Lecture 4 - Probability Foundation for Machine Learning

Prof. Xiaowei Huang

https://cgi.csc.liv.ac.uk/~xiaowei/

(Attendance Code: **566555**)

In the last week's lectures,

Module contents

A few applications of machine learning

Representation of instances as vectors

Data preprocessing

Learning basics

Supervised, unsupervised, semi-supervised learning

Topics of today -- Probability foundations

Random Variables

Joint and Conditional Distributions

Machine Learning is about the measurement of, and the reasoning about, probabilistic distributions.

Independence and Conditional Independence

Querying Joint Probability Distributions

Probability query

MAP query

Random Variables

Random Variable

Use RVs to express uncertainty etc

We have a population of students

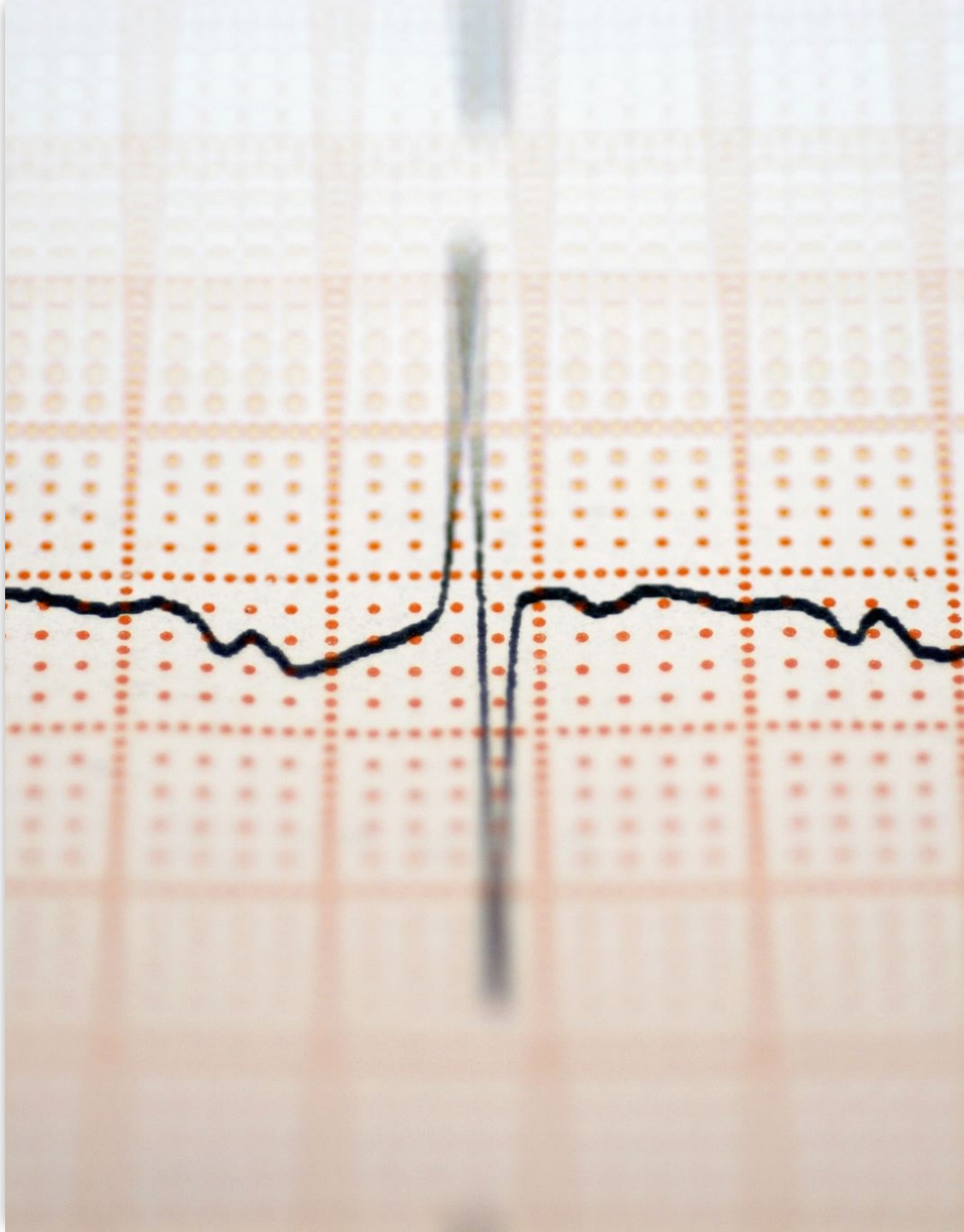
- We want to reason about their grades
- Random variable: *Grade*
- *P(Grade)* associates a probability with each outcome *Val(Grade)*={ *A, B, C* }

