(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 $I = (A \perp C \mid \{B, D\}, B \perp D \mid \{A, C\})$

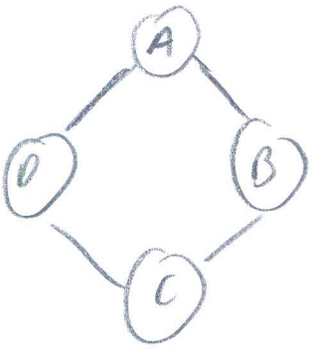
i.e. Are there C.Is. in the graph structure not in the listed C.Is.
- BN2 - via v-structure - D⊥B.

---

(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 $\psi_c$ : 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 $\psi_c$

UGM - using max cliques

- potential $\psi_c$ 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. $p(x_1, \ldots, x_n) = \frac{1}{Z} \prod_{c \in C} \psi_c(x_c)$

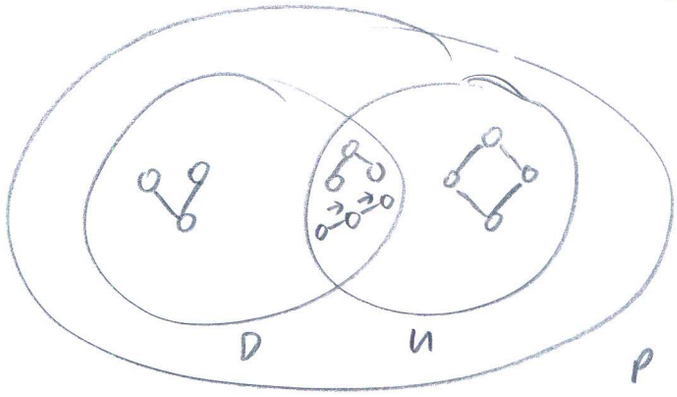
- (A2) (A3) + Theorem ↓
- (=) to factorisation law for BNs
- Can discover all cliques (max, sub) inside graph, write potential functions over cliques, multiply to get joint; captures all independences in a graph

---

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 $\psi_c$ as exponentiated energy functions $\phi_c$

- 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 (#) = $\begin{bmatrix} \alpha_1 & \theta_{12} & \theta_{13} & \cdots \\ \theta_{21} & \cdots & & \\ \vdots & & & \alpha_N \end{bmatrix}$
- $\mu$ is an offset → quadratic form for Eligy fn.

- C normalises the distri
- norm. constat

EX: Tabular expressions of pot. $\psi_c$
OR
EX: Physicists method
- directly define $\psi_c$ 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,