LECTURE 3: DIRECTED GRAPHICAL MODELS (a.k.a. Bayesian Networks)

1. Statistical independence

→ From Lecture 1:

Most general model for $p(x_1, \dots, x_d)$ is intractable

e.g. $x_i \in \{0, 1\}$, $\begin{cases} p(x_1, \dots, x_d) = \theta_{x_1, \dots, x_d} \geq 0 \\ \sum_{x_1, \dots, x_d} \theta_{x_1, \dots x_d} = 1 \end{cases}$

representation
inference
learning

} are cursed by dimension

→ Possible solutions:

- (Lecture 2) Forget everything except mean/variance
  $\Rightarrow$ Gaussian model, tractable

- Independence:
  - assume $p(x_1, \dots, x_d) = p(x_1) \cdots p(x_d)$

  for $x_i \in \{0, 1\}$, only need d parameters

  $p(x_i = 1) = \theta_i \in [0, 1]$

  But: very strong assumption!

- in-between: e.g.

$$p(x_1, \dots, x_d) = p(x_1, \dots, x_s) \cdots p(x_{d-s+1}, \dots, x_d)$$

$\Rightarrow$ $\frac{d}{s} \times 2^s$ parameters

More expressive, but still limited ...

- Conditional independence:

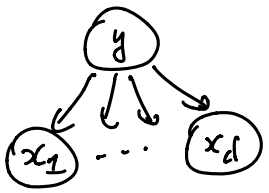
Def: $X \perp Y \mid Z$ if $p(x, y \mid z) = p(x \mid z)\, p(y \mid z)$

equivalently, $p(y \mid z, x) = p(y \mid z)$

Indeed, $p(y \mid z, x) = \frac{p(x, y \mid z)}{p(x \mid z)} = \frac{p(x \mid z)\, p(y \mid z)}{p(x \mid z)} = p(y \mid z)$

Note: $\neq$ than $X \perp Y$!!

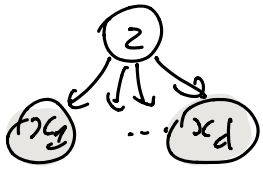
Ex: - Naive Bayes



$$p(x_1, \dots, x_d, y) = \underbrace{p(y)}_{\pi_y} \prod_i \underbrace{p(x_i \mid y)}_{\theta_{y,i}}$$

e.g. y = label, $\{x_i\}$ C.I. given label

- Latent variable / cause



$$p(x_1, \dots, x_d) = \int p(z) \prod_i p(x_i \mid z)\, dz$$

Note: always true if $(x_1, \dots, x_d)$ are "exchangeable"

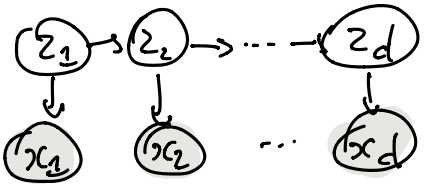
i.e. $(x_1, \dots, x_d) \overset{d}{=} (x_{\pi(1)}, \dots, x_{\pi(d)})$ for any permutation π

(De Finetti's thm)

but: $p(x_i \mid z)$ may be complicated ...

- Markov / Hidden-Markov Models

$$p(x_1, \dots, x_d) = p(x_1)\, p(x_2 | x_1) \cdots p(x_d | x_{d-1})$$



$$p(x_1, \dots, x_d) = \int \cdots \int p(z_1)\, p(z_2 | z_1) \cdots p(z_d | z_{d-1}) \cdot \prod_i p(x_i | z_i) \cdot dz_1 \cdots dz_d$$

e.g. • $z_i \in \{1, \dots, K\}$

$x_i | z_i = k \sim \text{Cat}(\theta_k)$

or

$x_i | z_i = k \sim \mathcal{N}(\mu_k, \Sigma_k)$

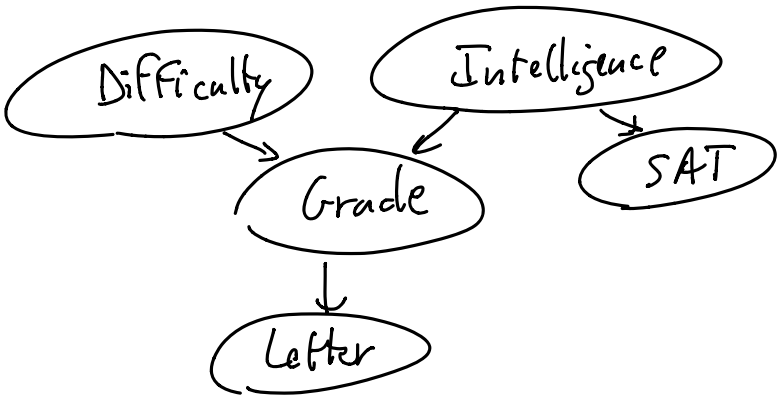
• $z_i | z_{i-1} \sim \mathcal{N}(A z_i + b, \Sigma_z)$

