Lecture 10 - scribbles

Tuesday, October 4, 2016
14:17

today: directed graph model

general issues in this class

A) representation → DGM
                  ↳ UGM
   parameterization → exponential family
   ] probabilities

B) inference computing $p(x_Q \mid x_E)$ ("query", "evidence")
   ↳ sum-product
     junction alg.

C) statistical estimation → MLE
   maximum entropy
   method of moments

Notation:

n discrete R.V. $X_1, \dots, X_n$ — V = set of vertices; one R.V. per node

joint $p(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = p(x_1, \dots, x_n)$

$= p(x_V) = p(x)$

for any $A \subseteq V$ (marginal on x_A)

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

(x_A: subset of "subscripts"; x_V; $x_{\{1,2,4\}} \leftrightarrow \{x_1, x_2, x_4\}$; $\sum_{x_{A^C}}$ ← summing over all possible values for $\{x_i\}_{i \in V \setminus A}$)

$1{:}n \triangleq \{1, \dots, n\}$

conditional independence: $X_A \perp\!\!\!\perp X_B \mid X_C$ $\quad A, B, C \subseteq V$

(F) $\Leftrightarrow$ $p(x_A, x_B \mid x_C) = p(x_A \mid x_C)\, p(x_B \mid x_C)$?

(C) $\Leftrightarrow$ $p(x_A \mid x_B, x_C) = p(x_A \mid x_C)$ $\quad \forall x_A, x_B, x_C$ s.t. $p(x_C) > 0$

↖ $p(x_B, x_C) > 0$

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

3 facts about C.I.:

3 facts about C.I.:

1) can repeat variables e.g. $X \perp\!\!\!\perp Y, Z \mid Z, W$ is fine to say

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

3) trick: extra conditioning on both side of equation maintains true statements

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

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

⊗ pairwise independence $\not\Rightarrow$ mutual independence

e.g. $X_3 = X_1 \text{ xor } X_2$ with $X_1 \perp\!\!\!\perp X_2 \sim \text{Bernoulli}(\frac{1}{2})$

chain rule: (always true) $p(x_V) = \prod_{i=1}^{n} p(x_i \mid x_{1:i-1})$

last conditional is $p(x_n \mid x_{1:n-1})$ → table with $\sim 2^n$ entries

graph model $p(x_V) = \prod_{i=1}^{n} p(x_i \mid x_{\pi_i})$

π_i: parents of i in graph → tables $2^{\max_i |\pi_i|+1}$

Directed graph model:

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

(DGM) a directed graphical model (aka Bayesian network) (associated with G) is a family (not necessarily unique) of distributions over X_V here $\mathcal{G}(G) \triangleq \{ p : \exists \text{ legal factors } f_i \text{ s.t.} \}$

dist. over X_V $p(x_V) = \prod_{i=1}^{n} f_i(x_i, x_{\pi_i})$

for potentials f_i s.t. 1) $f_i \geq 0$ $f_i : \Omega_{X_i} \times \Omega_{X_{\pi_i}} \to [0, \infty]$

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

is like a CPT (conditional prob. table)

$\pi_i \triangleq \{ j \in V : (j,i) \in E \}$

terminology is: if $p(x_V) = \prod_{i=1}^{n} f_i(x_i, x_{\pi_i})$ ← this is where G comes into play

terminology is: if $p(x_V) = \prod_{i=1}^{n} f_i(x_i, x_{\pi_i})$ ← this is where G comes into play (conditional prob table)

then say that "p factorizes according to G"

e.g. $p(x,y) = p(x)\,p(y|x)$ → 

$= p(y)\,p(x|y)$ →

"leaf plucking" property — fundamental property of DGM

let n be a leaf (i.e. n is not parent of anything)

then $p(x_{1:n-1}) \in \mathcal{G}(G - \{n\})$

$= \prod_{j \neq n} f_j(x_j, x_{\pi_j}) \in \mathcal{G}$ ($n \notin \pi_j \ \forall j$)

proof:

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

$$p(x_{1:n-1}) = \sum_{x_n} p(x_n, x_{1:n-1}) = \Big(\underbrace{\sum_{x_n} f_n(x_n, x_{\pi_n})}_{1 \text{ by definition}}\Big) \prod_{j \neq n} f_j(x_j, x_{\pi_j})$$

proposition: if $p \in \mathcal{G}(G)$; let $\{f_i\}$ be a factorization w.r.t. G for p

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

proof: WLOG, let $\{1, \dots, n\}$ be top sort

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

then $p(\underbrace{x_n}_{\text{leaf}}, x_{1:n-1})$

thus pluck it to get $p(x_{1:n-1}) = \prod_{j \neq n} f_j(x_j, x_{\pi_j})$ (factorization for DAG where x_{n-1} is now a leaf)

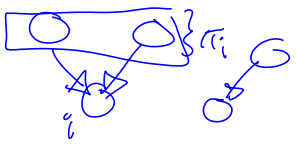
keep doing this [in reverse top sort order $n-1, \dots, i+1$] to get:

$$p(x_{1:i}) = f_i(x_i, x_{\pi_i}) \cdot \prod_{j < i} f_j(x_j, x_{\pi_j})$$

let $A \triangleq 1{:}i{-}1 \setminus \pi_i$ so that $1{:}i = \{i\} \cup \pi_i \cup A$ $\quad x_{1:i} \quad \triangleq q(x_{1:i-1})$

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

$$= f_i(x_i, x_{\pi_i})$$



equivalently, we can write: $\mathcal{G}(G) = \{p : p(x_V) = \prod_{i=1}^{n} p(x_i \mid x_{\pi_i})\}$

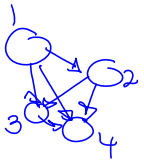
Note: adding edges $\Rightarrow$ more distributions

i.e. $G = (V, E)$ & $G' = (V, E')$ with $E' \supseteq E$

then $\mathcal{G}(G) \subseteq \mathcal{G}(G')$

examples: $E = \emptyset$ $\Rightarrow$ $\mathcal{G}(G)$ contains only fully independent dist. (trivial graph) i.e. $\prod_i p(x_i)$

- complete digraph: get WLOG $p(x_V) = \prod_i p(x_i \mid x_{1:i-1})$ (like chain rule)



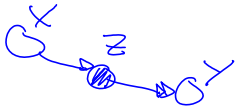
for all $i \neq j$ either $(i \to j) \in E$ or $(j \to i) \in E$ $\Rightarrow$ all distributions are in $\mathcal{G}(G_{\text{complete}})$

⊗ removing edges impose more factorization constraints

Basic 3 nodes graphs

factorization $\leftrightarrow$ cond. indep assumption

1) Markov chain:

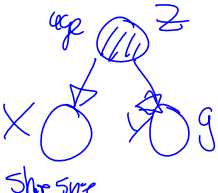


"future" "past" "present"

$Y \perp\!\!\!\perp X \mid Z$

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

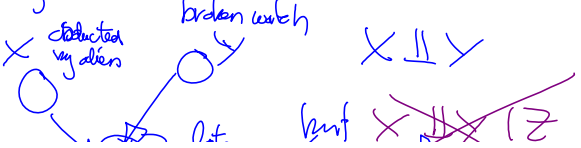
2) latent cause (hidden variable)



$X \perp\!\!\!\perp Y \mid Z$

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

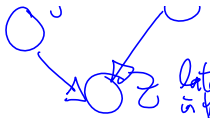
3) explaining away (v-structure)



$X \perp\!\!\!\perp Y$

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

(v-structure)



but ~~X ⊥ Z~~

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

non-monotonic property of conditioning

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

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