Probabilistic Reasoning/ Bayesian Networks

14 PROBABILISTIC REASONING

In which we explain how to build network models to reason under uncertainty according to the laws of probability theory.

The Bayesian Network
Why?

- Goal: Represent joint distribution P over some set of random variables $\mathcal{X} = \{X_1, \ldots, X_n\}$
- Worst case requires: $2^n - 1$ numbers
  - Computationally Expensive
  - Cognitively impossible for human experts
  - Costly Statistical data requirements

The Bayesian Network PLAN

- Exploit Islands of Tractability in High Dimensional Space Probability Distributions
  - Worst Case is Intractable
  - Real World Frequently NOT Worst Case
  - Efficiently Exploit properties in Real World Probability Distributions to induce Tractability

- Primary Tool:

INDEPENDENCE!

Independence Importance

- Independence w/ Probability Distribution enables a much more compact representation!
- I.E, FULLY INDEPENDENT
  - $P(X_1, \ldots, X_n)$ normally requires $2^n - 1$
  - $P(X_1, \ldots, X_n) = P(X_1)P(X_2)\ldots P(X_n)$ w/ INDEPENDENCE
    - Now Only 2n values!

Good News & Bad News

- Bad News:
  - Full Marginal Independence is rare!
- Good News:
  - Another type of independence is common!

Independence Intuition Example

- Acme Corporation
  - Needs to hire new analyst
  - High Intelligence is desired in analyst
- Problem:
  - Intelligence not directly measurable!?!?
- Solution:
  - Infer Intelligence w/ intelligence indicator!
  - SAT Scores are available from student applicants.

Independence Intuition Example

- Distribution Induced Included:
  - Intelligence: i^0 (low), i^1 (high)
  - SAT: s^0 (low), s^1 (high)
- Example Probability Distribution:

Intelligence (I)	SAT (S)	P(I, S)
i^0	s^0	0.665
i^0	s^1	0.035
i^1	s^0	0.06
i^1	s^1	0.24

Is this a legal probability distribution???

First: Let's Consider Chain Rule

- Chain Rule of Conditional Probabilities
- $P(X_1,\ldots,X_k)$ =
  - $P(X_1)P(X_2 \mid X_1)\cdots P(X_k \mid X_1,\ldots,X_{k-1})$.
- Yielding for our example:
  - $P(I,S) = P(I)P(S \mid I)$

- Who Cares?

Chain Rule… So What?

- First: Now w/ P(I) together w/ P(S|I)

P(Intelligence)	
i^0	i^1
0.7	0.3

P(SAT | Intelligence)		
I	s^0	s^1
i^0	0.95	0.5
i^1	0.2	0.8

- Now: P(I, S) = P(I)P(S|I)
- Still… So What?

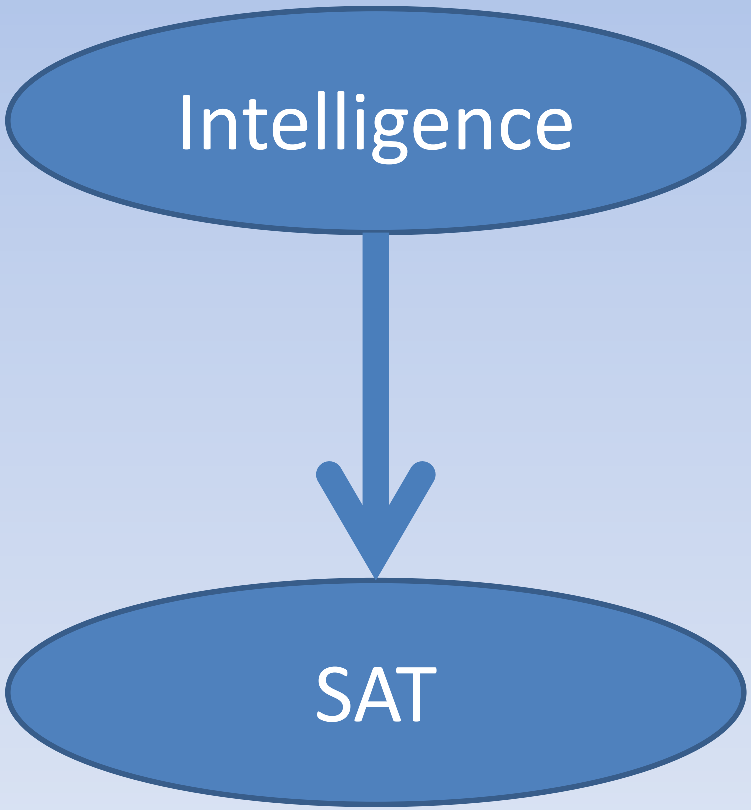
First: P(SAT | Intelligence)

P(SAT | Intelligence)		
I	s^0	s^1
i^0	0.95	0.5
i^1	0.2	0.8

- Intuitively, we are representing the process in a way that is more compatible w/ Causality.
- Various factors (genetics, upbringing, . . .) first determined (stochastically) the student's intelligence.
- His performance on the SAT is determined (stochastically) by his intelligence.
- We note that the models we construct are not required to follow causal intuitions, but they often do.

Peek @ Bayesian Net

- Have Not Defined Bayesian Networks yet… But:



Naïve Bayes Model

- Have all the tools needed to understand Naïve Bayes Model
- "Simplest example where a conditional parameterization is combined with conditional independence assumptions to produce a very compact representation of a high-dimensional probability distribution." pg. 48
- Start w/ Expanding Student Example

SAT & Grade

- Acme corporation has expanded their selection process.
  - Intelligence: i^0 (low), i^1 (high)
- Acme still has access to applicant SAT score:
  - SAT: s^0 (low), s^1 (high)
- Acme now has access to an applicant Grade in course:
  - Grade: $\{g^1, g^2, g^3\}$
- Now, How many independent parameters?

Full Joint Distribution

Intelligence (I)	SAT (S)	Grade (G)	P(I, S, G)
I^0	s^0	g^1	0.126
I^0	s^0	g^2	0.168
I^0	s^0	g^3	0.126
I^0	s^1	g^1	0.009
I^0	s^1	g^2	0.045
I^0	s^1	g^3	0.126
I^1	s^0	g^1	0.252
I^1	s^0	g^2	0.0224
I^1	s^0	g^3	0.0056
I^1	s^1	g^1	0.06
I^1	s^1	g^2	0.036
I^1	s^1	g^3	0.024

- How many independent parameters?

11 Independent Parameters

- Remember: Each Probability Distribution must SUM TO ONE.
- For each separate probability distribution we utilize, we can leave out One Parameter.
- One Parameter is fully determined by others.
  - Since the complete set must SUM TO ONE!

Independence w/ Student Example

- NOTE:
  - No Marginal Independencies!
- But:
  - Conditional Independencies??
- "If we know that the student has high intelligence, a high grade on the SAT no longer gives us information about the student's performance in the class."

Conditional Independence w/ Student Example

- "If we know that the student has high intelligence, a high grade on the SAT no longer gives us information about the student's performance in the class."
- Formally:
  - $P(g \mid i^1, s^1) = P(g \mid i^1)$.

Conditional Independence w/ Student Example

- Conditional Independence Assumed
  - $P \models (S \perp G \mid I)$
- Now we know:
  - $P(I,S,G) = P(S,G \mid I)P(I)$.
- Conditional Independence Yields:
  - $P(S, G \mid I) = P(S|I)\, P(G|I)$
- SO :
  - $P(I, S, G) = P(S|I)\, P(G|I)\, P(I)$

➢ So What ??

Full Joint Distribution

Intelligence (I)	SAT (S)	Grade (G)	P(I, S, G)
I^0	s^0	g^1	0.126
I^0	s^0	g^2	0.168
I^0	s^0	g^3	0.126
I^0	s^1	g^1	0.009
I^0	s^1	g^2	0.045
I^0	s^1	g^3	0.126
I^1	s^0	g^1	0.252
I^1	s^0	g^2	0.0224
I^1	s^0	g^3	0.0056
I^1	s^1	g^1	0.06
I^1	s^1	g^2	0.036
I^1	s^1	g^3	0.024

- $P(I, S, G) = P(S|I)\ P(G|I)\ P(I)$
- How many independent parameters?

Full Joint Distribution

Intelligence (I)	SAT (S)	Grade (G)	P(I, S, G)
I^0	s^0	g^1	0.126
I^0	s^0	g^2	0.168
I^0	s^0	g^3	0.126
I^0	s^1	g^1	0.009
I^0	s^1	g^2	0.045
I^0	s^1	g^3	0.126
I^1	s^0	g^1	0.252
I^1	s^0	g^2	0.0224
I^1	s^0	g^3	0.0056
I^1	s^1	g^1	0.06
I^1	s^1	g^2	0.036
I^1	s^1	g^3	0.024

- $P(I, S, G) = P(S|I)\ P(G|I)\ P(I)$
- How many independent parameters?

Full Joint Distribution: Parameterized

P(Intelligence)	
i^0	i^1
0.7	0.3

P(Grade | Intelligence)			
I	g^1 (A)	g^2 (B)	g^3(C)
i^0	0.2	0.34	0.46
i^1	0.74	0.17	0.09

P(SAT | Intelligence)		
I	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

- P(I, S, G) = P(S|I) P(G|I) P(I)
- How many independent parameters?
  - 7

Full Joint Distribution: Parameterized

P(Intelligence)	
i^0	i^1
0.7	0.3

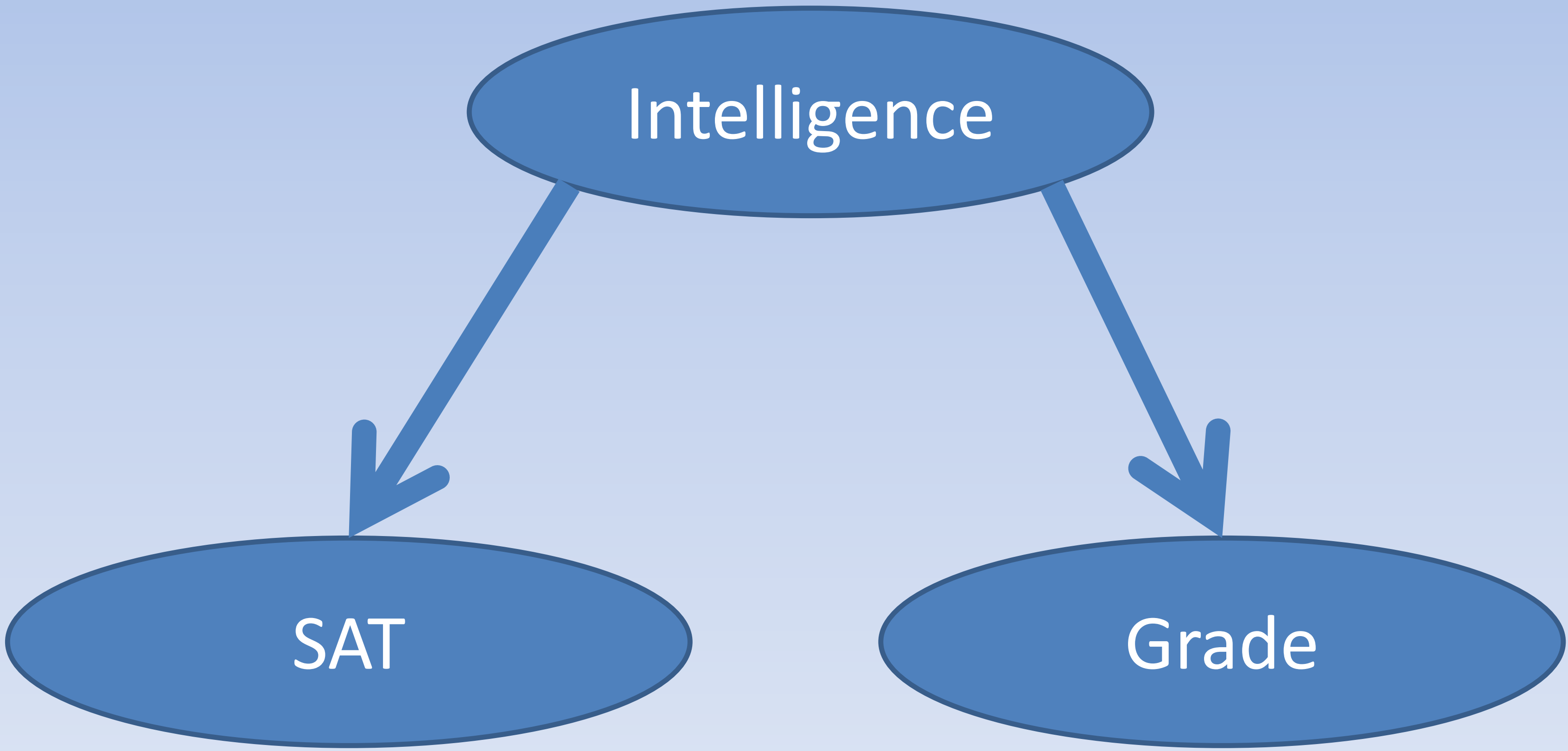
P(Grade | Intelligence)			
I	g^1 (A)	g^2 (B)	g^3(C)
i^0	0.2	0.34	0.46
i^1	0.74	0.17	0.09

P(SAT | Intelligence)		
I	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

- P(I, S, G) = P(S|I) P(G|I) P(I)
- How many independent parameters?
  - 7

Peek @ Bayesian Net

- Have Not Defined Bayesian Networks yet… But:



Queries

P(Intelligence)	
i^0	i^1
0.7	0.3

P(Grade | Intelligence)			
I	g^1 (A)	g^2 (B)	g^3(C)
i^0	0.2	0.34	0.46
i^1	0.74	0.17	0.09

P(SAT | Intelligence)		
I	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

- What is the probability of (i^1, s^1, g^1)
- $P(i^1, s^1, g^1) = P(i^1)\ P(s^1 \mid i^1)\ P(g^1 \mid i^1)$

P(SAT | Intelligence)		
I	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

Queries

P(Intelligence)	
i^0	i^1
0.7	0.3

P(Grade|Intelligence)			
I	g^1	g^2	g^3
i^0	0.2	0.34	0.46
i^1	0.74	0.17	0.09

G	P(Grade , i1, s1)
g^1 (A)	0.1776
g^2 (B)	0.0408
g^3 (C)	0.0216

- What is the probability of (i^1, s^1, g^1)
- $P(i^1, s^1, g^1) = P(i^1)\ P(s^1 \mid i^1)\ P(g^1 \mid i^1)$
- What is the probability of $(g^1 \mid i^1, s^1)$???

