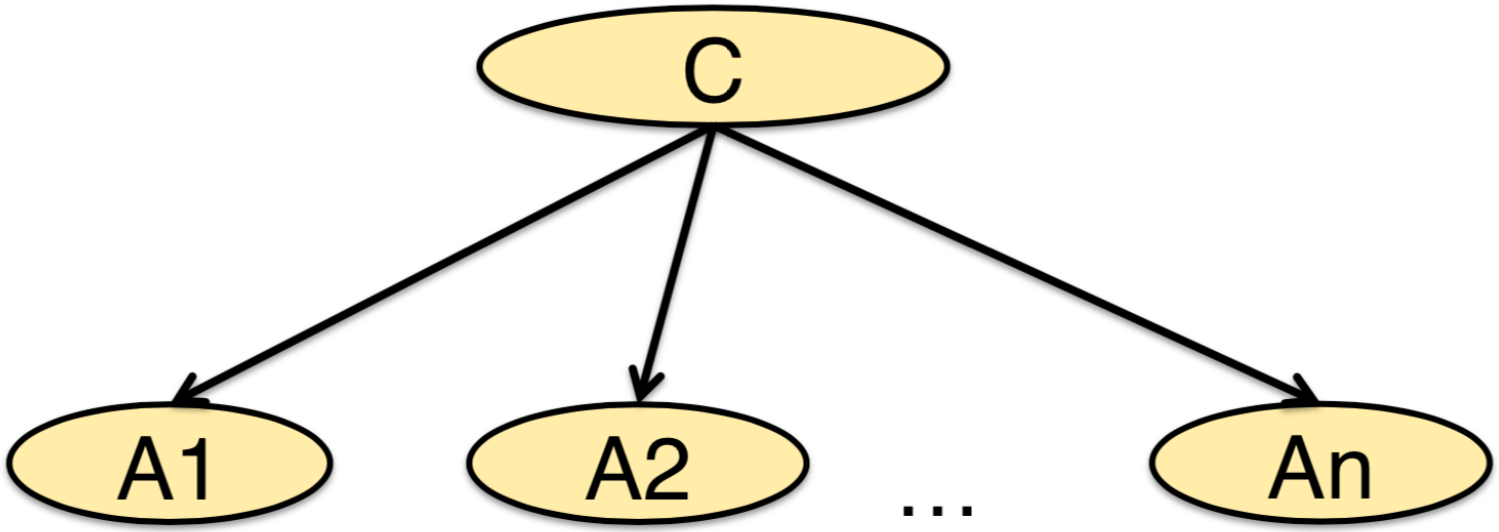
ISE 540
Text Analytics

Mayank Kejriwal
Research Assistant Professor/Research Lead
Department of Industrial and Systems Engineering
Information Sciences Institute
USC Viterbi School of Engineering
kejriwal@isi.edu

Naïve Bayes (NB)



Each item has a number of attributes

$A_1=a_1,\ldots,A_n=a_n$

We predict the class c based on

$c = \text{argmax}_c \prod_i P(A_i = a_i \mid C=c)\ P(C=c)$

Does the customer want sugar?

x1	x2	Y
A1: drink	A2: milk?	C: sugar?
coffee	no	yes
coffee	yes	no
tea	yes	yes
tea	no	no

