26/04/2020

- UGMs

- undirected gms - (Q) $P(x_1,\dots,x_8) = \frac{1}{2}\exp\{E(x_1)+\dots E(x_6,x_5,x_8)\}$ - (Q)
- UGMs → expert systems (complex causality webs)

HMM / dishonest casino

(*) key areas to focus on → 10708 learning objectives

- Areas to focus on:- - These should be included in 'lectures'
- BN definition
- formal specification of factorisation
- distinction between functional form of factorisation for DGM, UGMs
- understand significance of conditional independence in DGMs
- know the 3 local structures visually, what C.I relations they represent
- nuances of I-Maps ; appl. to example (*)
- local Markovian assumptions -
- d-separation: semantics, operation
- Global Markovian assumptions - Bayes ball.
- Equivalence theorem ; difference with I-maps (*)
- soundness + completeness in context of (*).

Jordan (2003) Ch 2 (read; annotate/supplement)

- very clear formal setup
- in general, not making use of additional 'structure', a joint probability distn of n random variables, each taking on r states requires full specification of r^n states in an n-dimensional probability table
- Graphical models represent this more economically with ass. about structure via 'local relationships'
- formal notation:-
- Directed graph $G(V,E)$ V-set of nodes E-edges G-acyclic
- 1-to-1 mapping r.v.s to nodes
- (*) For each node $i \in V \longrightarrow$ r.v. X_i
- $V = \{1,\dots,n\}$ and r.v.s $= \{X_1,\dots,X_n\}$

- each node has a set of parent nodes, may be empty
- for each node $i \in V$, we denote π_i as the set of parents of node i
- the set of random variables X_{π_i} are the parents of the r.v. X_i
- Part-child relations → locality → economical rep. of joint p.d.
- to each node i, associate function $f_i(x_i, x_{\pi_i})$ with properties:-

$$f_i(x_i, x_{\pi_i}) \geq 0 \quad \text{and} \quad \sum_{x_i} f(x_i, x_{\pi_i}) = 1$$

- the pre-conditional probability formalism is:
- let $V = \{1, \ldots, n\}$, given a set of functions $\{f_i(x_i, x_{\pi_i}) : i \in V\}$, define joint p.d:-

$$p(x_1, \ldots, x_n) \triangleq \prod_{i=1}^{n} f_i(x_i, x_{\pi_i}) \qquad (2.1)$$

- Have to verify definition obeys joint p.d. constraints
- choice of numerical values for $f_i \Rightarrow$ generation of specific joint p.d.
- Ranging over all numerical values $\Rightarrow$ family of joint prob distr. associated with graph G (*)
- This family can be characterised as products of local functions or graph-theoretically (edge patterns)
(*) The relationship between different ways of characterising the family of probability distributions associated with a graph that is key to PGM
- conditional probability formalism:

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i | x_{\pi_i}) \qquad (2.2)$$

- $p(x_i | x_{\pi_i})$ - local conditional prob. associated with graph G
- Building blocks wieldy joint p.d. synthesised associated with G.
- Example given: (see diag).

$$p(x_1, \ldots, x_6) = p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_3)p(x_6|x_2, x_5) \qquad (2.3)$$

- Product of LCD

Representational economy:

- r^n joint prob. table $\longrightarrow$ r^{m_i+1} table
  (no structure) (graph structure)

- m_i - no. of parent nodes of X_i
- local cond. distr. X_i - (m_i+1) dim table
  each node/r.v. has r values

(01): Exponential growth in n exchanged for exponential growth in m_i
(no. variables in domain) (no. of parents of indiv. nodes X_i (fan-in))

- small fan-in $\longrightarrow$ enormous reduction in complexity
- (02): But there is more; not just data structure promise, but inferential machinery

---

2.1 Conditional independence

- Independence:

x_A and x_B are independent, i.e. $x_A \perp\!\!\!\perp x_B$ if:-

$$p(x_A, x_B) = p(x_A)p(x_B) \tag{2.4}$$

- Conditional independence

x_A and x_C are conditionally independent, given x_B if:-

$$p(x_A, x_C | x_B) = p(x_A | x_B)\, p(x_C | x_B) \tag{2.5}$$

OR

$$p(x_A | x_B, x_C) = p(x_A | x_C) \tag{2.6}$$

↳ moving fluidly here is (S11)

for all x_B: $p(x_B) > 0$

(2.5) - independence; with addition of conditioning on $|x_B$ formula

(2.6) - "probability of x_A given x_B and x_C is equal to probability of x_A given x_C" i.e. conditioning on x_B and x_C is the same as omitting x_B (conditioning variable) in prob.

