model:



General first-order sequence model:



Two Kinds of Factors

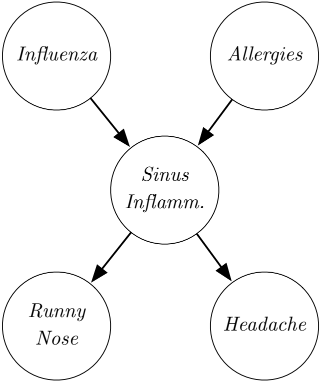
Conditional probability tables. E.g., if $I_j = \{1, 2, 3\}$:

$$f_j(v_1, v_2, v_3) = p(V_3 = v_3 \mid V_1 = v_1, V_2 = v_2)$$

Lead to **Bayesian networks** (with some constraints).

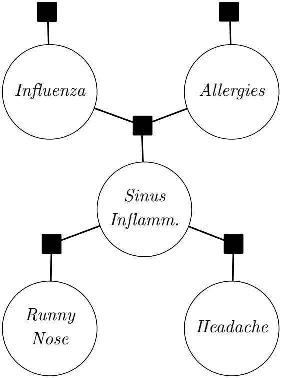
Potential functions (arbitrary nonnegative values).
Lead to **Markov random fields** (a.k.a. Markov networks).

Yucky Bayesian Network



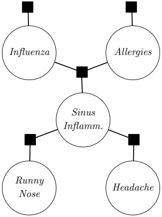
Sinus inflammation is caused by flu, but also by allergies.
Runny nose and headache are both caused by sinus inflammation.

Yucky Factor Graph



~~Sinus inflammation is caused by flu, but also by allergies.~~
~~Runny nose and headache are both caused by sinus inflammation.~~

Yucky Factor Graph



I	f_I
0	
1	

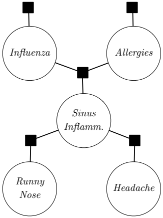
I	f_A
0	
1	

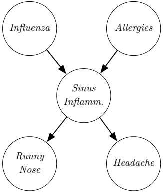
S	I	A	$f_{S,I,A}$
0	0	0	
0	0	1	
0	1	0	
0	1	1	
1	0	0	
1	0	1	
1	1	0	
1	1	1	

R	S	$f_{R,S}$
0	0	
0	1	
1	0	
1	1	

H	S	$f_{H,S}$
0	0	
0	1	
1	0	
1	1	

Yucky Factor Graph





I	f_I
0	
1	

I	f_A
0	
1	

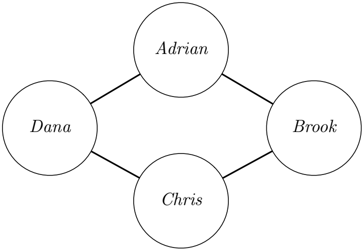
S	I	A	$f_{S,I,A}$
0	0	0	
0	0	1	
0	1	0	
0	1	1	
1	0	0	
1	0	1	
1	1	0	
1	1	1	

R	S	$f_{R,S}$
0	0	
0	1	
1	0	
1	1	

H	S	$f_{H,S}$
0	0	
0	1	
1	0	
1	1	

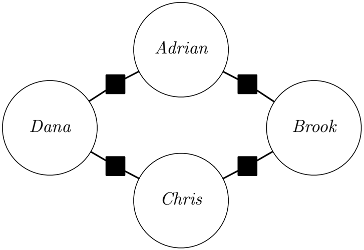
$$\begin{aligned} p(i,a,s,r,h) &= f_I(i) \cdot f_A(a) \cdot f_{S,I,A}(s,i,a) \cdot f_{R,S}(r,s) \cdot f_{H,S}(h,s) \\ &= p(i) \cdot p(a) \cdot p(s \mid i,a) \cdot p(r \mid s) \cdot p(h \mid s) \end{aligned}$$

Naughty Markov Random Field



Independencies: $A \perp C \mid B, D$; $\quad B \perp D \mid A, C$; $\quad \neg A \perp C$; $\quad \neg B \perp D$

Naughty Factor Graph



A	B	$f_{A,B}$
0	0	
0	1	
1	0	
1	1	

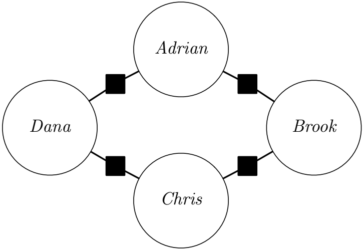
B	C	$f_{B,C}$
0	0	
0	1	
1	0	
1	1	

C	D	$f_{C,D}$
0	0	
0	1	
1	0	
1	1	

D	A	$f_{D,A}$
0	0	
0	1	
1	0	
1	1	

$$p(a, b, c, d) = \frac{f_{A,B}(a, b) \cdot f_{B,C}(b, c) \cdot f_{C,D}(c, d) \cdot f_{D,A}(d, a)}{\sum_{a' \in \mathrm{Val}(A)} \sum_{b' \in \mathrm{Val}(B)} \sum_{c' \in \mathrm{Val}(C)} \sum_{d' \in \mathrm{Val}(D)} f_{A,B}(a', b') \cdot f_{B,C}(b', c') \cdot f_{C,D}(c', d') \cdot f_{D,A}(d', a')}$$

