Probability Foundation for Machine Learning

Dr. Xiaowei Huang

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

In the last week's lectures,

- Module contents
- A few applications of machine learning
  - Representation of instances as vectors
- Learning basics
  - Supervised vs. unsupervised learning

Topics of today

- Probability foundations
  - Random Variables
  - Joint and Conditional Distributions
  - Independence and Conditional Independence
  - Querying Joint Probability Distributions
    - Probability query
    - MAP query

Random Variables

Random Variable

- 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 $P(A \cap B)$.
  - 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

Grade		Low	High
	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

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	

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

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

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

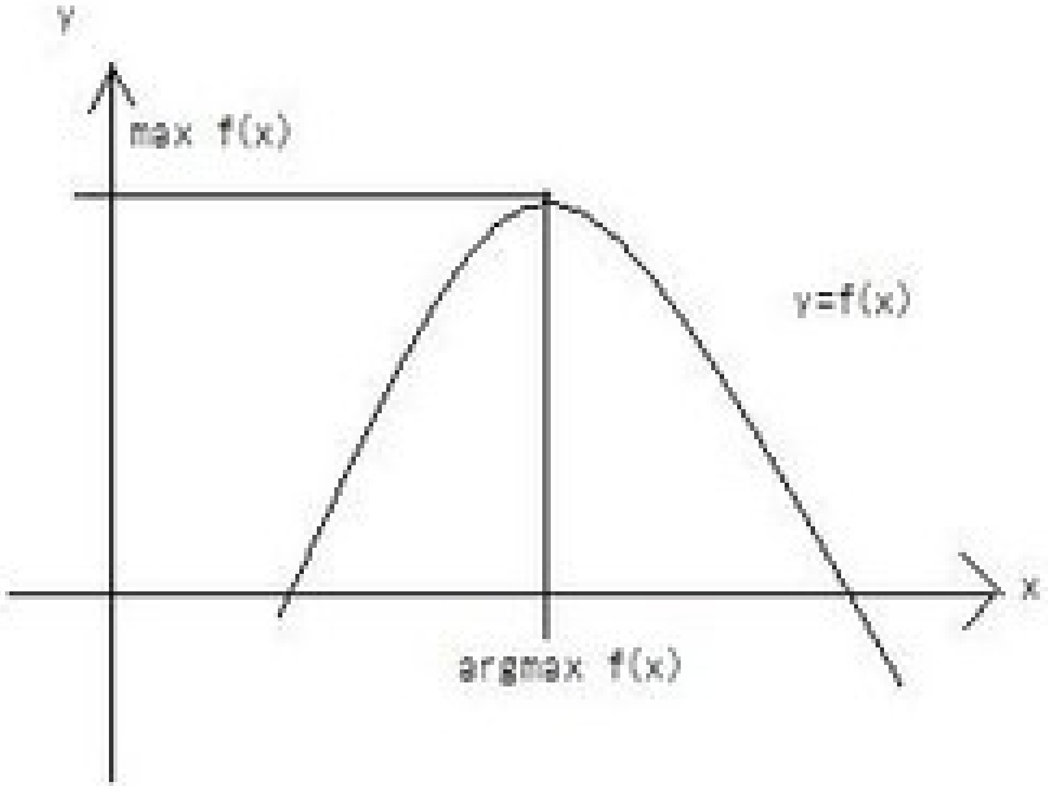
- Most common type of query is a probability query
- Query has two parts
  - *Evidence*: a subset E of variables and their instantiation e
  - *Query Variables*: a subset Y of random variables
- Inference Task: $P(Y|E=e)$
  - *Posterior probability distribution* over values y of Y
  - *Conditioned* on the fact $E=e$
  - Can be viewed as Marginal over Y in distribution we obtain by conditioning on e

