Lecture 5: January 28, 2015
cs 573: Probabilistic Reasoning
Professor Nevatia
Spring 2015

Review

- Assignment #2 (a) posted; 2 (b) to be posted next week; both due Feb 9.
- Previous lecture:
  - Global independences
    - d-separation
    - Markov Blanket
  - I-equivalence
  - Constructing I-maps
  - P-maps (concept only)
- Today's objective
  - Efficient represent of distributions, exploiting further structure in parameters

Next Steps

- We are finished with Chapter 3; Bayesian Networks
- Chapter 4 is about undirected networks (Markov networks)
- Chapter 5 is about specialized forms of conditional distributions that can be represented compactly
  - We will do this chapter before chapter 4 as it focuses on directed networks
  - Also introduces continuous variables
  - We will not cover chapter 5 in full detail

Local Models

- Size of a CPT depends on the number of values of a node X, the number of parents and the number of values of each parent
- For all binary values, the number is 2^n where n is the number of parents
- For 10 parents, this number is 1024. Easily achieved in medical diagnosis: 10 diseases may cause fever (high temperature)
  - It would not be reasonable to ask a physician to provide 1024 numbers!
  - It would also not be easy to *learn* these numbers from examples
    - How many examples needed for reliable estimation?
- Solution: Look for some structure in the CPD so we only have to provide a few numbers that can populate the full CPD.

Context-Specific Independences

- Consider $X_1 \rightarrow X_3 \leftarrow X_2$
- Consider following CPD

X_1	X_2	$P(X_3=1)$
1	0	.7
1	1	.7
0	1	.4
0	0	.2

- Is $X_3 \perp X_2 | X_1$?
- No, but $X_3 \perp X_2 | (X_1 = 1)$; only 3 parameters needed to represent the distribution.
- Notation: $X_3 \perp_c X_2 | c$; (c is the condition that $X_1 = 1$); context specific independence (CSI)