Assignment Probabilities: Examples



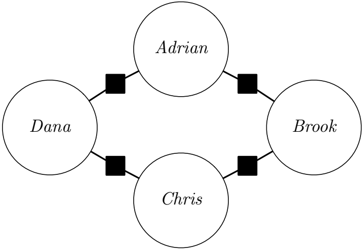
A	B	$f_{A,B}$
0	0	30
0	1	5
1	0	1
1	1	10

B	C	$f_{B,C}$
0	0	100
0	1	1
1	0	1
1	1	100

C	D	$f_{C,D}$
0	0	1
0	1	100
1	0	100
1	1	1

D	A	$f_{D,A}$
0	0	100
0	1	1
1	0	1
1	1	100

Assignment Probabilities: Examples



A	B	$f_{A,B}$
0	0	30
0	1	5
1	0	1
1	1	10

B	C	$f_{B,C}$
0	0	100
0	1	1
1	0	1
1	1	100

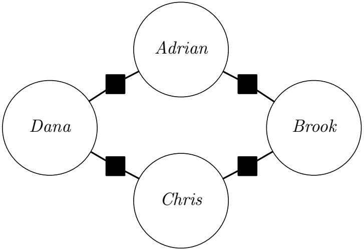
C	D	$f_{C,D}$
0	0	1
0	1	100
1	0	100
1	1	1

D	A	$f_{D,A}$
0	0	100
0	1	1
1	0	1
1	1	100

$$\sum_{a' \in \mathrm{Val}(A)} \sum_{b' \in \mathrm{Val}(B)} \sum_{c' \in \mathrm{Val}(C)} \sum_{d' \in \mathrm{Val}(D)} f_{A,B}(a', b') \cdot f_{B,C}(b', c') \cdot f_{C,D}(c', d') \cdot f_{D,A}(d', a')$$

$$= 7,201,840$$

Assignment Probabilities: Examples



A	B	$f_{A,B}$
0	0	30
0	1	5
1	0	1
1	1	10

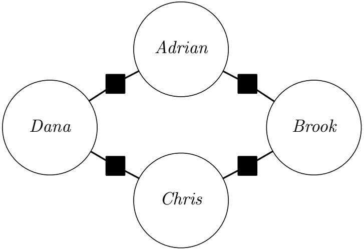
B	C	$f_{B,C}$
0	0	100
0	1	1
1	0	1
1	1	100

C	D	$f_{C,D}$
0	0	1
0	1	100
1	0	100
1	1	1

D	A	$f_{D,A}$
0	0	100
0	1	1
1	0	1
1	1	100

$$p(A=0, B=1, C=1, D=0) = \frac{5,000,000}{7,201,840} \approx 0.69$$

Assignment Probabilities: Examples



A	B	$f_{A,B}$
0	0	30
0	1	5
1	0	1
1	1	10

B	C	$f_{B,C}$
0	0	100
0	1	1
1	0	1
1	1	100

C	D	$f_{C,D}$
0	0	1
0	1	100
1	0	100
1	1	1

D	A	$f_{D,A}$
0	0	100
0	1	1
1	0	1
1	1	100

$$p(A = 1, B = 1, C = 0, D = 0) = \frac{10}{7,201,840} \approx 0.0000014$$

Structure and Independence

Bayesian networks:

- A variable is conditionally independent of its non-descendants given its parents.

Markov networks:

- Conditional independence derived from "Markov blanket" and separation properties.

Local configurations can be used to check *all* conditional independence questions; almost no need to look at the values in the factors!

Independence "Spectrum"

$\prod_{i=1}^{\ell} f_{V_i}(V_i)$	$f_{\boldsymbol{V}}(\boldsymbol{V})$
everything is independent	everything can be interdependent
minimal expressive power	arbitrary expressive power
fewer parameters	more parameters

Operations on Factors: Multiplication

Given two factors $f_{\boldsymbol{U}}$ and $f_{\boldsymbol{V}}$, we can create a new "product" factor such that:

$$f_{\boldsymbol{U} \cup \boldsymbol{V}}(\boldsymbol{u} \cup \boldsymbol{v}) = f_{\boldsymbol{U}}(\boldsymbol{u}) \cdot f_{\boldsymbol{V}}(\boldsymbol{v})$$

for all $\boldsymbol{u} \in \mathrm{Val}(\boldsymbol{U})$ and all $\boldsymbol{v} \in \mathrm{Val}(\boldsymbol{V})$.

A	B	$f_{A,B}$
0	0	30
0	1	5
1	0	1
1	1	10

$\cdot$

B	C	$f_{B,C}$
0	0	100
0	1	1
1	0	1
1	1	100

$=$

A	B	C	$f_{A,B,C}$
0	0	0	3,000
0	0	1	30
0	1	0	5
0	1	1	500
1	0	0	100
1	0	1	1
1	1	0	10
1	1	1	1,000

Operations on Factors: Multiplication

Given two factors $f_{\boldsymbol{U}}$ and $f_{\boldsymbol{V}}$, we can create a new "product" factor such that:

$$f_{\boldsymbol{U} \cup \boldsymbol{V}}(\boldsymbol{u} \cup \boldsymbol{v}) = f_{\boldsymbol{U}}(\boldsymbol{u}) \cdot f_{\boldsymbol{V}}(\boldsymbol{v})$$

for all $\boldsymbol{u} \in \mathrm{Val}(\boldsymbol{U})$ and all $\boldsymbol{v} \in \mathrm{Val}(\boldsymbol{V})$.

A	B	$f_{A,B}$
0	0	30
0	1	5
1	0	1
1	1	10

$\cdot$

B	C	$f_{B,C}$
0	0	100
0	1	1
1	0	1
1	1	100

$=$

A	B	C	$f_{A,B,C}$
0	0	0	3,000
0	0	1	30
0	1	0	5
0	1	1	500
1	0	0	100
1	0	1	1
1	1	0	10
1	1	1	1,000

This might remind you of a **join** operation on a database.

