10708 - L2 - Directed GMs - Bayesian Nets

24/04/2020

EX: Two representations

- DGM / Bayesian Networks

- Modern position: philosophical confrontation not as marked.
- GMs are not confined to any particular paradigm prob.; only require an optimality scoring
- Directed edges → causality

EX: Edges → directionality → factorisation law (product of traversal over graph)

- Hit node - define marginal / conditional (on parent nodes)

- Notation in graphical models is messy
- Notational structure in this class:-
- variable: V ; V_i (multivariate variable, i^{th} dimension) ; $V^{(i)}$ - instances

$\longrightarrow V^{(j)}_{i,R}$ index feature

- EX: Every variable has a block of characteristics
- variables are associated with realisations $v^{(j)}_{i,R}$
- Random variable → probability distn (stochasticity)
- Random vector $\underline{V} = \begin{bmatrix} V_1 \\ \vdots \\ V_N \end{bmatrix}$ - Random matrix - $\underline{V} = \begin{bmatrix} V_{1,1} & \cdots & V_{1,N} \\ \vdots & \ddots & \\ V_{N,1} & \cdots & V_{N,N} \end{bmatrix}$

- Parameters (special class) - Greek letters e.g. $\mu, \sigma^2, \alpha, \beta, \Phi, \Xi$

Dishonest Casino (example)

- Regime switching model with fair and loaded die

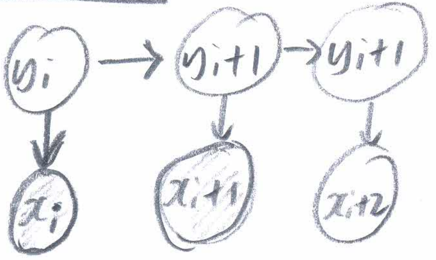
- Puzzles for dishonest casino
- note 'fishiness' of 6s: quantitative statements $\{1,2,4,\ldots 6,6,6\ldots\} = \underline{x}$
- (Q1) How probable is this sequence, given a fair die for whole role
  - Evaluation of: $p(\underline{X} = \underline{x})$?
- (Q2): what portion of sequence generated with fair, loaded
  - introduce y r.v. - choice of die (as an event you want to model)
  - Decoding of: $P(y|\underline{X} = \underline{x})$

- To aim; 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 $(i+1)$ indep. 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 phase $y = (y_1, \dots, y_T)$

$p(\underline{x}, y) = p(x_1, \dots, x_T, y_1, \dots, y_T)$ - 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.

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

Q: 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

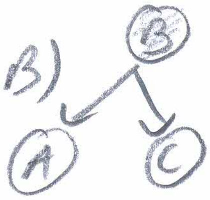
EX: will cover how to learn/estimate structure from data.

Formalism (LS & I)

- What do edges mean; other than guide to factorisation of joint distn
- independences - probabilistic definitions
- 3 key building blocks of GM (remember Bishop)

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

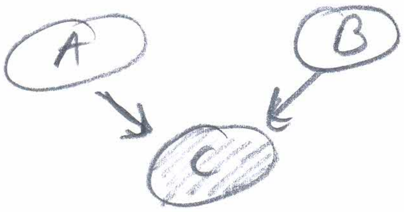
Fixing B decouples A, C $p(A,C|B) = p(A|B)p(C|B)$



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

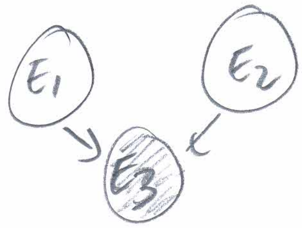
(N): Note language of decoupling, conditioning (see Bishop for refresher)

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

- 'EX being late'

- E_1 & 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

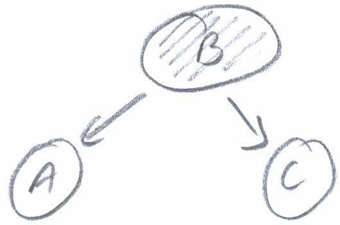
Given C $P(A,B|C) \neq P(A)P(B)$ } informal

---

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 (OT)(OT) $\boxed{50:00}$
- for every distribution defined on a domain X with r.v.s.
- can always define a set of ~~condit.~~ 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. (

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

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

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

- Which graph is I-map of distri P_1?
- Which graph is I-map of distri P_2?
- (W): (Ar) - clarify the logic of this - supplementary note
- (OT): 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 (express) write 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): (AS)

- systematic way of extracting all independences in bigger graphs (to get a definitive set of independences)
- 'd-seperation' - see slide definition
- (OT): 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) and remove directionality

Node of interest → ancestral graph → moralised ancestral graph

- 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)

W: 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.)

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

via d-sep (in a program)

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

- via inspection of c.p. tables

· Examples of above process (CPTs)

W: A7 review — introduction of Gaussians

· Summary of BN semantics
- Are there other independences that hold for every distri P that factorises G?
  - W A8: check you intuitively understand
  - via Proof — W: 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