Lecture 7: PGM — Representation

Qinfeng (Javen) Shi

8 September 2014

Intro. to Stats. Machine Learning
COMP SCI 4401/7401

Table of Contents I

Dice rolling game

Rolling a die (with numbers $1, ..., 6$).
Chance of getting a 5 $=?$

Dice rolling game

Rolling a die (with numbers $1, ..., 6$).
Chance of getting a $5 = ?$
$1/6$
Chance of getting a 5 or $4 = ?$

Dice rolling game

Rolling a die (with numbers $1, ..., 6$).
Chance of getting a 5 $=$?
$1/6$
Chance of getting a 5 or 4 $=$?
$2/6$

Events and confidence

Probability $\approx$ a degree of confidence that an outcome or an event (a number of outcomes) will occur.

Probability space (a.k.a Probability triple) $(\Omega, \mathcal{F}, P)$:

- Sample space or outcome space, denoted Ω (read "Omega") : the set of all possible outcomes (of the problem that you are considering).
  - roll a die: $\Omega = \{1, 2, 3, 4, 5, 6\}$. flip a coin: $\Omega = \{Head, Tail\}$.
- A set of events, a σ-Field (read "sigma-field") denoted $\mathcal{F}$: Each even $\alpha \in \mathcal{F}$ is a set containing zero or more outcomes (*i.e.* subset of Ω).
  - Event: roll a die to get 1: $\alpha = \{1\}$; to get 1 or 3: $\alpha = \{1, 3\}$
  - Event: roll a die to get an even number: $\alpha = \{2, 4, 6\}$
- Probability measure P: the assignment of probabilities to the events; *i.e.* a function returning an event's probability; *i.e.* a function P from events to probabilities

Probability measure

Probability measure (distribution) P over $(\Omega, \mathcal{F})$: a function from $\mathcal{F}$ (events) to $[0, 1]$ (range of probabilities), such that,

- $P(\alpha) \geq 0$ for all $\alpha \in \mathcal{F}$
- $P(\Omega) = 1$
- If $\alpha, \beta \in \mathcal{F}$ and $\alpha \cap \beta = \emptyset$, then $P(\alpha \cup \beta) = P(\alpha) + P(\beta)$

Probability measure

Probability measure (distribution) P over $(\Omega, \mathcal{F})$: a function from $\mathcal{F}$ (events) to $[0, 1]$ (range of probabilities), such that,

- $P(\alpha) \geq 0$ for all $\alpha \in \mathcal{F}$
- $P(\Omega) = 1$
- If $\alpha, \beta \in \mathcal{F}$ and $\alpha \cap \beta = \emptyset$, then $P(\alpha \cup \beta) = P(\alpha) + P(\beta)$

$\Downarrow$

- $P(\emptyset) = 0$
- $P(\alpha \cup \beta) = P(\alpha) + P(\beta) - P(\alpha \cap \beta)$

Interpretations of Probability

- Frequentist Probability: $P(\alpha) =$ frequencies of the event. *i.e.* fraction of times the event occurs if we repeat the experiment indefinitely.
  - A die roll: $P(\alpha) = 0.5$, for $\alpha = \{2, 4, 6\}$ means if we repeatedly roll this die and record the outcome, then the fraction of times the outcomes in α will occur is 0.5.

Interpretations of Probability

- Frequentist Probability: $P(\alpha) =$ frequencies of the event. *i.e.* fraction of times the event occurs if we repeat the experiment indefinitely.
  - A die roll: $P(\alpha) = 0.5$, for $\alpha = \{2, 4, 6\}$ means if we repeatedly roll this die and record the outcome, then the fraction of times the outcomes in α will occur is 0.5.
  - Problem: non-repeatable event *e.g.* "it will rain tomorrow morning" (tmr morning happens exactly once, can't repeat).
- Subjective Probability: $P(\alpha) =$ one's own degree of belief that the event α will occur.

Conditional probability

Event α: "students with grade A"
Event β: "students with high intelligence"
Event $\alpha \cap \beta$: "students with grade A and high intelligence"

Conditional probability

Event α: "students with grade A"
Event β: "students with high intelligence"
Event $\alpha \cap \beta$: "students with grade A and high intelligence"

Question: how do we update the our beliefs given new evidence? *e.g.* suppose we learn that a student has received the grade A, what does that tell us about the person's intelligence?