- Establish independence / conditional independence requires factorising the joint probability distribution

- PGM: representing a p.d. within a m formalism involves making indpence assumptions, which are embedded in the structure of the graph
  - This graphical structure, other indep. relations can be derived, reflecting that certain factorisations of joint p.d. imply other factorisations
  - factorisations - read off via graph search algorithms

- Graphical structure encodes conditional indpendce:-
- chain rule of prob: PMF in a general factored form, given an ordering on $\{x_1, \ldots, x_6\}$:-

$$p(x_1, \ldots, x_6) = p(x_1)p(x_2|x_1)p(x_3|x_1,x_2)p(x_4|x_1,x_2,x_3)p(x_5|x_1,x_2,x_3,x_4) \cdot p(x_6|x_1,x_2,x_3,x_4,x_5)$$

- under an arithmetic node ordering
- in general:

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i|x_1, \ldots, x_{n-1}) \qquad (2.7)$$

(Q): Compare (2.2) and (2.7) ⇒ conditioning variables dropped

(A1) missing variables in local c.p. functions ⇒ missing edges in underlying graph

(A2) It is this transfer of interpretation from missing variables → missing edges that underlies the probabilistic interpretation for missing edges in graph in terms of c.i.

- formally:
  - Define an ordering I of nodes in a graph to be topological if for every node $i \in V$ the nodes in π_i appear before i in the ordering.
  - E.g. $I = \{1,2,3,4,5,6\}$ (topological ordering for graph in 2.1)
  - (A3) Let v_i denote set of nodes that appear earlier than i in the ordering I, excluding the parent nodes π_i

    e.g. $v_5 = \{1,2,4\}$
  - v_i necessarily contains ancestors of node i (excluding parents π_i) and may contain other nondescendant nodes also

- Given a topological ordering Σ for a graph $\mathcal{G}$ we associate to the graph the following set of c.i. statements:-

$$\{X_i \perp\!\!\!\perp X_{v_i} \mid X_{\pi_i}) \quad \text{for } i \in V \tag{2.8}$$

- "Given the parents of a node (X_{π_i}), the node (X_i) is independent of all earlier nodes in the ordering (X_{v_i})"
- The example encodes following indep. ass:-

$$\begin{pmatrix} X_1 \perp\!\!\!\perp \emptyset \mid \emptyset \\ X_2 \perp\!\!\!\perp \emptyset \mid X_1 \end{pmatrix} \tag{2.9, 2.10}$$

$$X_3 \perp\!\!\!\perp X_2 \mid X_1 \tag{2.11}$$

$$X_4 \perp\!\!\!\perp \{X_1, X_3\} \perp X_2 \tag{2.12}$$

$$X_5 \perp\!\!\!\perp \{X_1, X_2, X_4\} \perp X_3 \tag{2.13}$$

$$X_6 \perp\!\!\!\perp \{X_1, X_3, X_4\} \perp \{X_2, X_5\} \tag{2.14}$$

- Interpretation of missing edges as c.i. consistent with (2.2) - nuance later
- verification example of (2.12) by direct calculation from (2.3):-

$$p(x_1, x_2, x_3, x_4) = \sum_{x_5} \sum_{x_6} p(x_1, \ldots, x_6) \quad \text{(marginalise)} \tag{2.15}$$

$$\text{marg. prob } \{X_1, \ldots, X_4\} = \sum_{x_5} \sum_{x_6} p(x_1) p(x_2|x_1) p(x_3|x_1) p(x_4|x_2) p(x_5|x_3) p(x_6|x_2, x_5) \quad \text{(subs.)} \tag{2.16}$$

$$= p(x_1) p(x_2|x_1) p(x_3|x_1) p(x_4|x_2) \sum_{x_5} p(x_5|x_3) \sum_{x_6} p(x_6|x_2, x_5) \quad \text{(sum.)} \tag{2.17}$$

($\sum_{x_5} p(x_5|x_3) = 1$, $\sum_{x_6} p(x_6|x_2, x_5) = 1$)

$$= p(x_1) p(x_2|x_1) p(x_3|x_1) p(x_4|x_2) \tag{2.18}$$

marg. prob $\{X_1, X_2, X_3\}$:-

$$p(x_1, x_2, x_3) = \sum_{x_4} p(x_1) p(x_2|x_1) p(x_3|x_1) p(x_4|x_2) \tag{2.19}$$

$$= p(x_1) p(x_2|x_1) p(x_3|x_1) \sum_{x_4} p(x_4|x_2)$$

$$= p(x_1) p(x_2|x_1) p(x_3|x_1) \tag{2.20}$$

hence: $p(x_4|x_1, x_2, x_3) = p(x_4|x_2)$ (recall all c.i definition!) (2.21)

- (Q1) above shows we can interpret missing edges in graph in terms of conditional independencies

(*) (Q): are there other conditional independence statements that are true of such joint probability distrib's, and do these have graphical interp.

- There are other conditional indepencies e.g. $X_1 \perp\!\!\!\perp X_6 \mid \{X_2, X_3\}$, not in list (2.14), but implied by the list.
- clarifiable by algebra (tedious)

(Q1): we want to write down ALL conditional indepencies implied by basic set.

