10-708 (L3)

Readings

- Kegequations; some intuition on thornier points

Kollie (2009) / Jordan (2003) (need to shorten + hence select + concise)

- BN cannot model a distri that satisfies $(A \perp C \mid \{B, D\})$ and $(B \perp D \mid \{A, C\})$ only.
- UGM → undirected edges (prob. interaction)
- similar to BN; parametrisation of MN — local interactions ↓ combine. — global model

- Product of local factors, normalised
- normalising constant → partition function (MRF in statistical physics)
- MN - connection between factorisation and independence properties

04.3 (Gibbs distri)

- An undirected graphical model represents a distri $p(x_1, \dots, x_n)$ defined by an undirected graph H, and a set of (positive) potential functions ψ_c associated with cliques of H s.t.

$$p(x_1, \dots, x_n) = \frac{1}{Z} \prod_{c \in C} \psi_c(x_c) \qquad Z = \sum_{x_1, \dots, x_n} \prod_{c \in C} \psi_c(x_c)$$

04.4 (MN factors.)

- we say a distri $p(x_1, \dots, x_n)$ with $\underline{\psi} = \{\psi_1(x_1), \dots, \psi_c(x_c)\}$ factorises over a MN H if each x_c $(c = 1, \dots, C)$ is a complete subgraph (*) of H
- Factors that parametrise MN are clique potentials

(*) - informally, subs. with idea of a maximal clique

(*) wlog, parametrisation using maximal cliques obscures structure in original set of factors

- informally: (Jordan)
  - clique (of a graph) → fully connected subset of nodes
  - maximal cliques → cliques that cannot be extended to include additional nodes without losing property of being fully connected

- formally :-
- For $G = \{E, V\}$, a complete subgraph (clique) is a subgraph $G' = \{V' \subseteq V, E' \subseteq E\}$ such that nodes V' are fully interconnected

- A maximal clique is a complete subgraph s.t. any superset $V'' \supset V'$ is not complete
- A sub-clique is not necessarily a maximal clique

! sub-cliques - can be edges, singletons

Koller (2009): see Box 4.B. - MN for CV

4.3. MN Independencies

- Similar to Kollerian pres. → Hows of prob. influence / active trails

D4.8

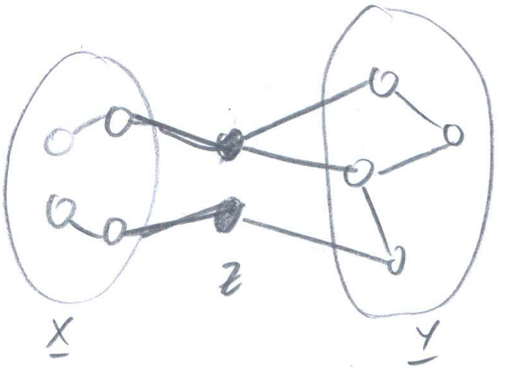
- Let H be an MN structure
- Let $X_1 - \dots - X_k$ be a path in H.
- Let $\underline{Z} \subseteq \underline{X}$ be a set of observed variables
- The path $X_1 - \dots - X_k$ is active given $\underline{Z}$ if none of the X_is $k=1,\dots,k$ is in $\underline{Z}$

Allows a definition of separation

D4.9 (separation/global independencies)

- We say a set of nodes $\underline{Z}$ separates $\underline{X}$ and $\underline{Y}$ in H, denoted $sep_H(\underline{X};\underline{Y}|\underline{Z})$, if there is no active path between any node $X \in \underline{X}$ and $Y \in \underline{Y}$ given $\underline{Z}$.
- We define the global independencies associated with H to be:-

$$\mathcal{I}(H) = \{(\underline{X} \perp \underline{Y} | \underline{Z}) : sep_H(\underline{X};\underline{Y}|\underline{Z})\}$$



Remark:

Independencies $\mathcal{I}(H)$ are precisely those guaranteed to hold for every distri P over H.

sep. criterion sound for detecting indep. properties over H.