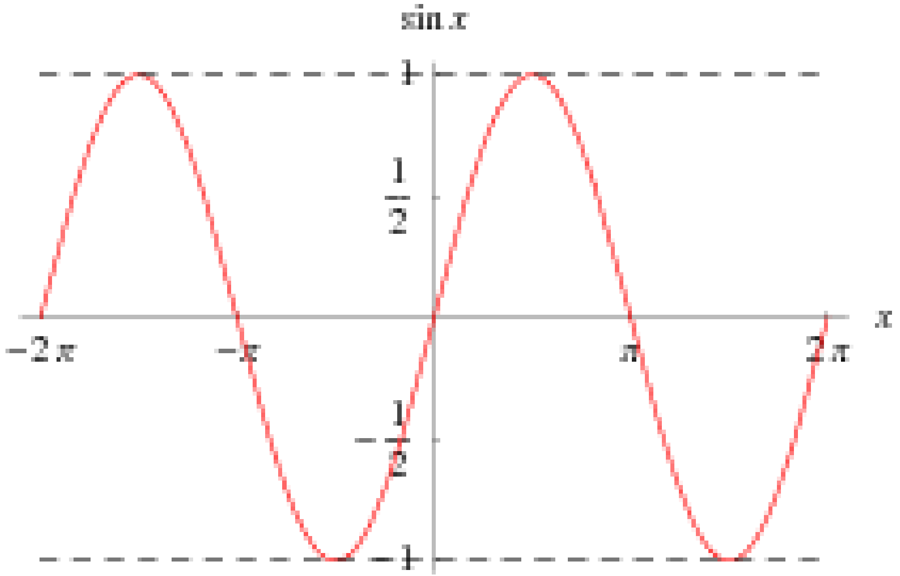
Probability Queries

- Marginal Probability Estimation

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

Recap: max and argmax

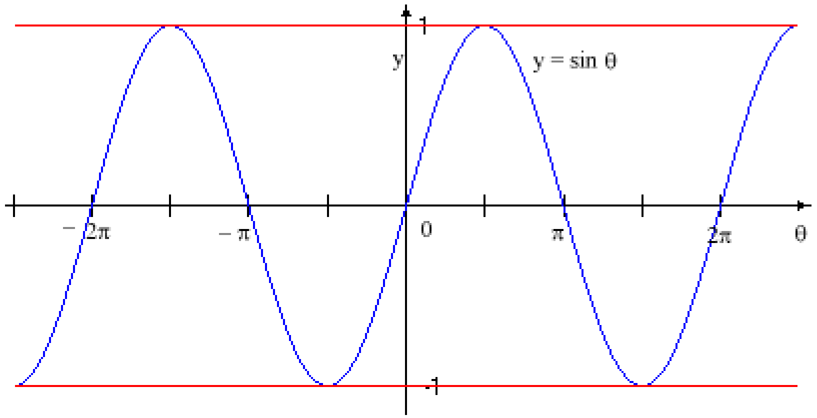




Recap: Max vs. argmax

- 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

- $\max_x \sin(x)$
  $= 1$
- $\mathrm{argmax}_x \sin(x)$
  = {(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

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

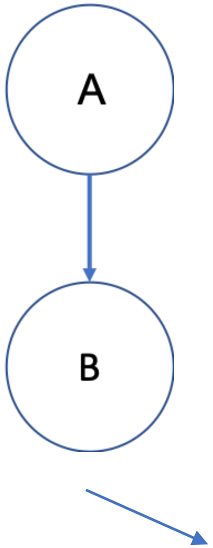
i.e., value of *w* for which *P(w,e)* is maximum

- 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



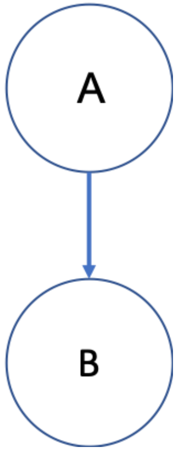
P(B|A)		
	b^0	b^1
a^0	0.1	0.9
a^1	0.5	0.5

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

Example of MAP Queries

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

- 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 *MAP*(*B*)?

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

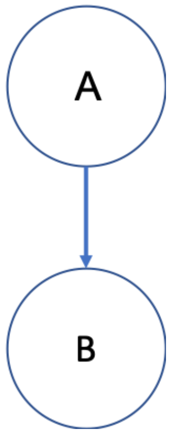
Example of MAP Queries

- 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)*?