Operations on Factors: Multiplication

Given two factors $f_{\boldsymbol{U}}$ and $f_{\boldsymbol{V}}$, we can create a new "product" factor such that:

$$f_{\boldsymbol{U} \cup \boldsymbol{V}}(\boldsymbol{u} \cup \boldsymbol{v}) = f_{\boldsymbol{U}}(\boldsymbol{u}) \cdot f_{\boldsymbol{V}}(\boldsymbol{v})$$

for all $\boldsymbol{u} \in \mathrm{Val}(\boldsymbol{U})$ and all $\boldsymbol{v} \in \mathrm{Val}(\boldsymbol{V})$.

A	B	$f_{A,B}$
0	0	30
0	1	5
1	0	1
1	1	10

$\cdot$

B	C	$f_{B,C}$
0	0	100
0	1	1
1	0	1
1	1	100

$=$

A	B	C	$f_{A,B,C}$
0	0	0	3,000
0	0	1	30
0	1	0	5
0	1	1	500
1	0	0	100
1	0	1	1
1	1	0	10
1	1	1	1,000

What happens if you multiply out all the factors in a factor graph?

Operations on Factors: Maximization

Given a factor $f_{\boldsymbol{U}}$ and a variable $V \notin \boldsymbol{U}$, we can transform $f_{\boldsymbol{U},V}$ into $f_{\boldsymbol{U}}$ by:

$$f_{\boldsymbol{U}}(\boldsymbol{u}) = \max_{v \in \mathrm{Val}(V)} f_{\boldsymbol{U},V}(\boldsymbol{u}, v)$$

for all $\boldsymbol{u} \in \mathrm{Val}(\boldsymbol{U})$.

A	C	$f_{A,C}$	
0	0	3,000	$B = 0$
0	1	500	$B = 1$
1	0	100	$B = 0$
1	1	1,000	$B = 1$

$$= \max_B$$

A	B	C	$f_{A,B,C}$
0	0	0	3,000
0	0	1	30
0	1	0	5
0	1	1	500
1	0	0	100
1	0	1	1
1	1	0	10
1	1	1	1,000

Operations on Factors: Marginalization

Given a factor $f_{\boldsymbol{U}}$ and a variable $V \notin \boldsymbol{U}$, we can transform $f_{\boldsymbol{U},V}$ into $f_{\boldsymbol{U}}$ by:

$$f_{\boldsymbol{U}}(\boldsymbol{u}) = \sum_{v \in \mathrm{Val}(V)} f_{\boldsymbol{U},V}(\boldsymbol{u}, v)$$

for all $\boldsymbol{u} \in \mathrm{Val}(\boldsymbol{U})$.

A	C	$f_{A,C}$
0	0	3,000 + 5
0	1	30 + 500
1	0	100 + 10
1	1	1 + 1,000

$$= \sum_{B}$$

A	B	C	$f_{A,B,C}$
0	0	0	3,000
0	0	1	30
0	1	0	5
0	1	1	500
1	0	0	100
1	0	1	1
1	1	0	10
1	1	1	1,000

Operations on Factors: Marginalization

Given a factor $f_{\boldsymbol{U}}$ and a variable $V \notin \boldsymbol{U}$, we can transform $f_{\boldsymbol{U},V}$ into $f_{\boldsymbol{U}}$ by:

$$f_{\boldsymbol{U}}(\boldsymbol{u}) = \sum_{v \in \mathrm{Val}(V)} f_{\boldsymbol{U},V}(\boldsymbol{u}, v)$$

for all $\boldsymbol{u} \in \mathrm{Val}(\boldsymbol{U})$.

A	C	$f_{A,C}$
0	0	3,000 + 5
0	1	30 + 500
1	0	100 + 10
1	1	1 + 1,000

$$= \sum_{B}$$

A	B	C	$f_{A,B,C}$
0	0	0	3,000
0	0	1	30
0	1	0	5
0	1	1	500
1	0	0	100
1	0	1	1
1	1	0	10
1	1	1	1,000

If you multiply out all the factors in a factor graph, then sum out each variable, one by one, until none are left, what do you get?

Factors are like numbers.

- Products are commutative: $f_1 \cdot f_2 = f_2 \cdot f_1$

Factors are like numbers.

- Products are commutative: $f_1 \cdot f_2 = f_2 \cdot f_1$
- Products are associative: $(f_1 \cdot f_2) \cdot f_3 = f_1 \cdot (f_2 \cdot f_3)$

Factors are like numbers.

- Products are commutative: $f_1 \cdot f_2 = f_2 \cdot f_1$
- Products are associative: $(f_1 \cdot f_2) \cdot f_3 = f_1 \cdot (f_2 \cdot f_3)$
- Sums are commutative: $\sum_X \sum_Y f = \sum_Y \sum_X f$

Factors are like numbers.

- Products are commutative: $f_1 \cdot f_2 = f_2 \cdot f_1$
- Products are associative: $(f_1 \cdot f_2) \cdot f_3 = f_1 \cdot (f_2 \cdot f_3)$
- Sums are commutative: $\sum_X \sum_Y f = \sum_Y \sum_X f$
- Maximizations are commutative: $\max_X \max_Y f = \max_Y \max_X f$

Factors are like numbers.

- Products are commutative: $f_1 \cdot f_2 = f_2 \cdot f_1$
- Products are associative: $(f_1 \cdot f_2) \cdot f_3 = f_1 \cdot (f_2 \cdot f_3)$
- Sums are commutative: $\sum_X \sum_Y f = \sum_Y \sum_X f$
- Maximizations are commutative: $\max_X \max_Y f = \max_Y \max_X f$
- Multiplication distributes over marginalization and maximization:

$$\sum_X (f_1 \cdot f_2) = f_1 \cdot \sum_X f_2$$

$$\max_X (f_1 \cdot f_2) = f_1 \cdot \max_X f_2$$

(assuming X is not in the scope of f_1).

