10708 - PGM - Readings 24/04/2020

(L1) - introduction

Jordan (2004):

(W): I don't understand the significance of kernels in the PGM specification

(W): It is not fully specified how the assertions of conditional independence in directed and undirected graphs differ (i.e. arrows/edge)

(W): I don't fully understand how we move from the formalism for cliques C → formalism for factors
(undirected graph) (factor graph)

(W): Conversion of directed → undirected formalism; work with (2) ~~it.~~

(W): 3.1 exact algorithms

- Not entirely sure now/needs investigation of distributive law for marginalisation of $p(x_1)$ i.e. how

$$p(x_1) = \sum_{x_2}\sum_{x_3}\sum_{x_4} \dots \sum_{x_6} \frac{1}{Z} \psi(x_1,x_2)\psi(x_1,x_3)\psi(x_2,x_4)\psi(x_3,x_5)\psi_6(x_2,x_5,x_6)$$

is simplified

(W): Key terminology :- elimination order, triangulation algorithm, tree width

(W): At a high-level, what is stated in this algorithm is an efficient way of reducing the computational complexity of marginalising a joint probability distn

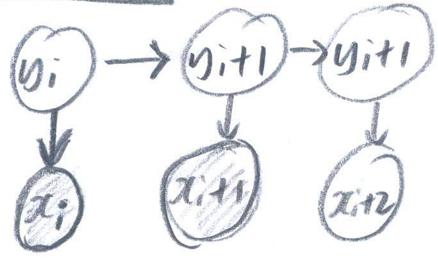
- Rest is details, machinery for doing so. (exact elimination algorithm)
- Elimination algorithm → sum-product → junction-tree algorithm.

• To sum; determine how 'loaded' die is.

Inference/learning: QB - How loaded is the die; how fair is the die, how often does the regime switch.

· These are 'layman' questions; think in terms of variable-structure-prob.

Let's formalise: - Remember r.v.s denote events (knowledge engineering)



(GM that reflects our story; but not exhaustively)

(one way of setting up)

· HMM ✓✓

· Shaded → observed; so observation x_i is observed; y_i latent (shown face of a die) (1/0 fair die)

· $x_i \in [1,2,3,4,5,6]$

· $y_i \in [0,1]$

Structure: causal, generative, coupling

· There are many ways ~~to~~ of 'cutting the cake' for an observer-relative specification e.g. specification of params./r.v.s. etc.; but have to deal appropriately with requisite complexity

· Need sequential evolution: add y_{i+1}, x_{i+1}

· How about selecting loaded, die - independent; or dependent event?

- If we keep choice of fair/loaded depended on previous choice of die; then we have about GM structure

- Markovian property: 1st order; immediate future $(i+1)$ independent of immediate past $(i-1)$ given present (i)

- HMM very widely used for modelling dependencies (Blackjack?)

Ex: Begin with joint distri:- $p(\underline{x}, y)$

- A sequence $\underline{x} = (x_1, \dots, x_T)$ and parse $y = (y_1, \dots, y_T)$

$p(\underline{x}, y) = p(x_1, \dots, x_T, y_1, \dots, y_T)$ - use factorisation using marg, cond. for parts

$= p(y_1)p(x_1|y_1) \dots p(y_T|y_{T-1})\,p(x_T|y_T)$

- only require:-

$p(y_1)$; $p(y_{i+1}|y_i)$; $p(x_i|y_i)$

- first marginal
- transmission
- emission

- can factorise in many ways; with assoc. interp.

$= p(y_1)p(y_2|y_1)\ldots p(y_T|y_{T-1})p(x_1|y_1)\ldots p(x_T|y_T)$

Ⓠ: check this/refresh memory

- evaluation question: $p(\underline{x}, y)$

- marginal: $p(\underline{x}) = \sum_{y} p(\underline{x}, y) = \sum_{y_1}\sum_{y_2}\ldots\sum_{y_N} \pi_{y_1} \prod_{t=2}^{T} a_{y_{t-1}, y_t} \prod_{t=1}^{T} p(x_t|y_t)$

(marginalise out y)

- $a_{y_{t-1}, y_t} = p(y_t|y_{t-1})$ - assume that this is constant every step

time homogeneous transmission probabilities

(assumption)

- known as stationarity assumption

- Posterior: (inference question) $p(y|\underline{x}) = \dfrac{p(\underline{x}, y)}{p(\underline{x})}$

EX: a way of representing the massive $p(\underline{x}, y)$ by simplifying.

- But even for a sequence of length T, examine computationally
- lots of summations over all possible values y_i can take.
- ∴ There are summations over 2^T possible y_i values - exponential ✗
- we want polynomial time (linear/quad/cubic)

- Bayesian Network
- see slides

EX: focus on what we mean by conditional independence assumptions

- specification of GM
  - qualitative (structure/topology from assumptions)
  - quantitative

Ex: will cover how to learn/estimate structure from data.

