Introduction to Bayesian Networks

Huizhen Yu

janey.yu@cs.helsinki.fi
Dept. Computer Science, Univ. of Helsinki

Probabilistic Models, Spring, 2010

Outline

Acknowledgment: Illustrative examples in this lecture are mostly from Finn Jensen's book, *An Introduction to Bayesian Networks*, 1996.

Outline

Directed Acyclic Graphs

Notation and definitions for a directed graph $G = (V, E)$:

- V: the set of vertices
- E: the set of directed edges; $(\alpha, \beta) \in E$ means there is an edge from vertex α to vertex β.
- If $(\alpha, \beta) \in E$, we say α is a *parent* of β and β is a *child* of α.
- We denote the set of parents of a vertex α by $\mathrm{pa}(\alpha)$, and the set of its children by $\mathrm{ch}(\alpha)$.
- cycle: a path that starts and ends at the same vertex.

A *directed acyclic graph* (DAG) is a directed graph that has no cycles.

Let G be a DAG and $X = \{X_v, v \in V\}$ discrete random variables associated with V. We say $P(X)$ *factorizes recursively* according to G if

$$p(x) = \prod_{v \in V} p\big(x_v \,|\, x_{\mathrm{pa}(v)}\big).$$

Examples we have seen: Markov chains, HMMs

Comparison between DAG and MRF

DAG:

- Dependence structure is specified hierarchically.
- Edges represent direct or causal influence.

Undirected graphs/MRF:

- Neighborhood relationship is symmetric.
- Edges represent interaction/association.

Graphs having both directed and undirected edges are called *chain graphs*; they can represent both kinds of dependence and are more general than DAG or MRF. (We will not study them, however.)

Building and Using Models

Three phases of developing a Bayesian network:

(i) Building models

- Specify random variables
- Specify structural dependence between variables
- Assign conditional probabilities to components of the model

(ii) Constructing inference engine

- Compile the model: representations convenient for computation (answering queries) are created internally.

(iii) Using the inference engine for case analysis

Parameter learning/adaptation based on data links (iii) partially to (i). Structural learning/adaptation is an active research topic.

In this lecture, we illustrate model building with simple examples.

Outline

Model Elements

Random variables:

- Hypothesis variables:
  their values are unobservable, but of interest in our problem.
- Information variables:
  their values are observed and informative about the hypothesis variables.
- Mediating variables:
  related to the underlying physical process, or introduced just for convenience.

Dependence structure: DAG

Conditional probabilities $\{p(x_v \mid x_{\mathrm{pa}(v)})\}$ for model components

Specifying variables in the model is the first step of model building and very important in practice, although in studying the theory we have taken it for granted.

Example I: Family Out?

Description:

When I go home at night, I want to know if my family is at home before I try the door. Often when the family leaves, an outdoor light is turned on. However, sometimes the light is turned on if the family is expecting a guest. Also, we have a dog. When nobody is at home, the dog is put in the back yard. The same is true if the dog has bowel trouble. Finally, if the dog is in the back yard, I will probably hear her barking, but sometimes I can be confused by other dogs barking.

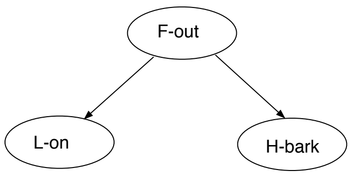
Hypothesis variable:

- Family out? (F-out)

Information variables:

- Light on? (L-on)
- Hear dog barking? (H-bark)

1st model of causal structure

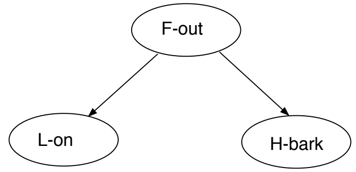


Example I: Family Out?

We now specify (subjective) conditional probabilities for the model.

For $P(\texttt{F-out})$:
One out of five week days, my family is out; so I set

F-out = y	F-out = n
0.2	0.8



For $P(\texttt{L-on} \mid \texttt{F-out})$:
The family rarely leaves home without turning the light on. But sometimes they may forget. So I set

	L-on = y	L-on = n
F-out = y	0.99	0.01
F-out =n	?	?