(\\) BN ⇒ connection with independence prop implied by MN structure + factorising a distri over the graph (*)

(\\) of both representation theorems for BN, equivalence of:- Gibbs fact. of distri P over a graph H ⟺ H is an I-map for P (that is P satisfies Markov ass. $\mathcal{I}(H)$)

4.3.1.1. (soundness)

(II) to T.3.2. i.e. Gibbs distri satisfies indep. associated with a graph.

i.e. soundness of separation — (factorisation $\overset{4.1/3.2}{\rightleftharpoons}$ (C.I.) $\overset{4.2/3.1}{\rightleftharpoons}$ accord. G

T.4.1.

- Let P be a distri over X, and H an MN structure over X
- If P is a Gibbs distri that factorises over H, then H is an I-map for P

- other direction i.e. C.I. of distri → factorisation: Hammersley-Clifford Theorem
- unlike for BN; HCT does not hold in general.
- require additional assumption that P is a positive distri

T.4.2

- Let P be a positive distri over X, and H a Markov network graph over X.
- If H is an I-map for P, then P is a Gibbs distri that factorises over H.

(*). for positive distris; the global independencies imply that distri factorises according to the network structure.

(*) For this class of distributions, we have that a distribution P factorises over a Markov Network if and only if H is an I-map of P.

(**) lecture def (*) an MN H is an I-map for a distri P if $I(H) \subseteq I(P)$ i.e. P entails $I(H)$

(*) If P is a Gibbs distri over H, then H is an I-map of P.

- soundness of separation as criterion for detecting independencies in MN.
- any distri that factorises over G satisfies the independence ass. implied by seperation
- completeness - strong version of completeness does not hold
- It is NOT the case that every pair of nodes X and Y that are not separated in H are dependent in every distribution P which factorises over H (use naive)

T.4.3

- Let H be a MN structure.
- If X and Y are not separated given Z in H, then X and Y are dependent given Z in some distri P that factorises over H.

- same arguments as 7.3.5 to conclude:-

(**) for almost all distri P that factorise over H (all distr. except for a set of measure 0 in space of factor param.), we have $I(P) = I(H)$.

- our defn of $I(H)$ is maximal one.
- for any <u>indpendence assertion</u> that is not a consequence of separation in H, we can always find a counterexample distri P that factorises over H.

<u>4.3.2. Indep. revisited</u>

BN: local indep. (each node is indep. of nondesc. given parents)
global indep (motived by d-sep)

- showed that these are equiv, in the sense of one implies the other.

(Q): (11) to BN; can we provide local indep. motived by MN, analagously to local indep of BN

(*)(*): 3 diff poss. defin. of indpendencies associated with <u>network structure</u>
two local, one global in def 4.9.

<u>4.3.2.1. - Local Markov ass.</u>

D.4.10 → intuitively, two variables directly connected; potential for direct correlation in an unmediated way.
- conversely, two vars not directly linked, some way of reading c.i.
- X and Y indep given all other nodes.

<u>D.4.10 (Pairwise indpendencies)</u>

Let H be a MN
we define pairwise indpendencies associated with H to be:-

$I_p(H) = \{(X \perp Y \mid \mathcal{X} - \{X, Y\} : \{X, Y\} \notin \mathcal{E}_H\}$

D.4.11 → analogue to local indep. associated with B.N.

<u>4.11. (Markov blanket)</u>

For a given graph H, we define the Markov blanket of X in H, $MB_H(X)$ to be the <u>neighbours</u> of X in H.

we define the <u>local indpendencies</u> associated with H to be:-

$I_l(H) = \{(X \perp \mathcal{X} - \{X\} - MB_H\{X\} \mid MB_H(X) : X \in \mathcal{X})\}$

i.e. local independencies state that X is indep. of nodes in graph given immed. neighbours.

- we will show that these local indep. ass. hold for any distri that factorises over H, so that X's Markov blanket in H truly does sep. it from all other variables.

