Probabilistic Reasoning

Unit # 6

Learning in BN

- Until the early 1990s the DAG in a Bayesian network was ordinarily hand-constructed by a domain expert.
- Then the conditional probabilities were assessed by the expert, learned from data, or obtained using a combination of both techniques.
- Eliciting Bayesian networks from experts can be a laborious and difficult process in the case of large networks.
- As a result, researchers developed methods that could learn the DAG from data.
- Furthermore, they formalized methods for learning the conditional probabilities from data.

MLE vs. MAP

- Suppose you are going to toss a coin.
- If you toss it 100 times and get 48 heads then the *maximum likelihood estimate (MLE)* is
  - P(heads) ≈ 0.48
- However, if you have a prior belief the relative frequency is around = 0.5, you might feel your prior experience is equivalent to having seen 50 heads in 100 tosses. In such case, your *maximum a posterior probability (MAP)* is
  - P(heads | 48, 52) = 98/200 = 0.49

Thumbtack Example

- Suppose you are going to repeatedly toss a thumbtack.
- Based on its structure, you might feel it should land heads about half the time, but you are not nearly so confident as you were with the coin from your pocket.
- So, you might feel your prior experience is equivalent to having seen 3 heads in 6 tosses.
- Then your prior probability of heads is
  - P(head) = 3/6 = 0.5
- After seeing 65 heads in 100 tosses, your posterior probability is
  - P(head | 65, 35) = 68 / 106 = 0.64

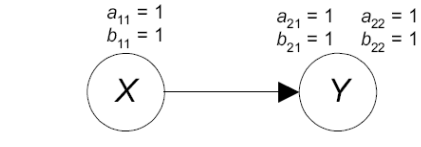
Dice Example

- Suppose we have an asymmetrical-, six-sided die, and we have little idea of the probability of each side coming up. However, it seems that all sides are equally likely. So, we assign 3 to each outcome.
- Suppose next we throw the die 100 times, with the following results:

Outcome	Number of Occurrences
1	10
2	15
3	5
4	30
5	13
6	27

Learning Parameters

- For each probability in the network there is a pair (aij, bij). The i indexes the variable; the j indexes the value of the parent(s) of the variable.
- For example, the pair (a11, b11) is for the first variable (X) and the first value of its parent (in this case there is a default of one parent value since X has no parent).
- The pair (a21, b21) is for the second variable (Y) and the first value of its parent, namely x1.
- We have attempted to represent prior ignorance as to the value of all probabilities by taking aij = bij = 1.



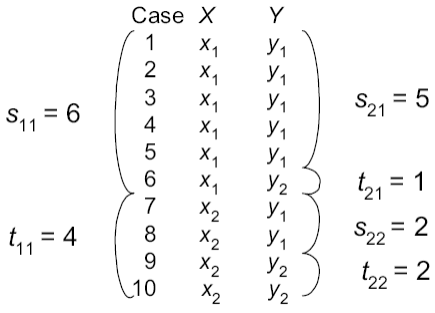
$P(x_1) = a_{11} / (a_{11}+b_{11}) = 1/2$

$P(y_1 \mid x_1) = a_{21} / (a_{21}+b_{21}) = 1/2$

$P(y_1 \mid x_2) = a_{22} / (a_{22}+b_{22}) = 1/2$

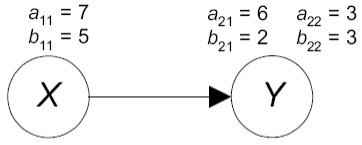
New Data

- When we obtain data, we use an (sij, tij) pair to represent the counts for the ith variable when the variable's parents have their jth value.



Updated Probabilities

- To determine the posterior probability distribution based on the data, we update each conditional probability with the counts relative to that conditional probability. Since we want an updated Bayesian network, we re-compute the values of the (aij, bij) pairs.



$P(x_1) = a_{11} / (a_{11}+b_{11}) = 7/12$

$P(y_1 \mid x_1) = a_{21} / (a_{21}+b_{21}) = 3/4$

$P(y_1 \mid x_2) = a_{22} / (a_{22}+b_{22}) = 1/2$

Right Prior?

Case	X	Y
1	x_1	y_1
2	x_1	y_1
3	x_1	y_1
4	x_1	y_1
5	x_1	y_1
6	x_1	y_2
7	x_2	y_1
8	x_2	y_1
9	x_2	y_2
10	x_2	y_2

- Should we assume a prior distribution?
- If yes, what values should be considered for the prior distribution?

Equivalent Sample Size for Prior

- First try
  - N(x1) = N(x2) = N(y1|x1) = N(y2|x1) = N(y1|x2) = N(y2|x2) = 1
  - where N(a) is the number of cases
  - Compute P(y1). Is it consistent with the prior?
- Then
  - Use the following data set
  - Now compute P(y1).