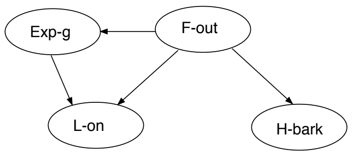
How to specify the rest of the probabilities for $P(\texttt{L-on} \mid \texttt{F-out} = n)$ and $P(\texttt{H-bark} \mid \texttt{F-out})$?

Example I: Family Out?

Introduce mediating variables for assessing probabilities:

- Expect guest? (Exp-g)

2nd model



For $P(\texttt{L-on} \mid \texttt{F-out} = n)$:
We have guests three times a month, so

$$P(\texttt{Exp-g} = y \mid \texttt{F-out} = n) = 0.1,$$

$$P(\texttt{L-on} = y \mid \texttt{Exp-g} = y, \texttt{F-out} = n) = 1,$$

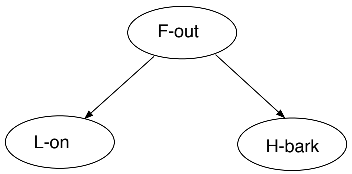
and

$$P(\texttt{L-on} = y \mid \texttt{F-out} = n) = 0.1.$$

I set

	L-on = y	L-on = n
F-out = y	0.99	0.01
F-out =n	0.1	0.9

Back to 1st model



Example I: Family Out?

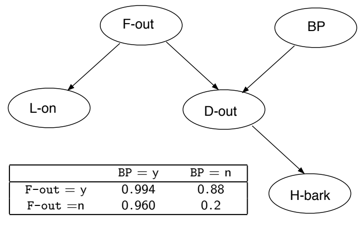
Introduce mediating variables for assessing probabilities $P(\texttt{H-bark} \mid \texttt{F-out})$:

- Dog out? (D-out)
- Bowel problem? (BP)

3rd model

For $P(\texttt{BP})$: I set

BP = y	BP = n
0.05	0.95



For $P(\texttt{D-out} \mid \texttt{F-out}, \texttt{BP})$:
Sometimes the dog is out just because she wants to be out:

$$P(\texttt{D-out} = y \mid \texttt{F-out} = n, \texttt{BP} = n) = 0.2.$$

After some reasoning ..., I set $P(\texttt{D-out} = y \mid \texttt{F-out}, \texttt{BP})$ to be

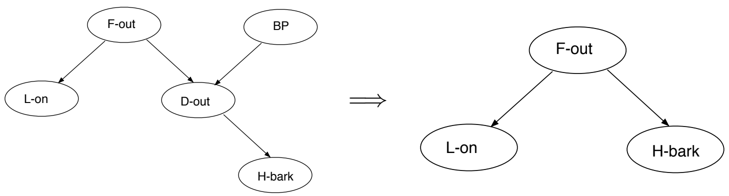
	BP = y	BP = n
F-out = y	0.994	0.88
F-out =n	0.960	0.2

For $P(\texttt{H-bark} \mid \texttt{D-out})$:
Sometimes I can confuse the barking of the neighbor's dog with that of mine. Without introducing another mediating variable, I take this into account in the probability assessment by setting

	H-bark = y	H-bark = n
D-out = y	0.6	0.4
D-out = n	0.2	0.8

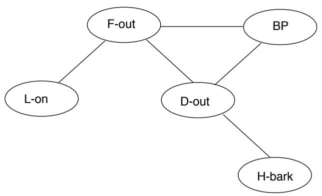
Example I: Family Out?

In this example, if mediating variables will never be observed, we can eliminate (marginalize out) them and get back to the 1st model with the corresponding probabilities:



We can verify this directly. Alternatively, we can also use the global Markov property on undirected graphs:
P of the larger model factorizes according to the graph on the right. The vertex set $\{\texttt{F-out}\}$ separates $\{\texttt{L-on}\}$ from $\{\texttt{H-bark}\}$, so

$$\texttt{L-on} \perp \texttt{H-bark} \mid \texttt{F-out}.$$



Example II: Insemination

Description:

Six weeks after insemination of a cow there are three tests for the result: blood test (BT), urine test (UT) and scanning (Sc). The results of the blood test and the urine test are mediated through the hormonal state (Ho) which is affected by a possible pregnancy (Pr). For both the blood test and the urine test there is a risk that a pregnancy does not show after six weeks because the change in the hormonal state may be too weak.

