Graphical models

Sunita Sarawagi
IIT Bombay
http://www.cse.iitb.ac.in/~sunita

Probabilistic modeling

- Given: several variables: $x_1, \ldots x_n$, n is large.
- Task: build a joint distribution function $\Pr(x_1, \ldots x_n)$
- Goal: Efficiently represent, estimate, and answer inference queries on the distribution
- Basic premise
  - Explicit joint distribution is dauntingly large
  - Queries are simple marginals (sum or max) over the joint distribution.

Example

- Variables are attributes are people.

Age	Income	Experience	Degree	Location
10 ranges	7 scales	7 scales	3 scales	30 places

- An explicit joint distribution over all columns not tractable: number of combinations: $10 \times 7 \times 7 \times 3 \times 30 = 44100$.

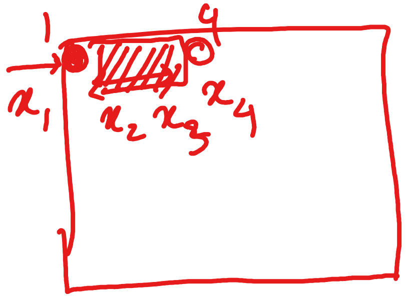
Alternatives to an explicit joint distribution

- Assume all columns are independent of each other: bad assumption
- Use data to detect pairs of highly correlated column pairs and estimate their pairwise frequencies
  - Many highly correlated pairs
    income $\not\perp\!\!\!\perp$ age, income $\not\perp\!\!\!\perp$ experience, age $\not\perp\!\!\!\perp$ experience
  - Ad hoc methods of combining these into a single estimate
- Go beyond pairwise correlations: conditional independencies
  - income $\not\perp\!\!\!\perp$ age, but income $\perp\!\!\!\perp$ age $\mid$ experience
  - experience $\perp\!\!\!\perp$ degree, but experience $\not\perp\!\!\!\perp$ degree $\mid$ income

Graphical models make explicit an efficient joint distribution from these independendencies

More examples of CIs

- The grades of a student in various courses are correlated but they become CI given attributes of the student (hard-working, intelligent, etc?)
- Health symptoms of a person may be correlated but are CI given the latent disease.
- Words in a document are correlated, but may become CI given the topic.
- Pixel color in an image become CI of distant pixels given near-by pixels.



$x_1 \perp\!\!\!\perp x_4 \mid x_2, x_3$

Graphical models

Model joint distribution over **several** variables as a product of smaller factors that is

1. *Intuitive* to represent and visualize
   - Graph: represent structure of dependencies
   - Potentials over subsets: quantify the dependencies
2. *Efficient* to query
   - given values of any variable subset, reason about probability distribution of others.
   - many efficient exact and approximate inference algorithms

Graphical models = graph theory + probability theory.

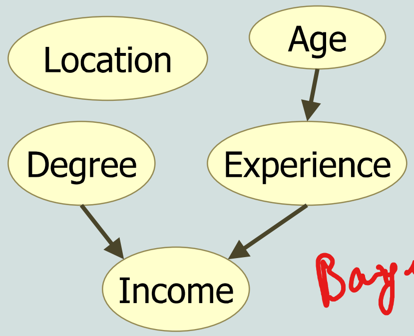
Representation

Structure of a graphical model: Graph + Potential

Graph

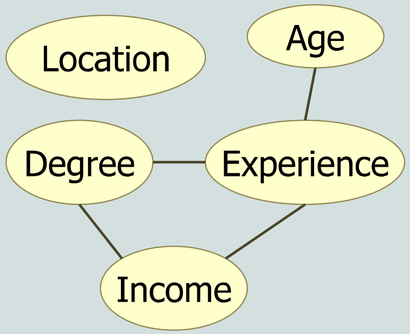
- Nodes: variables $\mathbf{x} = x_1, \ldots x_n$
  - Continuous: Sensor temperatures, income
  - Discrete: Degree (one of Bachelors, Masters, PhD), Levels of age, Labels of words
- Edges: direct interaction
  - Directed edges: Bayesian networks
  - Undirected edges: Markov Random fields

Directed



Bayesian NW

Undirected



MRF

Representation

Potentials: $\psi_c(\mathbf{x}_c)$ -

$V = \{1, 2, \ldots n\}$ $c \subseteq V$

eg: $\{1, 2\}$

$x_c = \{x_1, x_2\}$

- Scores for assignment of values to subsets c of directly interacting variables.
- Which subsets? What do the potentials mean?
  - Different for directed and undirected graphs

Probability

Factorizes as product of potentials

$$\Pr(\mathbf{x} = x_1, \ldots x_n) \propto \prod \psi_S(\mathbf{x}_S)$$

Directed graphical models: Bayesian networks

- Graph G: directed acyclic graphs (DAG)
  - Parents of a node: x_i $\text{Pa}(x_i)$ = set of nodes in G pointing to x_i
- Potentials: defined at each node in terms of its parents.