Inference

Most general definition: “reason about some variables, optionally given values of some others.” Let $\boldsymbol{O}$ be the observed variables and $\boldsymbol{U}$ be the unobserved ones; $\boldsymbol{V} = \boldsymbol{O} \cup \boldsymbol{U}$.

Three inference problems, all given $\boldsymbol{O} = \boldsymbol{o}$ . . .

Inference

Most general definition: "reason about some variables, optionally given values of some others." Let $\boldsymbol{O}$ be the observed variables and $\boldsymbol{U}$ be the unobserved ones; $\boldsymbol{V} = \boldsymbol{O} \cup \boldsymbol{U}$.

Three inference problems, all given $\boldsymbol{O} = \boldsymbol{o}$ ...

- **Marginal inference:** what is the marginal distribution over $\boldsymbol{Q} \subset \boldsymbol{U}$? ($p(\boldsymbol{Q} \mid \boldsymbol{o})$, marginalizing out the rest.)

Inference

Most general definition: "reason about some variables, optionally given values of some others." Let $\boldsymbol{O}$ be the observed variables and $\boldsymbol{U}$ be the unobserved ones; $\boldsymbol{V} = \boldsymbol{O} \cup \boldsymbol{U}$.

Three inference problems, all given $\boldsymbol{O} = \boldsymbol{o}$ ...

- **Marginal inference:** what is the marginal distribution over $\boldsymbol{Q} \subset \boldsymbol{U}$? ($p(\boldsymbol{Q} \mid \boldsymbol{o})$, marginalizing out the rest.)
  - Related: draw samples from that distribution.

Inference

Most general definition: "reason about some variables, optionally given values of some others." Let $\boldsymbol{O}$ be the observed variables and $\boldsymbol{U}$ be the unobserved ones; $\boldsymbol{V} = \boldsymbol{O} \cup \boldsymbol{U}$.

Three inference problems, all given $\boldsymbol{O} = \boldsymbol{o}$ ...

- **Marginal inference:** what is the marginal distribution over $\boldsymbol{Q} \subset \boldsymbol{U}$? ($p(\boldsymbol{Q} \mid \boldsymbol{o})$, marginalizing out the rest.)
  - Related: draw samples from that distribution.
- **Most probable explanation (MPE):** what is the most probable assignment to $\boldsymbol{U}$? ($\text{argmax}_{\boldsymbol{u}}\, p(\boldsymbol{u} \mid \boldsymbol{o})$)

Inference

Most general definition: "reason about some variables, optionally given values of some others." Let $\boldsymbol{O}$ be the observed variables and $\boldsymbol{U}$ be the unobserved ones; $\boldsymbol{V} = \boldsymbol{O} \cup \boldsymbol{U}$.

Three inference problems, all given $\boldsymbol{O} = \boldsymbol{o}$ ...

- **Marginal inference:** what is the marginal distribution over $\boldsymbol{Q} \subset \boldsymbol{U}$? ($p(\boldsymbol{Q} \mid \boldsymbol{o})$, marginalizing out the rest.)
  - Related: draw samples from that distribution.
- **Most probable explanation (MPE):** what is the most probable assignment to $\boldsymbol{U}$? ($\text{argmax}_{\boldsymbol{u}}\, p(\boldsymbol{u} \mid \boldsymbol{o})$)
  - Related: what is the most *dangerous* assignment to $\boldsymbol{U}$?

Inference

Most general definition: "reason about some variables, optionally given values of some others." Let $\boldsymbol{O}$ be the observed variables and $\boldsymbol{U}$ be the unobserved ones; $\boldsymbol{V} = \boldsymbol{O} \cup \boldsymbol{U}$.

Three inference problems, all given $\boldsymbol{O} = \boldsymbol{o}$ ...

- **Marginal inference:** what is the marginal distribution over $\boldsymbol{Q} \subset \boldsymbol{U}$? ($p(\boldsymbol{Q} \mid \boldsymbol{o})$, marginalizing out the rest.)
  - Related: draw samples from that distribution.
- **Most probable explanation (MPE):** what is the most probable assignment to $\boldsymbol{U}$? ($\operatorname{argmax}_{\boldsymbol{u}} p(\boldsymbol{u} \mid \boldsymbol{o})$)
  - Related: what is the most *dangerous* assignment to $\boldsymbol{U}$?
- **Maximum a posteriori (MAP):** what is the most probable assignment to $\boldsymbol{Q} \subset \boldsymbol{U}$? ($\operatorname{argmax}_{\boldsymbol{q}} p(\boldsymbol{q} \mid \boldsymbol{o})$)

Inference

Most general definition: "reason about some variables, optionally given values of some others." Let $\boldsymbol{O}$ be the observed variables and $\boldsymbol{U}$ be the unobserved ones; $\boldsymbol{V} = \boldsymbol{O} \cup \boldsymbol{U}$.

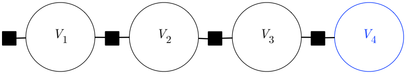
Three inference problems, all given $\boldsymbol{O} = \boldsymbol{o} \ldots$

- **Marginal inference:** what is the marginal distribution over $\boldsymbol{Q} \subset \boldsymbol{U}$? ($p(\boldsymbol{Q} \mid \boldsymbol{o})$, marginalizing out the rest.)
  - Related: draw samples from that distribution.
- **Most probable explanation (MPE):** what is the most probable assignment to $\boldsymbol{U}$? $(\operatorname{argmax}_{\boldsymbol{u}} p(\boldsymbol{u} \mid \boldsymbol{o}))$
  - Related: what is the most *dangerous* assignment to $\boldsymbol{U}$?
- **Maximum a posteriori (MAP):** what is the most probable assignment to $\boldsymbol{Q} \subset \boldsymbol{U}$? $(\operatorname{argmax}_{\boldsymbol{q}} p(\boldsymbol{q} \mid \boldsymbol{o}))$
  - Related: what values of $\boldsymbol{Q}$ have the lowest expected cost?