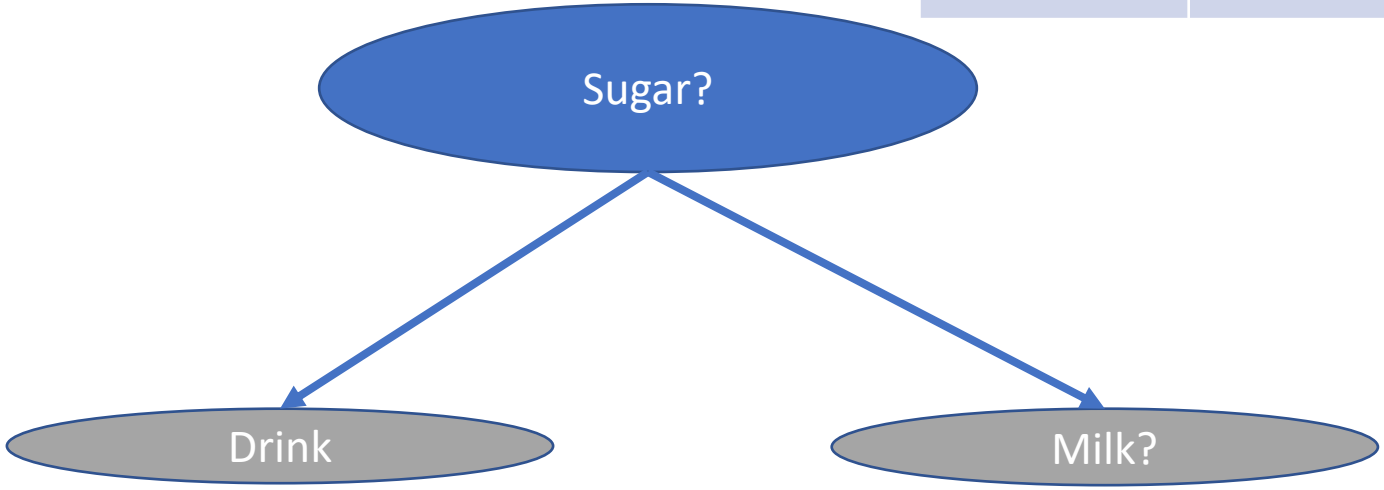
Can you train a Naïve Bayes classifier to predict whether the customer wants sugar or not?

What is P(coffee | sugar)?

Getting the NB parameters

- Use maximum likelihood!

Yes	No



Sugar?	Coffee	Tea
Yes		
No		

Sugar?	Yes	No
Yes		
No		

Does the customer want sugar?

x1	x2	Y
A1: drink	A2: milk?	C: sugar?
coffee	no	yes
coffee	yes	no
tea	yes	yes
tea	no	no

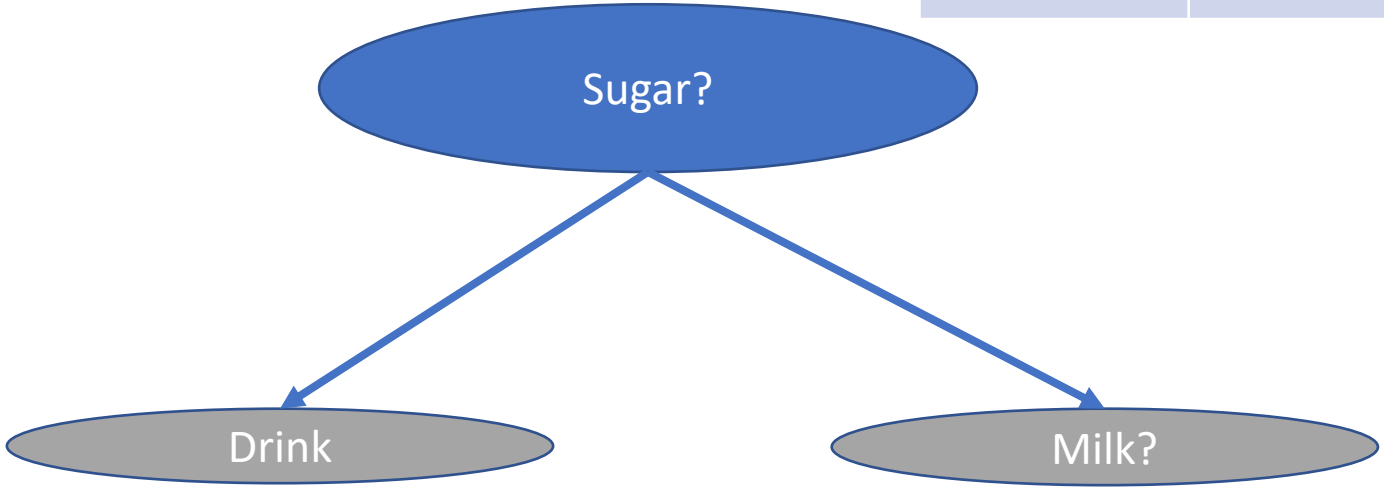
Can you train a Naïve Bayes classifier to predict whether the customer wants sugar or not?