Conditional probability

Event α: "students with grade A"
Event β: "students with high intelligence"
Event $\alpha \cap \beta$: "students with grade A and high intelligence"

Question: how do we update the our beliefs given new evidence? *e.g.* suppose we learn that a student has received the grade A, what does that tell us about the person's intelligence?

Answer: Conditional probability.
Conditional probability of β given α is defined as

$$P(\beta|\alpha) = \frac{P(\alpha \cap \beta)}{P(\alpha)}$$

Chain rule and Bayes' rule

- Chain rule: $P(\alpha \cap \beta) = P(\alpha)P(\beta|\alpha)$
  More generally,
  $P(\alpha_1 \cap \ldots \cap \alpha_k) = P(\alpha_1)P(\alpha_2|\alpha_1) \cdots P(\alpha_k|\alpha_1 \cap \ldots \cap \alpha_{k-1})$
- Bayes' rule:

$$P(\alpha|\beta) = \frac{P(\beta|\alpha)P(\alpha)}{P(\beta)}$$

Random Variables

Assigning probabilities to events is intuitive.

Assigning probabilities to attributes (of the outcome) taking various values might be more convenient.

- a patient's attributes such "Age", "Gender" and "Smoking history" ...
  "Age = 10", "Age = 50", ..., "Gender = male","Gender = female"
- a student's attributes "Grade", "Intelligence", "Gender" ...

$P(Grade = A)$ = the probability that a student gets a grade of A.

Random Variables

A random variable, such as Grade, is a function that associates with each outcome in Ω a value. *e.g.* Grade is defined by a function f_{Grade} that maps each person to his or her grade (say, one of A, B, C)

Grade $=$ A is a shorthand for the event $\{\omega \in \Omega : f_{Grade}(\omega) = A\}$

Intelligence $=$ high a shorthand for the event $\{\omega \in \Omega : f_{Intelligence}(\omega) = high\}$

Random Variables

Random Variable can take different types of values (*e.g.* discrete or continuous.

- random variable X, more formally $X(\omega)$
- $Val(X)$: the set of values that X can take
- x: a value $x \in Val(X)$

Shorthand notation:

- $P(x)$ short for $P(X = x)$ shorthand for

$$P(\{\omega \in \Omega : X(\omega) = x\})$$

- $\sum_x P(x)$ shorthand for $\sum_{x \in Val(X)} P(X = x)$

$$\sum_x P(x) = 1$$

Joint distribution

P(Grade, Intelligence).
Grade $\in \{A, B, C\}$
Intelligence $\in \{high, low\}$.

P(Grade = B, Intelligence = high) = ?
P(Grade = B) =?

		Intelligence		
		low	*high*	
Grade	*A*	0.07	0.18	0.25
	B	0.28	0.09	0.37
	C	0.35	0.03	0.38
		0.7	0.3	1

Marginal and Conditional distribution

Distributions:

- Marginal distribution $P(X) = \sum_{y \in Val(Y)} P(X, Y = y)$ or shorthand as $P(x) = \sum_y P(x, y)$
- Conditional distribution $P(X|Y) = \frac{P(X,Y)}{P(Y)}$

Rules for events carry over for random variables:

- Chain rule: $P(X, Y) = P(X)P(Y|X)$
- Bayes' rule: $P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$

Independence and conditional independence

Independences give factorisation.

- Independence
  $X \perp\!\!\!\perp Y \Leftrightarrow P(X,Y) = P(X)P(Y)$
  - Extension: $X \perp\!\!\!\perp Y, Z$ means $X \perp\!\!\!\perp H$ where $H = (Y, Z)$.
    $\Leftrightarrow P(X,Y,Z) = P(X)P(Y,Z)$
- Conditional Independence
  $X \perp\!\!\!\perp Y|Z \Leftrightarrow P(X,Y|Z) = P(X|Z)P(Y|Z)$
  - Independence: $X \perp\!\!\!\perp Y$ can be considered as $X \perp\!\!\!\perp Y|\emptyset$

Properties

For conditional independence:

- Symmetry: $X \perp\!\!\!\perp Y | Z \Rightarrow Y \perp\!\!\!\perp X | Z$
- Decomposition: $X \perp\!\!\!\perp Y, W | Z \Rightarrow X \perp\!\!\!\perp Y | Z$ and $X \perp\!\!\!\perp W | Z$
- Weak union: $X \perp\!\!\!\perp Y, W | Z \Rightarrow X \perp\!\!\!\perp Y | Z, W$
- Contraction: $X \perp\!\!\!\perp W | Z, Y$ and $X \perp\!\!\!\perp Y | Z \Rightarrow X \perp\!\!\!\perp Y, W | Z$
- Intersection:$X \perp\!\!\!\perp Y | W, Z$ and $X \perp\!\!\!\perp W | Y, Z \Rightarrow X \perp\!\!\!\perp Y, W | Z$

For independence: let $Z = \emptyset$ *e.g.*
$X \perp\!\!\!\perp Y \Rightarrow Y \perp\!\!\!\perp X$
$X \perp\!\!\!\perp Y, W \Rightarrow X \perp\!\!\!\perp Y$ and $X \perp\!\!\!\perp W$
...