Formalism (LS & I)

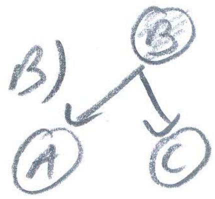
- what do edges mean; other than guide to factorisation of joint distr

- independences - probabilistic definitions

- 3 key building blocks of GM (remember Bishop)

1. common parent → children $A \perp C \mid B$ (A independent of C given B)

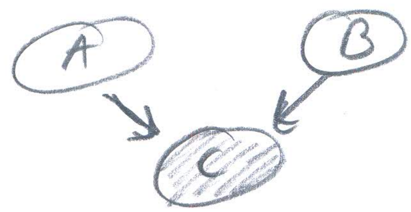
Fixing B decouples A, C $P(A,C \mid B) = P(A \mid B) P(C \mid B)$



2. Cascade (chain of r.v.s connected by directed edges)

- Fixing B decouples A, C

3. V-structure



Knowing C couples A and B

- because A can 'explain away' B wrt C

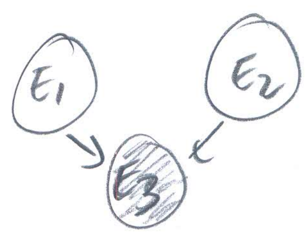
(N): Note language of decoupling, conditioning (see Bishop for refresher)

- Explaining away: E_1 'defective clock' E_2 - 'traffic jam to CMU' - define 2 events

- 'Ex being late'

- E_1 and E_2 are not dependent without a story ; but there is a further event E_3:-

E_3 - 'Ex being here' on time or not



- Then observe

Ex - not on time arrival (E_3)

- clock is slow (E_1)

(traffic situation)

- Depending on observation of E_3 and E_1 what are inferences about E_2? $p(E_2)$

- clock fine ⇒ p(traffic jam) is larger
- clock defective ⇒ p(traffic jam) is lower

} 2 irrelevant events become coupled as they jointly cause an observed event

(W): I want to further understand 'explaining away'

(A2) - clock is 'explaining away' probability assoc. with traffic jams.

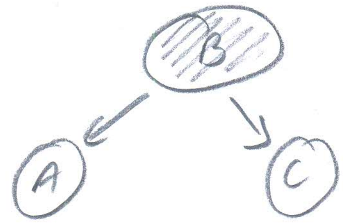
$P(AB) = \cancel{P(AB)}\ P(A)P(B)$ } check informal

Given C $P(A,B|C) \neq P(A)P(B)$

statistical just.

prove (is it a proof?)

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



- via factorisation/traversal:-

$P(A,B,C) = P(B)P(A|B)P(C|B)$ (by definition) using GM

$$P(A,C|B) = \frac{P(A,B,C)}{P(B)} = \frac{P(B)P(A|B)P(C|B)}{P(B)} = P(A|B)P(C|B)$$

(prob. definition)

knowing/ . using graph structure ⇔ knowing/ write down C.I relations (without lengthy proofs)

· I-map (07)(07) $\boxed{50:00}$

- for every distribution defined on a domain X with r.v.s.
- can always define a set of ~~conditi~~ independence assertions
(see slides for detail)
- 2 definitions
- Every set of independences can be associated with a graph object

(W)(A3): really have to clarify I-MAPS!

- I-set of 1st graph $X \perp Y$
I(G)
- I(G) of 2nd graph ϕ (no way to claim any independences)
- I(G) of 3rd -"- ϕ

· Now have joint distri of 2 binary r.v.s. (

$P_1 \to I(P_1) = X \perp Y$

(?) · How were these gained via inspection? (??)

$P_2 \to I(P_2) = \phi$

· Which graph is I-map of distri P_1?

· Which graph is I-map of distri P_2?

· (W): (A4) - clarify the logic of this - supplementary note

· (DT): A correspondence/relation between graph and distribution (can we establish a unique correspondence)?

- (Q): Given a graph G, can I write one distri or multiple distri?
{ OR given a distri P, can I write (express) one graph or multiple

↳ explainable A.I (future, but super interesting)

- There are multiple ways to systematically extract I maps from graph/distri. (earlier → local Markov ass.)
- Skip formal definitions EX (W): (A5)

- systematic way of extracting all independences in bigger graphs (to get a definitive set of independences)

'd-separation' - see slide definition

- (DT): Define node of interest; then find ancestral graph (remove non-ancestors descendants)

then find/manipulate by moralisation (i.e. couple/'marry' those nodes which are not connected, but have a common descendant (i.e. children) by connecting them)

Node of interest → ancestral graph → moralised ancestral graph

and remove directionality

- ways to travel from one node to another node through graph
- 2 nodes are not independent

more formally;

- In a moralised, ancestral graph ; $I(G) = \{ x \perp y \mid z \; : z \text{ d-separates } x \text{ and } y \}$

EX: How does d-separation capture 3 building blocks
(operation)

(h):(A6) - Have to read up on d-separation ; X and Y, Z can also be subsets

---

Bayes Ball - Skip in lecture ; there are recitations?

- Read in MJ -
- Homework/offline topic/reading

- Tie I(G) and I(P) together to formally establish equivalence
- equivalence theorem (Does it exist etc.)

$G \to I(G) \longrightarrow P_{(I(G))}$

via d-sep
(via program)

- via factorisation law ? ✓ (suggested ~~but not~~ quite)

- via inspection of c.p. tables

- examples of above process (CPTs)

(h):(A7) review

- introduction of Gaussians

- Summary of BN semantics
- Are there other independences that hold for every distri P that factorises G?
- (h)(A8): check you intuitively understand
- via Proof (h):(A8) - review nuances from hereon
1:10:11 → for nuances on soundness, completeness

- Reframe to question of soundness/completeness of d-sep. wrt BN factorisation law

-

- soundness (of d-sep): check you understand.
- completeness → trickier
  - use contrapositive statements about completeness
- very involved theoretical treatment about the equivalence theorem (theorem)
- review slides I(G), I(P), equivalence