Cours 4 - scribbles

Tuesday, October 20, 2015
23:19

today:
- finish DGM
- undirected GM
- exponential family

---

3 facts C.I.

1) can repeat variables $X \perp\!\!\!\perp Y, Z \mid Z, W$

2) decomposition: $X \perp\!\!\!\perp \underbrace{Y, Z}\mid W$

$$\Rightarrow \begin{cases} X \perp\!\!\!\perp Y \mid W \\ \text{and} \\ X \perp\!\!\!\perp Z \mid W \end{cases}$$

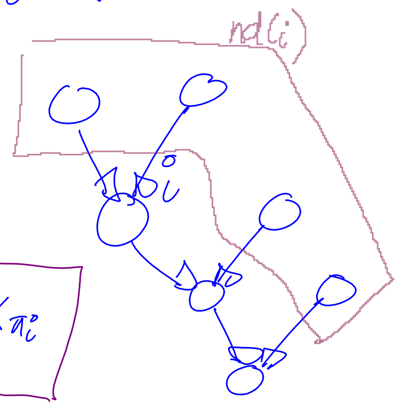
3) trick: extra-conditioning on both sides of eqn' doesn't change anything

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

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

---

let $nd(i) \triangleq \{ j : \text{no path from } i \to j \}$
"non-descendants"



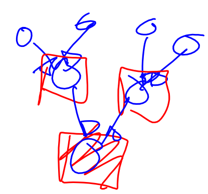
prop: $\boxed{p \in \mathcal{L}(G) \iff X_i \perp\!\!\!\perp X_{nd(i)} \mid X_{\pi_i} \quad \forall i \in V}$

proof:

$\Rightarrow$) key pt.: let i be fixed

$\exists$ a top. sort s.t. $nd(i)$ are just before i

i.e. $(\underline{nd(i)}, i, V \setminus (\{i\} \cup nd(i)))$



let $A \triangleq nd(i) \setminus \pi_i$

(plucking leaves) [marginalizing out $x_{V \setminus (\{i\} \setminus nd(i))}$]

$$p(x_i, x_{\pi_i}, x_A) = p(x_i | x_{\pi_i}) \prod_{j \in A} p(x_j | x_{\pi_j})$$

$$p(x_i | x_{nd(i)}) = p(x_i, \overbrace{x_{\pi_i}, x_A}^{x_{nd(i)}}) = p(x_i | x_{\pi_i}) \prod p(x_j | x_{\pi_j})$$

$$\overline{p(x_{\pi_i}, x_A)} \quad \underbrace{\sum_{x_i}[p(x_i|x_{\pi_i})]}_{1} \overbrace{\prod_{j\in nd(i)} \prod}^{j\in nd(i)} \underbrace{\quad}_{\text{no } x_i}$$

$$= p(x_i | x_{\pi_i})$$

$\Rightarrow \quad X_i \perp\!\!\!\perp X_{nd(i)} | X_{\pi_i}$

$\Leftarrow$ let $1{:}n$ is a top sort.

then $\{1, \dots, i-1\} \subseteq nd(i)$

(suppose $j \in \{1, \dots, i-1\}$ and $j \notin nd(i)$

$\Rightarrow \exists$ path from $i \rightsquigarrow j$ i.e. I have $(\dots, j, \dots, i, \dots, k \cdots \ell)$)

contradicts top-sort.

$$p(x_V) = \prod_{i=1}^{n} p(x_i | \underbrace{x_{1:i-1}}_{\subseteq nd(i)}) \quad \text{(chain rule)}$$

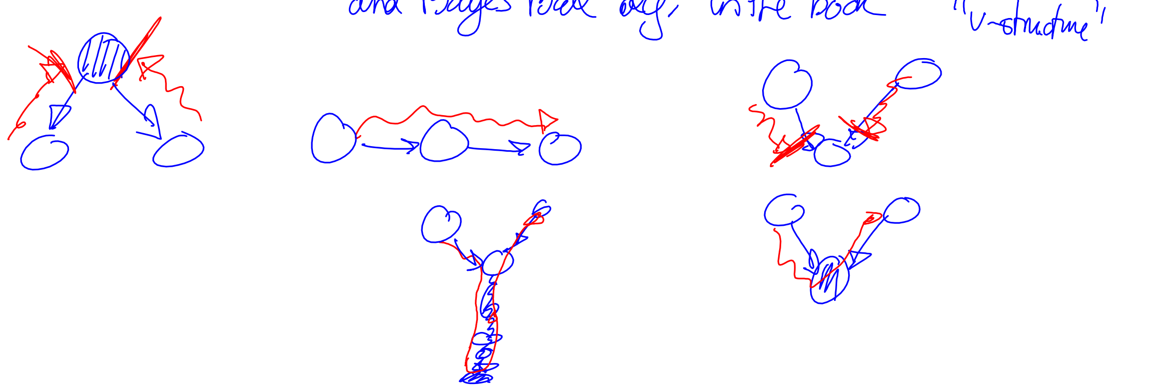
$$= \prod_{i=1}^{n} p(x_i | x_{\pi_i}) \quad \text{[by cond. indep.]}$$