Marginal and MAP Queries

Given joint distribution $P(Y, E)$, where

- Y, query random variable(s), unknown
- E, evidence random variable(s), observed *i.e.* E = e.

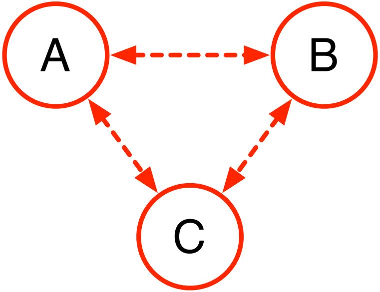
Two types of queries:

- Marginal queries (a.k.a. probability queries)
  task is to compute $P(Y|E = e)$
- MAP queries (a.k.a. most probable explanation)
  task is to find $y^* = \mathrm{argmax}_{y \in Val(Y)} P(Y|E = e)$

Break

Take a break ...

Scenario 1



Multiple problems $(A, B, \ldots)$ affect each other

Joint optimal solution of all $\neq$ the solutions of individuals

Scenario 2

Two variables X, Y each taking 10 possible values.
Listing $P(X, Y)$ for each possible value of X, Y requires specifying/computing 10^2 many probabilities.

Scenario 2

Two variables X, Y each taking 10 possible values.
Listing $P(X, Y)$ for each possible value of X, Y requires specifying/computing 10^2 many probabilities.

What if we have 1000 variables each taking 10 possible values?

Scenario 2

Two variables X, Y each taking 10 possible values.
Listing $P(X, Y)$ for each possible value of X, Y requires specifying/computing 10^2 many probabilities.

What if we have 1000 variables each taking 10 possible values?
$\Rightarrow 10^{1000}$ many probabilities

$\Rightarrow$ Difficult to store, and query naively.

Remedy

Structured Learning, specially Probabilistic Graphical Models (PGMs).

PGMs

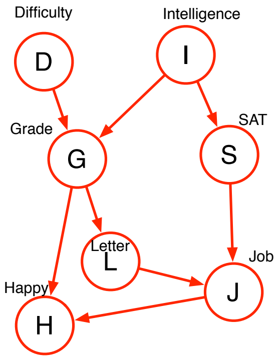
PGMs use graphs to represent the complex probabilistic relationships between random variables.

$$P(A, B, C, ...)$$

Benefits:

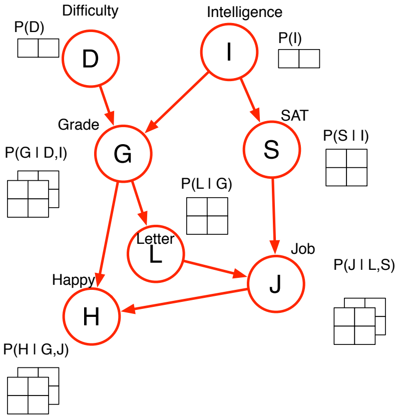
- compactly represent distributions of variables.
- Relation between variables are intuitive (such as conditional independences)
- have fast and general algorithms to query without enumeration. *e.g.* ask for $P(A|B = b, C = c)$ or $\mathbb{E}_P[f]$

An Example



Intuitive

An Example



Compact

History

- Gibbs (1902) used undirected graphs in particles
- Wright (1921,1934) used directed graph in genetics
- In economists and social sci (Wold 1954, Blalock, Jr. 1971)
- In statistics (Bartlett 1935, Vorobev 1962, Goodman 1970, Haberman 1974)
- In AI, expert system (Bombal *et al.* 1972, Gorry and Barnett 1968, Warner *et al.* 1961)
- Widely accepted in late 1980s. Prob Reasoning in Intelli Sys (Pearl 1988), Pathfinder expert system (Heckerman *et al.* 1992)