Marginal Inference

Given a factor graph with variables $\boldsymbol{V}$, find the marginal distribution over some $V_i \in \boldsymbol{V}$, $p(V_i)$.

Simple chain example, focusing on $i = 4$:



V_1	f_{V_1}
0	
1	

V_1	V_2	f_{V_1,V_2}
0	0	
0	1	
1	0	
1	1	

V_2	V_3	f_{V_2,V_3}
0	0	
0	1	
1	0	
1	1	

V_3	V_4	f_{V_3,V_4}
0	0	
0	1	
1	0	
1	1	

Observations

- If we had a single f_{V_4}, we could easily renormalize it to get $p(V_4)$.

Observations

- If we had a single f_{V_4}, we could easily renormalize it to get $p(V_4)$.
- Correct: $f_{V_4} = \sum_{V_1} \sum_{V_2} \sum_{V_3} f_{V_1} \cdot f_{V_1,V_2} \cdot f_{V_2,V_3} \cdot f_{V_3,V_4}$

Observations

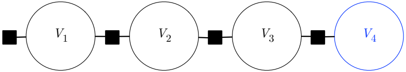
- If we had a single f_{V_4}, we could easily renormalize it to get $p(V_4)$.
- Correct: $f_{V_4} = \sum_{V_1} \sum_{V_2} \sum_{V_3} f_{V_1} \cdot f_{V_1,V_2} \cdot f_{V_2,V_3} \cdot f_{V_3,V_4}$
  - But that multiplied-out factor would have $\prod_i |\mathrm{Val}(V_i)|$ values!

Observations

- If we had a single f_{V_4}, we could easily renormalize it to get $p(V_4)$.
- Correct: $f_{V_4} = \sum_{V_1} \sum_{V_2} \sum_{V_3} f_{V_1} \cdot f_{V_1,V_2} \cdot f_{V_2,V_3} \cdot f_{V_3,V_4}$
  - But that multiplied-out factor would have $\prod_i |\mathrm{Val}(V_i)|$ values!
- Reorganize calculations:

$$\begin{aligned} &\sum_{V_1} \sum_{V_2} \sum_{V_3} f_{V_1} \cdot f_{V_1,V_2} \cdot f_{V_2,V_3} \cdot f_{V_3,V_4} \\ &= \sum_{V_3} f_{V_3,V_4} \cdot \left(\sum_{V_2} f_{V_2,V_3} \cdot \left(\sum_{V_1} f_{V_1,V_2} \cdot f_{V_1} \right) \right) \end{aligned}$$

Marginal Inference



V_1	f_{V_1}
0	
1	

V_1	V_2	f_{V_1,V_2}
0	0	
0	1	
1	0	
1	1	

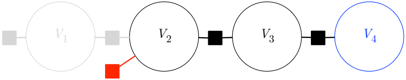
V_2	V_3	f_{V_2,V_3}
0	0	
0	1	
1	0	
1	1	

V_3	V_4	f_{V_3,V_4}
0	0	
0	1	
1	0	
1	1	

$$\sum_{V_1}\sum_{V_2}\sum_{V_3} f_{V_1} \cdot f_{V_1,V_2} \cdot f_{V_2,V_3} \cdot f_{V_3,V_4}$$

$$= \sum_{V_3} f_{V_3,V_4} \cdot \left(\sum_{V_2} f_{V_2,V_3} \cdot \left(\sum_{V_1} f_{V_1,V_2} \cdot f_{V_1}\right)\right)$$

Marginal Inference



V_2	f_{V_2}
0	
1	

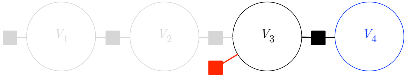
V_2	V_3	f_{V_2,V_3}
0	0	
0	1	
1	0	
1	1	

V_3	V_4	f_{V_3,V_4}
0	0	
0	1	
1	0	
1	1	

$$\sum_{V_1}\sum_{V_2}\sum_{V_3} f_{V_1} \cdot f_{V_1,V_2} \cdot f_{V_2,V_3} \cdot f_{V_3,V_4}$$

$$= \sum_{V_3} f_{V_3,V_4} \cdot \left(\sum_{V_2} f_{V_2,V_3} \cdot f_{V_2} \right)$$

Marginal Inference



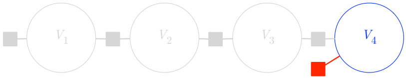
V_3	f_{V_3}
0	
1	

V_3	V_4	f_{V_3,V_4}
0	0	
0	1	
1	0	
1	1	

$$\sum_{V_1}\sum_{V_2}\sum_{V_3} f_{V_1} \cdot f_{V_1,V_2} \cdot f_{V_2,V_3} \cdot f_{V_3,V_4}$$

$$= \sum_{V_3} f_{V_3,V_4} \cdot f_{V_3}$$

Marginal Inference



V_4	f_{V_4}
0	
1	

$$\sum_{V_1}\sum_{V_2}\sum_{V_3} f_{V_1}\cdot f_{V_1,V_2}\cdot f_{V_2,V_3}\cdot f_{V_3,V_4}$$
$$= f_{V_4}$$

Variable Elimination

Given a factor graph with factors $\boldsymbol{f}$, eliminate variable V.