$$\in \mathcal{L}(G) \quad //$$

$\circledast$ any other cond. indep properties beyond $X_i \perp\!\!\!\perp X_{nd(i)} | X_{\pi_i}$. for DGM G?

yes? $\rightarrow$ see d-separation and Bayes Ball alg. in the book

"v-structure"



<u>other properties</u>

inclusion: $E \subseteq E'$ then $\mathcal{L}(G) \subseteq \mathcal{L}(G')$

reversal: if G is a directed tree [or forest]

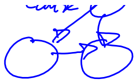
$(|\pi_i| \leq 1$, i.e. no v-structure)

then consider G' <u>any</u> other orientation of the tree

[take undirected tree version of G

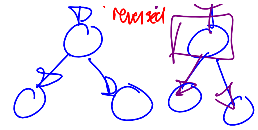
root new root

unobserved latent



and pick another root to orient

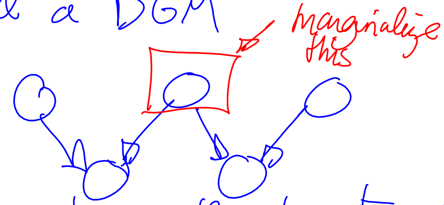
$$\mathcal{L}(G) = \mathcal{L}(G')$$



$\rightsquigarrow$ why direction of edge is <u>not</u> causal in a chain

marginalizing

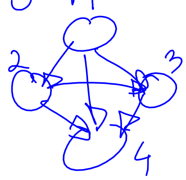
- marginalize leaves $\longrightarrow$ still a DGM
- it's not true in general



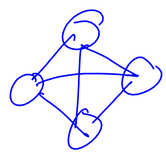
set of distributions obtained is <u>not</u> exactly characterized by a DGM

complete graph

Digraph:



undirected



(<u>clique</u>)
$\hookrightarrow$ set of of nodes with edge between every pair

i.e. C clique $\Leftrightarrow$ $i \neq j \in C$ $\exists \{i, j\} \in E$

undirected graphical model

$\rightarrow$ aka Markov random field or Markov networks

let $G = (V, E)$ be <u>undirected</u> graph

let $\mathcal{C}$ the set of <u>cliques</u> of G

then $\mathcal{L}(G) \triangleq \left\{ p : p(x) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C) \right\}$

where $\psi_C(x_C) \geq 0$ "potential"

$Z \triangleq \sum_x \left(\prod_{C \in \mathcal{C}} \psi_C(x_C) \right)$ "partition fct" or normalizer

<u>notes</u>:
- $\psi_C(x_C)$ is <u>not</u> related directly to $p(x_C)$ unlike in DGM
- $\psi'_C(x_C) = \text{constant} \cdot \psi_C(x_C)$ doesn't change anything
- sufficient to only consider $\mathcal{C}_{max}$ = set of maximal cliques

i.e. $C' \subseteq C$ can redefine $\psi_C^{new}(x_C) = \psi_C(x_C) \cdot \psi_{C'}(x_{C'})$

as before: $E \subseteq E' \Rightarrow \mathcal{F}(G) \subseteq \mathcal{F}(G')$

$E = \emptyset \rightsquigarrow \mathcal{F}(G) = \{p : \text{all } X_i \text{ are independent}\}$

$E = \underset{\text{(i.e. complete graph)}}{\text{all pairs}}$ $\quad \mathcal{F}(G) = \{\text{all distributions}\}$

if $\psi_c(x_c) > 0 \ \forall x_c \ \& \ c$ write $p(x) = \exp\Big(\underbrace{\sum_{c \in \mathcal{C}} \log \psi_c(x_c)}_{\text{negative energy function}} - \log Z\Big)$