What is P(coffee | sugar)?

Getting the NB parameters

- Use maximum likelihood!

Yes	No
0.5	0.5



Sugar?	Coffee	Tea
Yes	0.5	0.5
No	0.5	0.5

Sugar?	Yes	No
Yes	0.5	0.5
No	0.5	0.5

Another example

x1	x2	Y
A1: drink	A2: milk?	C: sugar?
coffee	no	yes
coffee	~~yes~~ no	no
tea	yes	yes
~~tea~~ coffee	no	no

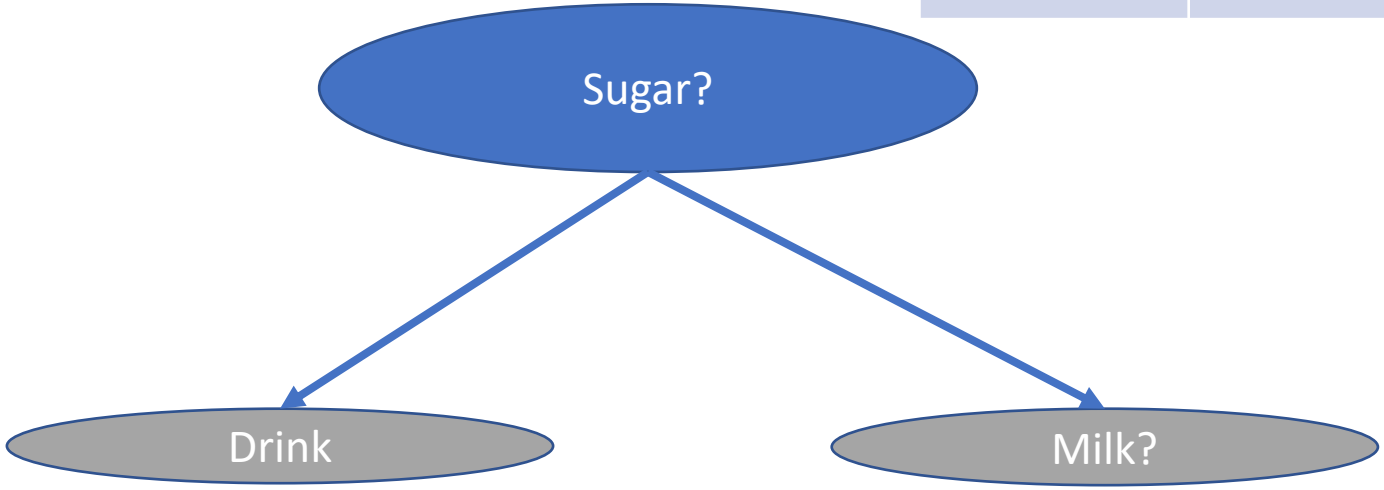
Can you train a Naïve Bayes classifier to predict whether the customer wants sugar or not?

What is P(coffee | sugar)?

Why is this a problem?

