Problem Session Week 7

Trevor Maxfield
maxfit@stanford.edu

Minae Kwon
minae@cs.stanford.edu

May 19th, 2023

CS 221 - Spring 2023, Stanford University

Reviewing Lecture Material

Reviewing Lecture Material

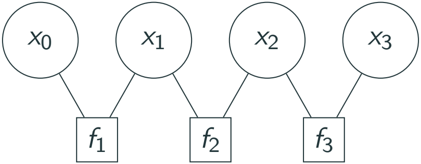
Markov Networks and Gibbs Sampling

Markov Network Motivation

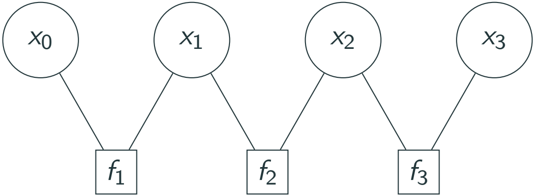
- Connecting factor graphs (CSPs, last week) with probability.
- In CSPs we found the **maximum weight assignment**:

$$\max_{x \in X} \text{Weight}(x) = \max_{x \in X} \prod_{j=1}^{m} f_j(x)$$

- Return a single x, no sense of how *likely* this x is.



Markov Network Motivation



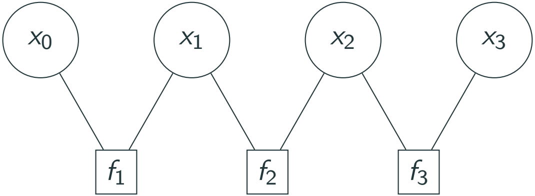
Let the domain of all X_j be $\{0, 1, 2, 3\}$. Define

$$f_i(x) = [x_i \geq i \ \wedge \ x_{i-1} \geq i]$$

so that any ($x_0 \geq 1$, $x_1 \geq 2$, $x_2 \geq 3$, $x_3 \geq 3$) is a satisfying assignment.

How many valid assignments?

Markov Network Motivation



Let the domain of all X_j be $\{0, 1, 2, 3\}$. Define

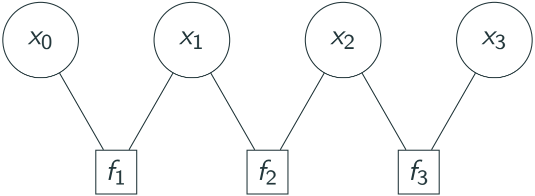
$$f_i(x) = [x_i \geq i \ \wedge \ x_{i-1} \geq i]$$

so that any ($x_0 \geq 1$, $x_1 \geq 2$, $x_2 \geq 3$, $x_3 \geq 3$) is a satisfying assignment.

How many valid assignments? 6

How many possible assignments?

Markov Network Motivation



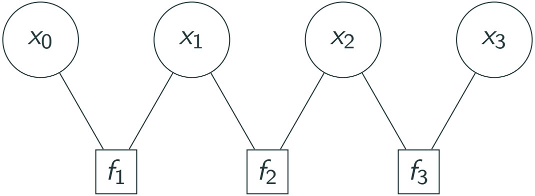
Let the domain of all X_j be $\{0, 1, 2, 3\}$. Define

$$f_i(x) = [x_i \geq i \ \wedge \ x_{i-1} \geq i]$$

so that any ($x_0 \geq 1$, $x_1 \geq 2$, $x_2 \geq 3$, $x_3 \geq 3$) is a satisfying assignment.

How many valid assignments? 6
How many possible assignments? 4^4

Markov Network Motivation



Let the domain of all X_j be $\{0, 1, 2, 3\}$. Define

$$f_i(x) = (4 - x_i)[x_i \geq i \ \wedge \ x_{i-1} \geq i]$$

Satisfying assignments aren't changed. But max weight is 6 with $x = (1, 2, 3, 3)$. How do we represent uncertainty about this assignment/weight?

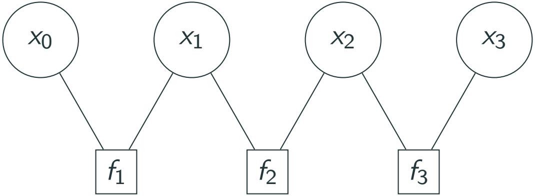
Definition
Markov Network: A factor graph which defines a joint distribution over random variables $X = (X_1, \ldots, X_n)$:

$$\mathbb{P}(X = x) = \frac{\text{Weight}(x)}{\sum_{x'} \text{Weight}(x')}$$

How much of the total weight possible weight is allocated in x?

We can **choose** to let this represent uncertainty.