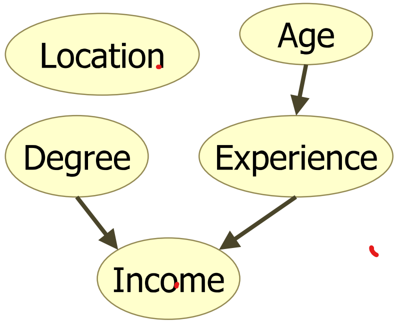
$$\psi_i(x_i, \text{Pa}(x_i)) = \Pr(x_i | \text{Pa}(x_i))$$

- Probability distribution

$$\Pr(x_1 \ldots x_n) = \prod_{i=1}^{n} \Pr(x_i | pa(x_i))$$

For the example in slide-3.

Example of a directed graph



$|P(L)| = 30-1$

$|P(A)| = 10-1$

$|P(D)| = 3-1$

$|P(E|A)| = 10 \times (7-1)$

$|P(I|D,E)| = 3 \times 7 \times (7-1)$

$\psi_1(L) = \Pr(L)$

NY	CA	London	Other
0.2	0.3	0.1	0.4

$\psi_2(A) = \Pr(A)$

20–30	30–45	$>$ 45
0.3	0.4	0.3

or, a Guassian distribution $(\mu, \sigma) = (35, 10)$

$\psi_3(E, A) = \Pr(E|A)$

Age	0–10	10–15	$>$ 15
20–30	0.9	0.1	0
30–45	0.4	0.5	0.1
$>$ 45	0.1	0.1	0.8

$\psi_5(I, E, D) = \Pr(I|D, E)$

3 dimensional table, or a histogram approximation.

Probability distribution

$$\Pr(\mathbf{x} = L, D, I, A, E) = \Pr(L)\Pr(D)\Pr(A)\Pr(E|A)\Pr(I|D,E)$$

Conditional Independencies

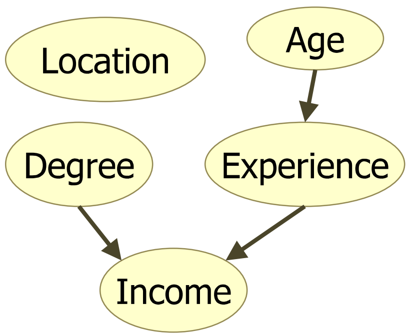
- Given three sets of variables X, Y, Z, set X is conditionally independent of Y given Z ($X \perp\!\!\!\perp Y|Z$) iff

$$\Pr(X|Y,Z) = \Pr(X|Z)$$

- Local conditional independencies in BN: for each x_i

$$x_i \perp\!\!\!\perp ND(x_i)|Pa(x_i)$$

- $L \perp\!\!\!\perp E, D, A, I$
- $A \perp\!\!\!\perp L, D$
- $E \perp\!\!\!\perp L, D|A$
- $I \perp\!\!\!\perp L, A|E, D$



CIs and Fractorization

Theorem

Given a distribution $P(x_1, \ldots, x_n)$ and a DAG G, if P satisfies Local-CI induced by G, then P can be factorized as per the graph. *Local-CI(P, G) $\implies$ Factorize(P, G)*

Proof.

- $x_1, x_2, \ldots, x_n$ topographically ordered (parents before children) in G.
- Local CI(P, G): $P(x_i | x_1, \ldots, x_{i-1}) = P(x_i | Pa_G(x_i))$
- Chain rule: $P(x_1, \ldots, x_n) = \prod_i P(x_i | x_1, \ldots, x_{i-1}) = \prod_i P(x_i | Pa_G(x_i))$
- $\implies$ Factorize(P, G)

$\square$

Also as Theorem 3.1 in KF book