1. Let $\boldsymbol{f}_{elim} \subset \boldsymbol{f}$ be the factors connected to V
2. Let $\boldsymbol{f}_{keep} = \boldsymbol{f} \setminus \boldsymbol{f}_{elim}$ be the rest
3. Let $f_{new} = \sum_{V} \prod_{f \in \boldsymbol{f}_{elim}} f$
4. Return $\boldsymbol{f}_{keep} \cup \{f_{new}\}$

Uses the graph structure to avoid exponential blowup; this is an example of dynamic programming.

Marginal Inference by Variable Elimination (No Evidence)

Given a factor graph with variables $\boldsymbol{V}$ and factors $\boldsymbol{f}$, find the marginal distribution over some $\boldsymbol{V}_{keep} \subset \boldsymbol{V}$.

1. Order the variables in $\boldsymbol{V} \setminus \boldsymbol{V}_{keep}$.
2. For each $V \in \boldsymbol{V} \setminus \boldsymbol{V}_{keep}$:
   - Eliminate V; i.e., remove factors connected to V and replace with the derived f_{new}.

The resulting factor graph is proportional to $p(\boldsymbol{V}_{keep})$.

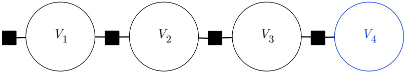
Marginal Inference by Variable Elimination (No Evidence)

Given a factor graph with variables $\boldsymbol{V}$ and factors $\boldsymbol{f}$, find the marginal distribution over some $\boldsymbol{V}_{keep} \subset \boldsymbol{V}$.

1. Order the variables in $\boldsymbol{V} \setminus \boldsymbol{V}_{keep}$.
   The ordering can make a huge difference!
2. For each $V \in \boldsymbol{V} \setminus \boldsymbol{V}_{keep}$:
   - Eliminate V; i.e., remove factors connected to V and replace with the derived f_{new}.

The resulting factor graph is proportional to $p(\boldsymbol{V}_{keep})$.

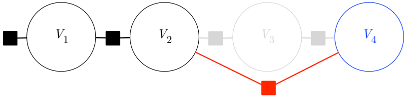
A Less Good Ordering



$$\sum_{V_1}\sum_{V_2}\sum_{V_3} f_{V_1}\cdot f_{V_1,V_2}\cdot f_{V_2,V_3}\cdot f_{V_3,V_4}$$

$$= \sum_{V_1} f_{V_1}\cdot\left(\sum_{V_2} f_{V_1,V_2}\cdot\left(\sum_{V_3} f_{V_2,V_3}\cdot f_{V_3,V_4}\right)\right)$$

A Less Good Ordering



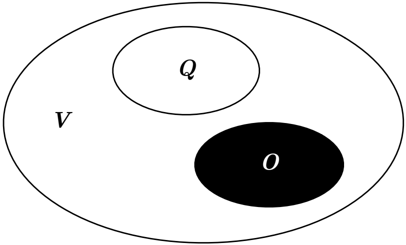
$$\sum_{V_1}\sum_{V_2}\sum_{V_3} f_{V_1}\cdot f_{V_1,V_2}\cdot f_{V_2,V_3}\cdot f_{V_3,V_4}$$

$$= \sum_{V_1} f_{V_1}\cdot\left(\sum_{V_2} f_{V_1,V_2}\cdot\left(\sum_{V_3} f_{V_2,V_3}\cdot f_{V_3,V_4}\right)\right)$$

$$= \sum_{V_1} f_{V_1}\cdot\left(\sum_{V_2} f_{V_1,V_2}\cdot f_{V_2,V_4}\right)$$

What About Evidence?

Original problem: given $\boldsymbol{O} = \boldsymbol{o}$, what is the marginal distribution over $\boldsymbol{Q} \subset \boldsymbol{U}$? (I.e., $p(\boldsymbol{Q} \mid \boldsymbol{O} = \boldsymbol{o})$.)



What About Evidence?

Original problem: given $\boldsymbol{O} = \boldsymbol{o}$, what is the marginal distribution over $\boldsymbol{Q} \subset \boldsymbol{U}$? (I.e., $p(\boldsymbol{Q} \mid \boldsymbol{O} = \boldsymbol{o})$.)

This adds a step at the beginning: **reduce** factors to "respect the evidence."

What About Evidence?

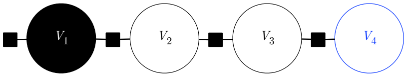
Original problem: given $\boldsymbol{O} = \boldsymbol{o}$, what is the marginal distribution over $\boldsymbol{Q} \subset \boldsymbol{U}$? (I.e., $p(\boldsymbol{Q} \mid \boldsymbol{O} = \boldsymbol{o})$.)

This adds a step at the beginning: **reduce** factors to "respect the evidence."

This will remind you of a **select . . . where** operation in a database.

Marginal Inference

Suppose V_1 is observed to take value 1.



V_1	f_{V_1}
0	
1	

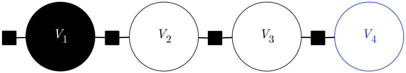
V_1	V_2	f_{V_1,V_2}
0	0	
0	1	
1	0	
1	1	

V_2	V_3	f_{V_2,V_3}
0	0	
0	1	
1	0	
1	1	

V_3	V_4	f_{V_3,V_4}
0	0	
0	1	
1	0	
1	1	

Marginal Inference

Suppose V_1 is observed to take value 1.