History

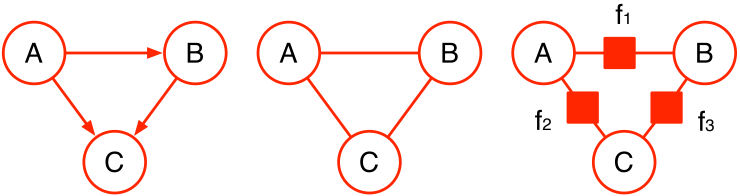
- Hot since 2001. Flexible features and principled ways of learning.
  CRFs (Lafferty *et al.* 2001), SVM struct (Tsochantaridis etal 2004), M^3Net (Taskar *et al.* 2004), DeepBeliefNet (Hinton *et al.* 2006)
- Super-hot since 2010. Winners of a large number of challenges with big data.
  Google, Microsoft, Facebook all open new labs for it.

History

Good books

- Chris Bishop's book "Pattern Recognition and Machine Learning" (Graphical Models are in chapter 8, which is available from his webpage) ≈ 60 pages
- Koller and Friedman's "Probabilistic Graphical Models" > 1000 pages
- Stephen Lauritzen's "Graphical Models"
- Michael Jordan's unpublished book "An Introduction to Probabilistic Graphical Models"
- $\cdots$

Representations



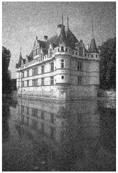
(a) Directed graph (b) Undirected graph (c) Factor graph

- Nodes represent random variables
- Edges reflect dependencies between variables
- Factors explicitly show which variables are used in each factor *i.e.* $f_1(A, B)f_2(A, C)f_3(B, C)$

Example — Image Denoising

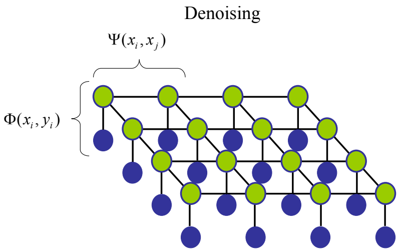
Denoising[1]

Original



Noisy

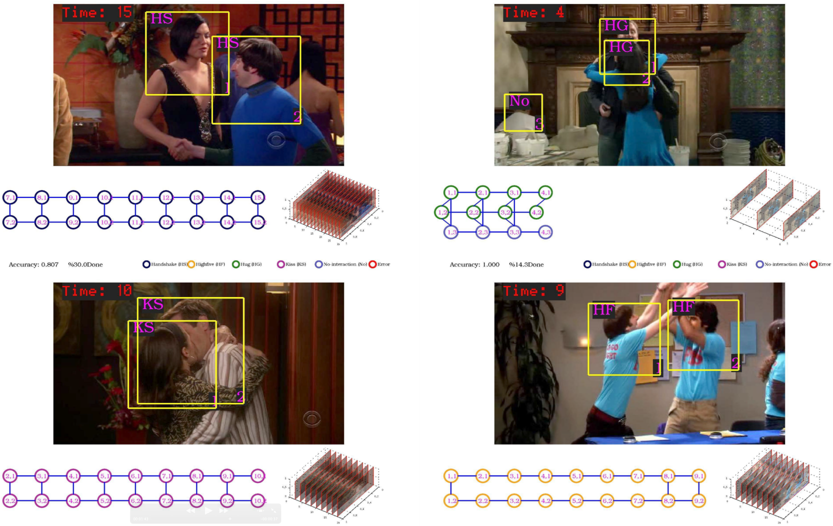
Corrected



$$X^* = \operatorname{argmax}_X P(X|Y)$$

[1] This example is from Tiberio Caetano's short course: "Machine Learning using Graphical Models"

Example —Human Interaction Recognition

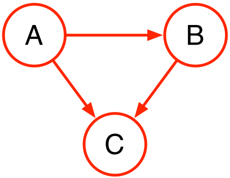


Factorisation for Bayesian networks

Directed Acyclic Graph (DAG):

Factorisation rule: $P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | Pa(x_i))$

$Pa(x_i)$ denotes parent of x_i. *e.g.* $(A, B) = Pa(C)$



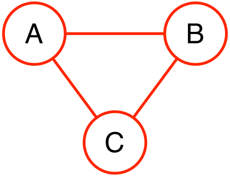
$\Rightarrow P(A, B, C) = P(A)P(B|A)P(C|A, B)$

Acyclic: no cycle allowed. Replacing edge $A \rightarrow C$ with $C \rightarrow A$ will form a cycle (loop *i.e.* $A \rightarrow B \rightarrow C \rightarrow A$), not allowed in DAG.

