

Lecture 11 - scribbles

Tuesday, October 9, 2018 14:15

today: directed graphical model

hwk 1 clarification:

Q2: $x_i \stackrel{iid}{\sim} \text{Mult}(1, \pi)$

\Rightarrow likelihood $p(x_1, \dots, x_n | \pi) = \prod_{i=1}^n \text{Mult}(x_i | 1, \pi)$

let $W = \sum_{i=1}^n x_i e_i$

different then $p(w_1=n_1, \dots, w_k=n_k | \pi) = \binom{n}{n_1, \dots, n_k} \prod_{j=1}^k \pi_j^{n_j}$

Graphical model

graph model \rightsquigarrow probability theory + C.S.

R.V.s \leftrightarrow graphs

graph \rightarrow efficient data structure

e.g. x_1, \dots, x_n R.V.s

$x_i \in \{0, 1\}$ $n \rightarrow 100$

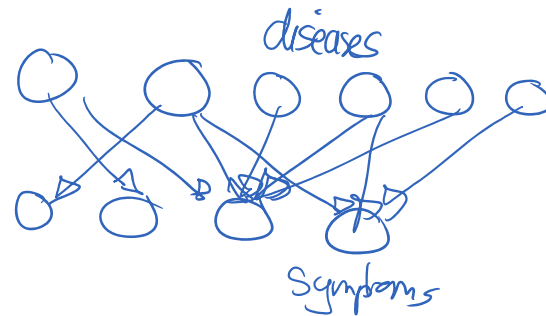
$\Rightarrow 2^{100}$ #'s table \rightsquigarrow untractable

graph theory review

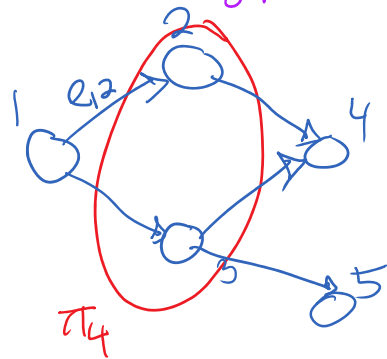
directed graph "digraph" $G = (V, E)$

$V = \{1, \dots, n\}$ "nodes/vertices"

QMR



directed graph $G = (V, E)$
 "digraph"



$\pi_i \triangleq \{j \in V : \exists (j,i) \in E\} =$
 set of parents of i

$V = \{1, \dots, n\}$ "nodes/vertices"

$E \subseteq V \times V$ "directed edges"

$e_{1,2} = (1,2)$

directed path $1 \rightsquigarrow 4 \rightarrow$ sequence of edges

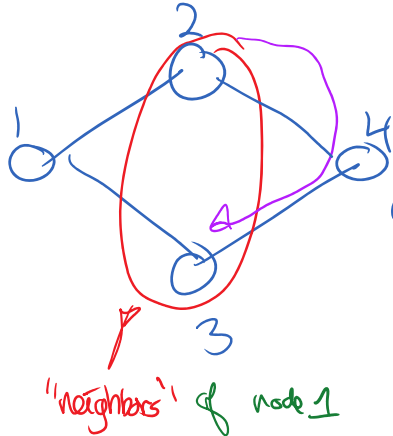
$(1,2), (2,4)$

or
 $(1,3), (3,4)$

undirected graph: $G = (V, E)$

elements of E are 2-sets
 (set of 2 elements)

(note: no self-loops)
 so $|e| = 2$



thus we have $\{i,j\} = \{j,i\}$

vs. $(i,j) \neq (j,i)$ [order matters]

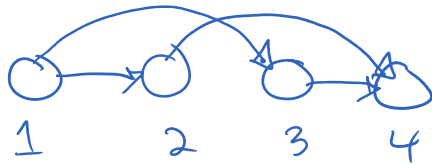
we have
 a (undirected) path from 2 to 3

neighbors replace the parents/children terminology from digraphs

def: DAG = directed acyclic graph = digraph with no cycle

def: an ordering $I: V \rightarrow \{1, \dots, n\}$ is said to be topological for G

def: an ordering $I: V \rightarrow \{1, \dots, n\}$ is said to be topological for G
 iff nodes in π_i appears before i in $I \quad \forall i$
 $(j \in \pi_i \Rightarrow I(j) < I(i))$