Conditional Independence: P(x_1, x_2 | y) = P(x_1 | y) P(x_2|y)
For example: P(drink=coffee,milk=yes|sugar=yes) = P(drink=coffee|sugar=yes) P(milk=yes|sugar=yes)

Yes	No
0.5	0.5



Sugar?	Coffee	Tea
Yes	0.5	0.5
No	1.0	0.0

Sugar?	Yes	No
Yes	0.5	0.5
No	0.0	1.0

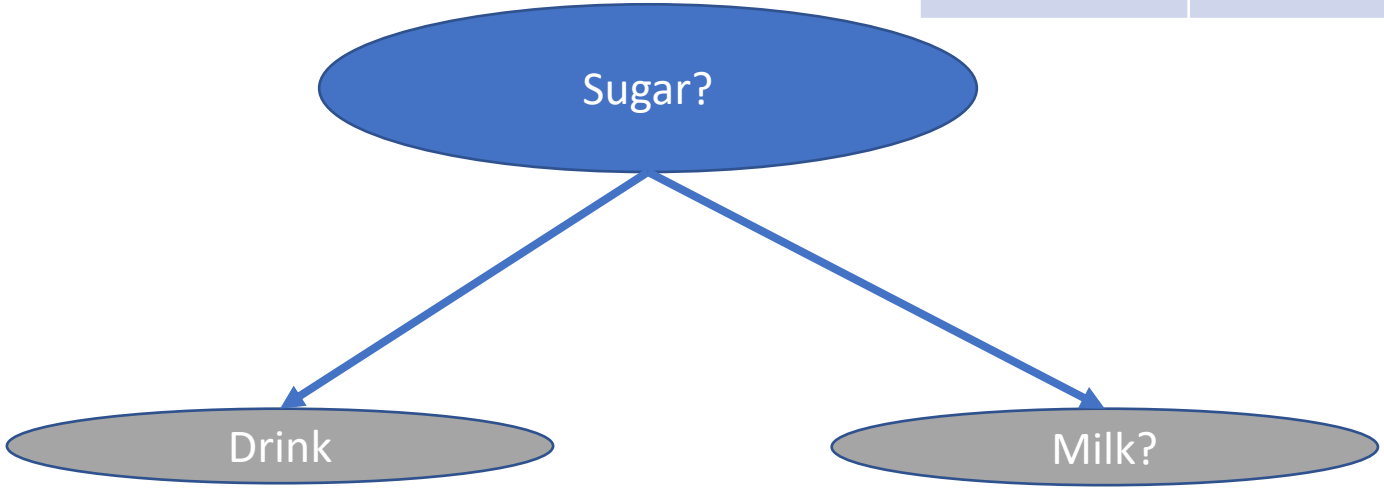
Prediction problem

- Given milk and drink, can you predict whether I want sugar or not?
- We want to pick the value for sugar (yes or no) that maximizes the probability P(sugar | milk, drink)
- From standard probability theory:
  - P(sugar | milk, drink) = P(milk, drink | sugar)P(sugar)
- Using conditional independence assumption:
  - P(milk, drink | sugar) = P(milk| sugar) P(drink | sugar)
- Putting the two together,
  - P(sugar | milk, drink) = P(milk| sugar) P(drink | sugar) P(sugar)
- But these are just the ‘cells’ in the NB table!
- What we want to do is to compute the ‘argmax’ for this expression over the ‘sugar’ variable

Another example

--What does NB predict given that I'm having coffee with milk? What is the probability of the prediction?
--What about tea without milk?