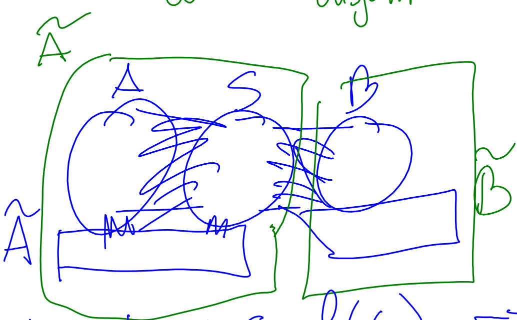
"exp. family"

conditional indep.

def: p satisfies global Markov property

iff $\forall \underset{\text{disjoint}}{A, B, S} \subseteq V$ s.t. S separates A from B in G (any path from A to B passes through S)

then $X_A \perp\!\!\!\perp X_B \mid X_S$



prop: $p \in \mathcal{F}(G) \Longrightarrow p$ satisfies global Markov property

proof: WLOG $A \cup B \cup S = V$ $\qquad$ $\& \triangleq \& \triangleq$ "and"

$[\ \tilde{A} \triangleq A \cup \{a \in V : a \ \& \ A \text{ are } \underline{\text{not}} \text{ separated by } S\} \setminus S$

$\tilde{B} \triangleq V \setminus (S \cup \tilde{A}) \Rightarrow \tilde{A} \ \& \ \tilde{B}$ are S-separated $]$

$\star$ let $C \in \mathcal{C}$; then we can't have both $C \cap A \neq \emptyset$ and $C \cap B \neq \emptyset$

$$p(x) = \frac{1}{Z} \underbrace{\prod_{\substack{C \subseteq A \cup S \\ C \in \mathcal{C}}} \psi_C(x_C)}_{} \ \underbrace{\prod_{\substack{C' \subseteq B \cup S \\ C' \not\subseteq S \\ C' \in \mathcal{C}}} \psi_{C'}(x_{C'})}_{}$$

$$= f(x_A, x_S)\, g(x_B, x_S)$$

$$p(x_A \mid x_S) \propto \underbrace{\sum_{x_B} f(x_{A \cup S})\, g(x_{B \cup S})}_{p(x_A, x_S)} = f(x_{A\cup S}) \overbrace{\sum_{x_B} g(x_{B \cup S})}^{\text{constant w.r.t. } x_A}$$

$$p(x_A \mid x_S) = \frac{f(x_A, x_S)}{\sum_{x_A'} f(x_A', x_S)} \qquad \underline{\text{similarly}}: \ p(x_B \mid x_S) = \frac{g(x_{B \cup S})}{\sum_{x_B'} g(x_B', x_S)}$$

$p(x_A \mid x_S)\, p(x_B \mid x_S) = \ldots f(x_A, x_S)\, g(x_B, x_S) \ldots = p(x_A, x_B \ldots$

$$p(x_A|x_S)\,p(x_B|x_S) = \frac{f(x_A,x_S)\,g(x_B,x_S)}{\sum_{x_A'}\sum_{x_B'} f(x_A',x_S)\,g(x_B',x_S)} \quad \left] \frac{p(x_A,x_B,x_S)}{p(x_S)} \right.$$

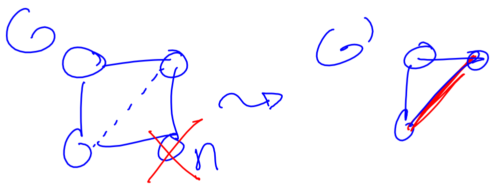
$$= p(x_A,x_B|x_S) \Rightarrow X_A \perp\!\!\!\perp X_B \mid X_S \quad //$$

thm.: Hammersley-Clifford

if $\forall x\ p(x)>0$, then $p \in \mathcal{L}(G) \Leftrightarrow p$ satisfies global Markov property

marginalization:

let $V' = V\setminus\{n\}$ E' = edges of G connecting all the neighbors of n and removing n



$$p \in \mathcal{L}(G) \Rightarrow p(x_{1:(n-1)}) \in \mathcal{L}(G')$$

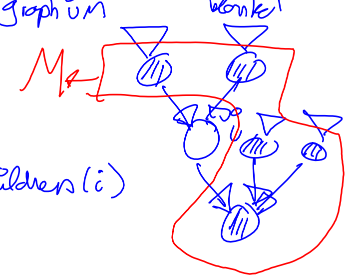
(true for any a $\Rightarrow$ we get closure)

directed vs. undirected GM

def: Markov blanket for i is the smallest set of nodes $M \subseteq V$ s.t. $X_i \perp\!\!\!\perp X_{\underbrace{V\setminus\{i\}}_{\text{rest of graph UM}}} \mid X_M$ (Markov blanket)