P(SAT | Intelligence)		
I	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

Queries

P(Intelligence)	
i^0	i^1
0.7	0.3

P(Grade|Intelligence)			
I	g^1	g^2	g^3
i^0	0.2	0.34	0.46
i^1	0.74	0.17	0.09

G	P(Grade | i^1, s^1)
g^1 (A)	0.74
g^2 (B)	0.17
g^3 (C)	0.09

- Normalize!
  - Sum over Grades for $P(G, i^1, s^1) = P(i^1, s^1)$
- What is the probability of $(g^1 \mid i^1, s^1)$

Queries

P(Grade | Intelligence, SAT)			
I, S	g^1 (A)	g^2 (B)	g^3(C)
i^0,s^0			
i^0,s^1			
i^1,s^0			
i^1,s^1			

- Normalize!
- What is the probability of (g^1 | i^1, s^1)

Queries

P(Grade | Intelligence, SAT)			
I, S	g^1 (A)	g^2 (B)	g^3(C)
i^0,s^0			
i^0,s^1			
i^1,s^0			
i^1,s^1	0.74	0.17	0.09

- Normalize!
- What is the probability of $(g^1 \mid i^1, s^1)$

Queries

P(Grade | Intelligence, SAT)			
I, S	g^1 (A)	g^2 (B)	g^3(C)
i^0,s^0	0.2	0.34	0.46
i^0,s^1	0.2	0.34	0.46
i^1,s^0	0.74	0.17	0.09
i^1,s^1	0.74	0.17	0.09

- Normalize!
- What is the probability of $(g^1 \mid i^1, s^1)$

Queries

- Turns out we really don't know intelligence.
- Want: $P(G \mid S)$

- What do we do?
  - $P(G, S)/P(S)$

- Marginalize!
  - $P(G, S, I) \Rightarrow P(G, S)$
  - $P(G, S) \Rightarrow P(S)$

Queries

P(Grade | Intelligence, SAT)			
I, S	g^1 (A)	g^2 (B)	g^3(C)
i^0,s^0	0.2	0.34	0.46
i^0,s^1	0.2	0.34	0.46
i^1,s^0	0.74	0.17	0.09
i^1,s^1	0.74	0.17	0.09

P(Grade | SAT)			
S	g^1 (A)	g^2 (B)	g^3(C)
s^0			
s^1			

Queries

P(Grade|Intelligence, SAT)			
I, S	g^1 (A)	g^2 (B)	g^3(C)
i^0,s^0	0.2	0.34	0.46
i^0,s^1	0.2	0.34	0.46
i^1,s^0	0.74	0.17	0.09
i^1,s^1	0.74	0.17	0.09

P(Grade|SAT)			
S	g^1 (A)	g^2 (B)	g^3(C)
s^0			
s^1			

Original Data

P(Intelligence)	
i^0	i^1
0.7	0.3

P(Grade | Intelligence)			
I	g^1 (A)	g^2 (B)	g^3(C)
i^0	0.2	0.34	0.46
i^1	0.74	0.17	0.09

P(SAT | Intelligence)		
I	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

- P(G|S) = P(G,S)|P(S)
  - P(I, S, G) = P(I)P(S|I)P(G|I)
  - Marginalize out I => P(S,G)
  - Marginalize out G => P(S)

i	s	g	P(I,S,G)	P(S,G)	P(S)
i0	s0	g1	0.133		
i1	s0	g1	0.0444	0.1774	
i0	s0	g2	0.2261		
i1	s0	g2	0.0102	0.2363	
i0	s0	g3	0.3059		
i1	s0	g3	0.0054	0.3113	0.725
i0	s1	g1	0.007		
i1	s1	g1	0.1776	0.1846	
i0	s1	g2	0.0119		
i1	s1	g2	0.0408	0.0527	
i0	s1	g3	0.0161		
i1	s1	g3	0.0216	0.0377	0.275
			1	1	1

P(G|S)

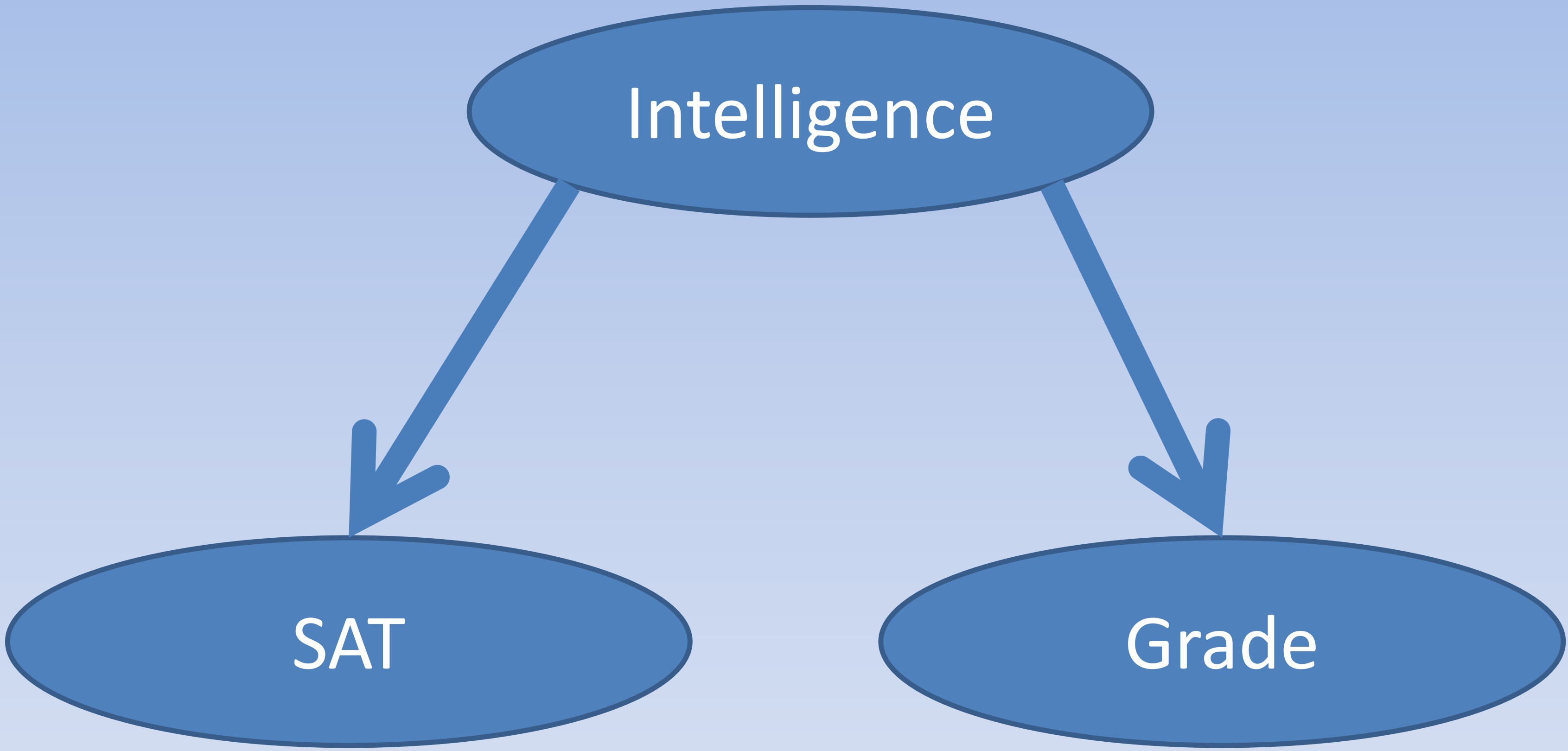
s	g	P(S,G)	P(S)	P(G|S)	
s0	g1	0.1774		0.244	
s0	g2	0.2363		0.326	
s0	g3	0.3113	0.725	0.429	
s1	g1	0.1846		0.671	
s1	g2	0.0527		0.192	
s1	g3	0.0377	0.275	0.137	
		1	1		

P(G|S)

s	g	P(S,G)	P(S)	P(G|S)	
s0	g1	0.1774		0.244	
s0	g2	0.2363		0.326	
s0	g3	0.3113	0.725	0.429	
s1	g1	0.1846		0.671	
s1	g2	0.0527		0.192	
s1	g3	0.0377	0.275	0.137	
		1	1		

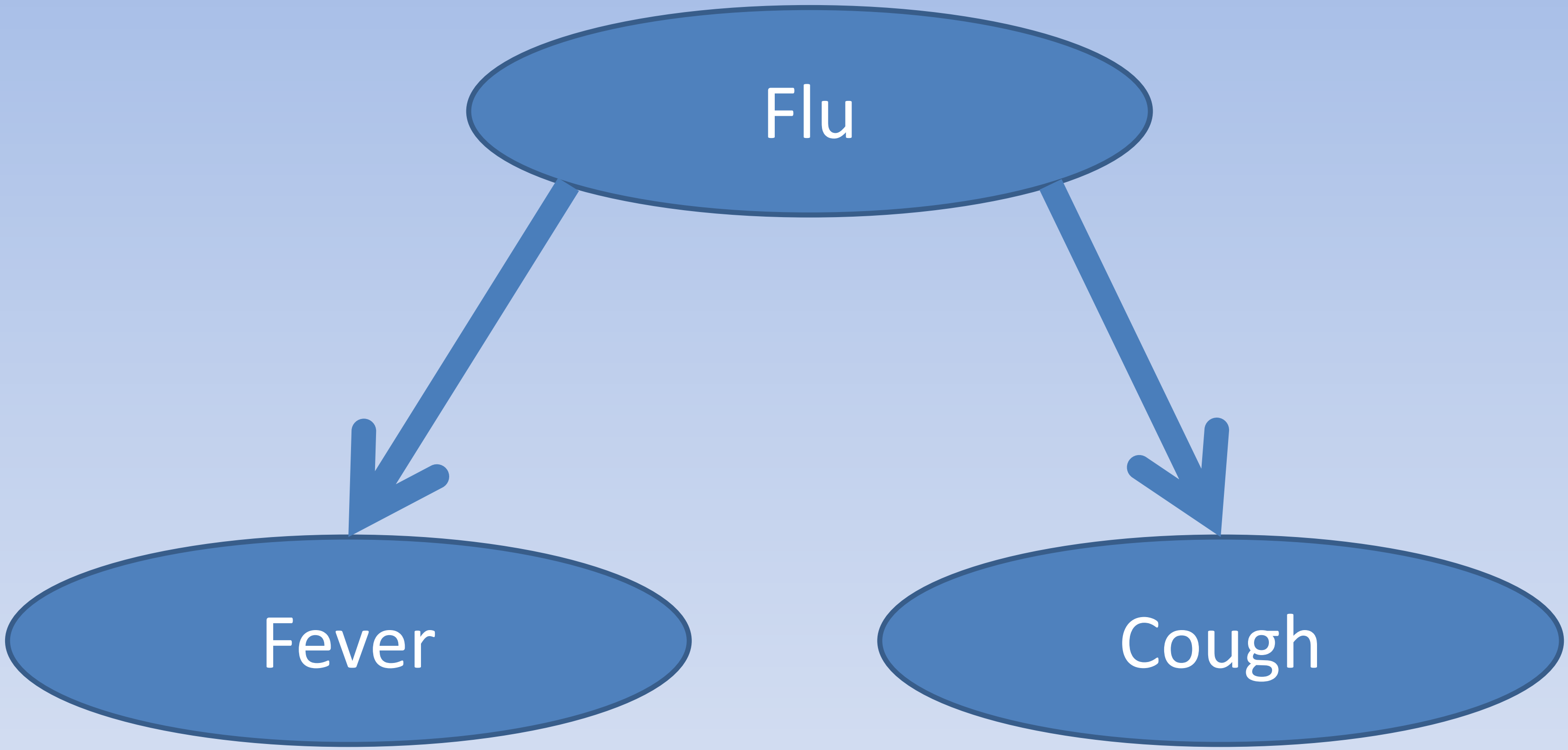
- Probability of an A (g1) given high SAT (s1) is 0.671!
- Probability of a C (g3) given low SAT (s0) is 0.429
  - It is MAP Assignment:
    - P(B|low SAT) = 0.326
    - P(A|low SAT) = 0.244