Yes	No
0.8	0.2



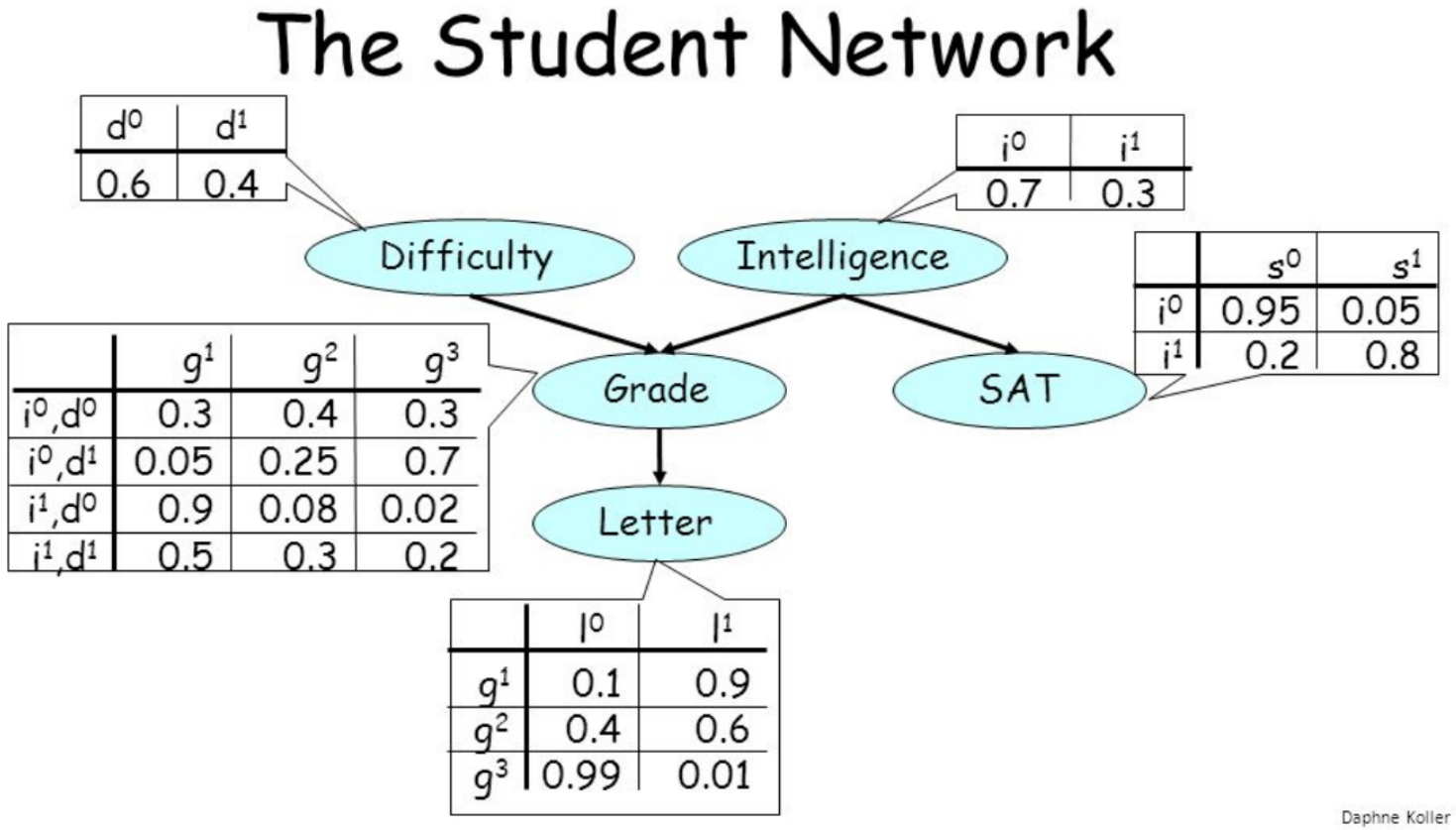
Sugar?	Coffee	Tea
Yes	0.6	0.4
No	0.9	0.1

Sugar?	Yes	No
Yes	0.3	0.7
No	0.6	0.4

What can go wrong ?

Think conditional independence assumption…

Can we make this model more ‘general’?