\rightarrow if topological ordering
 \Rightarrow all edges go from left to right
 ["no back edges"]

prop: digraph G is a DAG $\Leftrightarrow \exists$ a topological ordering of G

proof: \Leftarrow) trivial: no back edge \Rightarrow no cycle

\Rightarrow) use DFS algorithm to construct a top. sort. in $O(|E| + |V|)$

Notation for graph models:

n discrete R.v. X_1, \dots, X_n

discrete R.v. for simplicity

$V \leftarrow$ set of vertices
 one R.v. per node

conditional distribution concept.
 is tricky to formalize

see Borel-Kolmogorov paradox

joint $p(X_1=x_1, \dots, X_n=x_n) \stackrel{\text{shorthand}}{=} p(x_1, x_2, \dots, x_n) \quad x = x_{1:n}$
 $= p(x_V) \stackrel{(\cdot)}{=} p(x)$

for any $A \subseteq V$

$$p(x_A) = P\{X_i = x_i, i \in A\} = \sum_{x_{A^c}} p(x_A, x_{A^c})$$

subset of "subsets"

$$x_{\{1,2,4\}} \leftrightarrow \{x_1, x_2, x_4\}$$

summing over all possible values of $\{x_i\}_{i \in V \setminus A}$

revisit cond. indep.:

let $A, B, C \subseteq V$

$$\left(\sum_{B \subseteq A^c} \sum_{x_B} \dots \right)$$

* $X_A \perp\!\!\!\perp X_B \mid X_C$

$$(F) \Leftrightarrow p(x_A, x_B \mid x_C) = p(x_A \mid x_C) p(x_B \mid x_C) \quad \forall x_A, x_B, x_C \text{ st. } p(x_C) > 0$$

$$(C) \Leftrightarrow p(x_A \mid x_B, x_C) = p(x_A \mid x_C) \quad (\text{for } x_B, x_C \text{ st. } p(x_B, x_C) > 0)$$

* "marginal independence": $X_A \perp\!\!\!\perp X_B \mid \emptyset$

3 facts about C.I.:

1) can repeat variables in statement e.g. $X \perp\!\!\!\perp Y, Z \mid Z, W$ if you'd say (for convenience)

2) decomposition: $X \perp\!\!\!\perp (Y, Z) \mid W \Rightarrow X \perp\!\!\!\perp Y \mid W$
and $X \perp\!\!\!\perp Z \mid W$

and
 $X \perp\!\!\!\perp Z \mid W$

3) trick: ~~extra conditioning on both sides of equation~~
~~maintain true statement~~

(NOT REALLY!!!!)

NOTE: this is true for chain rule type of statements (like on the right)
BUT THIS TRICK IS NOT TRUE in general
 e.g. it does not work always on conditional independence statements

e.g. $p(x,y) = p(x|y)p(y)$ (always true)

$\Rightarrow p(x,y|z) = p(x|y,z)p(y|z)$ " "

Example:

$X \perp\!\!\!\perp Y \mid W$
 does not imply
 $X \perp\!\!\!\perp Y \mid W, Z$ (see hwk 1 Q1(d))

⊗ pairwise indep. $\not\Rightarrow$ mutual independence \rightarrow see lecture 3

⊗ chain rule
 (always true)