Hypothesis variable:

- Pregnant? (Pr)

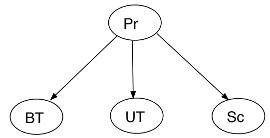
Information variables:

- Blood test result (BT)
- Urine test result (UT)
- Scanning result (Sc)

Mediating variable:

- Hormonal state (Ho)

1st naive Bayes model without the variable Ho

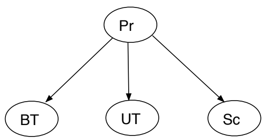


Example II: Insemination

Over-confidence of Naive Bayes model:

1st naive Bayes model without the variable Ho

Pr = y	Pr = n
0.87	0.13



	BT = y	BT = n
Pr = y	0.64	0.36
Pr =n	0.106	0.894

	UT = y	UT = n
Pr = y	0.73	0.27
Pr =n	0.107	0.893

	Sc = y	Sc = n
Pr = y	0.9	0.1
Pr =n	0.01	0.99

BT and UT results are counted as two independent pieces of evidence:

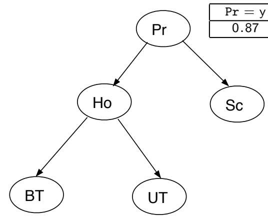
$P(\mathtt{BT} = n \mid \mathtt{Pr} = n)P(\mathtt{UT} = n \mid \mathtt{Pr} = n)P(\mathtt{Pr} = n)$
$= 0.894 \cdot 0.893 \cdot 0.13$
$P(\mathtt{BT} = n \mid \mathtt{Pr} = y)P(\mathtt{UT} = n \mid \mathtt{Pr} = y)P(\mathtt{Pr} = y)$
$= 0.36 \cdot 0.27 \cdot 0.87$

and

$$P(\mathtt{Pr} = n \mid \mathtt{BT} = n, \mathtt{UT} = n) = \frac{0.894 \cdot 0.893 \cdot 0.13}{0.894 \cdot 0.893 \cdot 0.13 + 0.36 \cdot 0.27 \cdot 0.87} \approx 0.55.$$

Example II: Insemination

2nd model with the mediating variable Ho



Pr = y	Pr = n
0.87	0.13

	Ho = y	Ho = n
Pr = y	0.9	0.1
Pr =n	0.01	0.99

	BT = y	BT = n
Ho = y	0.7	0.3
Ho =n	0.1	0.9

	UT = y	UT = n
Ho = y	0.8	0.2
Ho =n	0.1	0.9

We calculate $P(\mathtt{Pr} = n \mid \mathtt{BT} = n, \mathtt{UT} = n)$:

$$P(\mathtt{Pr} = n, \mathtt{BT} = n, \mathtt{UT} = n) = \sum_{x \in \{y,n\}} P(\mathtt{BT} = n \mid \mathtt{Ho} = x)P(\mathtt{UT} = n \mid \mathtt{Ho} = x)P(\mathtt{Ho} = x \mid \mathtt{Pr} = n)P(\mathtt{Pr} = n)$$
$$= 0.3 \cdot 0.2 \cdot 0.01 \cdot 0.13 + 0.9 \cdot 0.9 \cdot 0.99 \cdot 0.13 \approx 0.1043$$
$$P(\mathtt{Pr} = y, \mathtt{BT} = n, \mathtt{UT} = n) = \sum_{x \in \{y,n\}} P(\mathtt{BT} = n \mid \mathtt{Ho} = x)P(\mathtt{UT} = n \mid \mathtt{Ho} = x)P(\mathtt{Ho} = x \mid \mathtt{Pr} = y)P(\mathtt{Pr} = y)$$
$$= 0.3 \cdot 0.2 \cdot 0.9 \cdot 0.87 + 0.9 \cdot 0.9 \cdot 0.1 \cdot 0.87 \approx 0.1175$$

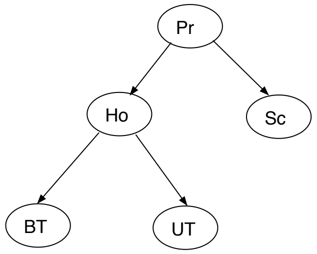
and $\quad P(\mathtt{Pr} = n \mid \mathtt{BT} = n, \mathtt{UT} = n) \approx \dfrac{0.1043}{0.1043 + 0.1175} \approx 0.47.$

(Naive Bayes prediction: 0.55)

Example II: Insemination

Mediating variable can play another important role, as shown here:

- There is dependence between the results of blood test (BT) and urine test (UT). But there is no causal direction in the dependence, and we are also unwilling to introduce a directed edge between the two variables.
- By introducing a mediating variable as their parent variable, we create association – like an undirected edge – between BT and UT when the mediating variable is not observed.