If $k=|Val\{X\}|$ then $\sum_{i=1}^{k} P(X = x^i) = 1$

- Distribution is referred to as a *multinomial*
- If *Val{X}={false,true}* then it is a *Bernoulli* distribution

$P(X)$ is known as the marginal distribution of X

Joint and Conditional Distributions



Recap: Marginal, joint, conditional probability

- **Marginal probability**: the probability of an event occurring P(A), it may be thought of as an unconditional probability. It is not conditioned on another event.
  - Example: the probability that a card drawn is red, i.e., P(red) = 0.5.
  - Another example: the probability that a card drawn is a 4, i.e., P(four)=1/13.

- **Joint probability**: P(A and B). The probability of event A **and** event B occurring. It is the probability of the intersection of two or more events. The probability of the intersection of A and B may be written P(A ∩ B).
  - Example: the probability that a card is a four and red, i.e., P(four and red) = 2/52=1/26. (There are two red fours in a deck of 52, the 4 of hearts and the 4 of diamonds).

- **Conditional probability**: P(A|B) is the probability of event A occurring, given that event B occurs.
  - Example: given that you drew a red card, what's the probability that it's a four, i.e., P(four|red)=2/26=1/13. So out of the 26 red cards (given a red card), there are two fours so 2/26=1/13.

Joint Distribution

- We are interested in questions involving several random variables
  - Example event: *Intelligence=high* and *Grade=A*
  - Need to consider joint distributions
  - Over a set $\chi=\{X_1,..,X_n\}$ denoted by $P(X_1,..,X_n)$
  - We use ξ to refer to a full assignment to variables χ, i.e. $\xi \varepsilon Val(\chi)$

