6.S896 - Algorithmic Statistics

Lecture 3: Introduction to Graphical Models

So far: product measures are nice (independence)

gaussians are nice (tails)

general dist'ns aren't nice (exponential sample lower bounds for easy tasks e.g testing uniformity)

Today: exploit conditional independence structure

Let us revisit multi-dimensional Gaussians

$$p(x_1,\dots,x_n) = \frac{1}{(2\pi)^{n/2}|\Sigma|}\cdot \exp\left(-\frac{1}{2}x^T\Sigma^{-1}x\right)$$ density of $\mathcal{N}(0,\Sigma)$

Σ: covariance matrix $\quad P=\Sigma^{-1}$: precision matrix

↳ term used by gauss

$G=(V,E)$ where $V=\{1,\dots,n\}$

& $\underset{i\neq j}{(i,j)}\in E$ iff $P_{ij}\neq 0$

Eg. suppose $Z_1, Z_2, Z_3 \overset{iid}{\sim} \mathcal{N}(0,1)$

$X = Z_1$, $Y = Z_1 + Z_2$, $W = Z_1 + Z_2 + Z_3$

$$\Sigma = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{pmatrix} \rightsquigarrow \Sigma^{-1} = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix}$$

$\rightsquigarrow$ resulting graph

$$X - Y - W$$

X_i & X_j are conditionally independent given $X_{[n]\setminus\{i,j\}}$

<u>Claim:</u> ① For $i \neq j$ suppose $i \not\sim j$ in G. Then: $\underbrace{X_i \perp\!\!\!\perp X_j \mid X_{[n]\setminus\{i,j\}}}$

② For disjoint sets A, B, S suppose all paths from A to B in G go through S - written $A \perp B \mid S$

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

<u>Proof:</u> ① denote $S = [n] \setminus \{i,j\}$

$$p(x) \propto \exp\Big(-\tfrac{1}{2} P_{ii} x_i^2 - \tfrac{1}{2} P_{jj} x_j^2 - x_i P_{iS} x_S - x_j P_{jS} x_S - \tfrac{1}{2} x_S^T P_{SS} x_S\Big)$$

so for some functions f, g: $p(x) = f(x_i, x_S) \cdot g(x_j, x_S)$

$\underset{\text{lemma 1}}{\Rightarrow} X_i \perp\!\!\!\perp X_j \mid X_S$

② similar ☒

Conditional Independence Cheat Sheet

[Whatever I say holds for discrete r.v.'s & their pmf'ns $p(\cdot)$
or continuous r.v.'s w/ densities $p(\cdot)$]

Def 1: The conditional density of X given Y is any f'n $p(x|y)$ such that

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

Def 2: Let X, Y, Z be r.v.'s. We say that X & Y are conditionally independent given Z if

$$p(x|y,z) = p(x|z) \quad \forall\, x, y, z \text{ s.t. } p(y,z) > 0$$

In this case we write $X \perp\!\!\!\perp Y \mid Z$

Lemma 1: Let X, Y, Z be random variables. Then the following are equivalent:

(a) $p(x,y,z) = f(x,z) \cdot g(y,z)$ for some f'ns f, g & all values x,y,z

(b) $p(x|y,z) = p(x|z) \quad \forall\, x, y, z$ s.t. $p(y,z) > 0$
(i.e. $X \perp\!\!\!\perp Y \mid Z$)

Proof: ((a) $\Rightarrow$ (b)) $p(y,z) = \int_x p(x\,y\,z)\,dx = \int_x f(x,z)\, g(y,z)\, dx$

$$= g(y,z) \cdot \tilde{f}(z)$$

Suppose y, z s.t. $p(y,z) > 0 \Rightarrow \tilde{f}(z) > 0 \Rightarrow g(y,z) = \dfrac{p(y,z)}{\tilde{f}(z)}$

$$\overset{b}{\Longrightarrow} \quad p(x,y,z) = p(y,z) \cdot \underbrace{\frac{f(x,z)}{\tilde{f}(z)}}_{\substack{\text{III by definition} \\ p(x|yz)}}$$

$\Rightarrow$ $p(x|yz)$ does not depend on y

$\Rightarrow$ $\quad p(x|yz) = p(x|z) \Rightarrow$ a

↳ why?

well... $p(xyz) = p(yz) \cdot p(x|yz)$

↳ $\underbrace{\int p(xyz)\,dy}_{\substack{\text{III} \\ p(xz)}} = \underbrace{\int p(yz)\, p(x|yz)\, dz}_{\substack{\text{II} \\ \underbrace{p(x|y_0 z)}_{\substack{\text{II} \\ p(x|z) \text{ by def.}}}\, p(z)}}$

(b $\Rightarrow$ a) $\quad p(xyz) = p(yz)\, p(x|yz) \overset{a}{=} p(yz)\, p(x|z)$

$\boxtimes$

Remark: Seems like we have been fooling around w/ symbols. But cond. independence non-trivial!

e.g. for r.v.'s X, Y, Z

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

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

think of examples!

