Probabilistic Machine Learning:
7. Belief Networks: representation

Tomasz Kajdanowicz, Przemysław Kazienko (substitution)

Department of Computational Intelligence
Wroclaw University of Technology

HR EXCELLENCE IN RESEARCH

Wrocław University of Science and Technology

The presentation was inspired by Chapter 3 of D. Koller and N.Friedman book "Probabilistic Graphical Models - Principles and Techniques", 2009.

Already covered

We have covered:

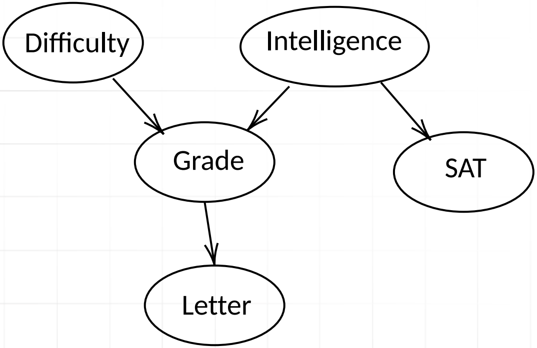
- inferring a distribution over a discrete variable drawn from a finite hypothesis space given a series of discrete observations
- inferring the probability that a coin shows up heads and dice has given value given a series of discrete observations
- basic notations used in graphical models

Now we will focus on graphical models:

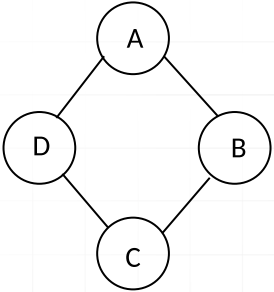
- **representation**: directed and undirected (template and plate models*)
- **inference**: exact and approximate, decision making
- **learning**: parameters and structure

Graphical models

Bayesian networks

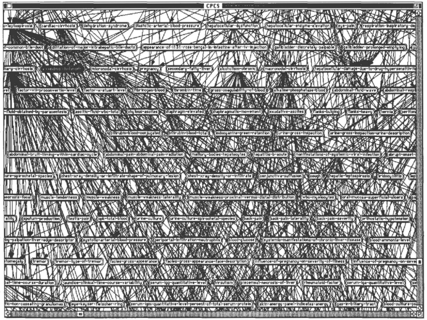


Markow networks

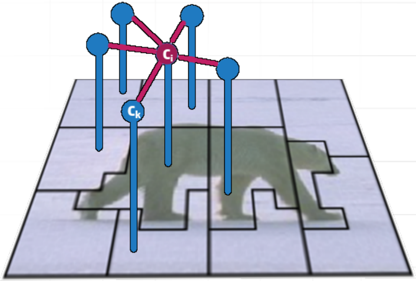


Graphical models

Computer-based Patient Case Study[1]
448 nodes, 908 edges



Markov Random Field over OpenCV example



[1]M. Pradhan, G. Provan, B. Middleton and M. H, Knowledge Engineering for Large Belief Networks, UAI, 1994

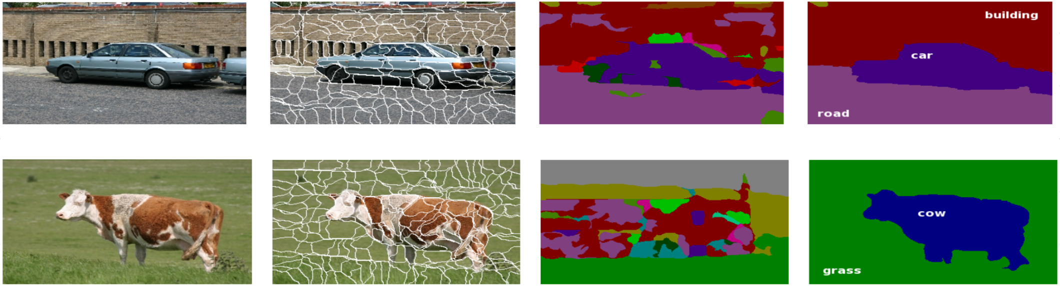
Why graphical representation?

- intuitive and compact data structure
- efficient reasoning using general-purpose algorithms
- sparse parametrization: by hand, learnt from data

Applications

- medical diagnosis
- natural language processing
- social network models
- computer vision
- speech recognition
- etc.