- develop critera for doing so without factorising joint into all possible ~~tripl~~ triples of var. subsets

(Q1) That is, a general graph search algorithm to so find all implied indepencies as well as explicit ones from graph structure.

2.12. conditional indepedence and Bayes ball

- Bayes ball - reachability notion a def. of sepection.
- we can not only derive C.I. assertions from (2.2); but other way around.

Recognition:
- To each graph we associate family of joint prob. distri.
- Family arises due to consideration of range over different choices of numerical values of local c.ps $p(x_i \mid x_{\pi_i})$. list of
- view c.i. statements gen by Bayes ball as constraints on ∉ joint p.d.
- Those joint p.d.s that meet → if not out. in the family

(Q): relationship between characterisation of a family of p.d.s in terms of conditional indepencies; and numerical characte. in terms of local cond. prob (5.2.1.3.)

---

3 canonical graphs

missing edges in PGM → c.i.

le

(2.4)

- missing edge: $X-Z$
- Hence C.I:-

① $X \perp\!\!\!\perp Z \mid Y$

② No other conditional independencies associated with this graph

- Justifying ①: (DAG structure ⇒ C.I.)

From DAG: $p(x,y,z) = p(x)p(y|x)p(z|y)$ (2.23)

(2.2)

$$\Rightarrow p(z|x,y) = \frac{p(x,y,z)}{p(x,y)} \quad (2.24)$$

$$= \frac{p(x)p(y|x)p(z|y)}{p(x)p(y|x)} \quad (2.25)$$

$$= p(z|y) \quad (2.26)$$

- establishing c.i.

- Justifying ②:- (611)

- "There are no further conditional independencies associated with this graph"
  - does NOT mean that no further conditional independencies can arise in any of the distributions in the family associated with this graph i.e. distributions with factorised form (2.23) (!)
  - There exist some distributions which exhibit additional independencies
  - e.g. free to choose any local c.p. $p(y|x)$ → choose distr in which $\forall$ prob. y same no matter value of x. then for this $p(y|x)$, $X \perp\!\!\!\perp Y$.

(611): Fig (2.4) $\not\Rightarrow$ X and Y are necessarily dependent (not independ).

(511): Edges do not necessarily imply dependence

BUT (611): missing edges do imply independence

- universally vs existentially qualified statements with respect to family of distributions associated with a graph.

- (611): Asserted conditional independencies always hold for these distributions
  - Implied/non-asserted conditional independencies sometimes fail to hold for distributions associated with a particular graph, but sometimes do hold.

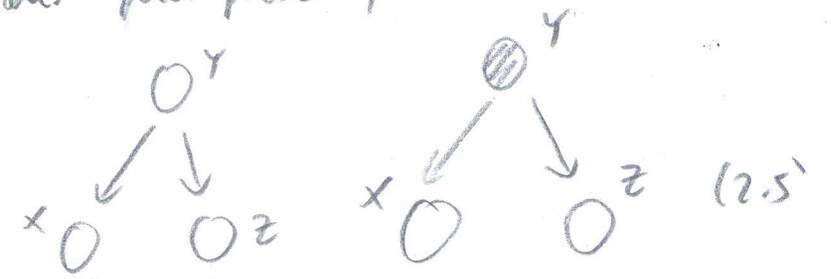
(51): • An algorithm based on conditional independencies will be correct for all distns associated with a graph
• An algorithm based on absence of c.i.s. will sometimes be correct, sometimes not.

Mental model for cascade:- past-present-future

common parent

- missing edge: X-Z

① $X \perp\!\!\!\perp Z \mid Y$



(2.5)

• Justification ①:- (DAG → C.I.)

$$p(x,y,z) = p(y)p(x|y)p(z|y) \quad (2.28)$$

$$p(x,z|y) = \frac{p(y)p(x|y)p(z|y)}{p(y)} \quad (2.29)$$

$$\Rightarrow p(x,z|y) = p(x|y)p(z|y) \Rightarrow X \perp\!\!\!\perp Z \mid Y \quad (2.30)$$

Mental model: 'Hidden variable'
- X-shoe size ; Z - 'grey hair' ; Y - age
- X, Z → marg., strong dep.
- But with Y-age, we might be willing to assert $X \perp\!\!\!\perp Z \mid Y$.
- Hidden 'Y' explains all of observed X-Z dep.

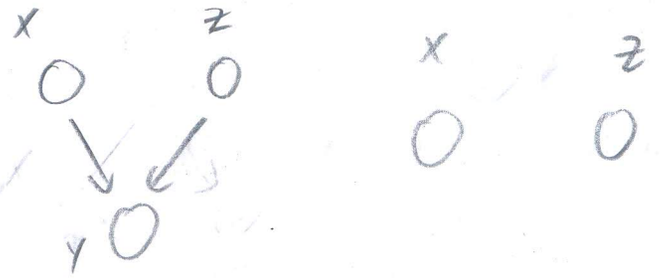
②: "No other c.i.s associated with this graph"
- No assertions of dependence, in particular we do not necess. assume X and Z dependt because they depend on Y.
- But we assert these at least some distns. in which such dependences hold.