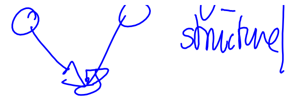
for UGM:
$M = \{j : \{i,j\} \in E\}$ (neighbors of i)

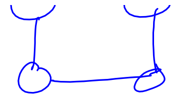
for DGM
$M = \pi_i \cup \text{children}(i) \cup \pi_j \mid j \in \text{children}(i)$



	DGM	UGM
factorization	$p(x) = \prod_i p(x_i \mid x_{\pi_i})$	$p(x) = \frac{1}{Z}\prod_{c\in\mathcal{C}} \psi_c(x_c)$
cond. indep.	d-separation [more than $X_i \perp\!\!\!\perp X_{nd(i)} \mid X_{\pi_i}$]	separation $X_A \perp\!\!\!\perp X_B \mid X_S$
marginalization	not closed in general [fine for a leaf]	closed

inference :





question: G is a DAG; when can we transform to equivalent UGM?

<u>def</u>: for G a DAG, call $\bar{G}$ the <u>moralized graph of G</u>

where $\bar{G}$ is undirected graph with same V

$$\bar{E} = \{ \{i,j\} : (i,j) \in E \} \quad \text{) undirected version of } E$$

$$\cup \{ \{k,l\} : k,l \in \pi_i \text{ for some } i \} \quad \text{) "moralization"}$$

no addition ($\emptyset$) when no "v-structure"

<u>prop</u>: for DAG G with <u>no v-structure</u> [forest]

then $\mathcal{L}(G) = \mathcal{L}(\bar{G}) = \mathcal{L}(\text{undirected}(G))$

but in general, only have $\mathcal{L}(G) \subseteq \mathcal{L}(\bar{G})$

<u>note</u>: not all subsets of cond. indep. statements can be precisely described by a graph model

[e.g. XOR example where pairwise indep. <u>but not</u> mutual indep]

$X_1 \perp\!\!\!\perp X_2$

$X_2 \perp\!\!\!\perp X_3$ $\quad X_1 \perp\!\!\!\perp X_3$ $\quad$ but not $X_1 \perp\!\!\!\perp X_2, X_3$

$[X_3 = X_1 \text{ xor } X_2]$

<u>general issues in this class</u>

A) representation / parametrization

↳ DGM, UGM   ↳ exp. family

B) inference (computing $p(x_A | x_B)$)

↳ sum-product alg.

C) statistical estimation

→ maximum likelihood

→ " entropy

<u>KL & MLE</u>

let $\mathcal{X}$ be discrete set

$$KL(p \| q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

n observations $x_1, \dots, x_n$

define empirical distribution $\hat{p}_n$ dist. over $\mathcal{X}$

with p.m.f. (proba. mass fct.) $\hat{p}_n(x) \triangleq \frac{1}{n} \sum_{i=1}^{n} \delta(x - x_i)$

Kronecker-delta fct.

$$\delta(z) = \begin{cases} 1 & \text{if } z = 0 \\ 0 & \text{o.w.} \end{cases}$$

prop.: $\{p_\theta\}_{\theta \in \Theta}$ parametric family on $\mathcal{X}$

then ML for p_θ $\Longleftrightarrow$ $\min KL(\hat{p}_n \| p_\theta)$

proof:

$$KL(\hat{p}_n \| p_\theta) = \sum_{x \in \mathcal{X}} \hat{p}_n(x) \log \frac{\hat{p}_n(x)}{p_\theta(x)}$$

$$= -H(\hat{p}_n) - \sum_x \hat{p}_n(x) \log p_\theta(x)$$

$$= -H(\hat{p}_n) - \frac{1}{n} \sum_x \sum_{i=1}^{n} \delta(x - x_i) \log p_\theta(x)$$

$$= -H(\hat{p}_n) - \frac{1}{n} \sum_{i=1}^{n} \underbrace{\left(\sum_{x \in \mathcal{X}} \delta(x - x_i) \log p_\theta(x) \right)}_{\log p_\theta(x_i)}$$

$$= \text{const.} - \ell(\theta)$$

//

max. entropy principle: (different than ML in general)

idea: consider $\mathcal{P}(\mathcal{X})$ a subset of distributions on $\mathcal{X}$ which satisfies some constraints (usually from data)