- Example of joint distribution
  - and marginal distributions

		Intelligence	
		Low	High
Grade	A	0.07	0.18
	B	0.28	0.09
	C	0.35	0.03

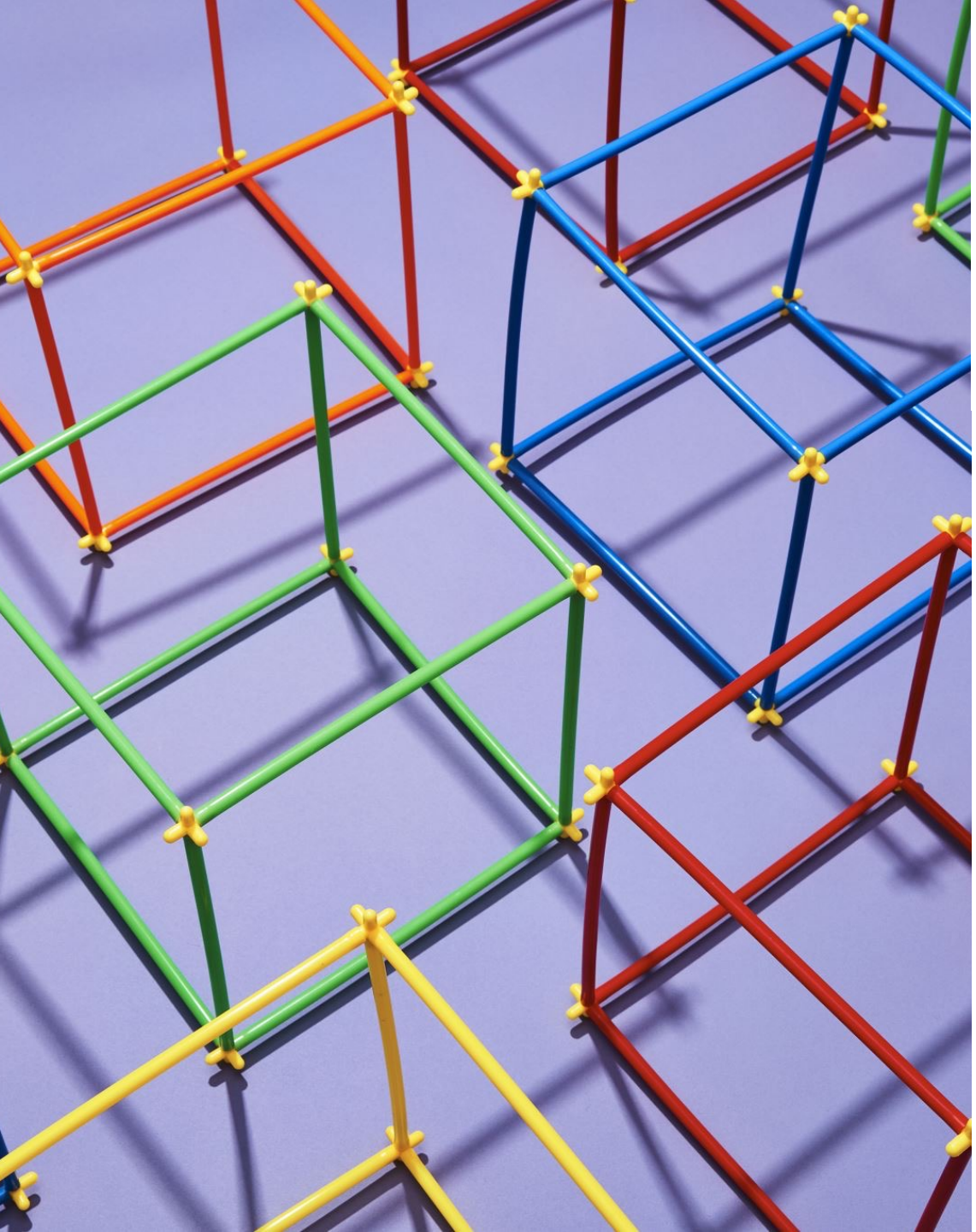
Conditional Probability

- *P(Intelligence|Grade=A)* describes the distribution over events describable by Intelligence given the knowledge that student's grade is *A*
  - It is not the same as the marginal distribution

$$P(Intelligence=high)=0.3$$

$$P(Intelligence=high|Grade=A)$$
$$=0.18/0.25$$
$$=0.72$$

	Intelligence		
Grade	Low	High	
A	0.07	0.18	0.25
B	0.28	0.09	0.37
C	0.35	0.03	0.38
	0.7	0.3	



Independence and Conditional Independence

Recap: Chain Rules

chain rule (also called the **general product rule**) permits the calculation of any member of the joint distribution of a set of random variables using only conditional probabilities.

$$\mathrm{P}(A_n, \ldots, A_1) = \mathrm{P}(A_n | A_{n-1}, \ldots, A_1) \cdot \mathrm{P}(A_{n-1}, \ldots, A_1)$$

$$\mathrm{P}(A_4, A_3, A_2, A_1) = \mathrm{P}(A_4 \mid A_3, A_2, A_1) \cdot \mathrm{P}(A_3 \mid A_2, A_1) \cdot \mathrm{P}(A_2 \mid A_1) \cdot \mathrm{P}(A_1)$$

Independent Random Variables

We expect $P(\alpha|\beta)$ to be different from $P(\alpha)$

- i.e., β is true changes our probability over α

Sometimes equality can occur, i.e, $P(\alpha|\beta)=P(\alpha)$

- i.e., learning that β occurs did not change our probability of α

We say event α is independent of event β, denoted

- $\alpha \perp \beta$, if $P(\alpha|\beta)=P(\alpha)$ or if $P(\beta)=0$

A distribution P satisfies $(\alpha \perp \beta)$ if and only if $P(\alpha \wedge \beta)=P(\alpha)P(\beta)$