Example III: Stud Farm

Description:

A stud farm has 10 horses. Their geneological structure is shown below. Ann is the mother of both Fred and Gwenn, but their fathers are unrelated and unknown. Every horse may have three genotypes: it may be sick (aa), a carrier (aA), or he may be pure (AA). None of the horses are sick except for John, who has been born recently. As the disease is so serious, the farm wants to find out the probabilities for the remaining horses to be carriers of the unwanted gene.

Hypothesis variables:

- genotypes of all the horses except for John

Information variable:

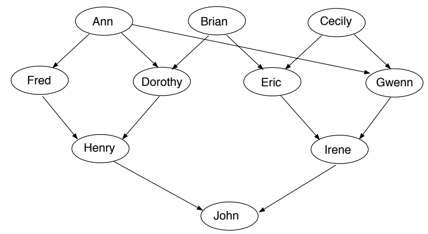
- John's genotype

Additional information:

- none of the other horses are sick

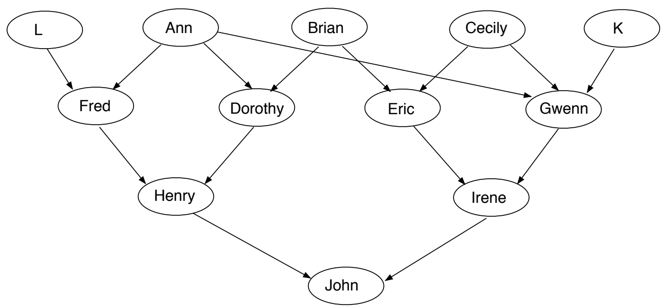
Mediating variables:

- genotypes of the two unknown fathers (L, K)



Example III: Stud Farm

Introduce mediating variables L and K:



The probabilities $P(\texttt{child} \mid \texttt{father}, \texttt{mother})$ of a child's genetic inheritance: numbers are probabilities for (aa, aA, AA)

	aa	aA	AA
aa	(1, 0, 0)	(0.5, 0.5, 0)	(0, 1, 0)
aA	(0.5, 0.5, 0)	(0.25, 0.5, 0.25)	(0, 0.5, 0.5)
AA	(0, 1, 0)	(0, 0.5, 0.5)	(0, 0, 1)

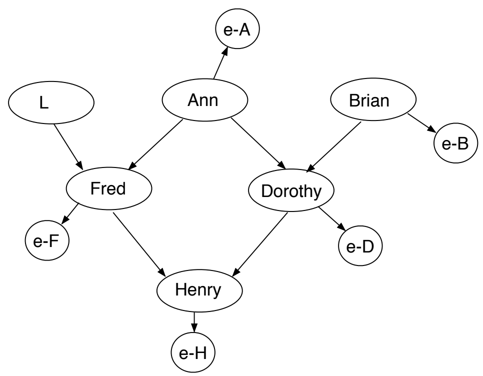
Prior probability of a horse being a carrier or pure:

aA	AA
0.01	0.99

Example III: Stud Farm

We can add evidence variables to represent the additional information that the other horses are not sick:

e−$*$ = {not *aa*}, graph partially shown



Now we have specified the model. (How to compute the probability of a horse being a carrier given all the evidence?)

Outline

- Bayesian Networks
  - Overview
  - Building Models
  - Modeling Tricks

Overview of the List

For handling undirected relations and constraints

- Mediating variables

For reducing the number of parameters in the model

- Noisy-or and its variants
- Divorcing (grouping) parents

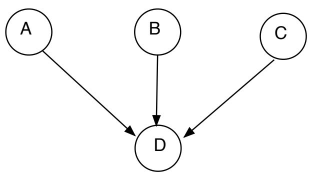
For handling expert disagreement and model adaptation

Handling Undirected Relations and Constraints

Suppose A, B, C are variables that do not have parents. They are marginally dependent with PMF $r(a, b, c)$, but it is undesirable to link them with directed edges.