Factorisation for Markov Random Fields

Undirected Graph:

Factorisation rule: $P(x_1, \ldots, x_n) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \psi_c(\mathbf{X}_c)$,
$Z = \sum_{\mathbf{X}} \prod_{c \in \mathcal{C}} \psi_c(\mathbf{X}_c)$,
where c is an index set of a clique (fully connected subgraph), $\mathbf{X}_c$ is the set of variables indicated by c.



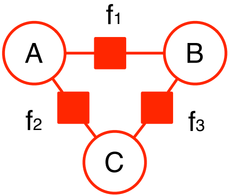
Consider $\mathbf{X}_{c_1} = \{A, B\}, \mathbf{X}_{c_2} = \{A, C\}, \mathbf{X}_{c_3} = \{B, C\}$
$\Rightarrow P(A, B, C) = \frac{1}{Z}\psi_{c_1}(A, B)\psi_{c_2}(A, C)\psi_{c_3}(B, C)$

Consider $\mathbf{X}_c = \{A, B, C\} \Rightarrow P(A, B, C) = \frac{1}{Z}\psi_c(A, B, C)$,

Factorisation for Markov Random Fields

Factor Graph:

Factorisation rule: $P(x_1, \ldots, x_n) = \frac{1}{Z} \prod_i f_i(\mathbf{X}_i)$, $Z = \sum_{\mathbf{X}} \prod_i f_i(\mathbf{X}_i)$



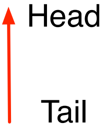
$$\Rightarrow P(A, B, C) = \frac{1}{Z} f_1(A, B) f_2(A, C) f_3(B, C)$$

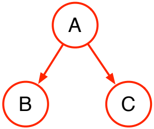
Independences

- Independence
  $A \perp\!\!\!\perp B \Leftrightarrow P(A, B) = P(A)P(B)$
- Conditional Independence
  $A \perp\!\!\!\perp B|C \Leftrightarrow P(A, B|C) = P(A|C)P(B|C)$

From Graph to Independences

Case 1:

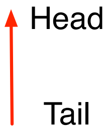


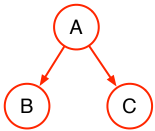


Question: $B \perp\!\!\!\perp C$?

From Graph to Independences

Case 1:



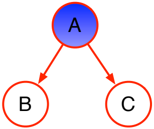


Question: $B \perp\!\!\!\perp C$?
Answer: No.

$$
\begin{aligned}
P(B, C) &= \sum_A P(A, B, C) \\
&= \sum_A P(B|A)P(C|A)P(A) \\
&\neq P(B)P(C) \text{ in general}
\end{aligned}
$$

From Graph to Independences

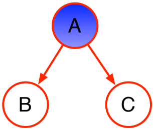
Case 2:



Question: $B \perp\!\!\!\perp C | A$?

From Graph to Independences

Case 2:

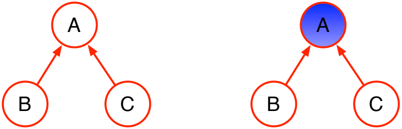


Question: $B \perp\!\!\!\perp C|A$?
Answer: Yes.

$$
\begin{aligned}
P(B, C|A) &= \frac{P(A, B, C)}{P(A)} \\
&= \frac{P(B|A)P(C|A)P(A)}{P(A)} \\
&= P(B|A)P(C|A)
\end{aligned}
$$

From graphs to independences

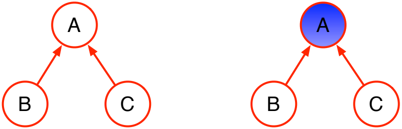
Case 3:



Question: $B \perp\!\!\!\perp C$, $B \perp\!\!\!\perp C | A$?

From graphs to independences

Case 3:

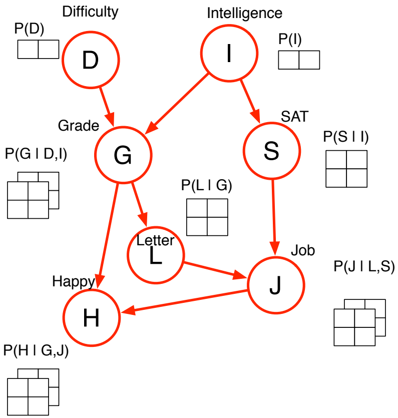


Question: $B \perp\!\!\!\perp C$, $B \perp\!\!\!\perp C | A$?

$$\because P(A,B,C) = P(B)P(C)P(A|B,C),$$
$$\therefore P(B,C) = \sum_A P(A,B,C)$$
$$= \sum_A P(B)P(C)P(A|B,C)$$
$$= P(B)P(C)$$

Parameters for bayesian networks

For bayesian networks, $P(x_1, \dots, x_n) = \prod_{i=1}^{n} P(x_i | Pa(x_i))$.
Parameters: $P(x_i | Pa(x_i))$.