V_1	f_{V_1}
[redacted]	
1	

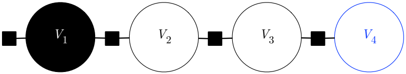
V_1	V_2	f_{V_1,V_2}
[redacted]		
1	0	
1	1	

V_2	V_3	f_{V_2,V_3}
0	0	
0	1	
1	0	
1	1	

V_3	V_4	f_{V_3,V_4}
0	0	
0	1	
1	0	
1	1	

Marginal Inference

Suppose V_1 is observed to take value 1.



V_1	f_{V_1}
[redacted]	
1	

V_1	V_2	f_{V_1,V_2}
[redacted]		
1	0	
1	1	

V_2	V_3	f_{V_2,V_3}
0	0	
0	1	
1	0	
1	1	

V_3	V_4	f_{V_3,V_4}
0	0	
0	1	
1	0	
1	1	

Note that f_{V_1} is now a constant; since we renormalize at the end, we can ignore it. Observed nodes may create a "separation" between variables of interest and some factors.

Marginal Inference by Variable Elimination with Evidence

Given a factor graph with variables $\boldsymbol{V}$ and factors $\boldsymbol{f}$, and given $\boldsymbol{O} = \boldsymbol{o}$ (where $\boldsymbol{O} \subset \boldsymbol{V}$), find the marginal distribution over $\boldsymbol{Q} \subseteq \boldsymbol{U} = \boldsymbol{V} \setminus \boldsymbol{O}$.

1. Reduce factors connected to $\boldsymbol{O}$ to respect the evidence.
2. Order the variables in $\boldsymbol{U} \setminus \boldsymbol{Q}$.
3. For each $V \in \boldsymbol{U} \setminus \boldsymbol{Q}$:
   - Eliminate V; i.e., remove factors connected to V and replace with the derived f_{new}.

The resulting factor graph is proportional to $p(\boldsymbol{Q} \mid \boldsymbol{O} = \boldsymbol{o})$.

Remarks on Computational Complexity

In general, denser graphs are more expensive.

Runtime and space depend on the size of the original and intermediate factors. (This is why ordering matters so much.)

Finding the best ordering is NP-hard.

Certain graphical structures allow inference in linear time with respect to the size of the *original* factors.

- Bayesian networks: polytrees
- Markov networks: chordal graphs

Return to Hidden Markov Models

- Hidden Markov models are not (quite) Bayesian networks.

Return to Hidden Markov Models

- Hidden Markov models are not (quite) Bayesian networks.
  - Given an observed sequence x, however, an HMM provides a pattern to construct a Bayesian network.

Return to Hidden Markov Models

- Hidden Markov models are not (quite) Bayesian networks.
  - Given an observed sequence $\boldsymbol{x}$, however, an HMM provides a pattern to construct a Bayesian network.
  - Sometimes called "dynamic graphical models."

Return to Hidden Markov Models

- Hidden Markov models are not (quite) Bayesian networks.
  - Given an observed sequence $\boldsymbol{x}$, however, an HMM provides a pattern to construct a Bayesian network.
  - Sometimes called "dynamic graphical models."
- Marginal inference for every Y_i in an HMM can be accomplished by variable elimination.

Return to Hidden Markov Models

- Hidden Markov models are not (quite) Bayesian networks.
  - Given an observed sequence $\boldsymbol{x}$, however, an HMM provides a pattern to construct a Bayesian network.
  - Sometimes called "dynamic graphical models."
- Marginal inference for every Y_i in an HMM can be accomplished by variable elimination.
  - All variables share some computation with those to their right and those to their left.

Return to Hidden Markov Models

- Hidden Markov models are not (quite) Bayesian networks.
  - Given an observed sequence $\boldsymbol{x}$, however, an HMM provides a pattern to construct a Bayesian network.
  - Sometimes called "dynamic graphical models."
- Marginal inference for every Y_i in an HMM can be accomplished by variable elimination.
  - All variables share some computation with those to their right and those to their left.
  - This is called the **forward-backward** algorithm.

Return to Hidden Markov Models

- Hidden Markov models are not (quite) Bayesian networks.
  - Given an observed sequence $\boldsymbol{x}$, however, an HMM provides a pattern to construct a Bayesian network.
  - Sometimes called "dynamic graphical models."
- Marginal inference for every Y_i in an HMM can be accomplished by variable elimination.
  - All variables share some computation with those to their right and those to their left.
  - This is called the **forward-backward** algorithm.
  - This is useful when we want to apply EM to HMMs (unsupervised sequence modeling).

Return to Hidden Markov Models

- Hidden Markov models are not (quite) Bayesian networks.
  - Given an observed sequence $\boldsymbol{x}$, however, an HMM provides a pattern to construct a Bayesian network.
  - Sometimes called "dynamic graphical models."
- Marginal inference for every Y_i in an HMM can be accomplished by variable elimination.
  - All variables share some computation with those to their right and those to their left.
  - This is called the **forward-backward** algorithm.
  - This is useful when we want to apply EM to HMMs (unsupervised sequence modeling).
  - It is also useful in supervised learning.

Related Topics

- Conditional random fields
- MPE inference
- MAP inference
- Inexact inference

Conditional Random Fields (Sequence Version)

Lafferty et al. (2001)

A nice confluence:

- Probabilistic graphical model-style reasoning, as in HMMs.
- Discriminative training, as with structured perceptron.