Example application: image segmentation



Distributions

Joint Distribution

- Intelligence (I)
  - i^0 (low), i^1 (high)
- Difficulty (D)
  - d^0 (easy), d^1 (hard)
- Grade (G)
  - g^1 (A), g^2 (B), g^3 (C)

I	D	G	P(I,D,G)
i^0	d^0	g^1	0.126
i^0	d^0	g^2	0.168
i^0	d^0	g^3	0.126
i^0	d^1	g^1	0.009
i^0	d^1	g^2	0.045
i^0	d^1	g^3	0.126
i^1	d^0	g^1	0.252
i^1	d^0	g^2	0.0224
i^1	d^0	g^3	0.0056
i^1	d^1	g^1	0.06
i^1	d^1	g^2	0.036
i^1	d^1	g^3	0.024

Distributions

Conditioning

condition on g^1

I	D	G	P(I,D,G)
i^0	d^0	g^1	0.126
i^0	d^0	g^2	0.168
i^0	d^0	g^3	0.126
i^0	d^1	g^1	0.009
i^0	d^1	g^2	0.045
i^0	d^1	g^3	0.126
i^1	d^0	g^1	0.252
i^1	d^0	g^2	0.0224
i^1	d^0	g^3	0.0056
i^1	d^1	g^1	0.06
i^1	d^1	g^2	0.036
i^1	d^1	g^3	0.024

Distributions

Conditioning: Reduction

condition on g^1

I	D	G	P(I,D,G)
i^0	d^0	g^1	0.126
i^0	d^1	g^1	0.009
i^1	d^0	g^1	0.252
i^1	d^1	g^1	0.06

Distributions

Conditioning: Renormalization

I	D	G	P(I,D,g^1)
i^0	d^0	g^1	0.126
i^0	d^1	g^1	0.009
i^1	d^0	g^1	0.252
i^1	d^1	g^1	0.06

$P(I, D, g^1)$

I	D	P(I,D$\mid g^1$)
i^0	d^0	0.282
i^0	d^1	0.02
i^1	d^0	0.564
i^1	d^1	0.134

$P(I, D|g^1)$

Distributions

Marginalization

Marginalize I

I	D	$P(I,D \mid g^1)$
i^0	d^0	0.282
i^0	d^1	0.02
i^1	d^0	0.564
i^1	d^1	0.134

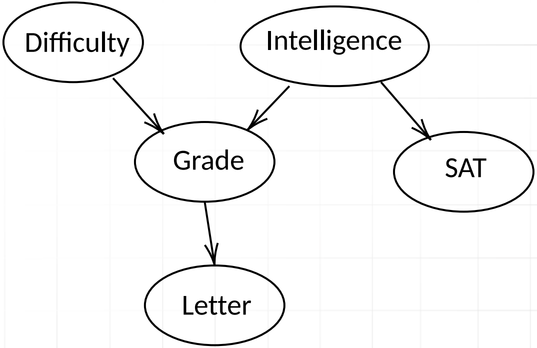
D	$P(D \mid g^1)$
d^0	0.846
d^1	0.154

Bayesian Network

Working example

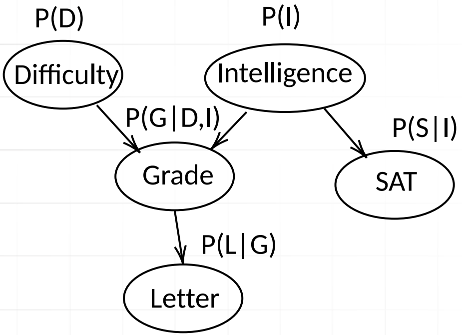
- **G**rade
- Course **D**ifficulty
- Student **I**ntelligence
- Student **S**AT
- Refference **L**etter