Conditional Independence

- While independence is a useful property, we don't often encounter two independent events
- A more common situation is when two events are independent given an additional event
  - Reason about student accepted at Stanford or MIT
  - These two are not independent
    - If student admitted to Stanford then probability of MIT is higher
    - If both based on GPA and we know the GPA to be A
    - Then the student being admitted to Stanford does not change probability of being admitted to MIT
    - $P(MIT|Stanford,Grade\ A)=P(MIT|Grade\ A)$
    - i.e., MIT is conditionally independent of Stanford given Grade A

Querying Joint Probability Distributions

Query Types

Probability Queries

- Given evidence (the values of a subset of random variables),
- compute distribution of another subset of random variables

MAP Queries

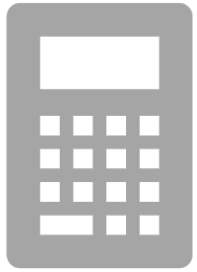
- Maximum a posteriori probability
- Also called MPE (*Most Probable Explanation*)
  - What is the most likely setting of a subset of random variables
- Marginal MAP Queries
  - When some variables are known

Probability Queries -- Most common type of queries

Query has two parts

Evidence: a subset *E* of variables and their instantiation *e*

Query Variables: a subset *Y* of random variables



Inference Task: $P(Y|E=e)$

Posterior probability distribution over values *y* of *Y*

Conditioned on the fact *E=e*

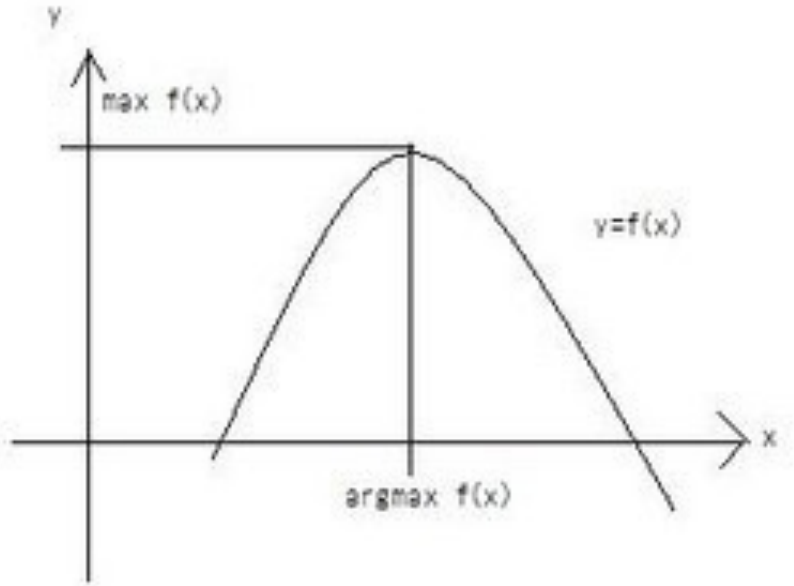
Can be viewed as Marginal over *Y* in distribution we obtain by conditioning on *e*