V-structure

$X \perp\!\!\!\perp Z$

- missing edge X-Y

(Statement about subgraph)



10 - Oct MS readings

Koller + Friedman (2009)

- (OH): use the intuitive examples here to supplement, reinforce understanding; they hold the key to applying the reasoning
- (W): The purposes of this reading → clarifying BN semantics, I-maps, equivalence, soundness, completeness
- (W): recall 36-705, 110 as cornerstone of statistical analysis
- (W): 10-708 → complementary through consideration of dependence

- Reason for parametrisation through cpd rather than full joint → modularity
- Adding node will not require new joint p.d.; modular in cpd representation.
- Local vs global

3.2.2.2. Bayesian Network semantics

Definition 3.1

- A Bayesian network structure G is a DAG whose nodes represent r.v.s. $X_1, \dots, X_n$.
- Let $Pa_{X_i}^G$ represent parents of X_i in G
- Let $NonDescendants_{X_i}$ denote variables that are not descendants of X_i
- (*) G encodes following set of conditional independence assumptions, called local independencies.
- These are denoted $I_l(G)$:-
- For each variable X_i : $(X_i \perp NonDescendants_{X_i} \mid Pa_{X_i}^G)$

"The local independencies state that each node X_i is conditionally independent of its nondescendants given parents"

3.2.3. Graphs and Distri

(OH):- 1.) Bayesian network graph with conditional independeccsetions

2.) Bayesian network graph with conditional probability distributions

These definitions are equivalent

(OH):

(*) distribution P satisfies local independencies associated with a graph G ⇔ P is representable as a set of CPDs associated with graph G

· 3.2.3.1. - I maps

- define set of independencies associated with distri P

Definitions 3.2. independencies in P

- Let P be a distri over X

(*) we define $I(P)$ to be the set of independence assertions of the form $(\underline{X} \perp \underline{Y} \mid \underline{Z})$ that hold in P.

(*) - we can rewrite statement that "P satisfies the local independencies associated with G" as $I_l(G) \subseteq I(P)$

(*) - semantically, "G is an I-map (independency map) for P"

---

definition 3.3 - I-map

- Let K be any graph object associated with a set of independencies $I(K)$
- We say "K is an I-map for a set of independencies I if $I(K) \subseteq I$"

(*) - semantically

"G is an I-map for P if G is an I-map for $I(P)$"

(?): do you understand these nuances?

- (*) Direction of subset inclusion ⇒ for G to be an I-map of P, it is necessary that G does not mislead us regarding independencies in P.

- steal at example to help with this abstract definition

- note that $I_l(G) \subseteq I(P)$ implies that $I(P)$ may contain independence assertions not present in $I_l(G)$!

- these definitions are thinicky → do some scribblings

write statements mathematically

Intuition from I-map to factorisation

- BN structure G encodes a set of c.i assumptions
- Every distri for which G is an I-map must satisfy the assumptions
- This is key to understanding factorised representation
- Consider any distri P for which student BN Gstudent is an I-map.

- Decompose joint distribution via chain rule:-

$$P(I,D,G,L,S) = P(I)P(D|I)P(G|I,D)P(L|I,D,G)P(S|I,D,G,L) \quad (3.15)$$

- No assumptions, but not helpful
- Apply c.i. assumptions, induced by B.N.
- Since Gstudent is an I-map for our distri. P

(3.14) → $(I \perp D) \in I(P) \Rightarrow P(D|I) = P(D)$

(3.10) → $(L \perp I, D | G) \in I(P) \Rightarrow P(L|I,D,G) = P(L|G)$

(3.11) → $(S \perp D, G, L | I) \in I(P) \Rightarrow P(S|D,G,L,I) = P(S|I)$

- Hence $P(I,D,G,L,S) = P(I)P(D)P(G|I,D)P(L|G)P(S|I)$ (3.16)

- Any entry in joint distri can be computed as a product of factors, one for each var.
- Each factor ⟶ a conditional prob. of variable given network parents
- This factorisation applies to any distribution P for which Gstudent is an I-map

---

Formally:

Def 3.4 (factorisation)

- Let G be a BN graph over $X_1, \dots, X_n$
- We say a distri P over same space factorises according to G if P can be expres:-

$$P(X_1, \dots, X_n) = \prod_{i=1}^{n} P(X_i | Pa_{X_i}^G)$$

- chain rule for BN
- Individual factors $P(X_i | Pa_{X_i}^G)$ are CPDs (conditional prob distri.)

03.5. Bayesian Network

- A Bayesian network is a pair $B = (G, P)$ where P factorises over G, and where P is specified as a set of CPDs associated with G's nodes
- The distri P is often annotated P_B.
- Phenomenon holds for G structure more generally.