Lemma 2: The following are true for r.v.'s $X\ Y\ Z\ W$

1. $X \perp\!\!\!\perp Y, W \mid Z \Rightarrow X \perp\!\!\!\perp Y \mid Z$ "decomposition"
2. $X \perp\!\!\!\perp Y, W \mid Z \Rightarrow X \perp\!\!\!\perp W \mid Y, Z$ "weak union"
3. $\left.\begin{matrix} X \perp\!\!\!\perp W \mid Y, Z \\ X \perp\!\!\!\perp Y \mid Z \end{matrix}\right\} \Rightarrow X \perp\!\!\!\perp Y, W \mid Z$ "contraction"
4. If $p(x,y,w,z) > 0$, $\forall x, y, w, z$, then "intersection"

$$X \perp\!\!\!\perp W \mid Y, Z \ \&\ X \perp\!\!\!\perp Y \mid W, Z \Rightarrow X \perp\!\!\!\perp WY \mid Z$$

Proof: 1. $p(x,y,w \mid z) = p(x \mid z)\, p(y,w \mid z) \xRightarrow{\int_W dw} p(x,y \mid z) = p(x \mid z)\, p(y \mid z)$

2. $p(x,y,w \mid z) = p(x \mid z) \cdot p(y,w \mid z) = \underbrace{p(x \mid z) \cdot p(y \mid z)}_{\overset{(\text{use } 1)}{\parallel}\ p(x,y \mid z)} \cdot p(w \mid y, z)$

thus $p(w \mid y, z) \equiv p(w \mid y, x, z)$ (by definition 1) ⊛

So $p(x,w \mid y,z) = p(x \mid y,z) \cdot \underbrace{p(w \mid x,y,z)}_{\overset{⊛}{\parallel}\ p(w \mid y,z)}$

3. $p(x\, y\, w \mid z) = p(y \mid z)\, p(x\, w \mid y, z)$

$$p(x, w \mid y, z) = \underbrace{p(x \mid y, z)}_{= p(x \mid z)}\, p(w \mid y, z)$$

4. Use lemma 1

$$p(x, y, w, z) = f(x, y, z) \cdot g(w, y, z) \text{ , for some } f, g$$
$$= \tilde{f}(x, w, z) \cdot \tilde{g}(y, w, z) \text{ , for some } \tilde{f}, \tilde{g}$$

by positivity $\Rightarrow$

$$f(x, y, z) = \frac{\tilde{f}(x, w, z)\, \tilde{g}(y, w, z)}{g(w, y, z)}$$

$$= \frac{\tilde{f}(x, w_0, z)\, \tilde{g}(y, w_0, z)}{g(w_0, y, z)}$$

for any fixed w_0 since LHS doesn't depend on z

thus $f(x, y, z) = \hat{f}(x, z)\, \hat{g}(y, z)$

$$\Rightarrow p(x, y, w, z) = \hat{f}(x, z)\, \hat{g}(y, z)\, g(w, y, z)$$

$$\xRightarrow{\text{lemma 1}} \quad x \perp\!\!\!\perp y w \mid z \quad \boxtimes$$

Undirected Graphical Models

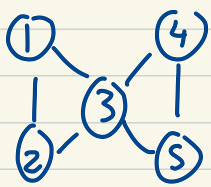
Recall: In multivariate gaussians structure of precision matrix implied cond. independence properties of the dist'n

generalize?

Def 3: given undirected graph $G = (V, E)$, a probability dist'n $p(x_V)$ satisfies the pairwise Markov property of G iff

$$(i,j) \notin E \implies X_i \perp\!\!\!\perp X_j \mid X_{V \setminus \{i,j\}}$$

e.g.