Difficulty (of the class): Takes values 0 (low difficulty) and 1 (high difficulty)

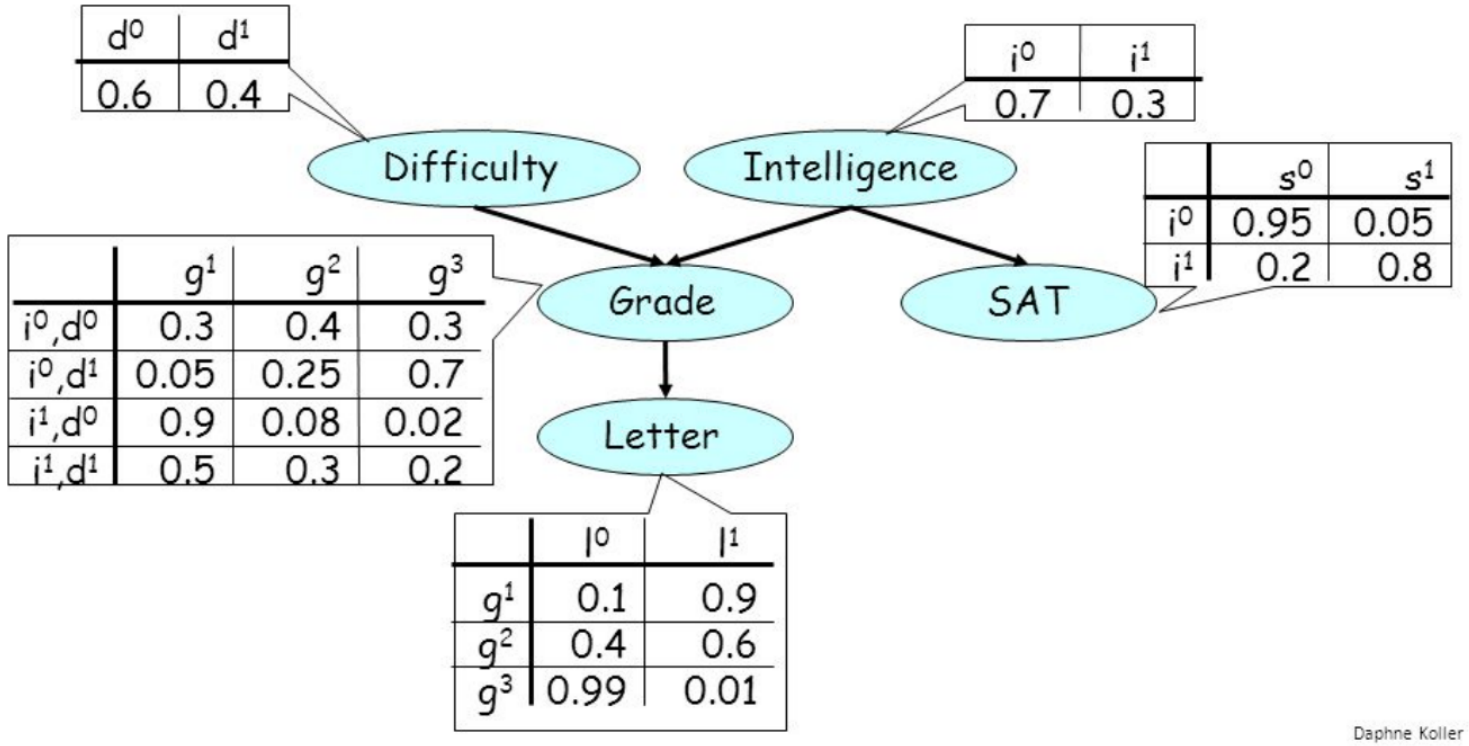
Intelligence (of the student): Takes values 0 (not intelligent) and 1 (intelligent)

Grade (the student gets in the class): Takes values 1 (good grade), 2 (average grade), and 3 (bad grade)

SAT (student's score in the SAT exam): Takes values 0 (low score) and 1 (high score)

Letter (quality of recommendation letter the student gets from the professor after completing the course): Takes values 0 (not a good letter) and 1 (a good letter)

The Student Network



Daphne Koller

What is the probability that a student gets a good letter (l1) given that the student's grade is g2?