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

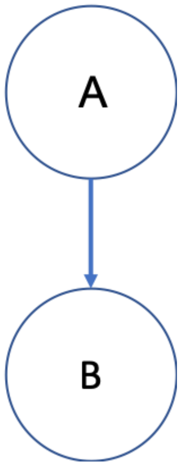
P(A)	
a^0	a^1
0.4	0.6



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

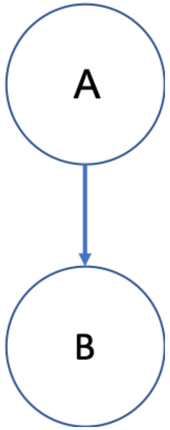
$P(A,B) = \quad P(B|A)\,P(A)$

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

Example of MAP Queries

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

- 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)*?

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

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

$$= \arg\max_{a,b}\{0.04, 0.36, 0.3, 0.3\}$$

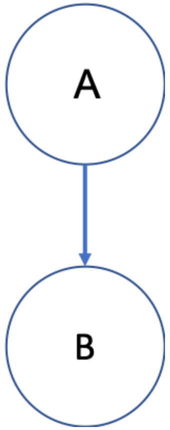
$$= a^0, b^1$$

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

Marginal MAP Query

- We looked for highest joint probability assignment of disease and symptom
- Can look for most likely assignment of disease variable only
- Query is not all remaining variables but a subset of them
  - *Y* is query, evidence is *E=e*
    Task is to find most likely assignment to *Y*:

$$MAP(Y \mid e) = \arg\max_y P(y \mid e)$$

  - If *Z=X-Y-E*

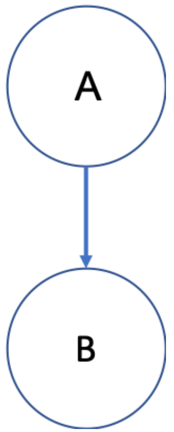
$$MAP(Y \mid e) = \arg\max_y \sum_z P(y, z \mid e)$$

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*)?

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



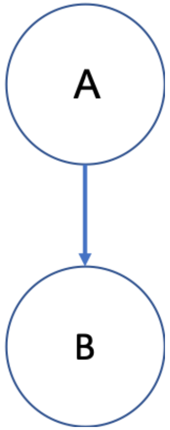
P(B|A)		
	b^0	b^1
a^0	0.1	0.9
a^1	0.5	0.5

Example of MAP Queries

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

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

$$MAP(B) = \arg\max_b P(b) = \arg\max_b \sum_a P(a,b)$$
$$= \arg\max_b \{0.34, 0.66\} = b^1$$



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

Marginal MAP Assignments

- They are not monotonic
- Most likely assignment MAP(Y_1|e) might be completely different from assignment to Y_1 in MAP({Y_1,Y_2}|e)
  - Q1: Most likely disease MAP(A)?
  - A1: Flu
  - Q2: Most likely disease and symptom MAP(A,B)?
  - A2: Mono and Fever
- Thus we cannot use a MAP query to give a correct answer to a marginal MAP query

Marginal MAP more Complex than MAP

- Contains both summations (like in probability queries) and maximizations (like in MAP queries)

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

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

Exercise 1

Joint distribution table as shown right

	B=1	B=2	B=3	B=4
A=1	0.12	0.18	0.24	0.08
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) =
P(A=2) =
P(B=3)=
P(B=4)=
P(A=1|B=2)=
P(B=3 | A = 3) =
MAP(A | B=2) =
MAP(B | A =2) =
MAP(A) =
MAP(B) =

Exercise 2

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	0.08	0.03	0.04	0.01

Can you calculate the following:

P(A=1) =
P(A=2) =
P(B=3)=
P(B=4)=
P(A=1|B=2)=
P(B=3 | A = 3) =
MAP(A | B=2) =
MAP(B | A =2) =
MAP(A) =
MAP(B) =