Introduce variable D, and define

$$P(D = y \mid A = a, B = b, C = c) = r(a, b, c),$$
$$P(D = n \mid A = a, B = b, C = c) = 1 - r(a, b, c).$$



Let $P(A), P(B), P(C)$ be uniform distributions. When using the network, we always enter the evidence $D = y$. Now

$$P(A = a, B = b, C = c \mid D = y) = r(a, b, c).$$

Constraints can be handled similarly. In this case A, B, C can also have parents. If they have to satisfy $f(A, B, C) = 0$, we let

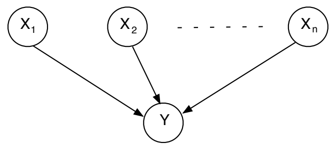
$$P(D = y \mid A = a, B = b, C = c) = \begin{cases} 1 & \text{if } f(a, b, c) = 0, \\ 0 & \text{otherwise.} \end{cases}$$

and we always let $D = y$. (For an example, see the washed-socks example in reference [1].)

Noisy-Or Gate

A typical network for modeling causes and consequences: e.g.,

- $\{X_i\}$: the presences of n possible diseases
- Y: the symptom



Difficulty in building/using the model:

- $\#$ parameters grows exponentially with the number of parents:

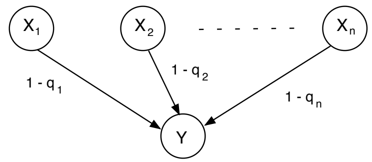
  Even if variables are all binary, the size of the conditional probability table $p(y \mid x_1, \dots, x_n)$ is 2^{n+1} with 2^n free parameters.

Noisy-or trick is useful for reducing the number of parameters, when each cause is thought to act independently.

Noisy-Or Gate

"Noisy-or" assumption (binary case):

- Each event $X_i = 1$ can cause $Y = 1$ unless an *inhibitor* prevents it.
- The inhibition probability is q_i, and the inhibitors are independent.



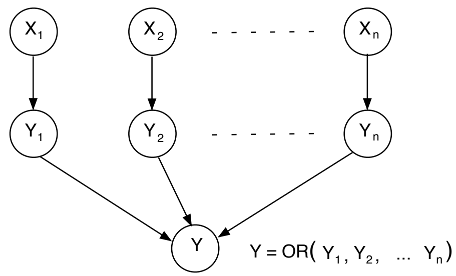
The conditional probabilities of $y_0 = 0,\ y_1 = 1$:

$$p(y_0 \mid x_1, \ldots, x_n) = \prod_{i: x_i = 1} q_i,$$

$$p(y_1 \mid x_1, \ldots, x_n) = 1 - \prod_{i: x_i = 1} q_i.$$

- # parameters is now linear in the number of parents.

Corresponding graphical model:



Generalization: "noisy-and," "noisy-max," "noisy" functional dependence.

Noisy-Or: Family-Out? Example

In the Family-out? example, to assign probabilities $P(\texttt{D-out} \mid \texttt{F-out}, \texttt{BP})$, I reason that there are three causes for the dog to be out, and if any one of them is present, the dog is out:

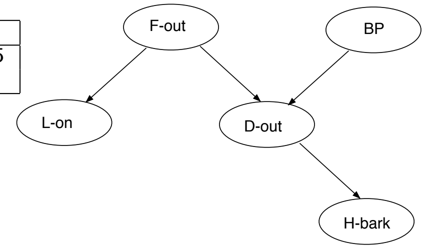
- the "background event" that the dog wants to be out: probability 0.2;
- $\texttt{F-out} = y$, which causes the dog to be out with probability 0.85;
- $\texttt{BP} = y$, which causes the dog to be out with probability 0.95.