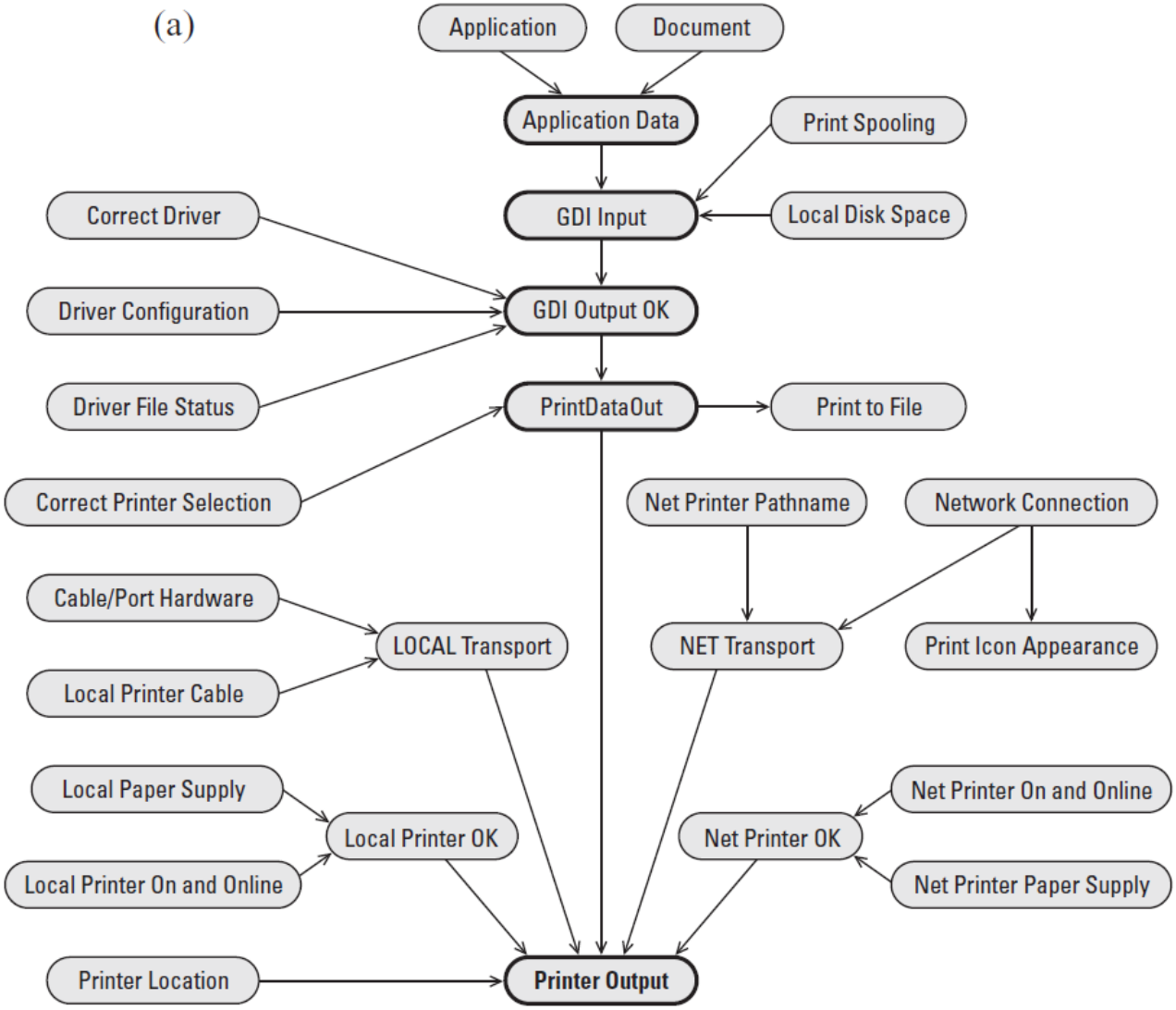
Context-Specific Independences

- Example: X_3 is Income (low, high), X_1 is weather (normal, drought), X_2 is profession (farmer, programmer)
  - Programmer's income does not depend on weather (assumption)
- How to represent such a CPD?
  - In the form of a tree (instead of a table). Draw one.
- Another example in the book, Ex 5.4, with three variables
- If there are no CSIs, tree CPD will be no more compact than table CPD
- Printer example (next page)
  - Note many nodes have 5-6 parents and some CSIs should be apparent.

Multiplexer CPD

- Child node takes on the value of one of its parents
- The selected parent depends on the value of another random variable, called the selector variable
- Analogous to a hardware multiplexer designed to choose one of its inputs to feed to the next level
- Formal definition in Def 5.3
- Used to define a printer function network in Windows 95
  - Claims that number of parameters reduced from 145 to 55

A Printer Diagnostic Network



Independencies for context-specific CPDs

- Context may induce additional dependencies, beyond those from the structure of the graph.
- Modified d-separation algorithm, called CSI-separation, is given in the book (Alg 5.2); skip for our course

Deterministic CPDs

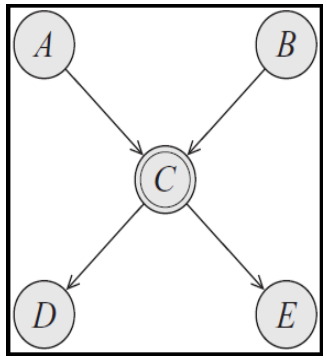
- X is a deterministic function of Pa_X

$$P(x \mid \mathrm{pa}_X) = \begin{cases} 1 & x = f(\mathrm{pa}_X) \\ 0 & \text{otherwise.} \end{cases}$$

- e.g. *f* could be the logical OR function of variables in Pa_X
  - Any function of parents could be used
- Realistic for any domains? Logical circuits, genetic tendencies…
- Flat tire example: one or more flat tires affect ride, steering, braking etc. Introduce a variable Flat-tire which is OR of all four tires to simplify.