4.3.2.2. - Relationships between Markov properties

- 3 sets of indep. assertions assoc. with a network structure H.
- For general distri $I_p(H)$ is weaker than $I_\ell(H)$ is weaker than $I(H)$.
- ~~All 3 are eq~~ (*) All 3 are equivalent for positive distri

Prop 4.3

- For any MN H, and any distri P, we have that if $P \models I_\ell(H)$ then $P \models I_p(H)$

Prop 4.4

- For any MN H, and any distri P, we have that if $P \models I(H)$ then $P \models I_\ell(H)$

T 4.4

- Let P be a positive distri. If P satisfies $I_p(H)$, then P satisfies $I(H)$.

Corollary 4.1

- The following statements are equivalent for a positive distri P.

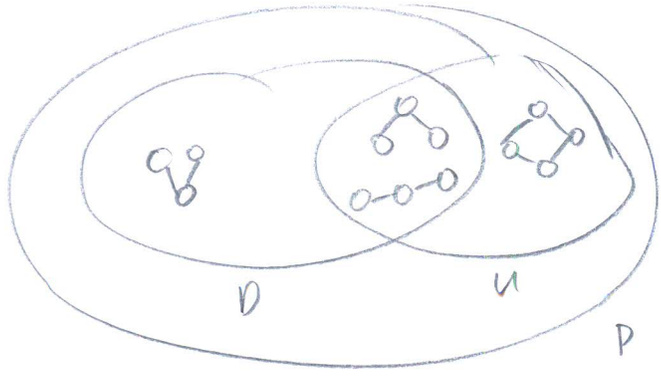
1. $P \models I_\ell(H)$
2. $P \models I_p(H)$
3. $P \models I(H)$

- ~~P~~ P-maps

- An MN H is a perfect-map for P if for any $\underline{X}, \underline{Y}, \underline{Z}$, we have that

$$sep_H(\underline{X} : \underline{Y} \mid \underline{Z}) \iff P \models (\underline{X} \perp \underline{Y} \mid \underline{Z})$$

- Theorem: not every distri has a perfect map as UGM



from hereon, Jordan (2003)

- exponential form

- constraining clique potentials to be +ve, unconstrained possibly

Q/H: what effect does this have on equivalence of local and global Markov properties?

- represent a clique potential as: $\psi_c(\underline{x}_c) = \exp\{-\phi_c(\underline{x}_c)\}$ (*)
- $\phi_c(\underline{x}_c)$ - a 'potential'
- Additive structure: $p(\underline{x}) = \frac{1}{Z}\prod_{c\in C}\exp\{-\phi_c(\underline{x}_c)\}$

$$= \frac{1}{Z}\exp\left\{-\sum_{c\in C}\phi_c(\underline{x}_c)\right\} = \frac{1}{Z}\exp\{-H(\underline{x})\}$$

'Boltzmann distri'

$H(\underline{x}) = \sum_{c\in C}\phi_c(\underline{x}_c)$

$H(\underline{x})$ - free energy

---

some notable forms / gen topologies:-

Boltzmann machines

fully connected graph, pairwise edge pot., binary valued nodes $(-1, 1)$

$$p(x_1,x_2,x_3,x_4) = \frac{1}{Z}\exp\left\{\sum_{i,j}\phi_{ij}(x_i,x_j)\right\} = \frac{1}{Z}\exp\left\{\sum_{i,j}\theta_{ij}x_ix_j + \sum_i \alpha x_i + C\right\}$$

call energy fn:-

$$H(\underline{x}) = \sum_{ij}(x_i-\mu)\theta_{ij}(x_j-\mu) = (\underline{x}-\underline{\mu})^T\underline{\underline{\Theta}}(\underline{x}-\underline{\mu})$$

new technique to learn this model, recover graph str. from data.

g models

odel energy of a physical system (atom interaction)

gular grid topology

ulti-state → Potts.