Markov Network Definition



Let the domain of all X_j be $\{0, 1, 2, 3\}$.

$$f_i(x) = (4 - x_i)[x_i \geq i \ \wedge \ x_{i-1} \geq i] \quad X_j \in \{0, 1, 2, 3\}$$

x_0	x_1	x_2	x_3	$W(x)$	$\mathbb{P}(x)$
1	2	3	3	6	0.33
2	2	3	3	4	0.22
3	2	3	3	2	0.11
1	3	3	3	3	0.166
2	3	3	3	2	0.11
3	3	3	3	1	0.055

Marginal Probabilities

Definition
The **marginal probability** of $X_i = v$ is given by:

$$\mathbb{P}(X_i = v) = \sum_{x : x_i = v} \mathbb{P}(X = x)$$

The probability that a given entry $X_i = v$ is just the sum over all assignments where $x_i = v$. Alternatively you might recognize this as:

$$\mathbb{P}(X_i = v) = \sum \mathbb{P}(X = x | X_i = v)$$

Allows us to focus on the *probability of single variables* in satisfying assignments.

Gibbs Sampling

What to do if too many possible assignments?

- Initialize x to a random complete assignment.
- Loop through variables $i = 1, \ldots, n$ until convergence.
  - Compute $\mathbb{P}(X_i = v | X_{-i} = x_{-i})$ for all $v \in \text{Domain}_i$
  - Set $x_i = v$ with probability $\mathbb{P}(X_i = v | X_{-i} = x_{-i})$
  - Increment $\text{count}_i(x_i$
- Estimate $\hat{\mathbb{P}}(X_i = x_i) = \frac{\text{count}_i(x_i)}{\sum_{v \in \text{Domain}_i} \text{count}_i(v)}$

Tricky step: "Compute $\mathbb{P}(X_i = v | X_{-i} = x_{-i})$ for all $v \in \text{Domain}_i$".

Gibbs Sampling

Note that:

$$\mathbb{P}(X_i = v | X_{-i} = x_{-i}) = \frac{\text{Weight}(x_{-i} \cup \{X_i : v\})}{Z\mathbb{P}(X_{-i} = x_{-i})}$$

Not helpful in this form! Still need Z and $\mathbb{P}(X_{-i} = x_{-i})$ (expensive)!

How is $\mathbb{P}(X_{-i} = x_{-i})$ calculated from weights though?

$$\mathbb{P}(X_{-i} = x_{-i}) = \frac{\sum_v \text{Weight}(x_{-i} \cup \{X_i : v\})}{Z}$$

Eureka! Fix x_{-i} and get proportion of weight made up by each v:

$$\mathbb{P}(X_i = v | X_{-i} = x_{-i}) = \frac{\text{Weight}(x_{-i} \cup \{X_i : v\})}{\sum_v \text{Weight}(x_{-i} \cup \{X_i : v\})}$$

Need to compute each $\text{Weight}(x_{-i} \cup \{X_i : v\})$ anyways, free!

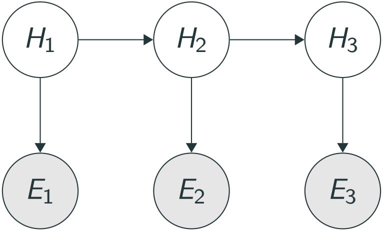
Reviewing Lecture Material

Bayesian Networks

Bayesian Networks

Markov networks but factors have more meaning. Both define joint probability distributions over assignments.

Markov Networks	**Bayesian Networks**
Arbitrary factors	Local conditional probabilities
Set of preferences	Generative process



Probability Review

Given $X = (X_1, \ldots, X_n)$

- Joint distribution: $\mathbb{P}(X = x)$
  - Maps each event x to its probability.
- Marginal distribution: $\mathbb{P}(X_i = x_i)$
  - Probability of observing i-th variable as x_i.
  - *Marginalize out* other X_{-i}.
- Conditional distribution: $\mathbb{P}(X_i = x_i \mid X_j = x_j)$
  - Given knowledge of x_j's occurrence (can be any number of other variables), what is the probability of observing $X_i = x_i$?

Example $X = (X_1, X_2, X_3, X_4)$

$$\mathbb{P}(X_1 \mid X_3 = x_3, X_4 = x_4)$$

Distribution of X_1 given that we **marginalize out** X_2 and condition on knowledge of X_3 and X_4.

Bayesian Network

Definition

Let $X = (X_1, \ldots, X_n)$ be random variables. A Bayesian network is a directed acyclic graph (DAG) that specifies a joint distribution over X as a product of local conditional distributions, one for each node:

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} p(x_i | x_{\mathrm{Parents}(i)})$$

Note that lowercase p is used to denote *local* conditional distribution (only conditioning on parents), which is specified as part of the network.