$$P(G, D, I, S, L)$$



Bayesian Network

Working example



- $P(D)$

d^0	d^1
0.6	0.4

- $P(I)$

i^0	i^1
0.7	0.3

- $P(S|I)$

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

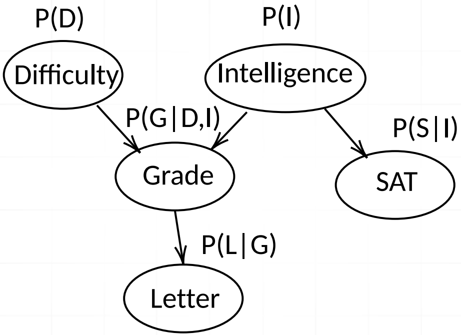
- $P(G|D,I)$

	$g^1(A)$	$g^2(B)$	$g^3(C)$
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

- $P(L|G)$

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Chain rule for Bayesian Networks



$$P(D, I, G, S, L) = P(D)P(I)P(G|I, D)P(S|I)P(L|G)$$

Distribution defined as a product of factors

Bayesian Network(Belief Network)

Bayesian Network

- is a directed acyclic graph (DAG) G whose nodes represents the random variables $X_1, \ldots, X_n$
- for each node X_i is given the conditional probability distribution (CPD) $P(X_i | pa_G(X_i))$

Joint distribution representation Bayesian Network represents a joint distribution via the chain rule for Bayesian Networks

$$P(X_1, \ldots, X_n) = \prod_i P(X_i | pa_G(X_i))$$

Does Bayesian Network represents a legal distribution?

- $P \geq 0$
- $\sum P = 1$

Does Bayesian Network represents a legal distribution?

$P \geq 0$

- P is a product of CPD
- CPD are non-negative
- thus P is non-negative

Does Bayesian Network represents a legal distribution?

$P = 1$

$$
\begin{aligned}
\sum_{D,I,G,S,L} P(D, I, G, S, L) &= \sum_{D,I,G,S,L} P(D)P(I)P(G|I, D)P(S|I)P(L|G) \\
&= \sum_{D,I,G,S} P(D)P(I)P(G|I, D)P(S|I) \sum_{L} P(L|G) \\
&= \sum_{D,I,G,S} P(D)P(I)P(G|I, D)P(S|I) \\
&= \sum_{D,I,G} P(D)P(I)P(G|I, D) \sum_{S} P(S|I) \\
&= \sum_{D,I} P(D)P(I) \sum_{G} P(G|I, D)
\end{aligned}
$$

When distribution P factorizes over G?

P factorizes over G

Let G be a graph over $X_1, \dots, X_n$. P factorizes over G if

$$P(X_1, \dots, X_n) = \prod_i P(X_i | Pa_G(X_i))$$

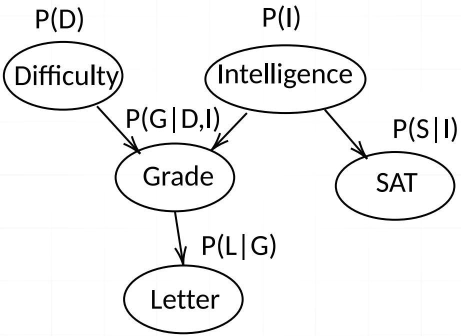
Reasoning patterns

Possible reasoning patterns:

- casual reasoning
- evidential reasoning
- intercasual reasoning

Reasoning patterns

Casual reasoning - top down



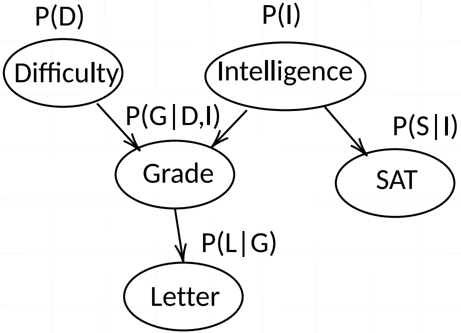
- $P(l^1) = ?$
- $P(l^1) \approx 0.5$

- $P(l^1|i^0)$
- $P(l^1|i^0) \approx 0.39$

- $P(l^1|i^0, d^0)$
- $P(l^1|i^0, d^0) \approx 0.51$

Reasoning patterns

Evidential reasoning - bottom up



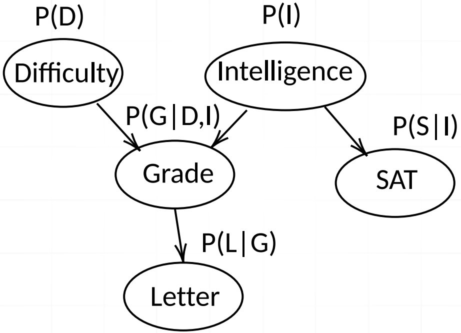
Initially we know that:

- $P(d^1) = 0.4$
- $P(i^1) = 0.3$
- but student got C grade (g^3)

- $P(d^1|g^3) \approx 0.63$
- $P(i^1|g^3) \approx 0.08$

Reasoning patterns

Intercausal reasoning



Initially we know that:

- $P(d^1) = 0.4$
- $P(i^1) = 0.3$
- but student got C grade (g^3)

- $P(d^1|g^3) \approx 0.63$
- $P(i^1|g^3) \approx 0.08$
- $P(i^1|g^3, d^1) \approx 0.11$

Independence

For events α, β, $P \models$(satisfies) $\alpha \perp \beta$ (independent) if:

- $P(\alpha, \beta) = P(\alpha)P(\beta)$ or
- $P(\alpha|\beta) = P(\alpha)$ or
- $P(\beta|\alpha) = P(\beta)$

For random variables X, Y, $P \models X \perp Y$ if:

- $P(X, Y) = P(X)P(Y)$ or
- $P(X|Y) = P(X)$ or
- $P(Y|X) = P(Y)$

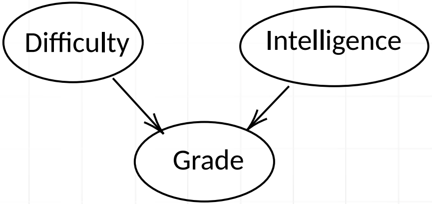
Independence

I	D	G	P(I,D,G)
i^0	d^0	g^1	0.126
i^0	d^0	g^2	0.168
i^0	d^0	g^3	0.126
i^0	d^1	g^1	0.009
i^0	d^1	g^2	0.045
i^0	d^1	g^3	0.126
i^1	d^0	g^1	0.252
i^1	d^0	g^2	0.0224
i^1	d^0	g^3	0.0056
i^1	d^1	g^1	0.06
i^1	d^1	g^2	0.036
i^1	d^1	g^3	0.024

I	D	P(I,D)
i^0	d^0	0.42
i^0	d^1	0.18
i^1	d^0	0.28
i^1	d^1	0.12

I	P(I)
i^0	0.6
i^1	0.4

D	P(D)
d^0	0.7
d^1	0.3



Conditional Independence

For sets of random variables X, Y, Z, $P \models (X \perp Y|Z)$ if:

- $P(X, Y|Z) = P(X|Z)P(Y|Z)$ or
- $P(X|Y, Z) = P(X|Z)$ or
- $P(Y|X, Z) = P(Y|Z)$

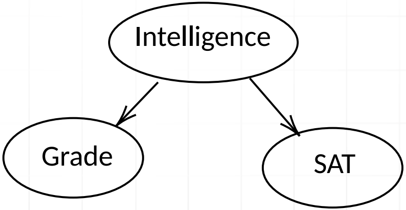
Conditional Independence

I	S	G	P(I,S,G)
i^0	s^0	g^1	0.114
i^0	s^0	g^2	0.1938
i^0	s^0	g^3	0.2622
i^0	s^1	g^1	0.006
i^0	s^1	g^2	0.0102
i^0	s^1	g^3	0.0138
i^1	s^0	g^1	0.252
i^1	s^0	g^2	0.0224
i^1	s^0	g^3	0.0056
i^1	s^1	g^1	0.108
i^1	s^1	g^2	0.0096
i^1	s^1	g^3	0.024

S	G	P(S,G $\mid i^0$)
s^0	g^1	0.19
s^0	g^2	0.323
s^0	g^3	0.437
s^1	g^1	0.01
s^1	g^2	0.017
s^1	g^3	0.023

S	P(S)
s^0	0.95
s^1	0.05

G	P(G)
g^1	0.2
g^2	0.34
g^3	0.46



Summary

We have covered today:

- joint distribution
- conditioning, reduction and renormalization
- marginalization
- chain rule for Bayesian Network
- Bayesian Network
- reasoning patterns
- conditional independence