Local factors: $f_i(\boldsymbol{x}, y, y') = \exp\left(\mathbf{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}, i, y, y')\right)$

Log loss, where the graphical model parameterizes the probability distribution:

$$\sum_{i=1}^{n} \underbrace{\log \sum_{\boldsymbol{y} \in \mathcal{L}^{\ell_i+1}} \exp\left(\mathbf{w} \cdot \sum_{j=1}^{\ell_i+1} \boldsymbol{\phi}(\boldsymbol{x}_i, j, y_j, y_{j-1})\right)}_{\text{fear}}$$

$$\underbrace{- \mathbf{w} \cdot \sum_{j=1}^{\ell_i+1} \boldsymbol{\phi}(\boldsymbol{x}_i, j, y_{i\ j}, y_{i\ j-1})}_{\text{hope}}$$

Conditional Random Fields (General Version)

Factor graph consisting of "input" variables $\boldsymbol{X}$ (always observed) and "output" variables $\boldsymbol{Y}$.

$$p(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{X} = \boldsymbol{x}) = \frac{\prod_j f_j(\boldsymbol{x}, \boldsymbol{y}_{I_j})}{\displaystyle\sum_{\boldsymbol{y}' \in \mathrm{Val}(\boldsymbol{Y})} \prod_j f_j(\boldsymbol{x}, \boldsymbol{y}'_{I_j})}$$

MLE:

$$\sum_{i=1}^{n} \underbrace{\log \sum_{\boldsymbol{y} \in \mathrm{Val}(\boldsymbol{Y})} \prod_j f_j(\boldsymbol{x}_i, \boldsymbol{y}_{I_j})}_{\text{fear}} - \underbrace{\log \prod_j f_j(\boldsymbol{x}_i, \boldsymbol{y}_{i\ I_j})}_{\text{hope}}$$

Marginal inference is required for calculating the left term and its gradient with respect to $\mathbf{w}$.

MPE Inference

$$\underset{\boldsymbol{u} \in \mathrm{Val}(\boldsymbol{U})}{\operatorname{argmax}}\, p(\boldsymbol{U} = \boldsymbol{u} \mid \boldsymbol{O} = \boldsymbol{o})$$

MPE Inference

$$\operatorname*{argmax}_{\boldsymbol{u} \in \mathrm{Val}(\boldsymbol{U})} p(\boldsymbol{U} = \boldsymbol{u} \mid \boldsymbol{O} = \boldsymbol{o})$$

Variable elimination and exact inference are identical to the marginal case!

MPE Inference

$$\underset{\boldsymbol{u} \in \mathrm{Val}(\boldsymbol{U})}{\operatorname{argmax}} \, p(\boldsymbol{U} = \boldsymbol{u} \mid \boldsymbol{O} = \boldsymbol{o})$$

Variable elimination and exact inference are identical to the marginal case!

Just replace each sum operation with a max operation, and add bookkeeping to recover the most probable assignment.

MPE Inference

$$\operatorname*{argmax}_{\boldsymbol{u} \in \mathrm{Val}(\boldsymbol{U})} p(\boldsymbol{U} = \boldsymbol{u} \mid \boldsymbol{O} = \boldsymbol{o})$$

Variable elimination and exact inference are identical to the marginal case!

Just replace each sum operation with a max operation, and add bookkeeping to recover the most probable assignment.

The Viterbi algorithm is, of course, an instance of this. Each "$s_i(*)$" is an intermediate factor.

MPE Inference

$$\underset{\boldsymbol{u} \in \mathrm{Val}(\boldsymbol{U})}{\operatorname{argmax}}\, p(\boldsymbol{U} = \boldsymbol{u} \mid \boldsymbol{O} = \boldsymbol{o})$$

Variable elimination and exact inference are identical to the marginal case!

Just replace each sum operation with a max operation, and add bookkeeping to recover the most probable assignment.

The Viterbi algorithm is, of course, an instance of this. Each "$s_i(*)$" is an intermediate factor.

Specifically for sequence models, it should be clear how factors/features that depend on the observed sequence $\boldsymbol{X}$ don't affect the asymptotics of exact inference.

Rocket Science: True MAP

Given a factor graph with variables $\boldsymbol{V}$ and factors $\boldsymbol{f}$, and given $\boldsymbol{O} = \boldsymbol{o}$ (where $\boldsymbol{O} \subset \boldsymbol{V}$), find the most probable assignment of $\boldsymbol{Q} \subset \boldsymbol{U} = \boldsymbol{V} \setminus \boldsymbol{O}$.

Let $\boldsymbol{R} = \boldsymbol{U} \setminus \boldsymbol{Q}$.

$$\begin{aligned}
&\operatorname*{argmax}_{\boldsymbol{q} \in \mathrm{Val}(\boldsymbol{Q})} p(\boldsymbol{Q} = \boldsymbol{q} \mid \boldsymbol{O} = \boldsymbol{o}) \\
&= \operatorname*{argmax}_{\boldsymbol{q} \in \mathrm{Val}(\boldsymbol{Q})} \sum_{\boldsymbol{r} \in \mathrm{Val}(\boldsymbol{R})} p(\boldsymbol{Q} = \boldsymbol{q}, \boldsymbol{R} = \boldsymbol{r} \mid \boldsymbol{O} = \boldsymbol{o})
\end{aligned}$$

Solution: first use marginal inference to eliminate $\boldsymbol{R}$, then use max inference to solve for $\boldsymbol{Q}$.

Alternative Inference Methods

Huge range of techniques!

Exact:

- Integer linear programming

Inexact:

- randomized (e.g., Gibbs sampling, importance sampling, simulated annealing)
- deterministic (e.g., mean field variational, loopy belief propagation, linear programming relaxations, dual decomposition, beam search)

Readings and Reminders

- Koller et al. (2007)
- Submit a suggestion for an exam question by Friday at 5pm.
- Your project is due March 9.

References I

Daphne Koller, Nir Friedman, Lise Getoor, and Ben Taskar. Graphical models in a nutshell, 2007. URL http://www.seas.upenn.edu/~taskar/pubs/gms-srl07.pdf.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML*, 2001.