Probabilistic Inference

How can we use Bayesian networks?

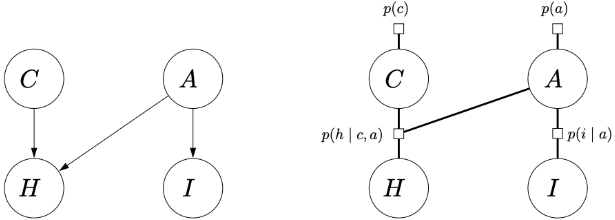
Remember that the network is just a computable representation of $\mathbb{P}(X_1, \dots, X_n)$, where we indicate interactions between variables as dependencies.

Given some evidence $e = E \subseteq X$, we query $Q \subseteq X$. Then we'd like:

$$\mathbb{P}(Q|E = e) \longleftrightarrow \mathbb{P}(Q = q|E = e) \text{ for all values } q$$

How do we compute? We'll see a similar trick as in Markov networks!

Probabilistic Inference

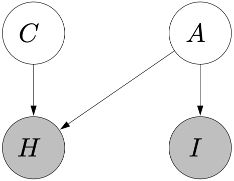


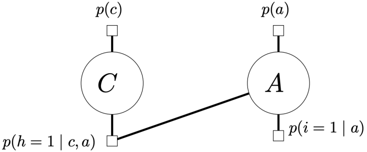
$$\mathbb{P}(C = c, A = a, H = h, I = i) = \frac{1}{Z}p(c)p(a)p(h \mid c, a)p(i \mid a)$$

> Bayesian network = Markov network with normalization constant $Z = 1$

Simply using definition of Bayesian networks.

Probabilistic Inference





$$\mathbb{P}(C = c, A = a \mid H = 1, I = 1) = \frac{1}{Z} p(c) p(a) p(h = 1 \mid c, a) p(i = 1 \mid a)$$

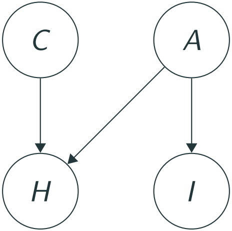
> Bayesian network with evidence = Markov network with $Z = \mathbb{P}(H = 1, I = 1)$

Choose normalization constant of 1/probability of evidence.

Alternatively remember:

$$\mathbb{P}(A|B) = \mathbb{P}(A)\mathbb{P}(B|A)/\mathbb{P}(B)$$

Probabilistic Inference



- Unobserved leaf? $\mathbb{P}(C = c, A = a | H = 1)$
  - Marginalize out I, can just ignore (sums to 1).
  - Ignore unobserved leaves.
- Independence? $\mathbb{P}(C = c | I = 1)$
  - Can ignore! $\mathbb{P}(C = c | I = 1) = p(c)$
  - Throw away disconnected components.

Reviewing Lecture Material

Summary

CSPs vs Markov Networks

Normalize weights (divide by total weight) and consider a probability distribution. Compute probability that a given variable takes on a certain assignment (marginal probability)

CSPs	**Markov Networks**
Variables	Random variables
Weights	Probabilities
Max weight assignments	Marginal probabilities

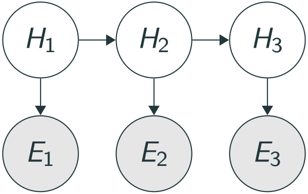
ICM vs Gibbs Sampling

Iterates on an assignment (local search), can be slow/inaccurate.

Iterated Conditional Modes	Gibbs Sampling
Maximum weight Assignment	Marginal probabilities
Choose best value	Sample a value
Converges to local optimum	Marginals converge to correct answer*

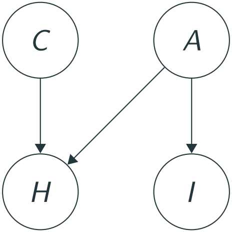
*when all weights are positive, can take a long time though.

Bayesian Network



- Random variables represent state of world.
- Directed edges dictate dependencies.
- Use local conditional distributions to get joint distributions.
- Probabilistic inference: given information ask questions about the world.

Inference on Bayesian Networks



- Condition on evidence (leaves)
- Ignore unobserved leaves.
- Discard disconnected components.
- Treat as Markov network, normalization constant of probability of evidence.

Next Week

More Bayesian networks, including forward backward and/or particle filtering.

Problems

Reviewing Lecture Material

Problems

Why Bayesian Networks

- Handle **heterogeneously** missing information.
- Incorporate **prior** knowledge.
- **Interpretable** intermediate variables.
- Precursor to **causal** models.