- pick $p \in \mathcal{P}(\mathcal{X})$ which maximize entropy

choose $\hat{p}$ by solving $\underset{p \in \mathcal{P}(\mathcal{X})}{\arg\max}\, H(p) = \underset{p \in \mathcal{P}(\mathcal{X})}{\arg\min}\, KL(p \,\|\, \text{uniform distribution on } \mathcal{X})$

Exponential family

a (flat/canonical) exponential family on $\mathcal{X}$

is a set of distributions defined by two quantities

I) $\underbrace{h(x)}_{\text{reference density}}\ \underbrace{d\mu(x)}_{\text{base measure}} \rightarrow$ reference measure on $\mathcal{X}$

base measure $\leftarrow$ counting / Lebesgue

II) $T: \mathcal{X} \to \mathbb{R}^p$ called "sufficient statistics" vector (aka feature vector)

members of family have dist.

$$p(x;\eta)\, d\mu(x) = \exp\left(\eta^T \boxed{T(x)} - A(\eta)\right) \boxed{h(x)\, d\mu(x)}$$

"canonical" parameter

"log-normalizer" aka, log-partition fct, cumulant generating fct.

defining pieces

* want $1 = \int_{\mathcal{X}} p(x;\eta)\, d\mu(x)$

$$= \int_{\mathcal{X}} \exp(\eta^T T(x))\, e^{-A(\eta)}\, h(x)\, d\mu(x)$$

$$\Rightarrow A(\eta) \triangleq \log\Big(\underbrace{\int_{x\in\mathcal{X}} \exp(\eta^T T(x))\, h(x)\, d\mu(x)}_{Z(\eta)} \Big)$$

domain $\Omega \triangleq \{\eta \in \mathbb{R}^p \mid A(\eta) < \infty\}$

* more generally, consider reparameterizations as a subset of the family by defining mapping

$$\eta: \Theta \to \Omega$$

consider $p(x;\theta) \triangleq p(x;\eta(\theta))$ for $\theta \in \Theta$

parameter

("curved exp. family" if $\eta(\Theta)$ is a curved manifold in Ω)

example: (multinomial family)

$X \sim \mathrm{Mult}(1,\pi)$; $\mathcal{X} = \{0,1\}^K$

$$\pi \in \Delta_K = \Big\{ \pi : \sum_{l=1}^{K} \pi_l = 1,\ \pi_l \geq 0 \Big\}$$

then for valid x (i.e. indicator vector)

$$p(x;\pi) = \prod_{c=1}^{K} \pi_c^{x_c} = \exp\left(\sum_{c=1}^{K} (\log \pi_c)\, x_c\right)$$ (suppose $\pi_c > 0$)

$$= \exp(\langle \eta(\pi), x\rangle) \quad -0 \;\; A(\eta(\pi))$$

where $\eta_c(\pi) = \log \pi_c$

$T(x) = x$

$d\mu(x) =$ counting measure on $\mathcal{X}$

$h(x) = \mathbb{1}\{x$ has exactly one entry equals to $1\}$

here $A(\eta(\pi)) = 0$ — here, $\Theta = \mathrm{int}(\Delta_K)$, $A(\eta(\theta)) = 0 \;\; \forall \theta \in \Theta$

$$A(\eta) = \log\left(\sum_{x\in\mathcal{X}} h(x) \exp(\langle \eta, x\rangle)\right)$$

$$= \log\left(\sum_{c=1}^{K} \exp(\eta_c)\right) < \infty \quad \forall \eta \in \mathbb{R}^K$$

$\Omega = \mathbb{R}^K$

$(\eta(\pi))_c = \log \pi_c$

thus $A(\eta(\pi)) = \log\Big(\underbrace{\sum_c \pi_c}_{1}\Big) = 0$

<u>remark:</u> $\Theta \to$ dimension $k-1$

$\Omega \to$ " K

for any x s.t. $h(x) \neq 0$ — $\sum_{c=1}^{K} x_c = 1$

i.e. $\underbrace{\sum_{c=1}^{K} T_c(x) - 1 = 0}$ affine linear dep. between components of T

multiple η's map to same distribution

(no identifiability) "overparameterized"

[not <u>minimal</u>]

for multinomial, minimal exp-family $T(x) = \begin{pmatrix} x_1 \\ \vdots \\ x_{K-1} \end{pmatrix}$