Case	X	Y
1	x_1	y_1
2	x_1	y_2
3	x_2	y_1
4	x_2	y_2

Theorem

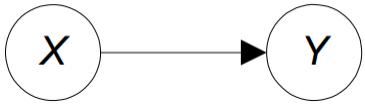
- Suppose we specify a Bayesian network for parameter learning in the case of binomial variables and assign for all i and j

$$aij = bij = N / 2qi$$

- where N is a positive integer and qi is the number of instantiations of the parents of the ith variable. Then the resultant Bayesian network has equivalent sample size N , and the joint probability distribution in the Bayesian network is uniform.

Equivalent Sample Size

- Suppose X and Y are binary variables.

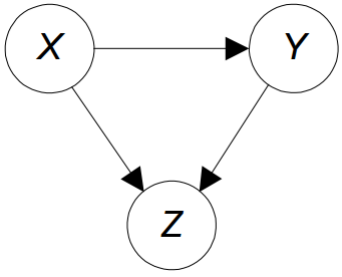


$$a_{11} = b_{11} = \frac{N}{2q_1} = \frac{2}{2 \times 1} = 1$$

$$a_{21} = b_{21} = a_{22} = b_{22} = \frac{N}{2q_2} = \frac{2}{2 \times 2} = .5.$$

Equivalent Sample Size (Cont'd)

- Suppose X, Y and Z re binary variables and we set N = 2.



$$a_{11} = b_{11} = \frac{N}{2q_1} = \frac{2}{2 \times 1} = 1$$

$$a_{21} = b_{21} = a_{22} = b_{22} = \frac{N}{2q_2} = \frac{2}{2 \times 2} = .5$$

$$a_{31} = b_{31} = a_{32} = b_{32} = a_{33} = b_{33} = a_{34} = b_{34} = \frac{N}{2q_3} = \frac{2}{2 \times 4} = .25.$$

Sa

Example (Source: Neapolitan)

- Assume that you feel your prior experience concerning the relative frequency of smokers in a particular bar is equivalent to having seen 14 smokers and 6 nonsmokers.
  - You then decide to poll individuals in the bar and ask them if they smoke. What is your probability of the first individual you poll being a smoker?
  - Suppose that after polling 10 individuals, you obtain these data (the value 1 means the individual smokes and 2 means the individual does not smoke):

    {1, 2, 2, 2, 2, 1, 2, 2, 2, 1}

    What is your probability that the next individual you poll is a smoker?

Example (Cont'd)

– Suppose that after polling 1000 individuals (it is a big bar), you learn that 312 are smokers. What is your probability that the next individual you poll is a smoker? How does this probability compare to your prior probability?

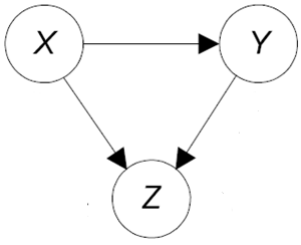
Example II (Source: Neapolitan)

- Suppose that you are going to sample individuals who have smoked two packs of cigarettes or more daily for the past 10 years. You will determine whether each individual's systolic blood pressure is $\leq$100, 101-120, 121-140, 141-160, or $\geq$161. Determine values of a1, a2, …, a5 that represent your prior probability of each blood pressure range.
- Next you sample such smokers. What is your probability of each blood pressure range for the first individual sampled?
- Suppose that after sampling 100 individuals, you obtain the following results:
- Compute your probability of each range for the next individual sampled.

Blood Pressure Range	# of Individuals in This Range
≤ 100	2
101-120	15
121-140	23
141-160	25
≥ 161	35

Example III
(Source: Neapolitan)

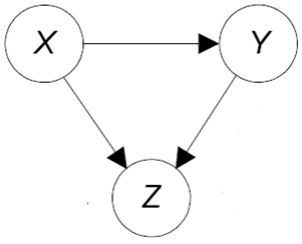
- Suppose that we have the following Bayesian network for parameter learning and the following data.
- Determine the updated BN for parameter learning.



Case	X	Y	Z
1	x_1	y_2	z_1
2	x_1	y_1	z_2
3	x_2	y_1	z_1
4	x_2	y_2	z_1
5	x_1	y_2	z_1
6	x_2	y_2	z_2
7	x_1	y_2	z_1
8	x_2	y_1	z_2
9	x_1	y_2	z_1
10	x_1	y_1	z_1
11	x_1	y_2	z_1
12	x_2	y_1	z_2
13	x_1	y_2	z_1
14	x_2	y_2	z_2
15	x_1	y_2	z_1

Homework Problem

- Develop Bayesian networks for parameter learning with equivalent sample sizes 1, 2, 4, and 8 for the following DAG.