$$p(x) = \frac{1}{Z}\exp\left\{\sum_{i,j\in N_i}\theta_{ij}x_ix_j + \sum_i \theta_{i0}x_i\right\}$$

Restricted Boltzmann Machines

$$p(x,h|\theta) = \exp\left\{\sum_i \theta_i \phi_i(x_i) + \sum_j \theta_j \phi_j(h_j) \sum_{ij} \theta_{ij} \phi_{ij}(x_i,h_j) - A(\theta)\right\}$$

- see paper

CRFs

- undirected graph rep. ; encode cond. distr $p(y|x)$ y_i - target x - obs.

$$p_\theta(y|x) = \frac{1}{Z(\theta,x)\exp\left\{\sum_c \theta_c f_c(x,y_c)\right\}}$$

10-708 - (L4)

04/05/2020

(L4) - Exact inference, variable Elimination - You have notes already, record intuitions

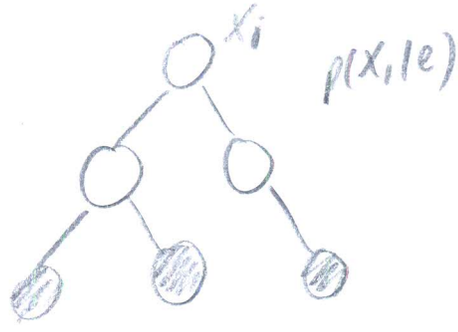
- Ex: 3 lectures on GM representation
- Ex: Focus on inference, then learning (which uses inference as a subroutine)

Query 1 - likelihood

- casino - quantit. specification of probability of fishy sequence of dice outcome - estimation/likelihood.
- marginal probability of evidence, likelihood
- marginalise over r.v.s. whose observations you do not have

Query 2 - C.P.

- use previous query as subsolution
- A posteior belief
- don't query all; interested in subset of hidden r.v.s.
- marginalise out hidden variables which we are not interested in



Applications of post. belief

Prediction: $P(C|A,B) = P(C|B)$

as $C \perp\!\!\!\perp A|B$

Diagnosis: $P(A|B,C) = P(A|B)$

as $C \perp\!\!\!\perp A|B$.

- Systematic way of using PGMs for potentially large GMs.
- E.g. social network
- Reduces necessity to deal with entire network due to C.I. properties

---

- PBN

- Optimisation semantics; representation learning/embedding is also an inference problem.

layers may correspond to different granularities of features. can be viewed as hidden r.v.s.

---

query 3 - MPA (maximum a posterior config. given evidence)
or (most probable assignment)

- Again previous query is subsoln.

(R): recall the distinctions made by JP for HMM

- which y gives highest C.P. mass

- Application - classification, explanation
- PGM may yield be useful in situations where simple linear class not ~~good~~ approp.

(*) Names

- MPA answer depends on framing

i) $\underset{y_1}{\operatorname{argmax}}\ p(y_1, y_2)$ $y_1^* = 1$ (expected value?)

ii) $\underset{y_1, y_2}{\operatorname{argmax}}\ p(y_1, y_2)$

$y_1^* = 0$ $y_2^* = 0$

- MPA is different depending on whether there is context in the form of y_2
- PGM makes this explicit → joint label or separate?
- Ex: y_1, y_2 connected in markov Blanket → 'use context'
  y_1, y_2 d-separated → 'no need for context'

Complexity of Inference

Ex. Proof of NP results ⟶ signpost of how to allocate your time

- computing $P(X = x \mid e)$ - NP hard in GM

(*) No assumptions about graphical models → no. of configuration increases exponentially.

(*) Need answer for any subset → need to truncate (without help of GM)

EX: Hardness does not mean not soluble

- 10-708 - In many cases, certain graph structures offer polynomial time solution methods, or approximate with poly. complexity

EX: class focuses on these generic (not special) cases ; tradeoff between rich graphical models and comput. feasibility

EX: deep learning paradigm ⟶ no thinking about algorithm ; not luxury in GM.