Independences in Deterministic CPDs

- Fig 5.1, C is given deterministically from A and B



- Follows that (D ⊥ E| A, B), would not be true for general CPD
- Deterministic Separation (X;**Y**| **Z**)
  - Algorithm 5.1 gives the modified d-separation algorithm
  - Omit details for our course

Noisy-OR

- Suppose that there are two causes that can cause similar symptoms, e.g. Flu and Malaria can cause high temp on their own with some probabilities
- What should be the probability of symptom if both causes are present?
- Noisy-OR provides a simple model
- Say, $P(Fever^1 | Flu^1, Malaria^0) = 0.7$ and $P(Fever^1 | Flu^0, Malaria^1) = .8$
  - What is a good value for $P(Fever^1 | Flu^1, Malaria^1)$?
- One way to view is that of "failure probabilities"

  $P(Fever^0 | Flu^1, Malaria^0) = 0.3$ and $P(Fever^0 | Flu^0, Malaria^1) = .2$

  $P(Fever^0 | Flu^1 \wedge Malaria^1) = .2 \times .3 = .06$, hence

  $P(Fever^1 | Flu^1, Malaria^1) = .94$
- Note: the "noisy OR" does ***not*** follow from laws of probability but is a simple model that may apply to many situations.

Another Example (from KF Book)

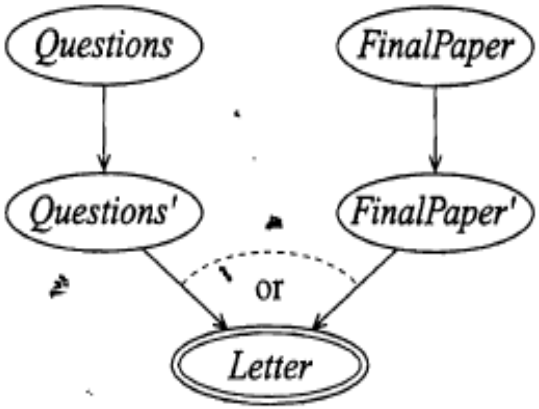
- A student can get a good letter by participating in class (asking good questions) OR by writing a good Final Paper but Professor is forgetful so there is some noise in the process.
- Let $P(l^1|q^1, f^0)=.8$ and $P(l^1|q^0, f^1)=.9$, then $P(l^0 \mid q^1, f^1)=.2x.1=.02$
- Noisy-OR CPD is

Q, F	l^0	l^1
$q^0 f^0$	1	0
$q^0 f^1$	0.1	0.9
$q^1 f^0$	0.2	0.8
$q^1 f^1$	0.02	0.98

- Prob that Q causes L is called the noise parameter, λ_Q(.8 in example); similar definition for λ_F (.9 in example);
- Can add a third parent with *leak probability* for Prof writing a good letter anyway (λ_0, say .0001)

Another View of the Example

- Another view is as a graph with deterministic OR



- *Questions'* is true if professor remembers student's participation
- *FinalPaper'* is true if professor reads and appreciates the paper
- Introducing auxiliary variables can often simplify the representation. P (Q′|Q) = .8 ; P (F′|F) =.9
- Can add a third parent with leak for Prof writing a good letter anyway (λ_0, say .0001)
- Also used: Noisy-AND, Noisy-Max (aggregating function is a deterministic AND/MAX)

Formal Definition

Let Y be a binary-valued random variable with k binary-valued parents $X_1, \ldots, X_k$. The CPD $P(Y \mid X_1, \ldots, X_k)$ is a noisy-or *if there are $k+1$* noise parameters *$\lambda_0, \lambda_1, \ldots, \lambda_k$ such that*

$$P(y^0 \mid X_1, \ldots, X_k) = (1 - \lambda_0) \prod_{i \,:\, X_i = x_i^1} (1 - \lambda_i) \tag{5.2}$$

$$P(y^1 \mid X_1, \ldots, X_k) = 1 - [(1 - \lambda_0) \prod_{i \,:\, X_i = x_i^1} (1 - \lambda_i)]$$

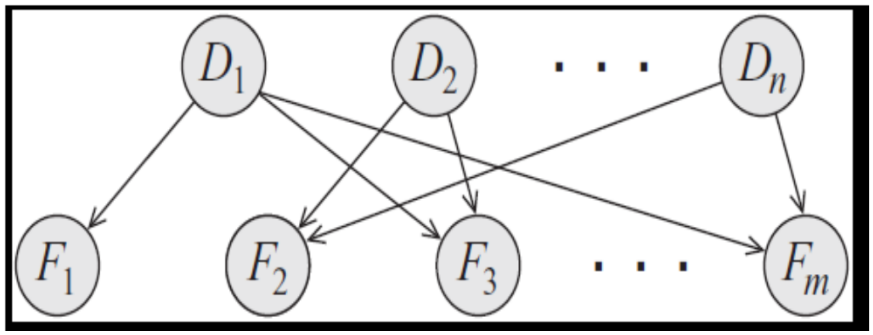
We note that, if we interpret x_i^1 as 1 and x_i^0 as 0, we can rewrite equation (5.2) somewhat more compactly as:

$$P(y^0 \mid x_1, \ldots, x_k) = (1 - \lambda_0) \prod_{i=1}^{k} (1 - \lambda_i)^{x_i}. \tag{5.3}$$