Structure Learning

- Structure Learning consists of learning the DAG in a Bayesian network from data.
- We need to learn a DAG that satisfies the Markov condition with the probability distribution P that is generating the data.
- *We do not know P, all we know are the data.*

Score-based Structure Learning

- In score-based structure learning, we assign a score to a DAG based on how well the DAG fits the data.
- Two popular scores are:
  - Bayesian Score
  - Bayesian Information Criterion Score

Probability of Data

- Suppose that we are about to repeatedly toss a thumbtack (or perform any repeatable experiment with two outcomes).
- Suppose further that we assume exchangeability, and we represent our prior belief concerning the probability of heads using Dirichlet distribution with parameters a and b, where a and b are positive integers and m = a + b.
- Let D be data that consists of s heads and t tails in n trials.
- Then

$$P(\mathrm{D}) = \frac{(m-1)!}{(m+n-1)!} \times \frac{(a+s-1)!(b+t-1)!}{(a-1)!(b-1)!}.$$

Example 8.2 (Source: Neapolitan)

- Suppose that, before tossing a thumbstack, we assign a=3 and b=5 to model the slight belief that tails is more probable than heads.
- We then toss the thumbstack ten times and obtain four heads and six tails.
- The probability of obtaining these data D is given by

$$P(D) = \frac{(8-1)!}{(8+10-1)!} \times \frac{(3+4-1)!\,(5+6-1)!}{(3-1)!\,(5-1)!}$$

Example 8.2 (Source: Neapolitan)

- Sometimes the following equation

$$P(\mathrm{D}) = \frac{(m-1)!}{(m+n-1)!} \times \frac{(a+s-1)!(b+t-1)!}{(a-1)!(b-1)!}.$$

- is written as

$$P(\mathrm{D}) = \frac{\Gamma(m)}{\Gamma(m+n)} \times \frac{\Gamma(a+s)\Gamma(b+t)}{\Gamma(a)\Gamma(b)}.$$

- Γ denotes the gamma function. When n is an integer ≥ 1, we have

$$\Gamma(n) = (n-1)!$$

Learning DAG Models using Bayesian Score

- We can score a DAG model G based on data D by determining how probable the data are given the DAG model. That is, we compute (D|G).
- The formula for this probability is the same as discussed in the previous slides, except there is a term for each probability in the network.

Learning DAG Models using Bayesian Score (Cont'd)

$$P(\mathrm{D}|\mathbb{G}_1) = \frac{\Gamma(m_{11})}{\Gamma(m_{11}+n_{11})} \times \frac{\Gamma(a_{11}+s_{11})\Gamma(b_{11}+t_{11})}{\Gamma(a_{11})\Gamma(b_{11})} \times \frac{\Gamma(m_{21})}{\Gamma(m_{21}+n_{21})} \times \frac{\Gamma(a_{21}+s_{21})\Gamma(b_{21}+t_{21})}{\Gamma(a_{21})\Gamma(b_{21})} \times \frac{\Gamma(m_{22})}{\Gamma(m_{22}+n_{22})} \times \frac{\Gamma(a_{22}+s_{22})\Gamma(b_{22}+t_{22})}{\Gamma(a_{22})\Gamma(b_{22})}.$$

$$P(\mathrm{D}|\mathbb{G}_2) = \frac{\Gamma(m_{11})}{\Gamma(m_{11}+n_{11})} \times \frac{\Gamma(a_{11}+s_{11})\Gamma(b_{11}+t_{11})}{\Gamma(a_{11})\Gamma(b_{11})} \times \frac{\Gamma(m_{21})}{\Gamma(m_{21}+n_{21})} \times \frac{\Gamma(a_{21}+s_{21})\Gamma(b_{21}+t_{21})}{\Gamma(a_{21})\Gamma(b_{21})}.$$

Examples

- Compute P(G_1 |D) and P(G_2 |D) for the following cases. Assume P(G_1) = P(G_2) = 0.5

Case	J	F
1	j_1	f_1
2	j_1	f_1
3	j_1	f_1
4	j_1	f_1
5	j_1	f_2
6	j_2	f_1
7	j_2	f_2
8	j_2	f_2

Case	J	F
1	j_1	f_1
2	j_1	f_1
3	j_1	f_1
4	j_1	f_1
5	j_2	f_2
6	j_2	f_2
7	j_2	f_2
8	j_2	f_2

Case	J	F
1	j_1	f_1
2	j_1	f_1
3	j_1	f_2
4	j_1	f_2
5	j_2	f_1
6	j_2	f_1
7	j_2	f_2
8	j_2	f_2

- (0.517, 0.483) (0.959, 0.041) (0.331,0.661)

Learning DAG Patterns

- The DAG F → J is Markov equivalent to the DAG J → F.
- As long as we use a prior equivalent sample size, they will have the same scores.
- In general, we cannot distinguish Markov equivalent DAGs based on data.
- So, we are actually learning Markov equivalence classes (DAG patterns) when we learn a DAG model from data.