$p(x_V) = \prod_{i=1}^n p(x_i | x_{1:i-1})$

\rightarrow issue last cond. $p(x_n | x_{1:n-1})$
 table with 2^n entries

directed graph model

assume
 $\prod_{i=1}^n p(x_i | \pi_i)$

π_i parents of i in graph G

\rightarrow tables of $2^{\max_i |\pi_i| + 1}$

15h27

Directed graphical model

Let $G=(V,E)$ be a DAG

a directed graphical model (DGM) (associated with G) (aka Bayesian network)

is a family of distributions over X_V

here $\mathcal{L}(G) \triangleq \{ p \text{ is a dist. over } X_V : \exists \text{ local factors } f_i \text{'s s.t. } p(x_V) = \prod_{i=1}^n f_i(x_i | x_{\pi_i}) \}$ (not necessarily unique)

for potentials f_i s.t. 1) $f_i \geq 0$

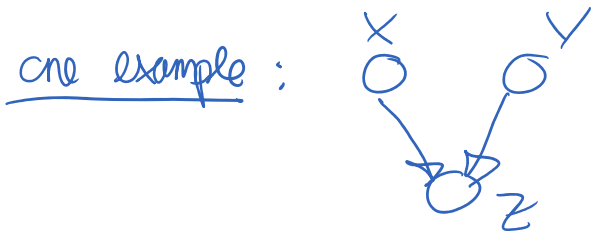
$f_i: \Omega_{x_i} \times \Omega_{x_{\pi_i}} \rightarrow [0, 1]$

2) $\forall x_{\pi_i}, \sum_{x_i} f_i(x_i | x_{\pi_i}) = 1$

f_i is like a CPT (conditional probability table)

terminology: if we can write $p(x_V) = \prod_i f_i(x_i | x_{\pi_i})$ this is where G comes into play

then we say that " p factorizes according to G "



$\exists f_i \text{ s.t.}$

$$p \in \mathcal{P}(G) \iff p(x, y, z) = f_x(x) f_y(y) f_z(z | x, y)$$

[very soon: we show that $p(x_i | x_{\pi_i}) = f_i(x_i | x_{\pi_i})$]

"leaf picking" property fundamental property of DGM

$p(x) \in \mathcal{P}(G)$ let n be a leaf in G (i.e. n is not a parent of anything)
 a) then $p(x_{1:n-1}) \in \mathcal{P}(G - \{n\})$
 remove n from G

b) if $p(x_{1:n}) = \prod_{i=1}^n f_i(x_i | x_{\pi_i}) \Rightarrow p(x_{1:n-1}) = \prod_{i=1}^{n-1} f_i(x_i | x_{\pi_i})$

proof:

$$p(x_n, x_{1:n-1}) = f_n(x_n | x_{\pi_n}) \prod_{j \neq n} f_j(x_j | x_{\pi_j})$$

no n in any π_j

$$p(x_{1:n-1}) = \sum_{x_n} p(x_n, x_{1:n-1}) = \left(\sum_{x_n} f_n(x_n | x_{\pi_n}) \right) \left(\prod_{j \neq n} f_j(x_j | x_{\pi_j}) \right)$$

by definition no x_n there

proposition: if $p \in \mathcal{P}(G)$

let $\{f_i\}$ be a factorization w.r. to G for p

then $\forall i, p(x_i | x_{\pi_i}) = f_i(x_i | x_{\pi_i}) \quad \forall x_{\pi_i}$

i.e. factors are correct conditionals

proof:

WLOG, let $\xi_{1, \dots, n}$ be a top sort. (possible since G is a DAG)

Let i be fixed, we want to prove that $p(x_i | x_{\pi_i}) = f_i(x_i | x_{\pi_i})$

• n is a leaf \rightarrow pluck it to get $p(x_{1:n-1}) = \prod_{j=1}^{n-1} f_j(x_j | x_{\pi_j})$

factorization for a DAG
where x_{n-1} is now a leaf

pluck $n-1$ as well!