Naïve Bayes



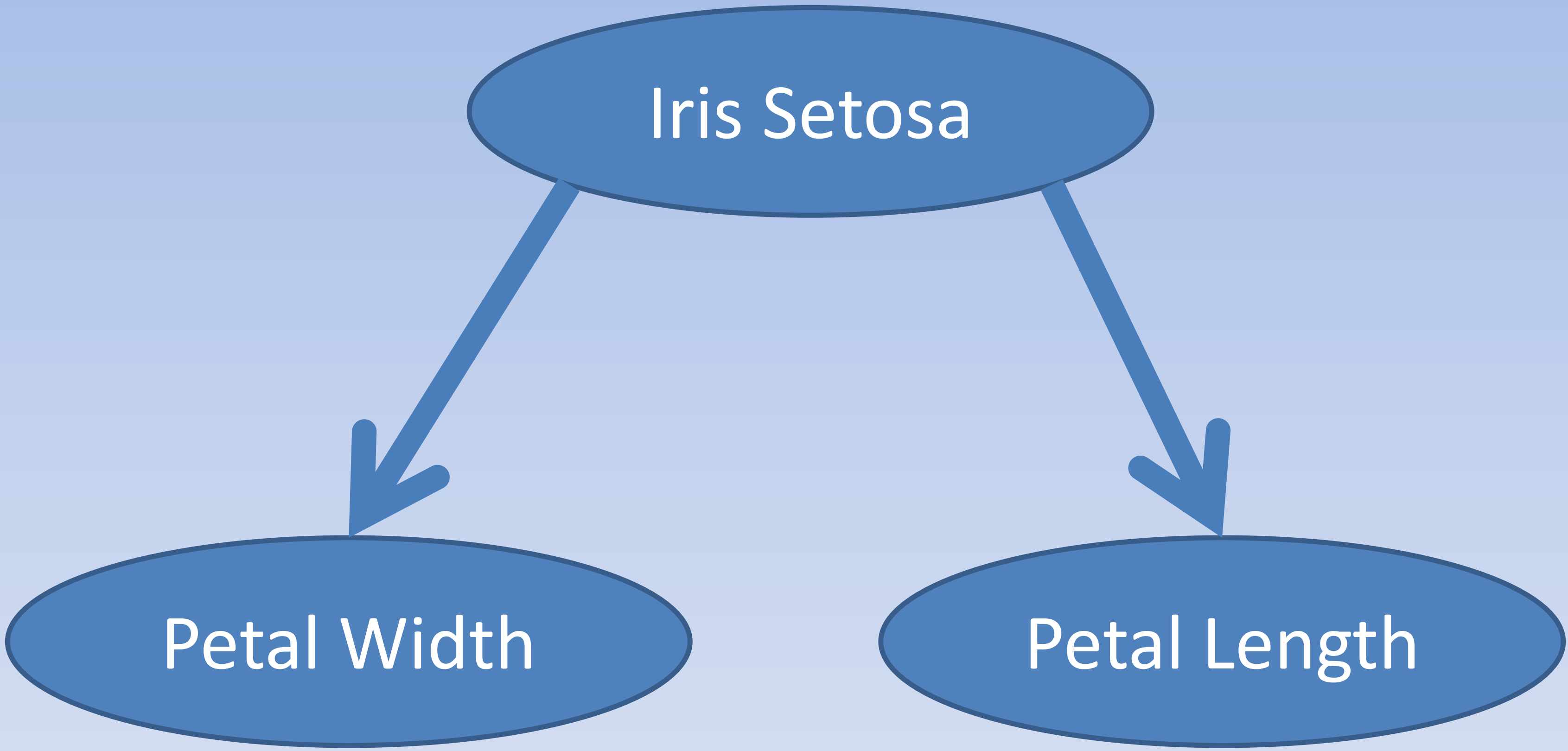
- P(Intelligence| Grade, SAT)

Naïve Bayes



- P(Flu| Fever, Cough)

Naïve Bayes



- P(Iris Setosa| Petal Width, Petal Length)

Naïve Bayes: Generally

50 *Chapter 3. The Bayesian Network Representation*

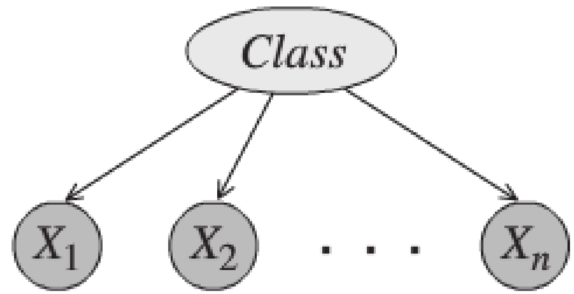


Figure 3.2 **The Bayesian network graph for a naive Bayes model**

- AKA: Idiot Bayes
- Naive Bayes Assumption:
  - Conditionally Independent features given class.
  - $(X_i \perp X_{-i} \mid C)$ for all i

Naïve Bayes: Generally

- Naive Bayes Assumption:
  - Conditionally Independent features given class.
  - $(X_i \perp X_{-i} \mid C)$ for all i
    - $X_{-i} = \{X_1,\ldots,X_n\} - \{X_i\}$

- $P(C, X_1, \ldots, X_n) = P(C)P(X_1|C)P(X_2|C)\ldots P(X_n|C)$
- Full Joint would require (Assuming Booleans)?
  - 2^N-1 parameters
- Parameterized under Naïve Bayes Assumption Requires:
  - 2N+1 Parameters!!!!

Naïve Bayes Model
Used Frequently !

- Naïve Bayes Model is used frequently because of the simplicity!
- Easily used to choose between two classes given features:
  - Odds of c^1 versus c^2
  - Does not require normalization

$$\frac{P(c^1 \mid x_1, \ldots, xn)}{P(c^2 \mid x_1, \ldots, xn)}$$

Naïve Bayes Model
Used Frequently !

- Naïve Bayes Model is used frequently because of the simplicity!
- Easily used to choose between two classes given features:
  - Odds of c^1 versus c^2
  - Does not require normalization

$$\frac{P(c^1 \mid x_1, \ldots, xn)}{P(c^2 \mid x_1, \ldots, xn)} = \frac{P(c^1)}{P(c^2)} \prod \frac{P(x_i \mid c^1)}{P(xi \mid c^2)}$$

Naïve Model: Issues

- This model was used in the early days of medical diagnosis.
- Small number of parameters needed.
- Experts Easily Elicited for parameters.
- Several early systems were shown to provide better diagnoses than those made by expert physicians.

Naïve Model: Issues

- Model makes several strong assumptions that are not generally true
- Patient can have at most one disease
- Given the patient's disease, symptoms & test results all independent.
- In particular, the model tends to overestimate the impact of certain evidence by "overcounting" it.

Naïve Model: Issues

- Both hypertension (high blood pressure) and obesity are strong indicators of heart disease.
  - However, these two symptoms are themselves highly correlated!
- Naïve Bayes Model, which contains a multiplicative term for each of them, double-counts the evidence they provide about the disease.
- Studies show that the diagnostic performance of a naive Bayes model can degrade as features increase!
- Degradation often traced to violations of the strong conditional independence assumption.

- This phenomenon led to the use of more complex Bayesian networks, with more realistic independence assumptions…….

Introducing: Bayesian Networks

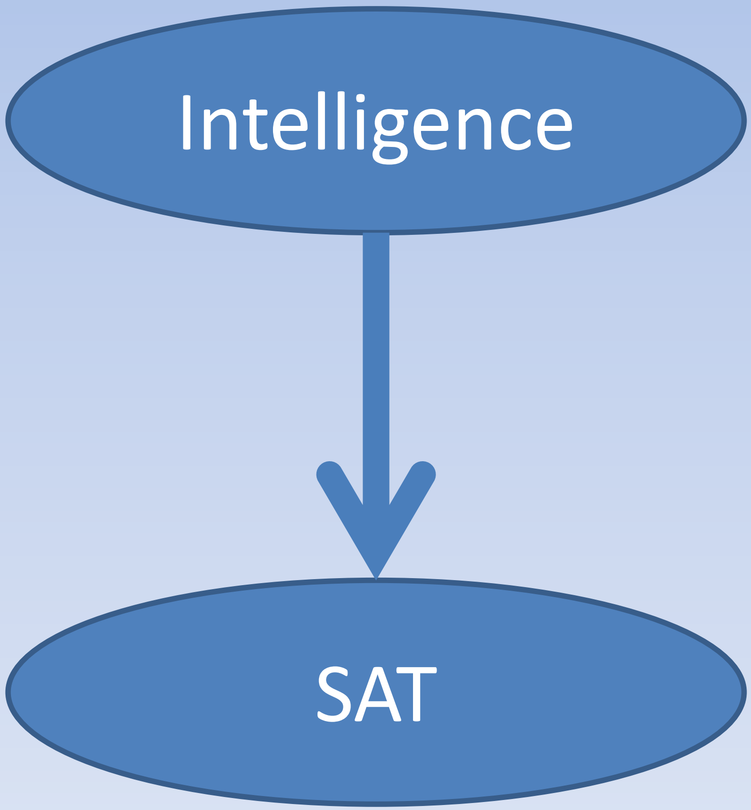
- Intuitions similar to Naïve Bayes Model
- Conditional Independencies exploited to allow representation that is Compact & Natural.
- But Not Restricted w/ Naïve independence assumptions of Naïve Bayes Model.
- Tailoring allowed so our representation of the distribution only include reasonable independencies!

Bayesian Networks : Finally

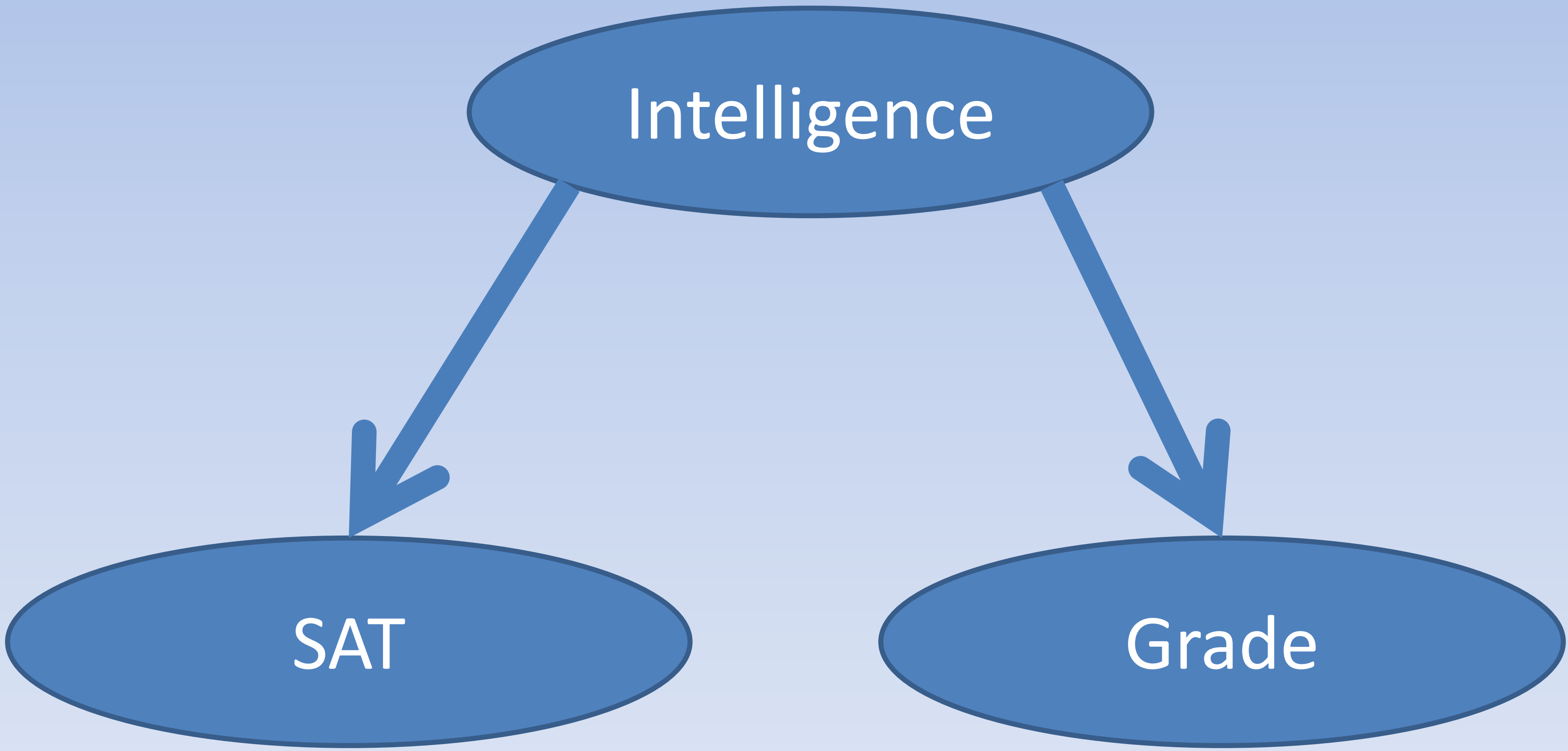
- Core Idea:
  - Directed Acyclic Graph (DAG)
  - Nodes represent Random Variables in our Domain
  - Edges represent a direct influence from one variable to another.