Note: notation "trick" above: expand for a single parent

BN2O Network

- Top layer corresponds to a set of *causes* (diseases), second layer to *findings* (symptoms or test results)
  - Second layer variables have noisy-OR models



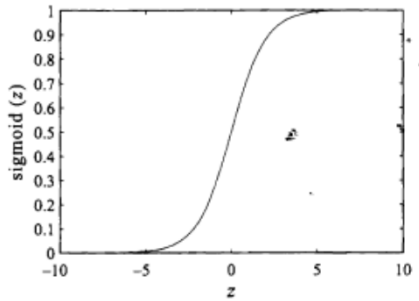
$$P(f_i^0 \mid \mathrm{Pa}_{F_i}) = (1-\lambda_{i,0}) \prod_{D_j \in \mathrm{Pa}_{F_i}} (1-\lambda_{i,j})^{d_j}$$

- A patient usually has only a small number of findings (most other symptoms are false). This is equivalent to removing parts of network that correspond to these findings (this requires modifying the parameters of the new network appropriately).

Generalized Linear Models

- First consider case where Y is Binary-valued, parents can take on numerical values. We want to specify $P(Y| X_1, X_2 \ldots X_k)$
- Define $f(X_1, X_2 \ldots X_k) = \sum_i w_i X_i$ (note x_i is 1 if T, 0 otherwise)
- We can say $Y=y^1$ if $f(\ldots) > \tau$
- Eliminate τ by adding w_0 to *f* above to give $w_0 + \sum_i w_i X_i > 0$
- Instead of a hard threshold, we can define a smoother function, called *sigmoid* or *logit* function

$$\text{sigmoid}(z) = \frac{e^z}{1+e^z}$$



Let Y be a binary-valued random variable with k parents $X_1, \ldots, X_k$ that take on numerical values. The CPD $P(Y \mid X_1, \ldots, X_k)$ is a logistic CPD *if there are $k+1$* weights *$w_0, w_1, \ldots, w_k$ such that:*

$$P(y^1 \mid X_1, \ldots, X_k) = \text{sigmoid}(w_0 + \sum_{i=1}^{k} w_i X_i).$$

■

Multi-valued Variables

- Let Y be multi-valued with *m* possible values. Parents are numerical valued as before. We can specify a logistic combination for each value y^i of Y.

Let Y be an m-valued random variable with k parents $X_1, \ldots, X_k$ that take on numerical values. The CPD $P(Y \mid X_1, \ldots, X_k)$ is a multinomial logistic *if for each $j = 1, \ldots, m$, there are $k+1$*

weights $w_{j,0}, w_{j,1}, \ldots, w_{j,k}$ such that:

$$\ell_j(X_1, \ldots, X_k) = w_{j,0} + \sum_{i=1}^{k} w_{j,i} X_i$$

$$P(y^j \mid X_1, \ldots, X_k) = \frac{\exp(\ell_j(X_1, \ldots, X_k))}{\sum_{j'=1}^{m} \exp(\ell_{j'}(X_1, \ldots, X_k))}.$$

Continuous Variables

- Difficult to define in general, an often used case is where all continuous variables have Gaussian distributions
- Variable Y, one parent X: Mean of Y is a linear function of value of X, variance of Y does not depend on X
- Generalize to multiple parents: mean of Y is a linear function of means of parents, variance of Y does not depend on variances of parents. Called *Linear Gaussian CPD*.