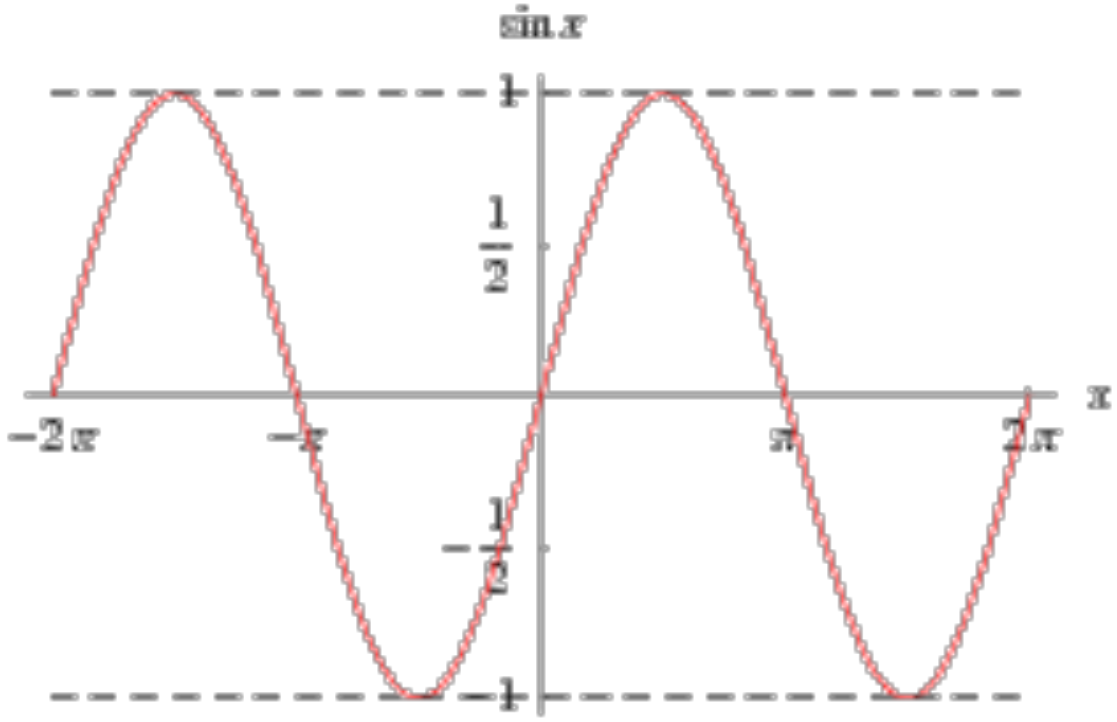
Probability Queries

- Marginal Probability Estimation

$$P(Y = y_i \mid E = e) = \frac{P(Y = y_i, E = e)}{P(E = e)}$$

Recap: max and argmax

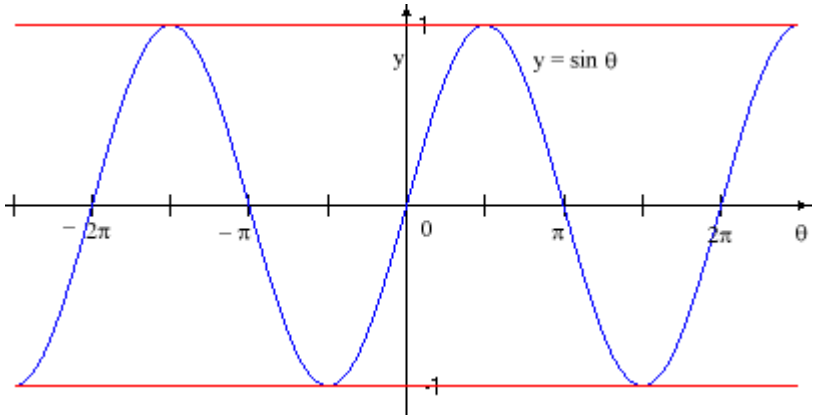




Recap: Max vs. argmax

- Let x be in a range [a,b] and f be a function over [a,b], we have
  - max f(x) to represent the maximum value of f(x) as x varies through [a,b]
  - argmax f(x) to represent the value of x at which the maximum is attained

- $\max_x \sin(x)$
  = 1
- $\text{argmax}_x \sin(x)$
  = {(0.5+2n)*pi | n is integer }
  = {…, -1.5pi, 0.5pi, 2.5pi, …}



MAP Queries (Most Probable Explanation)

- Finding a high probability assignment to some subset of variables
- Most likely assignment to all non-evidence variables $W = V - E$

$$MAP(W \mid e) = \arg\max_{w} P(w, e)$$

i.e., value of *w* for which *P(w,e)* is maximum

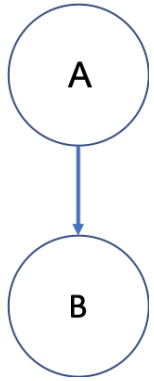
An example case. Other cases will be introduced later