- Let's Revisit a Few….

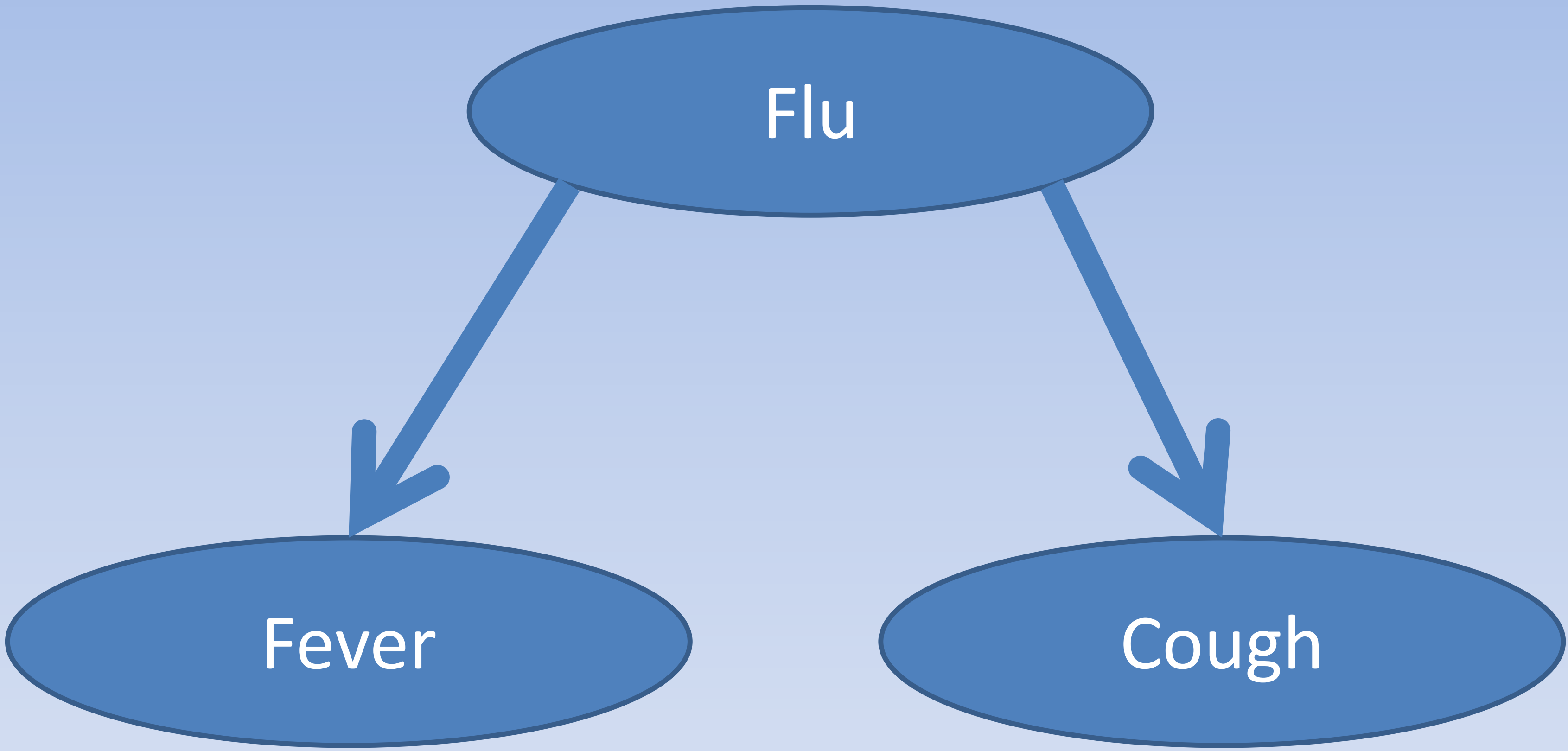
Bayesian Network Graph



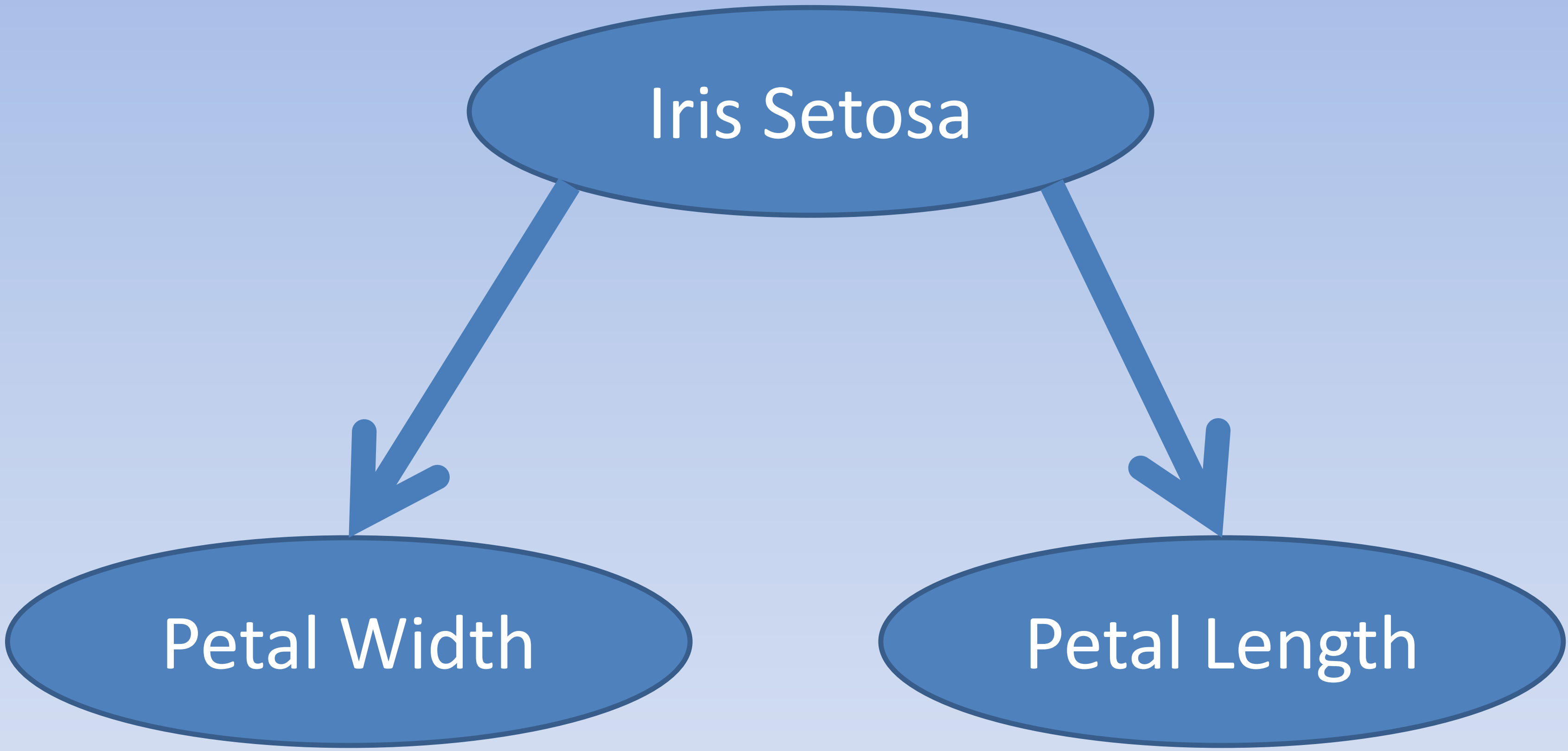
Bayesian Network Graph



Bayesian Network Graph



Bayesian Network Graph



Bayesian Net Graph

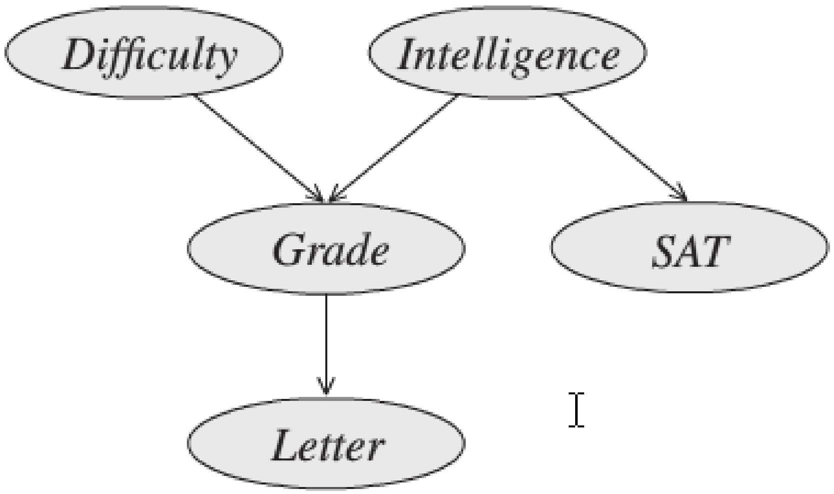


Figure 3.3 The Bayesian Network graph for the Student **example**

Bayesian Network Graph

- View 1: Data structure that provides the skeleton for representing a joint distribution compactly in a factorized way.

- View 2: Compact representation of Conditional Independence Assumptions.

Enhanced Example

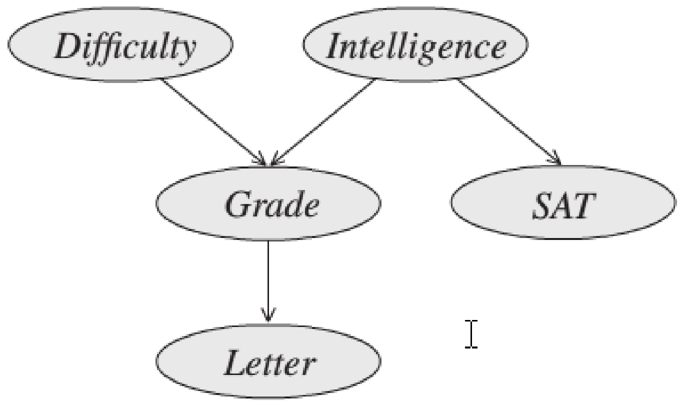
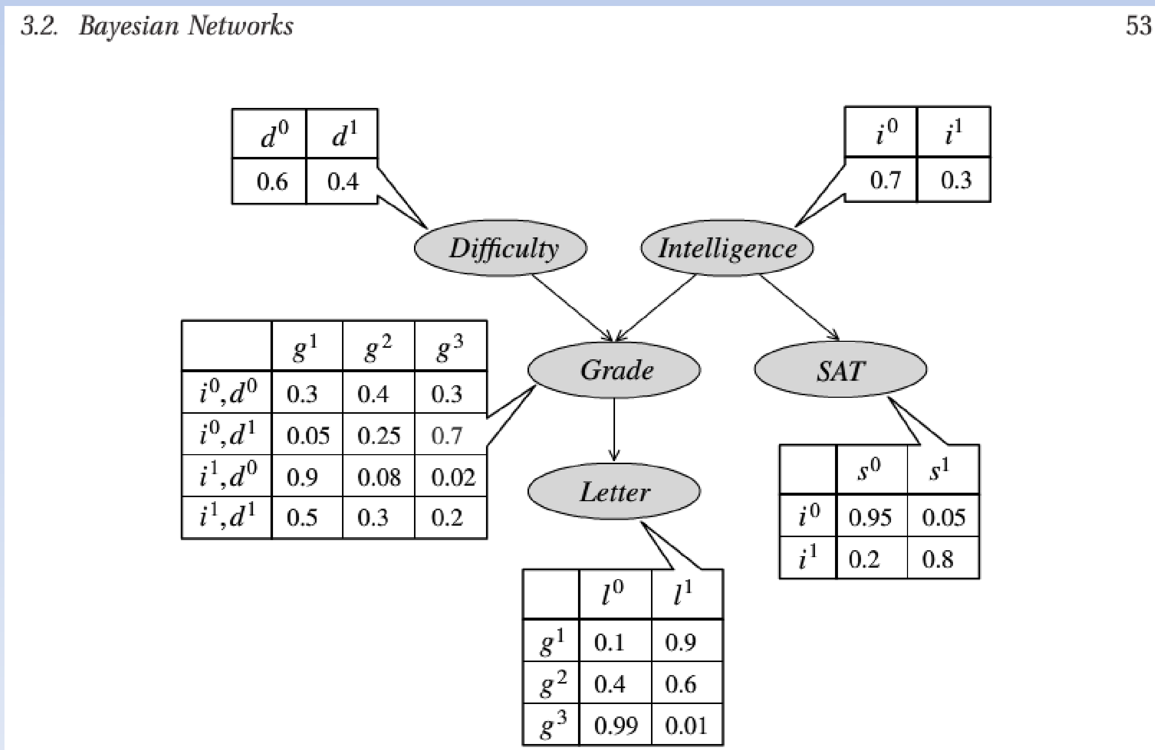


Figure 3.3 The Bayesian Network graph for the Student **example**

- Intelligence (I): Val(I)={i^0 (low), i^1 (high)}
- SAT (S): Val(S)={s^0 (low), s^1 (high)}
- Grade (G): Val(G)={g^1 (A), g^2 (B), g^3 (C)}
- ADD:
  - Course Difficulty (D): Val(D)={d^0 (easy), d^1 (hard)}
  - Letter of Recommendation (L): Val(L) = {l^0 (weak), l^1 (strong)}

Bayesian Networks : (CPD’s)

- 2nd Component of Bayesian Network are Local Probability Models that describe Parent’s influence on a Variable.