keep doing this [in reverse order $n-1, n-2, \dots, i+1$] to get

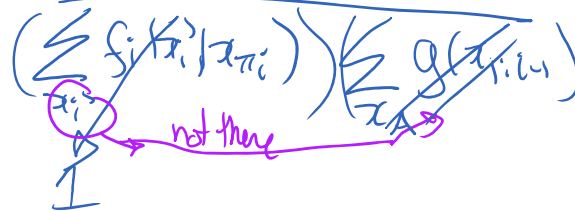
$$p(x_{1:i}) = f_i(x_i | x_{\pi_i}) \prod_{j=i}^{n-1} f_j(x_j | x_{\pi_j})$$

$$\stackrel{\Delta}{=} g(x_{1:i-1}) \leftarrow \text{ie. no } x_i \text{ by top sort property?}$$

partition $1:i$ as $\{i\} \cup \pi_i \cup A$ [ie. $A \stackrel{\Delta}{=} 1:i-1 \setminus \pi_i$]

$$p(x_{1:i}) = p(x_i, x_{\pi_i}, x_A)$$

$$p(x_i | x_{\pi_i}) = \frac{\sum_{x_A} p(x_i, x_{\pi_i}, x_A)}{\sum_{x_A} \sum_{x_i} p(x_i, x_{\pi_i}, x_A)} = f_i(x_i | x_{\pi_i}) \frac{\sum_{x_A} g(x_{1:i-1})}{\sum_{x_A} g(x_{1:i-1})} = f_i(x_i | x_{\pi_i})$$



\rightarrow we can substitute into

⊥

* we can safely write

len 11

$$\mathcal{J}(G) = \left\{ p : p(x_V) = \prod_{i=1}^n p(x_i | x_{\pi_i}) \right\}$$

Properties:

adding edges \Rightarrow more distributions

ie. $G = (V, E)$ and $G' = (V, E')$ with $E' \supseteq E$
then $\mathcal{J}(G) \subseteq \mathcal{J}(G')$

examples:

- $E = \emptyset$
(trivial graph)

$\Rightarrow \mathcal{J}(G)$ contains only fully independent distributions

$$\text{ie. } p(x_V) = \prod_{i=1}^n p(x_i)$$

\uparrow
 $p(x_i | \emptyset)$

- complete digraph (DAG)

for all $i \neq j$ either $(i, j) \in E$
or $(j, i) \in E$



get wlog $p(x_V) = \prod_{i=1}^n p(x_i | x_{1:i-1})$

(like chain rule)
 \Rightarrow all distributions are in $\mathcal{J}(G_{\text{complete}})$
on x_V

~~*~~ removing edges impose more factorization constraints
(and thus cond. indep. assumptions)

Basic 3 nodes graphs

Basic 3 nodes graphs

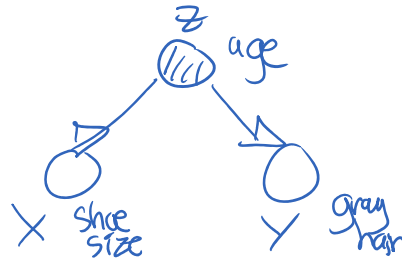
factorization \Leftrightarrow cond. indep assumption

1) Markov chain



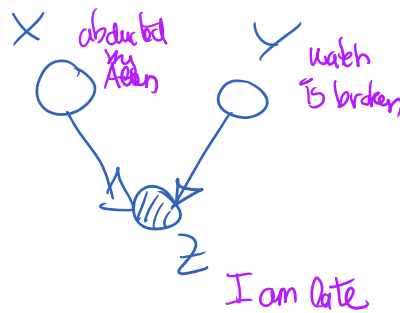
here "future" "past" "present"
 $Y \perp\!\!\!\perp X \mid Z$
 but $X \not\perp\!\!\!\perp Y$ (for some $p \in \mathcal{P}(G)$)

2) latent cause
 (hidden variable)



$X \perp\!\!\!\perp Y \mid Z$
 but $X \not\perp\!\!\!\perp Y$

3) explaining away
 competing effect
 (V-structure)



$X \perp\!\!\!\perp Y$
 but $X \not\perp\!\!\!\perp Y \mid Z$

non-monotonic property of conditioning:

$$p(\text{alien}) \text{ is tiny}$$

$$p(\text{alien} \mid \text{late}) > p(\text{alien})$$

$$p(\text{alien} \mid \text{late, broken watch}) < p(\text{alien} \mid \text{late})$$