$x_i | z_i \sim \mathcal{N}(C z_i + d, \Sigma_x)$

(state-space model / Kalman Filtering)

2. Directed Graphical Models

Ex:



(From Koller & Friedman)

- $D \perp I$? yes
- $D \perp I \mid G$? no e.g. $p(I \mid G, D = \text{hard}) > p(I \mid G)$ ("explaining away")
- $G \perp S$? no
- $G \perp S \mid I$? yes
- $D \perp I \mid L$? no

Factorize: $P(D, I, G, S, L) = P(D)\, P(I)\, P(G \mid D, I)\, P(S \mid I)\, P(L \mid G)$

More generally, consider a DAG $G = (V, E)$

$pa(i) = \{j \text{ s.t. } (j, i) \in E\}$ (parents of node i)

We want:

$$p(x_1, \dots, x_d) = \prod_{i=1}^{d} p(x_i \mid x_{pa(i)}) \qquad (*)$$

Def. If this holds, we say p factorizes according to G

Def: A Directed graphical model/structure $G = (V, E)$ encodes the following (local) independencies, denoted $\mathcal{I}_\ell(G)$:

$$\forall i, \quad (X_i \perp X_{\text{non-descendants}(i)} \mid X_{pa(i)})$$

Fact:
If p obeys the C.I.s in $\mathcal{I}_p(G)$, it factorizes as (*)

proof: use the chain rule with any topological ordering of the variable relative to G
(i.e. $(i,j) \in E \Rightarrow i < j$) ∎

- Independencies and I-maps

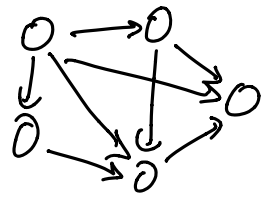
Def: For a prob. distribution $p(x)$, we denote by $\mathcal{I}(p)$ the set of C.I. of the form $(X \perp Y | Z)$ that hold in p.

Def: (I-map)
We say the K (e.g. graph) is an I-map for $\mathcal{I}(p)$ if K "encodes" a set of CIs $\mathcal{I}(K)$ with $\mathcal{I}(K) \subseteq \mathcal{I}(p)$

Fact: If G is an I-map for p, then p factorizes according to G

→ indeed, G encodes at least the CIs in $\mathcal{I}_p(G)$, which is sufficient for factorization.

Ex:



Fully-connected graph

- G: G is always an I-map

- G: o o o : I(G) is large, requires independence!

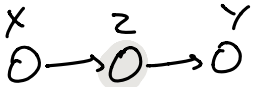
→ Do we have the converse? Yes!

[Fact: If p factorizes according to G, then G is an I-map for p.

exercise: check on Student example!

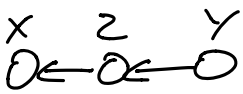
■ D-separation

→ Three node graphs: for (X, Y, Z), when do we have $X \perp Y \mid Z$?

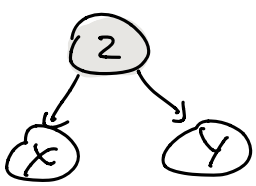
- chain: 

("causal trail")

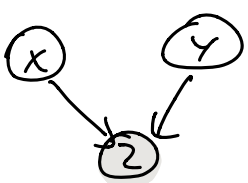
$p(y|x,z) = p(y|z)$ ✓



✓

- "common cause": 

✓

- "v-structure": 

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

(explaining away)

→ d-separation (directed separation)

Def: A path $X \overset{Z}{\rightsquigarrow} Y$ is active in the following scenarios:

- $X \to Z \to Y$ and Z is not observed
- $X \leftarrow Z \leftarrow Y$ " "
- $X \overset{Z}{\leftarrow} \to Y$ " "
- $X \to Z \overset{Z}{\leftarrow} Y$ and Z or any descendant of Z is observed

Def: We say X and Y are d-separated given Z, denoted $d\text{-sep}_G(X, Y | Z)$ if there is no active path path between any $X_i \in X$ and $Y_i \in Y$ given Z.

We then write $\mathcal{I}(G) = \{(X \perp Y | Z) : d\text{-sep}_G(X; Y | Z)\}$

Note: we have $\mathcal{I}(G) \subseteq \mathcal{I}(P)$ iff P factorizes over G.

(Thm) ⇔ any d-sep statement is satisfied by such P.

Q: What about $\mathcal{I}(G) = \mathcal{I}(P)$?

Not true in general:

	$y=0$	$y=1$
$x=0$	0.4	0.6
$x=1$	0.4	0.6

→ $X \perp Y$, yet we have that

p factorizes over $\textcircled{X} \to \textcircled{Y}$

Thm: For almost all distributions p that factorize over G, we have $I(G) = I(p)$

(i.e. except a set of measure 0)

Remark: Generally, no unicity in the choice of graph

e.g. $x \to z \to y$ $\qquad$ z with $x \leftarrow z \rightarrow y$

$x \leftarrow z \leftarrow y$

Note: this is different for causal models!