$X_1 \perp\!\!\!\perp X_4 \mid X_{235}$

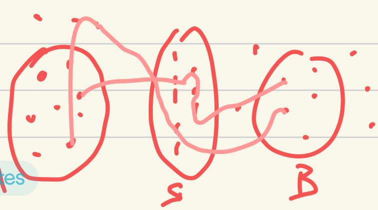
$X_2 \perp\!\!\!\perp X_5 \mid X_{134}$

$X_2 \perp\!\!\!\perp X_4 \mid X_{135}$

$X_1 \perp\!\!\!\perp X_5 \mid X_{234}$

Def 4: given G, dist'n p satisfies the global Markov property of G iff

for all disjoint sets $A, B, S \subseteq V$:
if $A \perp B \mid S$ (i.e. A & B disconnected in $G \setminus S$)
then $X_A \perp\!\!\!\perp X_B \mid X_S$



e.g. in above graph $X_{12} \perp\!\!\!\perp X_5 \mid X_3$

Def 5: given G, we say p factorizes according to G iff

$$p(x_V) = \prod_{\underbrace{C \in \mathcal{C}(G)}} \psi_C(x_C)$$

some function called clique potential

cliques of G

Theorem: If $p(x_V)$ factorizes according to G, then p satisfies global Markov property wrt G.
(& thus also local Markov property)

Proof: take disjoint $A, B, S \subseteq V$ such that $\underbrace{A \perp B \mid S}$

A & B are disconnected in $G_{V \setminus S}$

$\tilde{A}$: vertices connected to A via paths in $G_{V \setminus S}$

$\tilde{B} = V \setminus (S \cup \tilde{A})$

Clearly $A \subseteq \tilde{A}$, $B \subseteq \tilde{B}$

also every clique C is either $C \subseteq \tilde{A}$ ← call set of these cliques $\mathcal{C}_A \subseteq \mathcal{C}$

or $C \subseteq \tilde{B}$ ← $\mathcal{C}_B \subseteq \mathcal{C}$

thus $p(x) = \prod_{c \in \mathcal{C}} \psi_c(x_c) = \prod_{c \in \mathcal{C}_A} \psi_c(x_c) \cdot \prod_{c \in \mathcal{C}_B} \psi_c(x_c)$

$= f(x_{\tilde{A}}, x_S) \cdot f(x_{\tilde{B}}, x_S)$

Lemma 1

$\Rightarrow \quad X_{\tilde{A}} \perp\!\!\!\perp X_{\tilde{B}} \mid X_S$

Lemma 2

$\Rightarrow \quad X_A \perp\!\!\!\perp X_B \mid X_S$

$\boxtimes$

p factorizes wrt G $\Rightarrow$ p obeys global Markov property wrt G

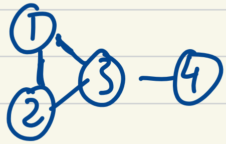
$\Downarrow$

how about this? Can we close the circle? --- p obeys local Markov property wrt G

Theorem [Hammersley-Clifford '71]

If $p(x_v) > 0, \forall x_v$, and p obeys local Markov property wrt. graph $G = (V, E)$, then p factorizes wrt G!

Remark: Conditions in H-C theorem are necessary!

Consider graph G: 

suppose X_3, X_4 are ind. Bernoulli$(1/2)$

and $X_1 = X_2 = X_4$ wpr 1

then $X_1 \perp\!\!\!\perp X_4 \mid X_2, X_3$

$X_2 \perp\!\!\!\perp X_4 \mid X_1, X_3$

But $X_4 \perp\!\!\!\perp X_1, X_2 \mid X_3$ (so p can't factorize wrt G as otherwise this would also be implied)

Nomenclature: A prob. dist'n that factorizes wrt G is called

a Markov Random Field

a Markov Network

an undirected graphical model

When $p(x_V) > 0$ it's also called a

Gibbs Random Field

Eg. Applications of Undirected Graphical Models

- used a ton in Statistical Physics
  Probability
  Machine Learning
  Statistics
  Social Science
  Biology
  ⋮

- commonly, they are described in term of
  - an undirected graph: $G=(V,E)$
  - an energy function: $\mathcal{E}(x_V)$
  - a temperature parameter: T

  with respect to which: $p(x_V) = \frac{1}{Z} \exp\left(-\frac{1}{T} \cdot \mathcal{E}(x_V)\right)$

called "partition function"
[hard to appx in general]