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

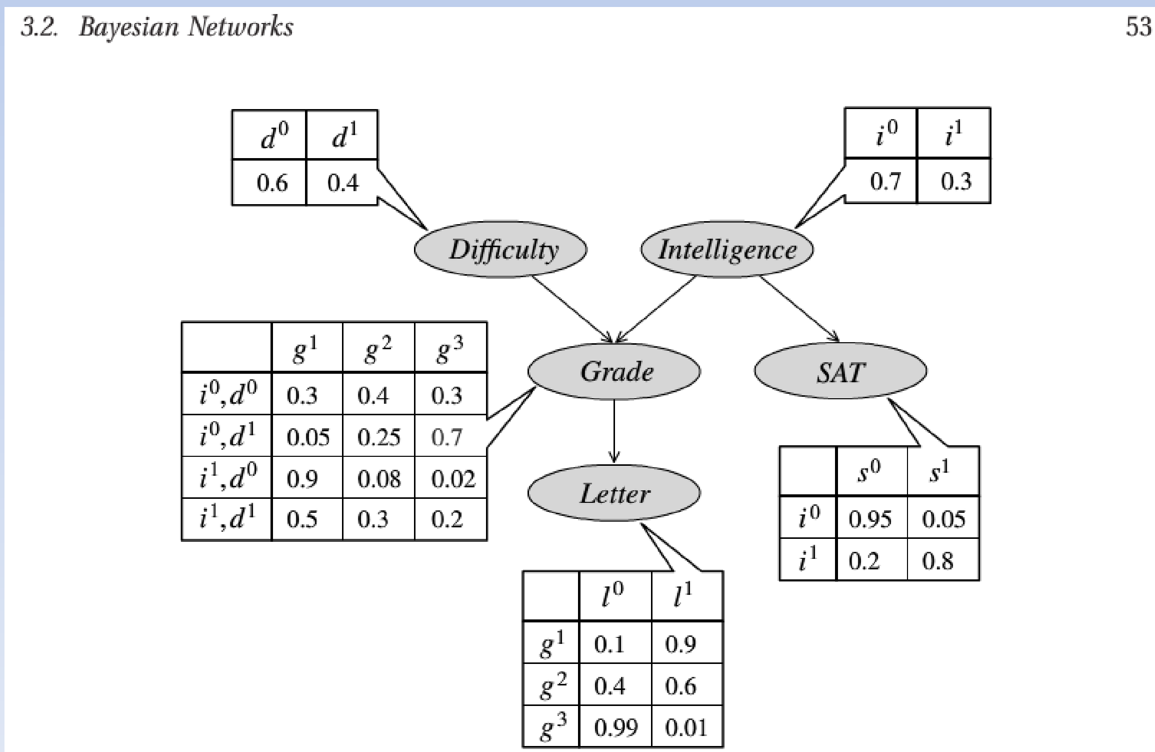
	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Bayesian Networks : (CPD’s)

- Each variable is associated with a conditional probability distribution (CPD) that specifies a distribution CPD over the values of X given each possible joint assignment of values to its parents in the model.
- For a node with no parents, the CPD is conditioned on the empty set of variables.



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Bayesian Network

- The network structure together with its CPDs is a Bayesian network $\mathcal{B}$;
- Book uses $\mathcal{B}^{\text{student}}$ to refer to the Bayesian network for the student example.

- How do we use $\mathcal{B}^{\text{student}}$ to compute parameters from the full joint distribution?

Let's Query w/ $\mathcal{B}^{\text{student}}$

- What's the probability:
  - An intelligent student
  - With High SAT Score
  - Taking an easy class
  - Get's a B
  - Resulting in a Weak Letter of Recommendation

Let's Query w/ $\mathcal{B}^{student}$

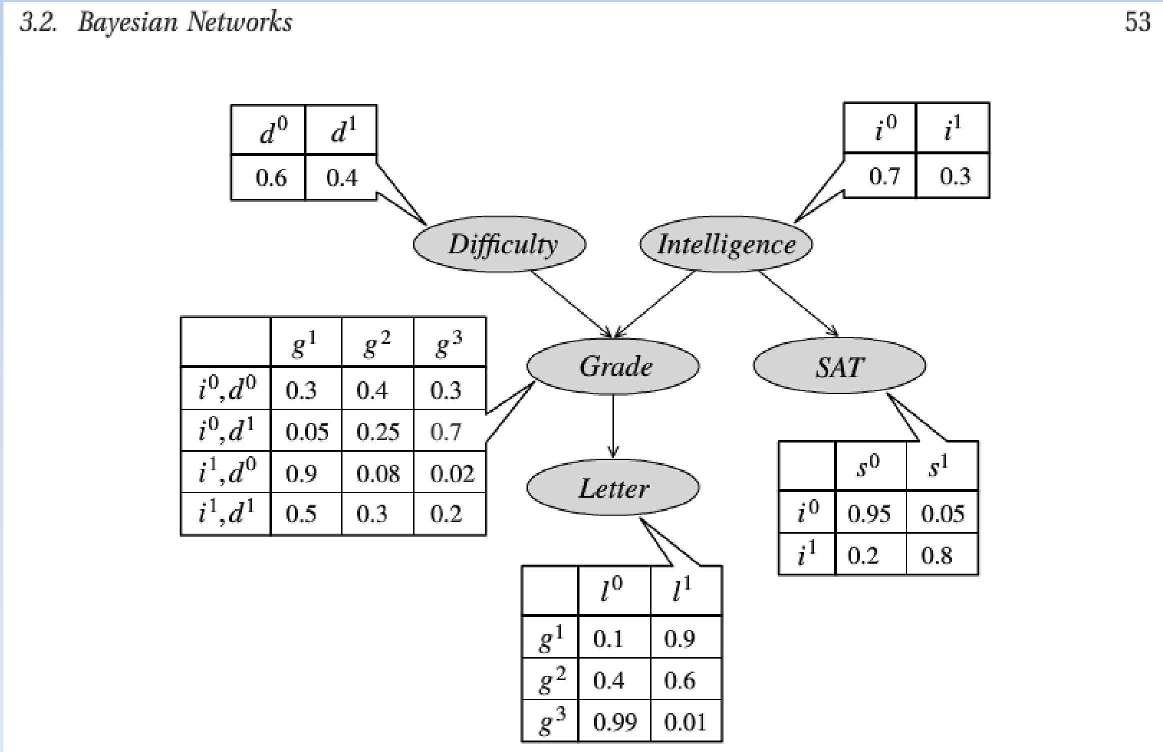
- What's the probability:
  - An intelligent student: $I=i^1$
  - With High SAT Score: $S=s^1$
  - Taking an easy class: $D=d^0$
  - Get's a B: $G=g^2$
  - w/ Weak Letter of Recommendation: $L = l^0$
- $P(i^1, d^0, g^2, s^1, l^0) = ???$

Let's Query w/ $\mathcal{B}^{\text{student}}$

- $P(i^1, d^0, g^2, s^1, l^0) =$
  - ➢ $P(i^1)P(d^0)P(g^2 | i^1, d^0)P(s^1 | i^1)P(l^0 | g^2)$

Let's Query w/ $\mathcal{B}^{\text{student}}$

- $P(i^1, d^0, g^2, s^1, l^0) =$
  - ➢ $P(i^1)P(d^0)P(g^2 \mid i^1, d^0)P(s^1 \mid i^1)P(l^0 \mid g^2)$



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

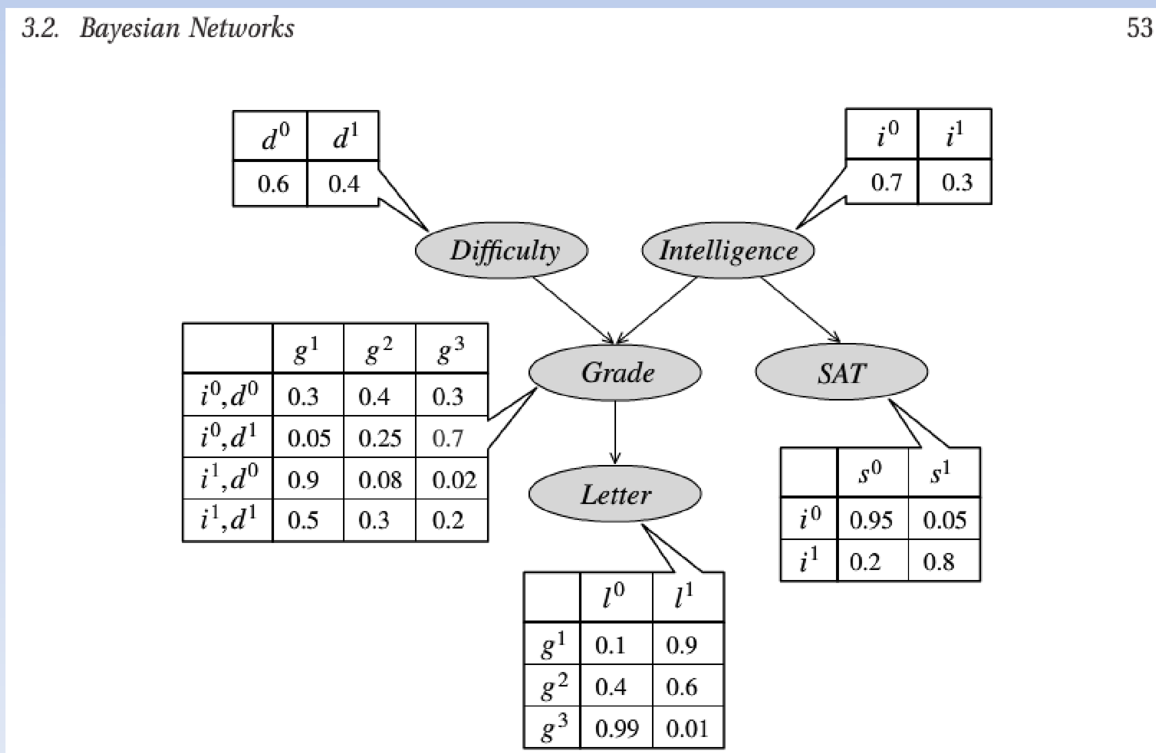
	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Let’s Query w/ $\mathcal{B}^{\text{student}}$

- $P(i^1, d^0, g^2, s^1, l^0)$ =
  - $P(i^1)P(d^0)P(g^2 | i^1, d^0)P(s^1 | i^1)P(l^0 | g^2)$
  - 0.3



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

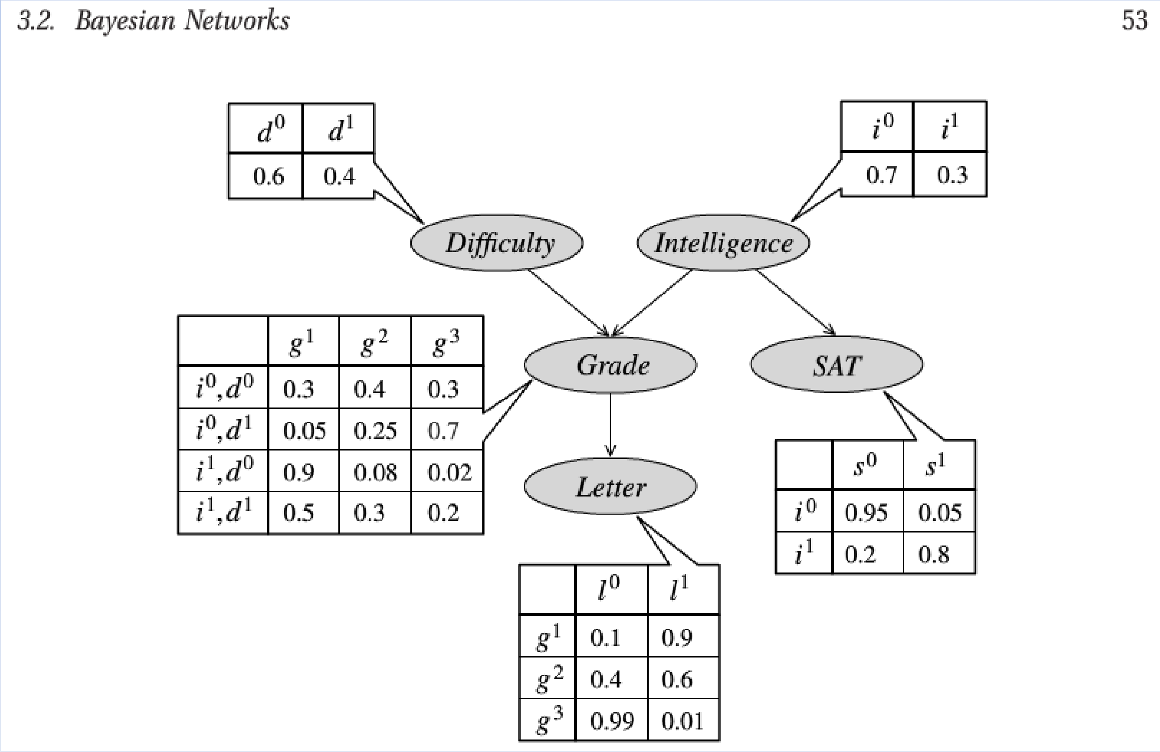
	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Let’s Query w/ $\mathcal{B}^{student}$