Theorem (3.1)

- Let G be a BN structure over a set of r.v.s. X, and let P be a joint distribution over the same space.

(**) If G is an I-map for P, then P factorises according to G.

Proof: - see Koller & Friedman (may use for exercises)

Comments: - C.I. assumptions implied by BN structure G allow factorisation of a distri P for which G is an I-map into small CPDs.

- (Ctd): 1.3.1. Shows direction of fundamental connection between the c.independencies encoded by the BN structure and the factorisation of distri into local probabilistic models / CPDs.

- C.I. $\Rightarrow$ factorisation

- Theorem 3.2 (converse: factorisation according to $G \Rightarrow$ C.I.)

- Let G be a BN structure over a set of r.v.s. X and let P be a joint distri. over same space.

(**) If P factorises according to G, then G is an I-map for P.

- 3.3 Independencies in graphs

- Graph structure $G \rightarrow$ encodes a set of c.i. assumptions $I_c(G)$
- Knowing only that a distribution factorises over G, we can conclude that distri satisfies $I_c(G)$. (i.e. $I_c(G) \subseteq I(P)$)
- (Q): Recall Jordan :- Are there any other independencies that we can 'read off' directly from G.
  - Are there other independences that hold for every distribution P that factorises over G?
  (a distri family)

(Q): A question about implied independencies i.e. $I(P) - I_c(G)$ (?)

D-separation

- When can we guarantee that an independence $(\underline{X} \perp \underline{Y} | \underline{Z})$ holds in a distri associated with a BN structure G.
- In interests of concision, I'm going to leave comments
- The next section is on the canonical local structures/independences; in Jordan, cascade, common parent, v-structure.
- Koller (2009) adds additional info/intuition and categorises as:-

1. Direct connection $X \to Y$
2. Indirect connection, (3-node networks)
   - i) Indirect causal
   - ii) Indirect evidential
   - iii) common cause
   - iv) common effect

   } each contains intuition

- Key is to note that the key event is whether a node is observed or unobserved (around which C.I statements, (dependencies)/decoupling is generated)
- Koller (2009): (D): when probabilistic influence can flow from X to Y via Z, then we have an active trail / blocking
- 10708 L2 notes: Better summary linking Bayes Ball to Koller's idea of active trails.
- (W): essentially: { Bayes Ball - reachability - ball - blocked (Jordan)
  - probabilistic influence - active-trails - blocked (Koller)

  ↳ Both encode intuitions about what leads to d-separation

General case:
- Longer trail $X_1 \rightleftharpoons \dots \rightleftharpoons X_n$
- For influence to flow from X_1 to X_n, it needs to flow through every node on trail.
- X_1 can influence X_n if every two edge trail $X_{i-1} \rightleftharpoons X_i \rightleftharpoons X_{i+1}$ along trail allows influence to flow.

Formally:

D.3.6.

Let G be a BN structure, and $X_1 \rightleftharpoons X_2 \rightleftharpoons \dots \rightleftharpoons X_n$ a trail in G.
Let $\underline{Z}$ be a subset of observed r.v.s.
The trail $X_1 \rightleftharpoons X_2 \rightleftharpoons \dots \rightleftharpoons X_n$ is active given $\underline{Z}$ if:-

- Whenever we have a v-structure $X_{i-1} \to X_i \leftarrow X_{i+1}$ then X_i or one of its descendants are in $\underline{Z}$.
- No other node along the trail is in $\underline{Z}$

If X_1 or X_n are in $\underline{Z}$, the trail is not active.

- Need to account for graphs with more than one trail between nodes

-d-separation (✱): d-separation / i) Bayes ball algorithm
\ ii) separation on moralised ancest. graph
\ iii) Active trails (iv). Multiple algos for same intuition

D.3.7 (d-separation)

-Let $\underline{X}, \underline{Y}, \underline{Z}$ be three sets of nodes in G.

-We say $\underline{X}$ and $\underline{Y}$ are d-separated given $\underline{Z}$ i.e. $\text{dsep}_G(\underline{X}; \underline{Y} \mid \underline{Z})$, if there is no active trail between any nodes $X \in \underline{X}$ and $Y \in \underline{Y}$ given $\underline{Z}$.

I(G) is the set of independencies that correspond to d-separation

$$I(G) = \{(\underline{X} \perp \underline{Y} \mid \underline{Z}) : \text{d-sep}_G(\underline{X}; \underline{Y} \mid \underline{Z})\}$$

This set is called the set of global Markov independences

The independencies in I(G) are precisely those that are guaranteed to hold for every distri over G.

---

some formalities (mathematical) regarding d-sep. (i.e. proof style)
esirable properties of d-separation

3. (Soundness of d-sep)

f a distri P factorises according to G, then $I(G) \subseteq I(P)$

ark: -Any independence reported by d-separation is satisfied by the underlying distri

- If two nodes X and Y are d-separated given some $\underline{Z}$, we are guaranteed they are c.i. given $\underline{Z}$.