Then $P(\texttt{D-out} = y \mid \texttt{F-out}, \texttt{BP})$ is given by

	BP = y	BP = n
F-out = y	$1 - 0.8 \cdot 0.05 \cdot 0.15$	$1 - 0.8 \cdot 0.15$
F-out =n	$1 - 0.8 \cdot 0.05$	$1 - 0.8$

which gives

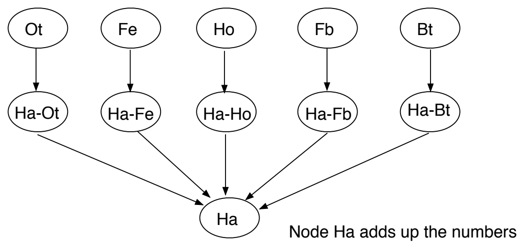
	BP = y	BP = n
F-out = y	0.994	0.88
F-out =n	0.960	0.2



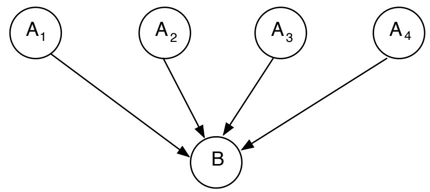
Noisy Functional Dependence

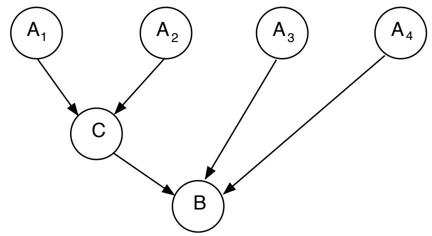
Example: Headache

Headache (Ha) may be caused by fever (Fe), hangover (Ho), fibrositis (Fb), brain tumor (Bt), and other causes (Ot). Let Ha has states *no, mild, moderate, severe*. The various causes support each other in the effect. We still feel however that the impacts of the causes are independent: if the headache is at level ℓ, and we add an extra cause for headache, then we expect the result to be a headache at level q independent of how the initial state has been caused. We want to combine the effects of various causes.



Mediating Variables for "Divorcing" Parents



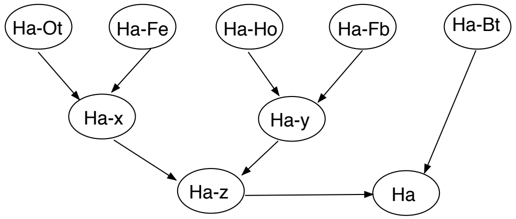


Example: Bank Loan

To help the bank decide when a customer applies for a mortgage, the customer is asked to fill in a form giving information on: type of job, yearly income, other financial commitments, number and type of cars in the family, size and age of the house, price of the house, number of previous addresses during the last five years, number of children in the family, number of divorces, and number of children not living in the family.

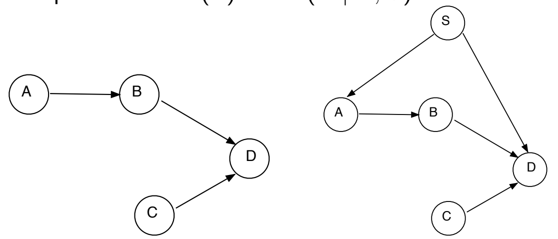
We can partition the 10 variables into three mediating variables describing: economic potentials, stability, and security of the mortgage.

Headache Example with "Divorcing"



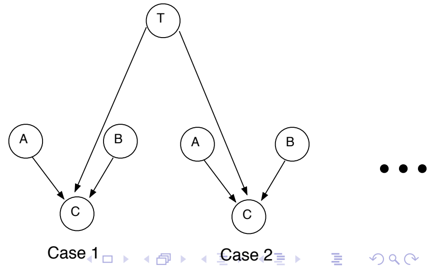
Expert Disagreement and Model Adaptation

Suppose two experts agree on the model structure for A, B, C, D, but disagree on the probabilities $P(A)$ and $P(D \mid B, C)$.



We can add a node S representing the experts, $s \in \{1, 2\}$, and express our confidence in the experts via $P(S)$.

Similarly, we can prepare for adaptation of model parameters based on data by introducing a type variable T and copying other variables for each case:



Further Readings

For an introduction to building models:

1. Finn V. Jensen. *An Introduction to Bayesian Networks*. UCL Press, 1996. Chap. 3.
2. Finn V. Jensen and Thomas D. Nielsen. *Bayesian Networks and Decision Graphs*. Springer, 2007. Chap. 3.

[2] describes more advanced models such as object-oriented Bayesian networks.