- $P(i^1, d^0, g^2, s^1, l^0) =$
  - $P(i^1)P(d^0)P(g^2 | i^1, d^0)P(s^1 | i^1)P(l^0 | g^2)$
  - $0.3 \cdot 0.6$



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

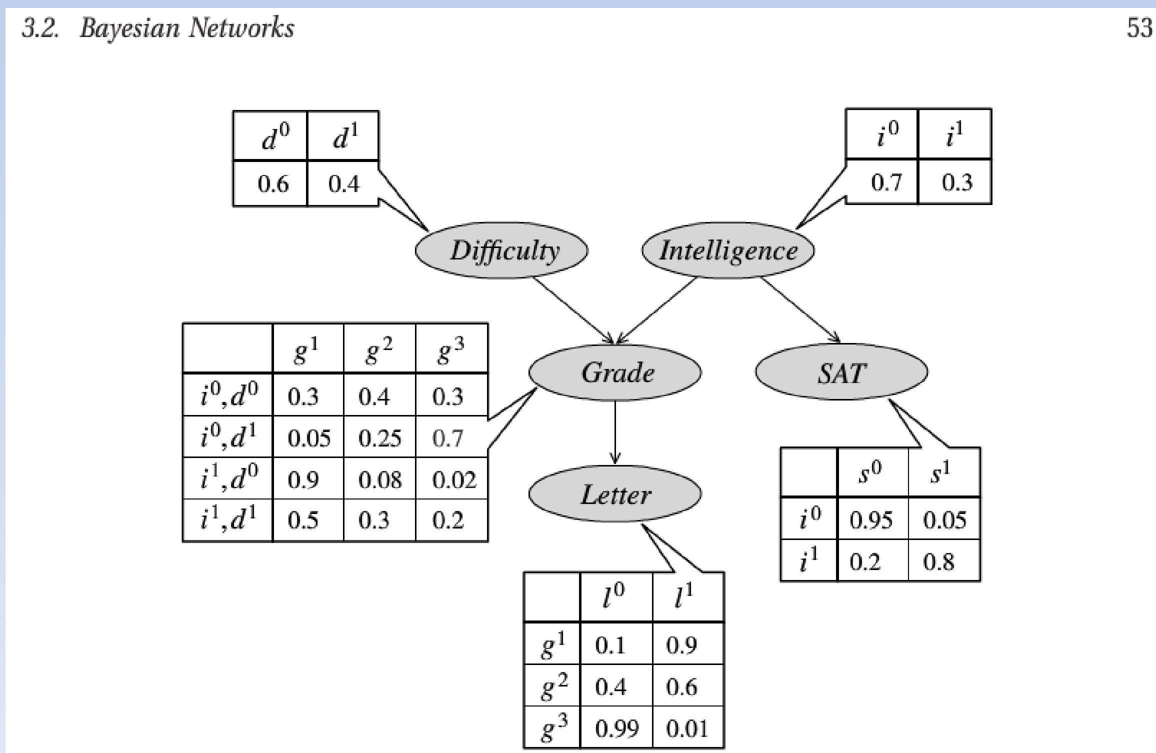
	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Let’s Query w/ $\mathcal{B}^{\text{student}}$

- $P(i^1, d^0, g^2, s^1, l^0) =$
  - ➢ $P(i^1)P(d^0)P(g^2 | i^1, d^0)P(s^1 | i^1)P(l^0 | g^2)$
  - ➢ $0.3 \cdot 0.6 \cdot 0.08$



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

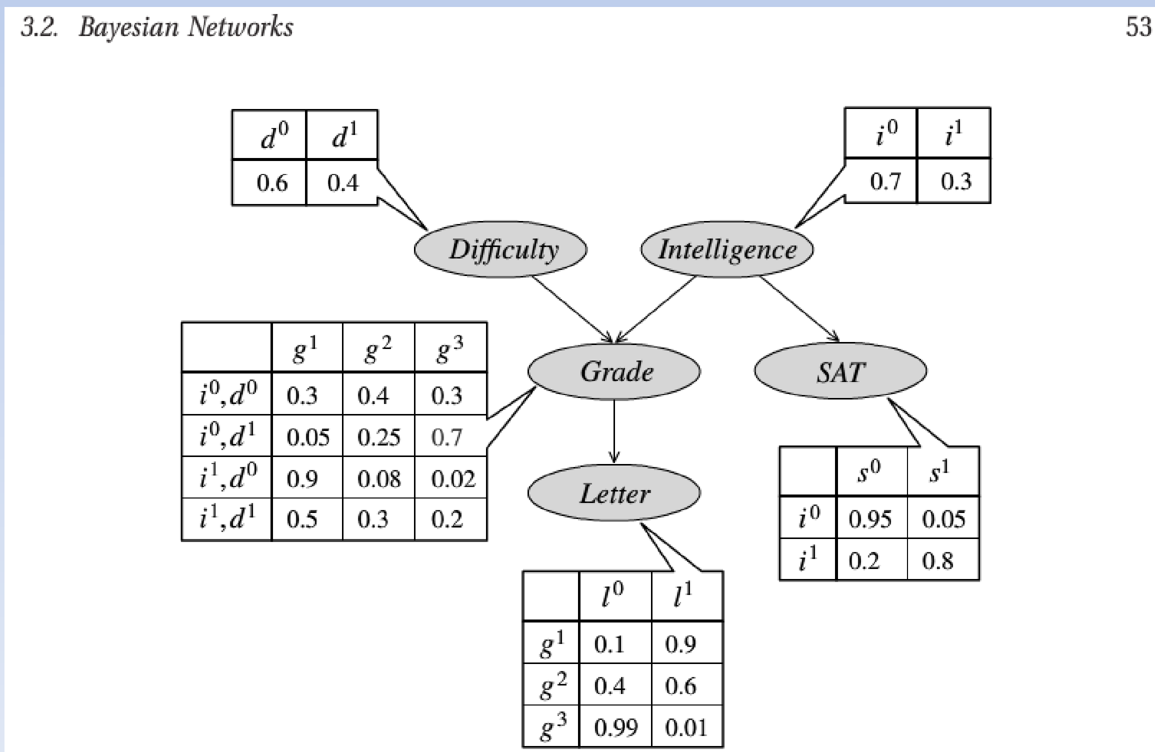
	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Let's Query w/ $\mathcal{B}^{student}$

- $P(i^1, d^0, g^2, s^1, l^0) =$
  - $P(i^1)P(d^0)P(g^2 | i^1, d^0)P(s^1 | i^1)P(l^0 | g^2)$
  - $0.3 \cdot 0.6 \cdot 0.08 \cdot 0.8$



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

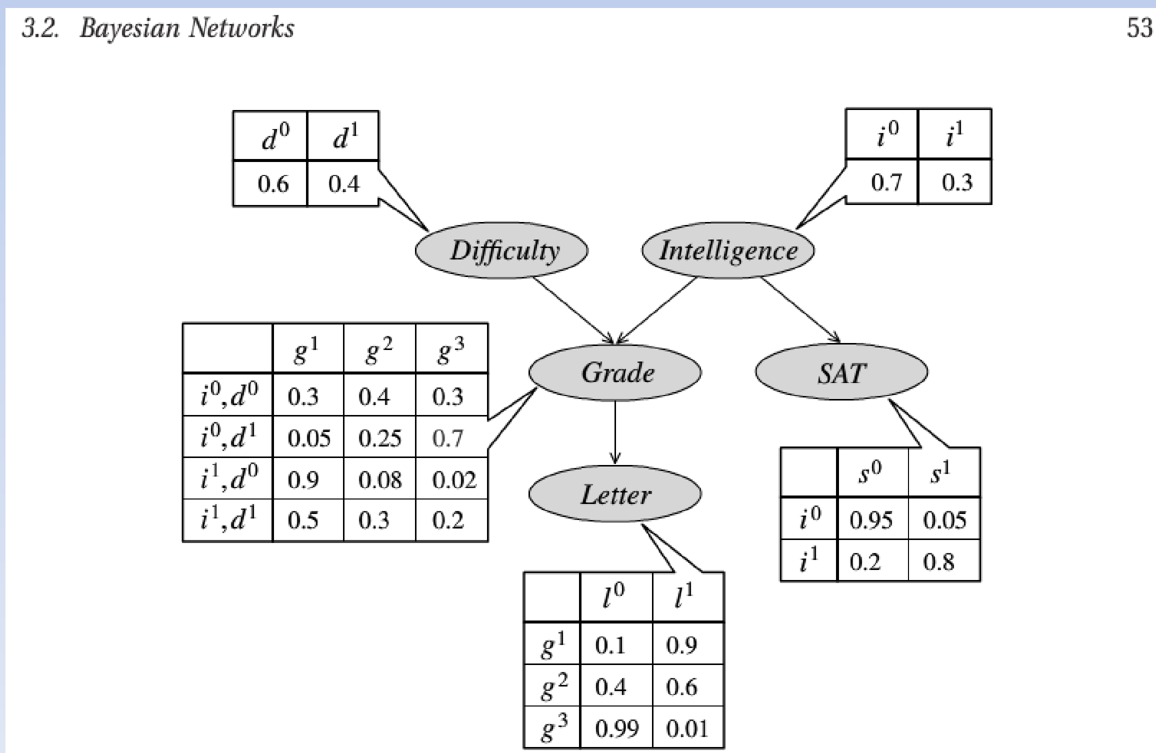
	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Let's Query w/ $\mathcal{B}^{\text{student}}$

- $P(i^1, d^0, g^2, s^1, l^0) =$
  - ➢ $P(i^1)P(d^0)P(g^2 | i^1, d^0)P(s^1 | i^1)P(l^0 | g^2)$
  - ➢ $0.3 \cdot 0.6 \cdot 0.08 \cdot 0.8 \cdot 0.4 =$



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

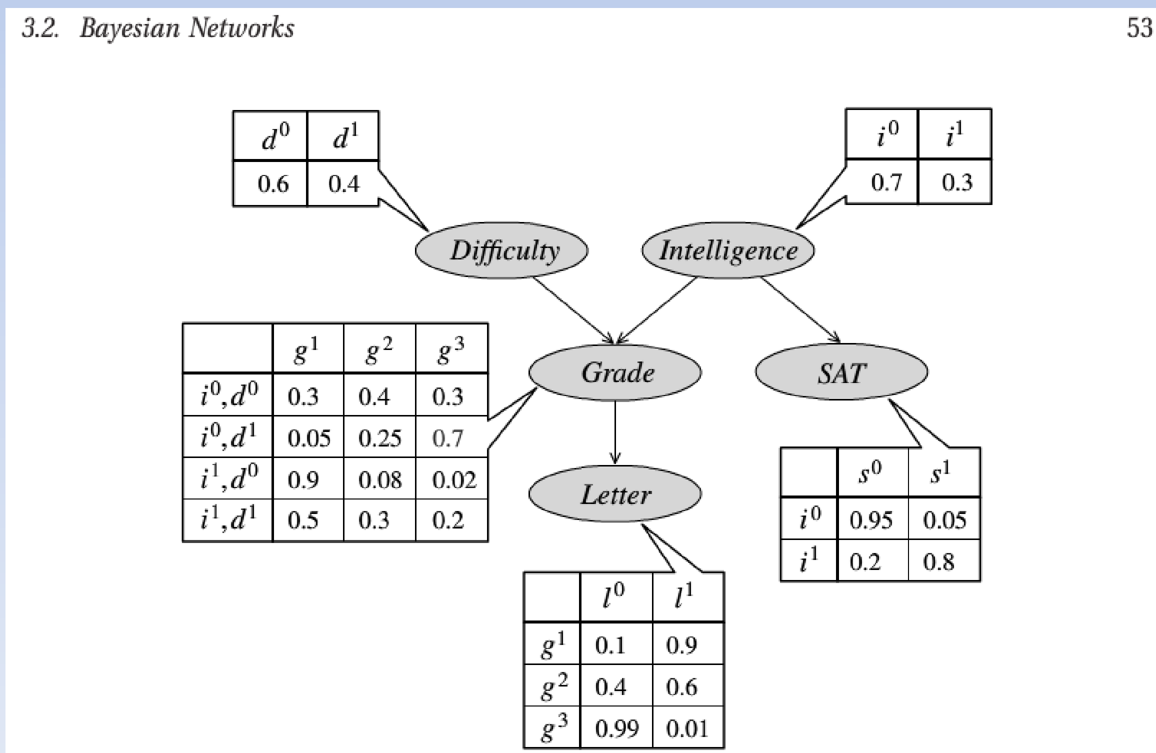
	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Let's Query w/ $\mathcal{B}^{\text{student}}$