---

pleteness - d-separation detects all possible independencies

: If two variables X and Y are independent given $\underline{Z}$, they are d-separated

defined in this form ⟶ no specification of distribution in which X and Y are independent

i.e. actually a statement which make a set of distri / distri family?

D 3.8. (Faithful)

- A distr P is faithful to G if, whenever $(X \perp Y \mid \underline{Z}) \in I(P)$, then $dsep_G(X; Y \mid \underline{Z})$.
- Any independence in P is reflected in the d-sep. properties of graph (?)
- A candidate formalisation of completeness is in terms of converse/contrapositive (Koller) ~ (Xing)

(*) for any distr P that factorises over G, we have that P is faithful to G

(*) i.e. if X and Y are not d-separated given $\underline{Z}$ in G; then X and Y are dependent in all distribut. P that factorise over G. ($L2_2$)

(CII) :- converse to soundness (Koller)
- If true, two together (which?) imply that for any P that factorises over G, we have $I(P) = I(G)$.

→ - Highly desirable property is false (?) which (?)

(*) :- Even if a distr factorises over G, it can still contain additional independencies

(*) not reflected in structure

- completeness property does not hold for this candidate definition of completeness

- settle for a weaker definition:-
  - If $(X \perp Y \mid \underline{Z})$ in all distrib P that factorise over G, then $dsep_G(X; Y \mid \underline{Z})$.
  - contrapositive :- If X and Y are not d-separated given $\underline{Z}$ in G; then X and Y are dependent in some distribution that factorises over G.

- Formally construct theorem:

1.3.4

- Let G be a BN structure.
- If X and Y are not d-separated given $\underline{Z}$ in G, then X and Y are dependent given $\underline{Z}$ in some distribution that factorises over G.

Proof :- See Koller (2009).

- Remark: completeness result tells us that our definition of I(G) is maximal. for any independence assertion not a consequence of d-separation in G, we can find a counterexample distr P that factorises over G.

1.3.5 (measure theoretic qual.)

- For almost all distributions P that factorise over G, that is for all distributions except for a set of measure zero in the space of CPD parametrisations, we have $I(P) = I(G)$.

- stronger than 1.3.4

(*)(*) - Results state that for almost all parametrisations P of the graph G (!!) (that is for almost all possible choices of CPDs for the variables), the d-separation test precisely characterise the independences that hold for P.

3.3.4 - I-equivalence

- $I(G)$ specifies a set of c.i. assertions associated with a graph.
- can abstract away details of graph structure and view as specification of indep. properties

(*) - very different BN structures can be equivalent in that they encode precisely the same set of c.i assertions (e.g. local structures/canonical maps).

D.3.9. (I-equivalence)

- 2 graph structures K_1 and K_2 over $\mathcal{X}$ are I-equivalent if $I(K_1) = I(K_2)$
- The set of all graphs over $\mathcal{X}$ is partitioned into a set of mutually exclusive and exhaustive I-equivalence classes.
- Which are the set of equivalence classes induced by the I-equivalence relation

---

(*) I-equivalence of 2 graphs $\Rightarrow$ any distri P that can be factorised over one of these graphs can be factorised over the other.

(*) - no intrinsic property of P that would allow us to associate it with one graph rather than an equivalent one.

(*) - important implications on our ability to determine dir. of influence

(L3) - 10-708 PGMs

30/04/2020

- Undirected Graphical Models
- Record intuitions, questions - be selective

(A) Review of independence properties of DAGs.

- The properties of DAGs require more reading of Koller.

(*): A fully connected DAG G is an I-map for any distr. $I(G) = \emptyset \subseteq I(P) \; \forall P$.

- Minimal I-map (A) (A1)
- Other characteristics → readings

(Q): How to find d-separation was a key question. Multiple ways.

1. Moralised ancestral graph

Above:

- Can be solve algorithmically

Ex: Bayes ball algorithm - mechanical protocol for detecting C.I.

- (W): Note that d-separation / Bayes ball is an encoding of obs. that simple appeals to 'blocking' do not hold for V-structures (conditioning on a variable)

- Key exception is V-structure
- Using Bayes ball algorithm, we build up d-separation an entire graph through consideration of each of these canonical graph structures / local motifs/patches.

I-Maps → see Koller (Q) (A2)

- 2 sets I(P), I(G) are equivalent?

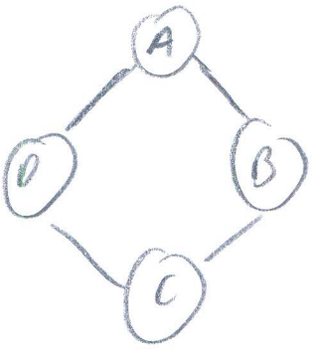
(Q): For any graph G and distr. P, do we always expect a P-map / equivalence.