Approaches to inference

- Exact inference - guaranteed theoretically to get exact answer entailed by model
- Approx inference - approx 'true answer' ; more heavily used practically

EX: many deep learning methods/model architectures may be good; but inference algorithm not → hamper performance

---

Marginalisation/Elimination. - trivial statistically

- Q: Likelihood of protein E being active (binary state)

- $P(e)$

- marginalisation

- Q: How expensive is marginalisation, exponential⁴ (?) A: clarity

- computation difficult with very long chain (trad. summation, enum; non PGM)
- use chain decomp (PGM):-

- Two strategies:

1) Exponential cost - Hold node together, 2^4 configurations ; change only a record values of joint probability node (?) A: clarity complexity: $O(k^n)$ (naive)

$P(e) =$

2) (*) connection savings → not all c.p.s. function of a

(*) Note $\sum_a p(a)p(b|a) = \phi(b) = p(b)$ (i.e. a function of b)

K - no. of states
n - no. of r.v.s.

- more systematic method than 1)
- Repeat until we have $p(e) = \sum_d p(e|d)p(d)$
- one of summation has cost(K^2)
  cost of e.g. $\sum_a p(a)p(b|a)$ is $|b| \times |a|$ i.e. 4 (as a = 2 states)
- If for n eliminations → $O(nK^2)$ (quadratic complexity (quadratic) in biggest config of nodes K (linear in no. of nodes))

GM
- structure gives opportunities

HMM
- C.P. :- $p(y_i|x_1,\dots,x_T) = \sum_{\{y_t\}_{t=1}^T \setminus y_i} \cdots \sum p(y_1,\dots,y_T,x_1,\dots,x_T)$ → prob hidden state given entire seq.

- Viterbi law :→ How does this affect complexity of inference.

Ex: Illustrates inference complex. reduction for HMM (⑫⑬ - key junction for full underst.)

(*) $\sum_{y_2}\sum_{y_3}\cdots\sum_{y_T}\cdots\sum_{y_1} p(y_1)p(x_1|y_1)p(y_2|y_1)$

excl. y_i (query)

$m(x_1,y_2) = p(x_1,y_2)$ (?) ⑫⑬ - check

(*) $\sum_{y_3}\cdots\sum_{y_T}\cdots\sum_{y_2} p(x_2|y_2)p(y_3|y_2)f(x_1,y_2)$

$m'(x_1,x_2,y_3)$ (*)

- Repeat
- Algorithm is linear in no. of nodes, quadratic in no. of states (same as)
- This is the HMM forward algorithm (elimination algorithm)

- HMM - different elimination (sequence)

(*) $\sum_{y_1} \cdots \sum_{y_T} \underbrace{p(x_T|y_T)\,p(y_T|y_{T-1})}$

$= n(x_T, y_{T-1}) = p(x_T|y_{T-1})$

- ultimately: - $n' = p(x_{1:T}|y_{T-1})$ ... (w)
- Backward algorithm: Eliminating nodes from the tail. (15)
- n' takes semantics — item (prob. of late 1/2 of sequence and given latent states)
  of condit. prob. of partial sequence of observed given 1 hidden state.
- n item takes semantics: -
  of joint prob of 1st half of sequence given n hidden states.

(w)(15): Have to solidify understanding of HMM forward-backward
- major invention of the 70s (figuring it out algebraically not trivial)
- ordering the nodes for elimination → highly specialised c.i. insight.
- PGMS allow determination of poss. /feasibility via simplified ag.

- undirected chains } chain models have recurring pattern
- CRFS ↓

(*)
- Sum-product operation

$\sum_{z} \prod_{\phi \in F} \phi$

F - set of factors
ϕ - factors

- Ex: A key vehicle for understanding complexity for inference on GMS: -
  (*) count no. of summations, multiplications

(*) Appl. of sum-product to GMS beyond chains (*)

(*) write Query: $P(X_i, e) = \sum_{x_n} \cdots \sum_{x_2} \prod_i P(x_i|pa_i)$ (outline an ordering of summation signs)