- $P(i^1, d^0, g^2, s^1, l^0) =$
  - ➢ $P(i^1)P(d^0)P(g^2 | i^1, d^0)P(s^1 | i^1)P(l^0 | g^2)$
  - ➢ $0.3 \cdot 0.6 \cdot 0.08 \cdot 0.8 \cdot 0.4 = 0.004608$



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

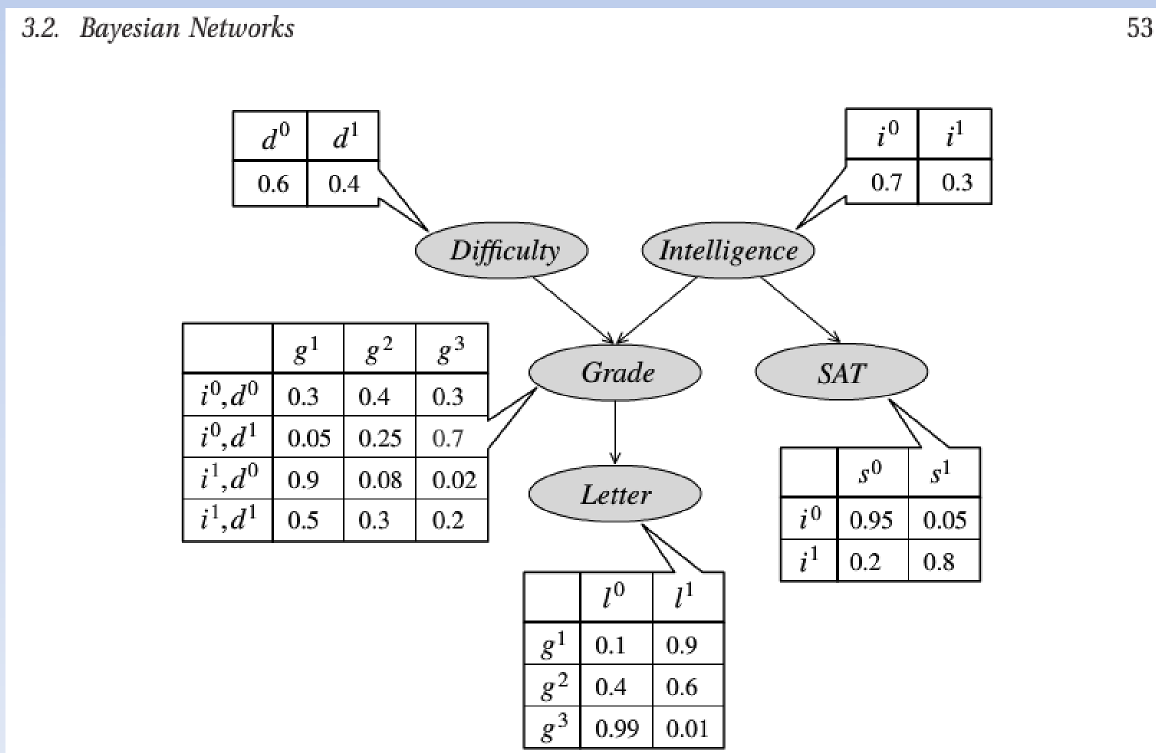
	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Let's Query w/ $\mathcal{B}^{\text{student}}$

- $P(i^1, d^0, g^2, s^1, l^0) =$
  - ➢ $P(i^1)P(d^0)P(g^2 | i^1, d^0)P(s^1 | i^1)P(l^0 | g^2)$
  - ➢ $0.3 \cdot 0.6 \cdot 0.08 \cdot 0.8 \cdot 0.4 = 0.004608$????



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

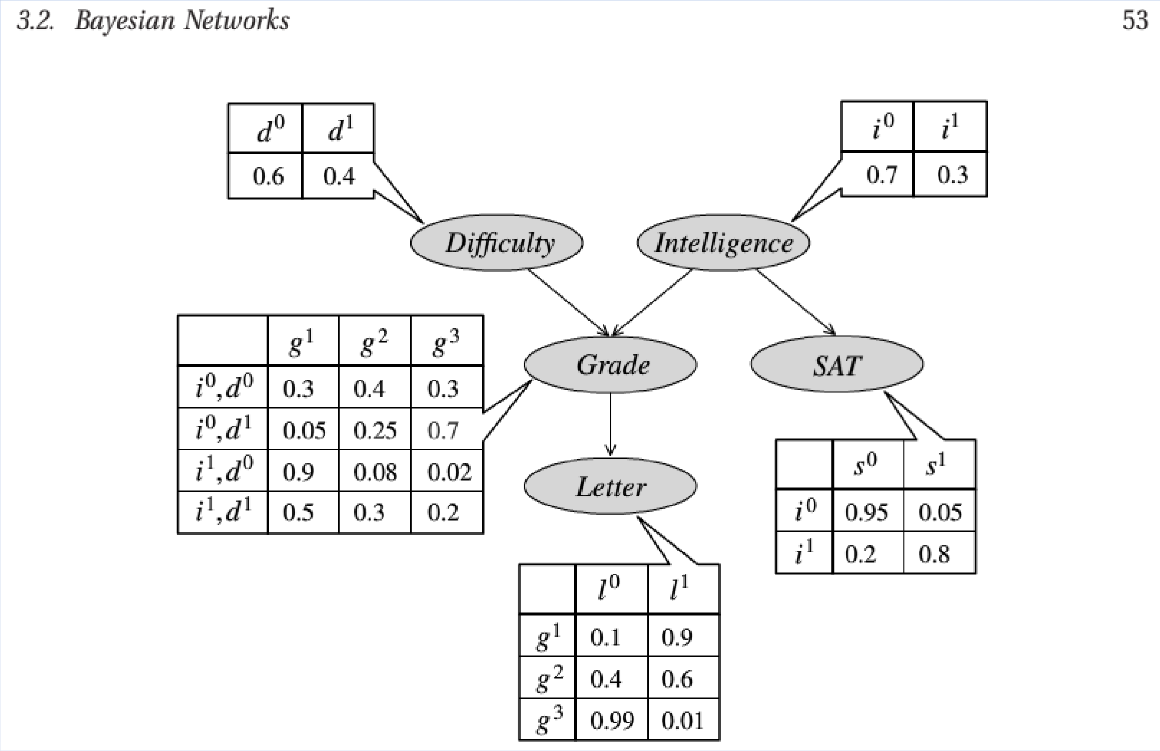
	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

First Example w/ Chain Rule for Bayesian Networks

- P(I, D, G, S, L)=
  - P(I)P(D)P(G|I,D)P(S|I)P(L|G)

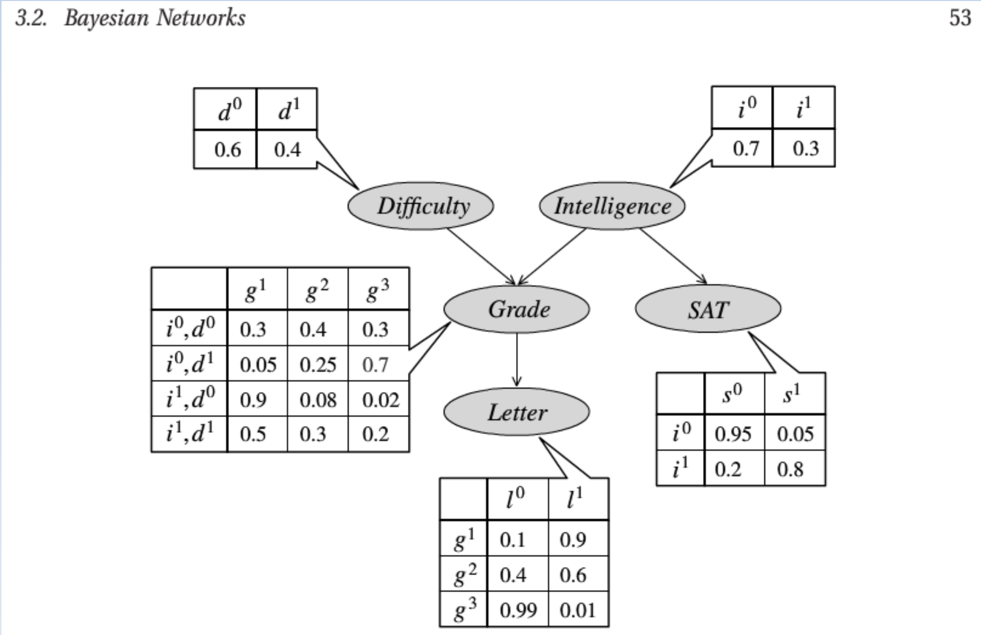


Are we SURE
Result is Probability Distribution?

- All values must be greater than 1
- All values must sum to 1

Are we SURE
Result is Probability Distribution?

- All values must be greater than 1
  - Table values taken from a CPD, so each greater than or equal to 1.
  - These values when multiplied only yield a value greater than 1



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

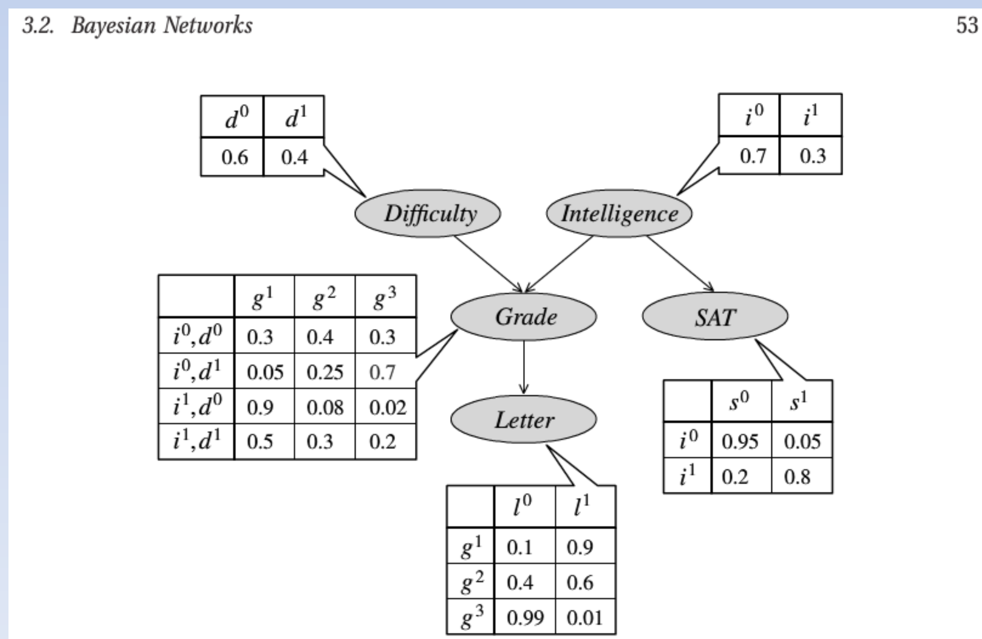
	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Are we SURE
Result is Probability Distribution?

- All values must sum to 1
  - ➢ $\sum_I \sum_D \sum_G \sum_S$ P(I)P(D)P(G|I,D)P(S|I) $\sum_L$ P(L|G)

- NOTE:
  - – $\sum_L$ P(L|G) $= 1$
  - – $\sum_S$ P(S|I) $= 1$
  - – $\sum_G$ P(G|I,D) $= 1$



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Probability Queries & Bayesian Networks

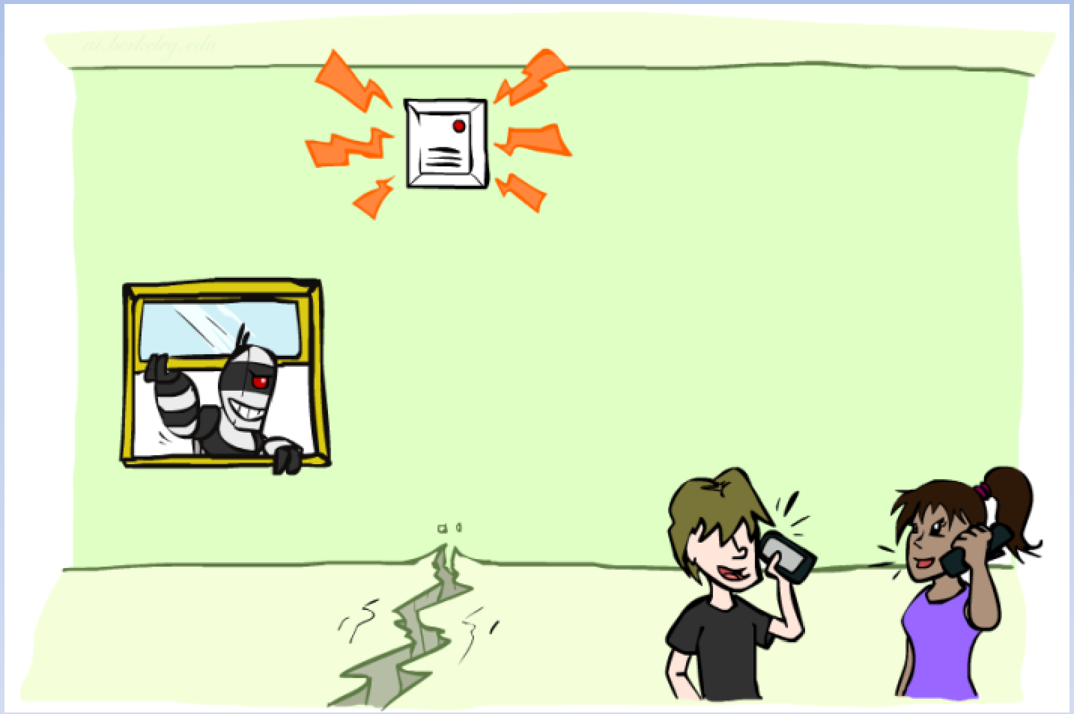
- Conditional Probability Queries
  - Evidence: E=e
  - Query: a subset of variables Y
  - Task: computer P(Y | E=e)
  - P(NoGas|Gauge=empty,Lights=on,Starts=false)
- Conjunctive queries: $P(X_i,X_J|E=e)=P(X_i|E=e)P(X_J|X_i,E=e)$
- Optimal decisions: decision networks include utility information; probabilistic inference required for P(outcome|action,evidence)
- Value of information: which evidence to seek next?
- Sensitivity analysis: which probability values are most critical?
- Explanation: why do I need a new starter motor?