- Theorem - NO!
- By counterexample
- Try and define CPD that satisfies these two C.Is. $A \perp C \mid \{B, D\}$; $B \perp D \mid \{A, C\}$.
- Constant: you must draw with DAGs.
- Are 2 maps BN1 and BN2 equivalent to the $I = (A \perp C \mid \{B, D\}, B \perp D \mid \{A, C\})$

i.e. Are there C.Is. in the graph structure not in the listed C.Is.
- BN2 - via v-structure - D⊥B.

(Ex.) There are no DAGs which encodes exactly the same set of conditional indpendences in this set. ⇒ not every distri has a P-map in DAG.

(Ex.): But $A \perp C \mid \{B, D\}$ and $B \perp D \mid \{A, C\}$ can be captured by:-



initiates

- Graph separation in UGMs

UGMs

- More expressive, precise about conditional indpendences.
- Pairwise relationships, no causality (no parents/ancestors).
- Assign 'scores' to configurations
- Not generative, i.e. using conditional probability distri.

CV problems → pixel labelling

Ex: seems like C.S. believe that all intuitions can be encoded algorithmically

- Grid model
- Ising model (magnetism)
- Probabilistic model with symmetrically connected r.vs.
  ↳ contingency tables to express preferences over patterns
- DGM have semantics of directionality, causality, temporality
- UGMs do not
- Maybe synchronic vs diachronic use?

- canonical model for Go → PGM before AlphaGo.
- information retrieval

$\psi(\cdot)$ is a table here

- canonical definitions → readings notes
- undirected graph H
- potential function ψ_c : mapping from config → no.
- potential functions
- are pre-probabilistic (do not need to conform to rules of probability)

- Gibbs distribution

x_1	x_2	$\psi_c(\underline{x})$
1	1	10
1	0	1
0	1	1
0	0	2

$\underline{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$

- (W): But later normalise (?) using a partition function
- (✡): Partition function - sum of product of all potentials over all possible configurations
  - i.e. across all subsets over which you have defined potential
- (✡) - define c appropriately i.e. not entire set of r.v.s.
- statistical physics → probabilistic graphical models !

- quantitative specifications - cliques, potentials

Ex: potential function defined over cliques (W): formal defi

- max clique - maximal fully connected subset of graph
- sub-clique - pairs of nodes and singletons

(?): why are cliques important?

- clique - totally connected - no condit. independence within the clique
- every configuration is possible and has to be 'honoured'
- large configuration with certain disconnections → there may be independences, certain configurations may not deserve attention.
- cliques - every configurations has to be associated with a characterisation (e.g. potential number or prob. mass)

Ex: only define potential functions on cliques

clique potentials - interp.

- (X)–(Y)–(Z) - not max. clique
- 2 potentials over (X,Z) and (Y,Z) ; or over singletons X, Y, Z

- via A number of ways breaking down joint P(X,Y,Z)
- Ex: Not sure what point about lack of correspondence between marginal, conditional probabilities, and potential functions
- (physics interp to of potential (e.g. singleton spin; pairwise magnetism)
- is useful for intuition
- (*) Forget the probabilistic semantics from DGMs for ψ_c

UGM - using max cliques

- potential ψ_c is a mapping from a tuple $\psi_c(x_{1,2,4})$
- $p(x_1, \ldots, x_4)$ - 4 dim table, 2 states for each r.v.
- represent $\psi_c(x_{1,2,4})$, $\psi_c(x_{2,3,4})$ → 2 3D tables

UGMs - using subcliques

- using pairwise potentials
- Pairwise Markov models (lazy)

UGM - canonical rep.

- (Q) is canonical used?
  - canonical repres. → potential functions for all cliques in the graph
- (Q) is overrep used?
  - (Ex) Not necessarily; think about NN as over-represented which have everything possible as placeholder to anticipate complexity; even though spec. of sparse model.

EX (Q): (L) (E3) HW: P_1 max clique, P_2 pairwise clique, $I(P_1) = I(P_2)$?
- same set of c.i.s. for 2 dists?
- Are no. ways dists specified in the 2 ways $P(P_1) = P(P_2)$ the same? ?

(11) - Independence properties

- conditional independence properties via graph separation
- global Markov independencies = global Markov property (A4) (A5) } - pick up on notes.
- weak Markov independencies (A6) (A6)

- in Markov networks, CCI are more intuitive - neighbour based
- Markov blankets

Ex: Graphical models unpopular nowadays due to deep learning; but certain properties are useful.

- Markov blankets in DGMS ⟶ drawing samples cond. on evidence (deep gen models)
  - difficult - uses $P(X_i | X_{i-1})$
  - gigantic, conditioning on many many things.

- If DGM can be modelled as a MN :-

$P(X_i | X_{-i}) \equiv P(X_i | MB_i)$ (a lot smaller) ⟶ (511) energy states in atomic bomb (!)