Difficulty (of the course): Takes values 0 (not difficult) and 1 (difficult)

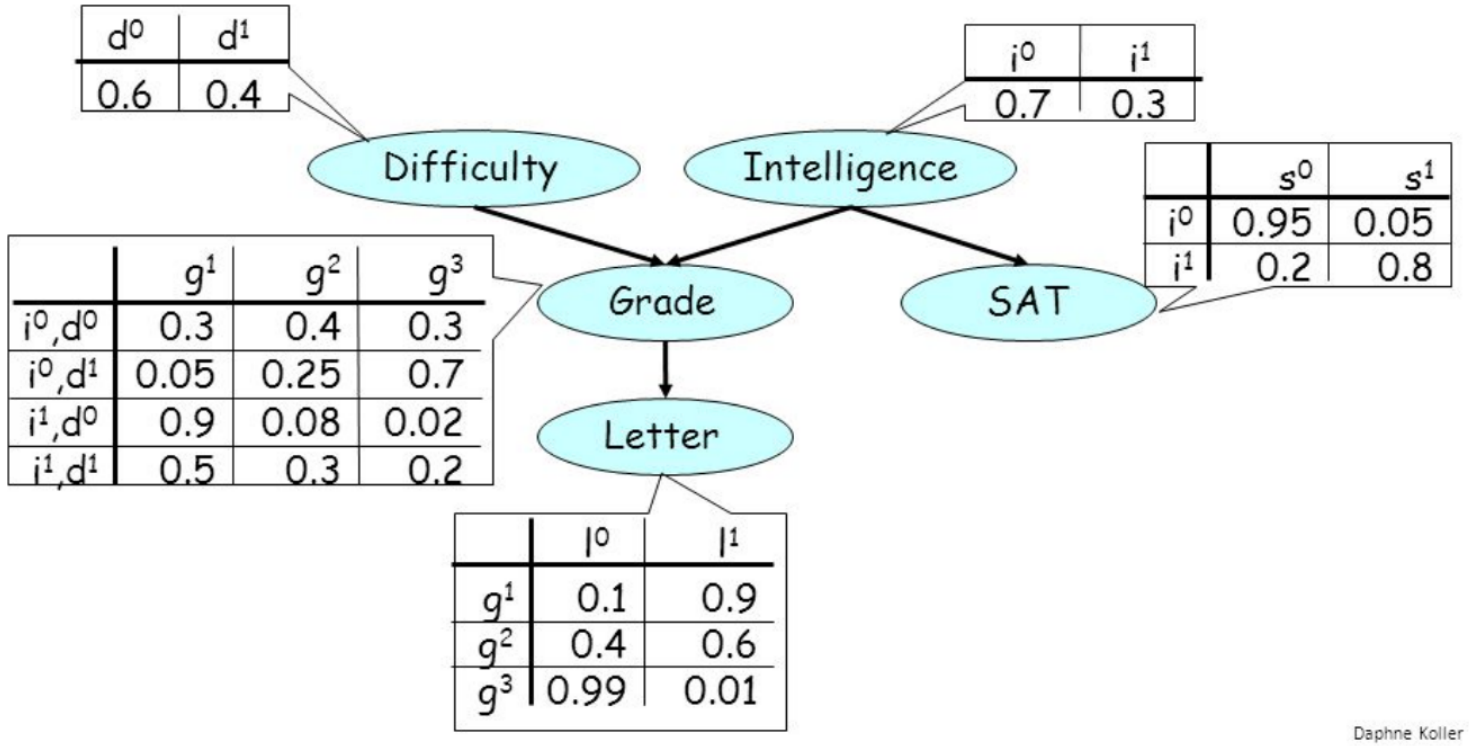
Intelligence (of the student): Takes values 0 (not intelligent) and 1 (intelligent)