Let Y be a continuous variable with continuous parents $X_1, \ldots, X_k$. We say that Y has a linear Gaussian *model if there are parameters $\beta_0, \ldots, \beta_k$ and σ^2 such that*

$$p(Y \mid x_1, \ldots, x_k) = \mathcal{N}\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k; \sigma^2\right).$$

In vector notation,

$$p(Y \mid \boldsymbol{x}) = \mathcal{N}\left(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}; \sigma^2\right).$$

■

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$$

ε is N $(0,\sigma^2)$

Hybrid Models

- Child node, X, is continuous. Some parents may be discrete, some continuous.
- CPD of X is a linear Gaussian function of continuous parents' means but the function depends on value of discrete parents.

Let X be a continuous variable, and let $\boldsymbol{U} = \{U_1, \ldots, U_m\}$ be its discrete parents and $\boldsymbol{Y} = \{Y_1, \ldots, Y_k\}$ be its continuous parents. We say that X has a conditional linear Gaussian *(CLG) CPD if, for every value $\boldsymbol{u} \in Val(\boldsymbol{U})$, we have a set of $k+1$ coefficients $a_{\boldsymbol{u},0}, \ldots, a_{\boldsymbol{u},k}$ and a variance $\sigma^2_{\boldsymbol{u}}$ such that*

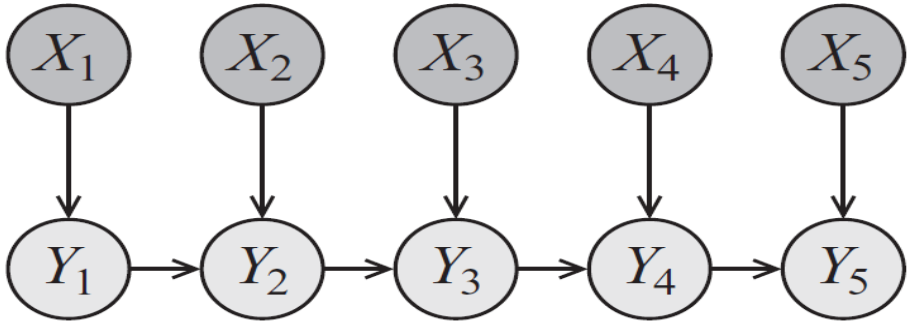
$$p(X \mid \boldsymbol{u}, \boldsymbol{y}) = \mathcal{N}\left(a_{\boldsymbol{u},0} + \sum_{i=1}^{k} a_{\boldsymbol{u},i} y_i; \sigma^2_{\boldsymbol{u}}\right)$$

A Bayesian network is called a CLG network *if every discrete variable has only discrete parents and every continuous variable has a CLG CPD.*

- Resulting distribution is a *mixture* of Gaussians, weights of different Gaussians depend on the probabilities of discrete parent assignments.

Conditional BN (CBN)

- Consider following network: Let X_i be *evidence* variables



- Suppose we specify all CPDs but no priors for Xi; then we will not be able to compute P(X) but can compute P(Y|X) and related marginals. Such a network is called Conditional BN (CBN).
- More generally:

$$P_{\mathcal{B}}(\boldsymbol{Y}, \boldsymbol{Z} \mid \boldsymbol{X}) = \prod_{X \in \boldsymbol{Y} \cup \boldsymbol{Z}} P(X \mid \mathrm{Pa}_X^{\mathcal{G}}).$$

The distribution $P_{\mathcal{B}}(\boldsymbol{Y} \mid \boldsymbol{X})$ is defined as the marginal of $P_{\mathcal{B}}(\boldsymbol{Y}, \boldsymbol{Z} \mid \boldsymbol{X})$:

$$P_{\mathcal{B}}(\boldsymbol{Y} \mid \boldsymbol{X}) = \sum_{\boldsymbol{Z}} P_{\mathcal{B}}(\boldsymbol{Y}, \boldsymbol{Z} \mid \boldsymbol{X}).$$

- Better known is related undirected version: Conditional Random Field (CRF); we will study in detail later.

Summary

- Considered some structures that help specify a CPD with fewer parameters
  - The appropriate structure (deterministic, tree, rules) depends on the problem domain
  - Noisy-OR has proven to be a very valuable approximation tool
- Linear models can be a convenient approximation for many problems
- Undirected models (ch.4) next

Next Class

- Read sections 4.1 through 4.4 of the KF book