Probability Queries & Bayesian Networks

- The following are all NP-Hard
  - Given a PGM P_Φ, a variable X, and a value $x \in val(X)$
    - Compute $P_\Phi(X=x)$
      - Or even $P_\Phi(X=x) > 0$
  - Let $\varepsilon < 0.5$. Given a PGM P_Φ, a variabe X, and a value $x \in val(X)$ and an observation $e \in val(E)$
    - Find a number p that has $|P_\Phi(X=x|E=e) - p| < \varepsilon$

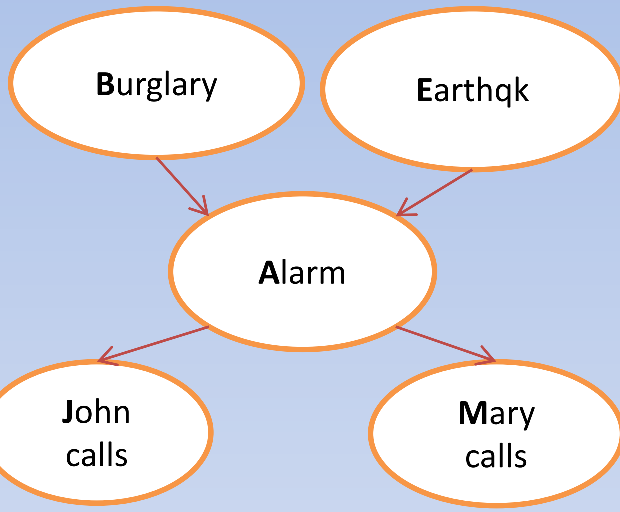
Example: Alarm Network

- Variables
  - B: Burglary
  - A: Alarm goes off
  - M: Mary calls
  - J: John calls
  - E: Earthquake!

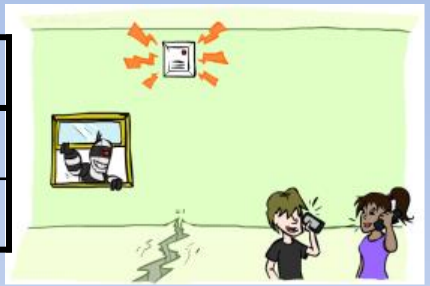


Example: Alarm Network

B	P(B)
+b	0.001
-b	0.999



E	P(E)
+e	0.002
-e	0.998



B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

Exact Inference: Enumeration

512 Chapter 14. Probabilistic Reasoning

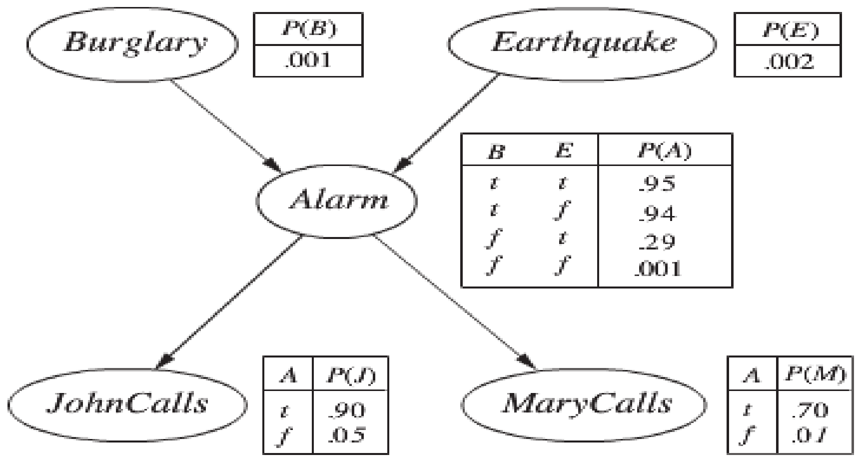


Figure 14.2 A typical Bayesian network, showing both the topology and the conditional probability tables (CPTs). In the CPTs, the letters B, E, A, J, and M stand for *Burglary*, *Earthquake*, *Alarm*, *JohnCalls*, and *MaryCalls*, respectively.

Exact Inference: Enumeration

- Simple query on the burglary network:
  - What is the probability of a burglary if both John and Mary Call?
- P(B|j,m)
  - = P(B,j,m)/P(j,m)
  - = αP(B,j,m)
  - $\alpha \sum_{e} \sum_{a} \mathrm{P}(\mathrm{B}, \mathrm{e}, \mathrm{a}, \mathrm{j}, m)$

Exact Inference: Enumeration.

- P(B|j,m) =

  $=\alpha\sum_e \sum_a$ P(B, e, a, j, m)

  $=\alpha\sum_e \sum_a$ P(B)P(e)P(a|B,e)P(j|a)P(m|a)

  $=\alpha$P(B) $\sum_e$ P(e) $\sum_a$ P(a|B,e)P(j|a)P(m|a)

Exact Inference: Enumeration.

- P(B|j,m) =

  $=\alpha P(B)\sum_e P(e)\sum_a P(a|B,e)P(j|a)P(m|a)$

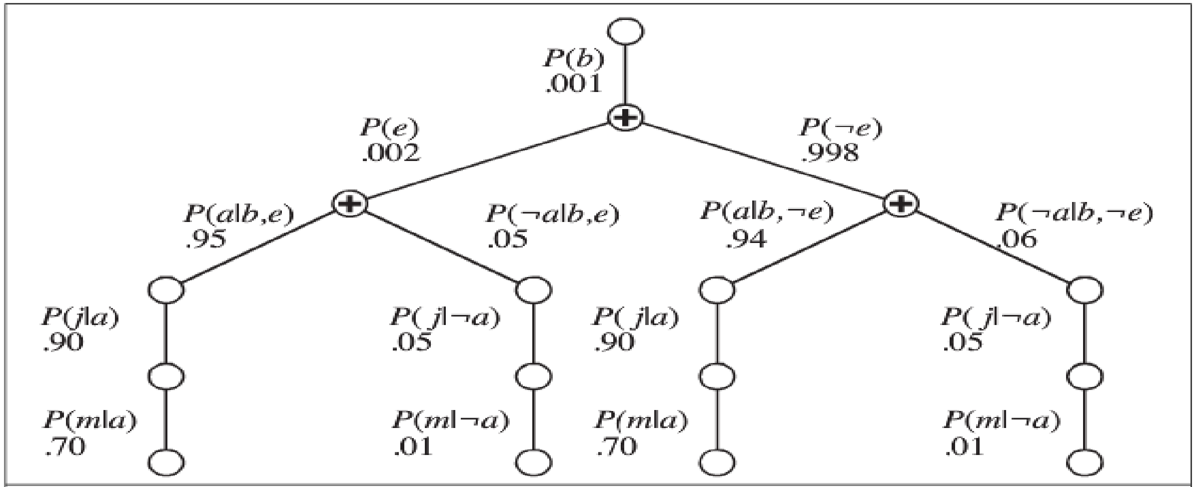


Figure 14.8 The structure of the expression shown in Equation (14.4). The evaluation proceeds top down, multiplying values along each path and summing at the "+" nodes. Notice the repetition of the paths for j and m.

Exact Inference: Enumeration w/ Recursion

```
function ENUMERATION-ASK(X, e, bn) returns a distribution over X
  inputs: X, the query variable
          e, observed values for variables E
          bn, a Bayes net with variables {X} ∪ E ∪ Y   /* Y = hidden variables */

  Q(X) ← a distribution over X, initially empty
  for each value x_i of X do
      Q(x_i) ← ENUMERATE-ALL(bn.VARS, e_{x_i})
          where e_{x_i} is e extended with X = x_i
  return NORMALIZE(Q(X))
```

```
function ENUMERATE-ALL(vars, e) returns a real number
  if EMPTY?(vars) then return 1.0
  Y ← FIRST(vars)
  if Y has value y in e
      then return P(y | parents(Y)) × ENUMERATE-ALL(REST(vars), e)
      else return Σ_y P(y | parents(Y)) × ENUMERATE-ALL(REST(vars), e_y)
          where e_y is e extended with Y = y
```

Figure 14.9 The enumeration algorithm for answering queries on Bayesian networks.

Exact Inference: Enumeration.

- P(B|j,m) =

  $=\alpha P(B) \sum_e P(e) \sum_a P(a|B,e)P(j|a)P(m|a)$

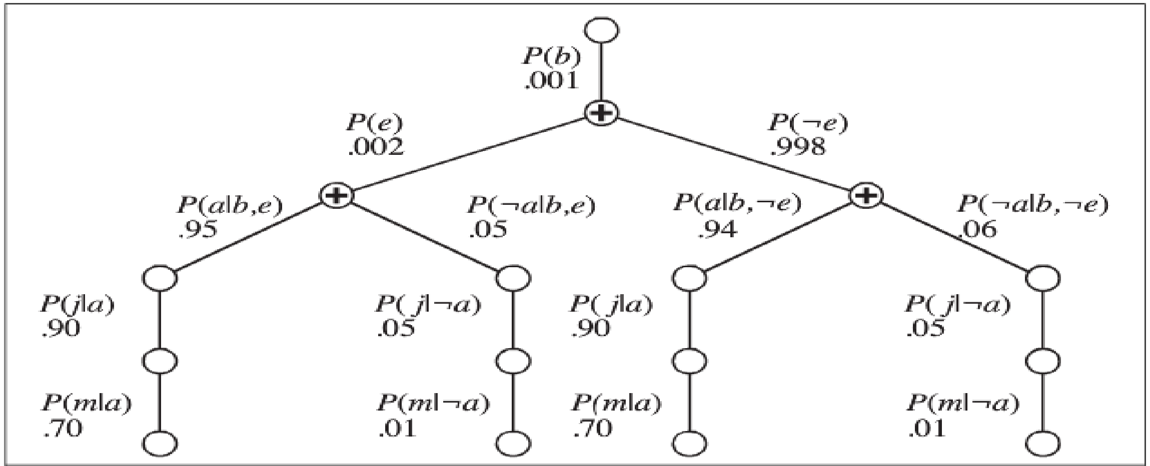


Figure 14.8 The structure of the expression shown in Equation (14.4). The evaluation proceeds top down, multiplying values along each path and summing at the "+" nodes. Notice the repetition of the paths for j and m.

Exact Inference: Enumeration

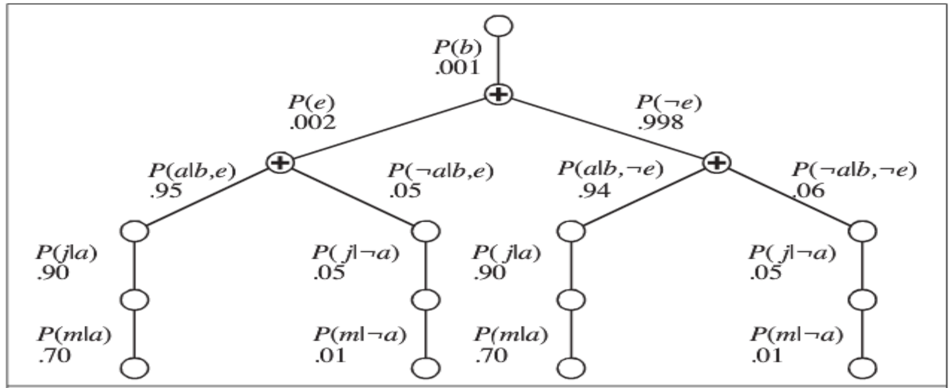


Figure 14.8 The structure of the expression shown in Equation (14.4). The evaluation proceeds top down, multiplying values along each path and summing at the "+" nodes. Notice the repetition of the paths for j and m.

- Recursive depth-first enumeration:
  - O(n) space,
  - O(dn) time

- Lots of repeated calculations
- Maybe Dynamic Programming!