Scoring Larger DAG Models

- Suppose we have a DAG G = (V, E) where V is a set of binomial random variables, we assume exchangeability, and we use a Dirichlet distribution to represent our prior belief for each conditional probability distribution of every variable in V.
- Suppose further we have data D consisting of a set of data items such that each data item is a vector of values of all the variables in V. Then

$$P(\mathsf{D}|\mathbb{G}) = \prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(N_{ij})}{\Gamma(N_{ij}+M_{ij})}\frac{\Gamma(a_{ij}+s_{ij})\Gamma(b_{ij}+t_{ij})}{\Gamma(a_{ij})\Gamma(b_{ij})}$$

Scoring Larger DAG Models (Cont'd)

- n is the number of variables
- q_i is the number of different instantiations of the parents of X_i.
- a_{ij} is our ascertained prior belief concerning the number of *times* X_i *took*
- *its first value when the parents of* X_i *had their* j_{th} *instantiation.*
- b_{ij} is our ascertained prior belief concerning the number of *times* X_i *took*
- *its first value when the parents of* X_i *had their* j_{th} *instantiation.*
- s_{ij} is *the number of times in the data that* X_i *took its first value when the*
- *parents of* X_i *had their* j_{th} *instantiation.*
- t_{ij} is *the number of times in the data that* X_i *took its first value when the*
- *parents of* X_i *had their* j_{th} *instantiation.*
- *N_{ij} and M_{ij} are as follows:*
  - $N_{ij} = a_{ij} + b_{ij}$
  - $M_{ij} = s_{ij} + t_{ij}$

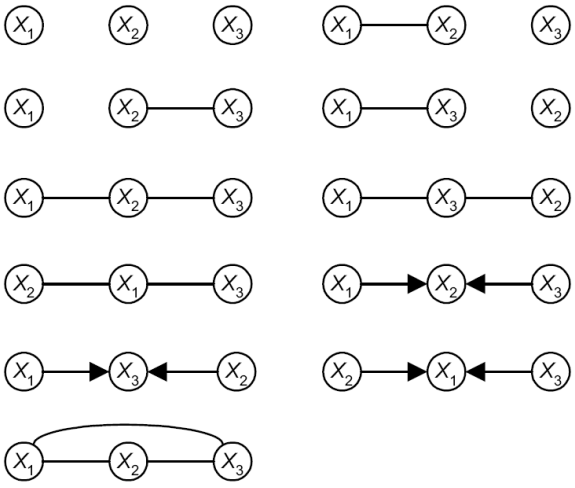
Number of Possible DAGs

- When there are not many variables, we can exhaustively score all possible DAGs. We then select the DAG(s) with the highest score.
- However, when the number of variables is not small, it is computationally unfeasible to find the maximizing DAGs by exhaustively considering all DAG patterns.
- Number of DAGs containing n nodes is:

$$
\begin{aligned}
f(n) &= \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} 2^{i(n-i)} f(n-i) \qquad n > 2 \\
f(0) &= 1 \\
f(1) &= 1
\end{aligned}
$$

- f(2) = 3, f(3) = 25, f(5) = 29,000 and f(10) = 4.2×10^{18}

DAG Pattern with 3 Variables



Number of Possible DAGs (Source: Jensen)

Nodes	Number of DAGs	Nodes	Number of DAGs
1	1	13	$1.9 \cdot 10^{31}$
2	3	14	$1.4 \cdot 10^{36}$
3	25	15	$2.4 \cdot 10^{41}$
4	543	16	$8.4 \cdot 10^{46}$
5	29281	17	$6.3 \cdot 10^{52}$
6	$3.8 \cdot 10^{6}$	18	$9.9 \cdot 10^{58}$
7	$1.1 \cdot 10^{9}$	19	$3.3 \cdot 10^{65}$
8	$7.8 \cdot 10^{11}$	20	$2.35 \cdot 10^{72}$
9	$1.2 \cdot 10^{15}$	21	$3.5 \cdot 10^{79}$
10	$4.2 \cdot 10^{18}$	22	$1.1 \cdot 10^{87}$
11	$3.2 \cdot 10^{22}$	23	$7.0 \cdot 10^{94}$
12	$5.2 \cdot 10^{26}$	24	$9.4 \cdot 10^{102}$