- Difference from probability query
  - Instead of a probability we get the most likely value for all remaining variables

Example of MAP Queries

- Medical Diagnosis Problem
  - Diseases (*A*) cause Symptoms (*B*)
  - Two possible diseases: Mono and Flu
  - Two possible symptoms: Headache and Fever

P(A)	
a^0	a^1
0.4	0.6



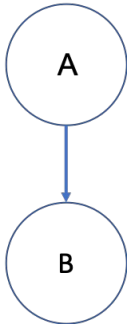
P(B|A)		
	b^0	b^1
a^0	0.1	0.9
a^1	0.5	0.5

Notation for probabilistic graphical models, to be introduced in later part of this module

Example of MAP Queries

- Medical Diagnosis Problem
  - Diseases (*A*) cause Symptoms (*B*)
  - Two possible diseases: Mono and Flu
  - Two possible symptoms: Headache and Fever

- Q1: Most likely disease *MAP (A)*?
- Q2: Most likely disease and symptom *MAP(A,B)*?
- Q3: Most likely symptom *MA*P(*B*)?



P(A)	
a^0	a^1
0.4	0.6

P(B|A)		
	b^0	b^1
a^0	0.1	0.9
a^1	0.5	0.5

Example of MAP Queries

- Medical Diagnosis Problem
  - Diseases (*A*) cause Symptoms (*B*)
  - Two possible diseases: Mono and Flu
  - Two possible symptoms: Headache and Fever

- Q1: Most likely disease MA*P(A)*?

$$\mathrm{MAP(A)} = \arg\max_a A \quad = a^1$$

P(A)	
a^0	a^1
0.4	0.6

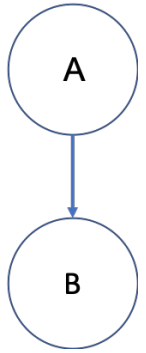
Example of MAP Queries

- Medical Diagnosis Problem
  - Diseases (*A*) cause Symptoms (*B*)
  - Two possible diseases: Mono and Flu
  - Two possible symptoms: Headache and Fever

- Q2: Most likely disease and symptom *P(A,B)*?

$$MAP(A,B) = \arg\max_{a,b} P(A,B)$$

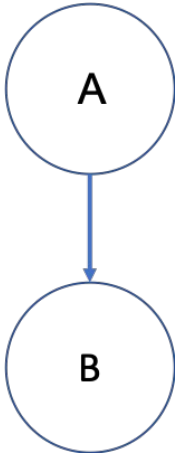
P(A)	
a^0	a^1
0.4	0.6



P(B|A)		
	b^0	b^1
a^0	0.1	0.9
a^1	0.5	0.5

Example of MAP Queries

P(A)	
a^0	a^1
0.4	0.6



P(B|A)		
	b^0	b^1
a^0	0.1	0.9
a^1	0.5	0.5

P(A,B) = P(B|A) P(A)

P(A,B)		
	b^0	b^1
a^0	0.04	0.36
a^1	0.3	0.3

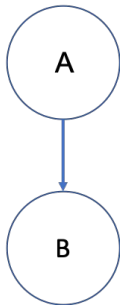
Example of MAP Queries

- Medical Diagnosis Problem
  - Diseases (*A*) cause Symptoms (*B*)
  - Two possible diseases: Mono and Flu
  - Two possible symptoms: Headache and Fever

- Q2: Most likely disease and symptom *P(A,B)*?

$$
\begin{aligned}
MAP(A,B) &= \arg\max_{a,b} P(A,B) \\
&= \arg\max_{a,b} P(B \mid A)P(A) \\
&= \arg\max_{a,b} \{0.04, 0.36, 0.3, 0.3\} \\
&= a^0, b^1
\end{aligned}
$$

P(A)	
a^0	a^1
0.4	0.6

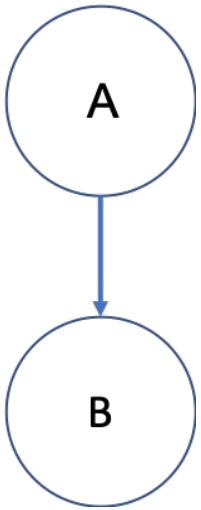


In this case, Y = {A,B}, E = {}, Z = {}, where Y, Z, E are defined in the next slide.