Grade (the student gets in the class): Takes values 1 (good grade), 2 (average grade), and 3 (bad grade)

SAT (student's score in the SAT exam): Takes values 0 (low score) and 1 (high score)

Letter (quality of recommendation letter the student gets from the professor after completing the course): Takes values 0 (not a good letter) and 1 (a good letter)

The Student Network



Daphne Koller

Does it matter whether the student is 'intelligent' or not?
NO iff P(I1|i1) = P(I1|i0), YES iff P(I1|i1) != P(I1|i0)

Difficulty (of the course): Takes values 0 (easy) and 1 (difficult)

Intelligence (of the student): Takes values 0 (not intelligent) and 1 (intelligent)

Grade (the student gets in the class): Takes values 1 (good grade), 2 (average grade), and 3 (bad grade)

SAT (student's score in the SAT exam): Takes values 0 (low score) and 1 (high score)

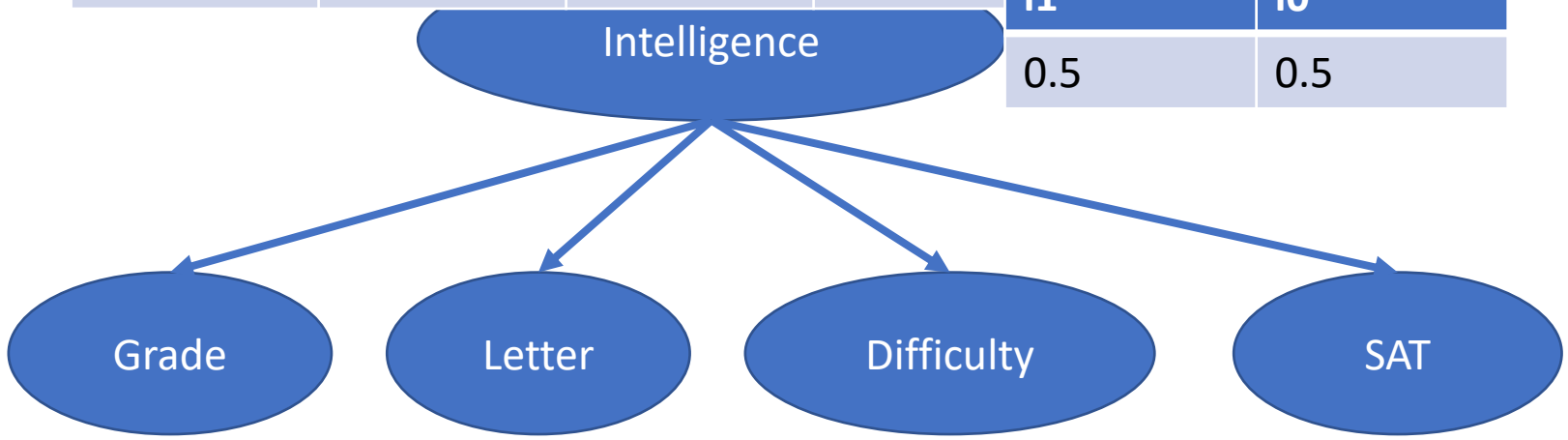
Letter (quality of recommendation letter the student gets from the professor after completing the course): Takes values 0 (not a good letter) and 1 (a good letter)

How do we represent the student's network as a naïve Bayes?

- First question you must ask is what is the 'class' variable (usually the variable you will be trying to 'predict')…
  - Suppose it is Intelligence

I	G	L	D	SAT
i0	g1	l1	d0	s1
i1	g1	l1	d0	s1
i0	g1	l1	d0	

i1	i0
0.5	0.5



Intelligence	g1
i1	1.0
i0	1.0

What if the target class was ‘Letter’? What would be the NB model? Can the two models give ‘different’ results?

If I gave you a ‘table of observations’ just like the coffee/sugar example, would you be able to infer the CPTs?