- statistical sampling algorithms used M.B. property to speed up comp.

- soundness & completeness → (A2) (A7)
- Analogies

Hammersley-Clifford

- for arbitrary potentials in Gibbs distr.
- Then the functional form of Gibbs distr is the ONLY WAY to write prop distrs on graph i.e. $p(x_1, \ldots, x_n) = \frac{1}{Z} \prod_{c \in C} \psi_c(x_c)$

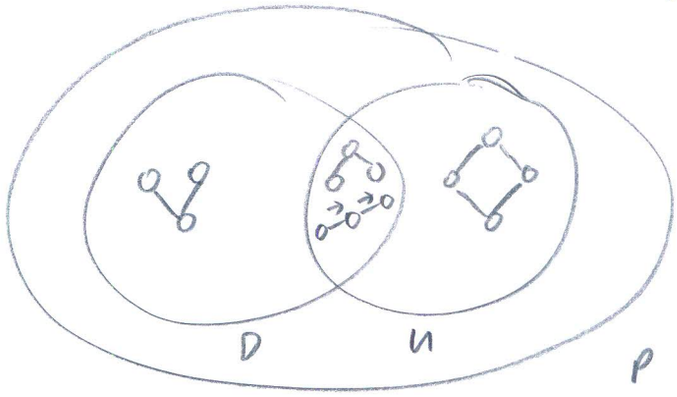
- (A2) (A3) + Theorem ↓
- (=) to factorisation law for BNs
- Can discover all cliques (max, sub) inside graph, write potential functions over cliques, multiply to get joint; captures all independences in a graph

---

Perfect maps

- High-level
- Ex: we know there are certain independences in a distr for which we cannot find a P-map in DGMS.

- similarly, there are certain c.i.s. that could exist in a distri P that we cannot find/express in UGM.
- Perfect-map is not an entitlement for arbitrary graphs
- see venn diagram (#X) (#Y)



- D cannot capture (diamond)
- U cannot capture V-structure

---

ex: specification of potential fns:
- when we specify CPDs (in DGM), we have probability constraints
- specifying potential fns in Gibb distribution → must be non-zero constraint

- there may be symmetries over 0 e.g. -1, 1
- Gibbs distri inadequate
↳ define potential functions ψ_c as exponentiated energy functions ϕ_c

- free energy / Boltzmann distri (stat phys)
- log-linear (statistics)
- allows specification of -ve nos (getting around constraint)

Boltzmann Machines
- Pairwise and singleton potentials ☑

(#): $P(x_1, \ldots, x_N) = \frac{1}{Z} \exp\left\{ \sum_{i,j} \phi_{ij}(x_i, x_j) + \sum_i \phi_i \right\}$

- Turn state variables → nos → allow direct definition of pairwise potentials, singleton pot.
- Allows linear algebraic specification (#) = $\begin{bmatrix} \alpha_1 & \theta_{12} & \theta_{13} & \cdots \\ \theta_{21} & \cdots & & \\ \vdots & & & \alpha_N \end{bmatrix}$
- μ is an offset → quadratic form for Eligy fn.

- C normalises the distri
- norm. constat

EX: Tabular expressions of pot. ψ_c
OR
EX: Physicists method
- directly define ψ_c as simple linear function over state values x

- Ising model
- Quadratic and singleton states

- RBMs
- important in ML/DL
- bipartite connection of hidden, visible
- Globalist defn (*) - singleton, singleton, pairwise
- very rich, deep connection to DNN
- Interface on RBM; computational graph very related to backprop in DNN

- computational and model analogies with DNN.

Properties of RBM

- Ex: why no directionality?
- Ex: ML scientists vs statisticians → after putting forward model, have to compute

(*) computational algorithm as important as model

- directionality has significant computational consequences.
- creates tons of V-structures → coupling → difficult

- Inuse Gibbs sampling possible due to this.

Ex: modelling mathematics tightly coupled with efficiency, comput. of algorithms

---

UGM semantics → constructive defn.

- RBM text modelling (topic modelling in UGM space)
- localist definitions give intuitive specifications to get globalist defn.

- Reference points:-

CRFs - Lafferty. very famous (difficult paper)
- HMM model with counterpart in undirected space (no directionality)

- Potential functions → very interesting interp.
- H-potentials → spell-checker e.g. g2 co-occurrence frequency/prob.
- Potentials can capture interesting global effects
- CRFs great template for feature engineering, language modelling

---

ex: Graphical models with c.i properties - neighbours - Markov property
- 3 special UGMs: 1. Ising model
  2. RBM
  3. CRFs.

(*) @: where does graph structure come from / arbitrarily? OR / data-driven?
- Graph structure itself can be learnt from data
- data-driven causality inference?

(EX): algorithms for unique discovery of structure exist (GOOD)
- Provably correct under certain conditions

(EX) For most conditions, cannot uniquely discover a structure
- many structures will give you same score you want to optimise
- causality is a statistical effect,