Example 1: Ising Model

$x_V \in \{\pm 1\}^V$

direct interaction between i,j

external field on i

$$p(x_V) = \frac{1}{Z} \exp\left(\sum_{\substack{i,j \\ i \neq j}} \theta_{ij} x_i x_j + \sum_i \theta_i x_i \right)$$

ex 2: ERGMs (exponential family random graph models)

- dist'n over graph (a.k.a. adjacency matrix)
- $V = [n] \times [n]$

$x_V \in \{0,1\}^{[n]\times[n]}$

↳ $x_{ij} \in \{0,1\}$ depending on whether node i & j are connected

$$p(x_V) \propto \exp\left(\sum_{S \in \mathcal{M}} \theta_S \cdot \left(\begin{array}{c} \#\text{ copies of } S \\ \text{in graph} \\ \text{defined by } x_V \end{array} \right) \right)$$

↑ some family of small graphs e.g. $\mathcal{M} = \{ -, \Delta, \vee, \square \}$

Closing Fun: recall testing if a dist'n q over $\{\pm 1\}^n$ is uniform i.e.

(P): $q = \mathcal{U}(\{\pm 1\}^n)$ vs $d_{TV}(q, \mathcal{U}) > \varepsilon$

requires $\Omega\left(\frac{2^{n/2}}{\varepsilon^2}\right)$ samples

What if I know that q is nice, eg.

suppose I know q is an Ising model

$$q(x) = \frac{1}{Z} \exp\left(\sum_{i,j} \theta_{ij} x_i x_j + \sum_i \theta_i x_i \right)$$

with unknown θ_{ij}'s ?

whose $|\theta|_\infty \leq O(1)$

<u>Claim</u>: If q is Ising model can solve problem P w/ poly$(n, \frac{1}{\varepsilon})$ samples.

[Daskalakis-Dikkala-Kamath'19]

<u>Proof Idea</u>: take two Ising models θ, θ'

$$SKL(q_\theta, q_{\theta'}) = KL(q_\theta \| q_{\theta'}) + KL(q_{\theta'} \| q_\theta)$$

$$= \sum_x q_\theta \log \frac{q_\theta}{q_{\theta'}} + \sum_x q_{\theta'} \log \frac{q_{\theta'}}{q_\theta}$$

$$= \sum_x q_\theta(x) \log \frac{\frac{1}{Z_\theta} \exp(\cdots)}{\frac{1}{Z_{\theta'}} \exp(\cdots)} + \sum q_{\theta'}(x) \log \frac{\frac{1}{Z_{\theta'}} \exp(\cdots)}{\frac{1}{Z_\theta} \exp(\cdots)}$$

$$= \sum_x q_\theta(x)\left(\sum_{i,j}(\theta_{ij}-\theta'_{ij})x_ix_j + \sum_i(\theta_i-\theta'_i)x_i\right)$$

cancel $\log\frac{Z_\theta}{Z_{\theta'}}$ w/ $\log\frac{Z_{\theta'}}{Z_\theta}$

$$+ \sum_x q_{\theta'}(x)\left(\sum_{i,j}(\theta'_{ij}-\theta_{ij})x_ix_j + \sum_i(\theta'_i-\theta_i)x_i\right)$$

$$= \sum_{i,j}(\theta_{ij}-\theta'_{ij})\cdot\left(\mathbb{E}_\theta x_ix_j - \mathbb{E}_{\theta'} x_ix_j\right)$$

$$+ \sum_i(\theta_i-\theta'_i)\left(\mathbb{E}_\theta x_i - \mathbb{E}_{\theta'} x_i\right)$$

now take $\theta' = \vec{0}$ then $q_{\theta'}$ = uniform dist'n

$$SKL(p_\theta, \text{uniform}) = \sum_{ij}\theta_{ij}\cdot\mathbb{E}_\theta x_ix_j + \sum_i \theta_i\,\mathbb{E}x_i$$

if p_θ = uniform $SKL(p_\theta, \text{uniform}) = 0$

If $TV(p_\theta, u) > \varepsilon \rightsquigarrow SKL(p_\theta, u) > 4\varepsilon^2$

$\rightsquigarrow \exists\ i,j$ s.t. $\mathbb{E}_\theta x_ix_j > \frac{4\varepsilon^2}{n^2}$

or $\exists\ i$ s.t. $\mathbb{E}_\theta x_i \geq \frac{4\varepsilon^2}{n}$

completing proof is left as exercise $\square$

Made with Goodnotes