P(B|A)		
	b^0	b^1
a^0	0.1	0.9
a^1	0.5	0.5

Example of MAP Queries

- Q3: Most likely symptom *MAP(B)*?



P(A)	
a^0	a^1
0.4	0.6

P(B|A)		
	b^0	b^1
a^0	0.1	0.9
a^1	0.5	0.5

MAP Query

- *Y* is query, evidence is *E=e.* Task is to find most likely assignment to *Y*:
  - If *Z=X-Y-E is empty*

$$MAP(Y \mid e) = \arg\max_{y} P(y \mid e)$$

  - If *Z=X-Y-E is not empty*

$$MAP(Y \mid e) = \arg\max_{y} \sum_{z} P(y, z \mid e)$$

The set X of features can be split into three subsets Y, E, and Z, where

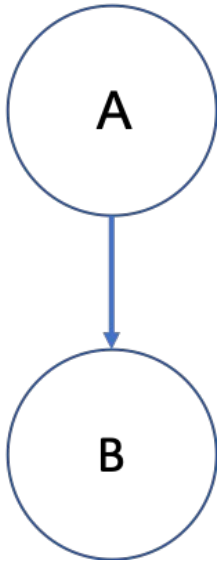
- Y is the set of variables to reason about,
- E is the set of variables that we observe, and
- Z is the set of variables we don't care.

Example of MAP Queries

- Medical Diagnosis Problem
  - Diseases (*A*) cause Symptoms (*B*)
  - Two possible diseases: Mono and Flu
  - Two possible symptoms: Headache and Fever

- Q3: Most likely symptom P(*B*)?

In the case, Y = {B},
E = {}, and Z = {A}



P(A)	
a^0	a^1
0.4	0.6

P(B|A)		
	b^0	b^1
a^0	0.1	0.9
a^1	0.5	0.5

Example of MAP Queries

P(A)	
a^0	a^1
0.4	0.6

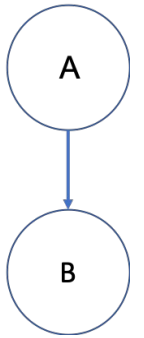
marginalisation

- Q3: Most likely symptom P(*B*)?

definition

$$MAP(B) = \arg\max_b P(b) = \arg\max_b \sum_a P(a,b)$$

$$= \arg\max_b \{0.34, 0.66\} = b^1$$



P(A,B) = P(A)P(B|A)

P(A,B)		
	b^0	b^1
a^0	0.04	0.36
a^1	0.3	0.3

P(B|A)		
	b^0	b^1
a^0	0.1	0.9
a^1	0.5	0.5

Exercise 1

Joint distribution table as shown right

Check the exercises of the lecture notes for answer

	B=1	B=2	B=3	B=4
A=1	0.12	0.18	0.24	0.06
A=2	0.06	0.09	0.12	0.03
A=3	0.02	0.03	0.04	0.01

Can you calculate the following:

P(A=1) = 0.6

P(A=2) = 0.3

P(B=3)= 0.4

P(B=4)= 0.1

P(A=1|B=2)=0.6

P(B=3 | A = 3) =0.4

MAP(A | B=2) = 1

MAP(B | A =2) =3

MAP(A) = 1

MAP(B) = 3

Exercise 2

Check the exercises of the lecture notes for answer

Joint distribution table as shown right

	B=1	B=2	B=3	B=4
A=1	0.12	0.18	0.24	0.02
A=2	0.06	0.09	0.12	0.03
A=3	0.06	0.03	0.04	0.01

Can you calculate the following:

P(A=1) = 0.56

P(A=2) = 0.3

P(B=3)= 0.4

P(B=4)= 0.06

P(A=1|B=2)= 0.6

P(B=3 | A = 3) = 2/7

MAP(A | B=2) = 1

MAP(B | A =2) = 3

MAP(A) = 1

MAP(B) = 3