Parameters for MRFs

For MRFs, let $\mathcal{V}$ be the set of nodes, and $\mathcal{C}$ be the set of clusters c.

$$P(\mathbf{x};\theta) = \frac{\exp(\sum_{c\in\mathcal{C}} \theta_c(\mathbf{x}_c))}{Z(\theta)}, \tag{1}$$

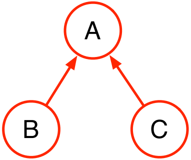
where normaliser $Z(\theta) = \sum_{\mathbf{x}} \exp\{\sum_{c''\in\mathcal{C}} \theta_{c''}(\mathbf{x}_{c''})\}$.
Parameters: $\{\theta_c\}_{c\in\mathcal{C}}$.
Inference:

- MAP inference $\mathbf{x}^* = \operatorname{argmax}_{\mathbf{x}} \sum_{c\in\mathcal{C}} \theta_c(\mathbf{x}_c)$
  $\log P(\mathbf{x}) \propto \sum_{c\in\mathcal{C}} \theta_c(\mathbf{x}_c)$
- Marginal inference $P(\mathbf{x}_c) = \sum_{\mathbf{x}_{\mathcal{V}/c}} P(\mathbf{x})$

Learning (parameter estimation): learn θ and the graph structure.

- Often assume $\theta_c(\mathbf{x}_c) = \langle \mathbf{w}, \Phi_c(\mathbf{x}_c)\rangle$.
- $\theta \leftarrow$ empirical risk minimisation (ERM).

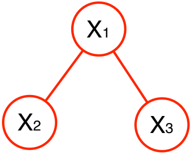
Inference - variable elimination



What is $P(A)$, or $\operatorname{argmax}_{A,B,C} P(A,B,C)$?

$$
\begin{aligned}
P(A) &= \sum_{B,C} P(B)P(C)P(A|B,C) \\
&= \sum_{B} P(B) \sum_{C} P(C)P(A|B,C) \\
&= \sum_{B} P(B)m_1(A,B) \quad (C \text{ eliminated}) \\
&= m_2(A) \quad (B \text{ eliminated})
\end{aligned}
$$

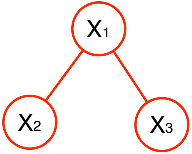
Inference - variable elimination



$$P(x_1, x_2, x_3) = \frac{1}{Z}\psi(x_1, x_2)\psi(x_1, x_3)\psi(x_1)\psi(x_2)\psi(x_3)$$

$$\begin{aligned} P(x_1) &= \frac{1}{Z}\sum_{x_2, x_3}\psi(x_1, x_2)\psi(x_1, x_3)\psi(x_1)\psi(x_2)\psi(x_3) \\ &= \frac{1}{Z}\psi(x_1)\sum_{x_2}\Big(\psi(x_1, x_2)\psi(x_2)\Big)\sum_{x_3}\Big(\psi(x_1, x_3)\psi(x_3)\Big) \\ &= \frac{1}{Z}\psi(x_1)m_{2\to 1}(x_1)m_{3\to 1}(x_1) \end{aligned}$$

Inference - variable elimination



$$
\begin{aligned}
P(x_2) &= \frac{1}{Z}\psi(x_2)\sum_{x_1}\left(\psi(x_1,x_2)\psi(x_1)\sum_{x_3}[\psi(x_1,x_3)\psi(x_3)]\right)\\
&= \frac{1}{Z}\psi(x_2)\sum_{x_1}\psi(x_1,x_2)\psi(x_1)m_{3\to 1}(x_1)\\
&= \frac{1}{Z}\psi(x_2)m_{1\to 2}(x_2)
\end{aligned}
$$

Inference - Message Passing

In general,

$$P(x_i) = \frac{1}{Z}\psi(x_i) \prod_{j \in Ne(i)} m_{j \to i}(x_i)$$

$$m_{j \to i}(x_i) = \sum_{x_j} \left(\psi(x_j)\psi(x_i, x_j) \prod_{k \in Ne(j) \setminus \{i\}} m_{k \to j}(x_j) \right)$$

That's all

Thanks!