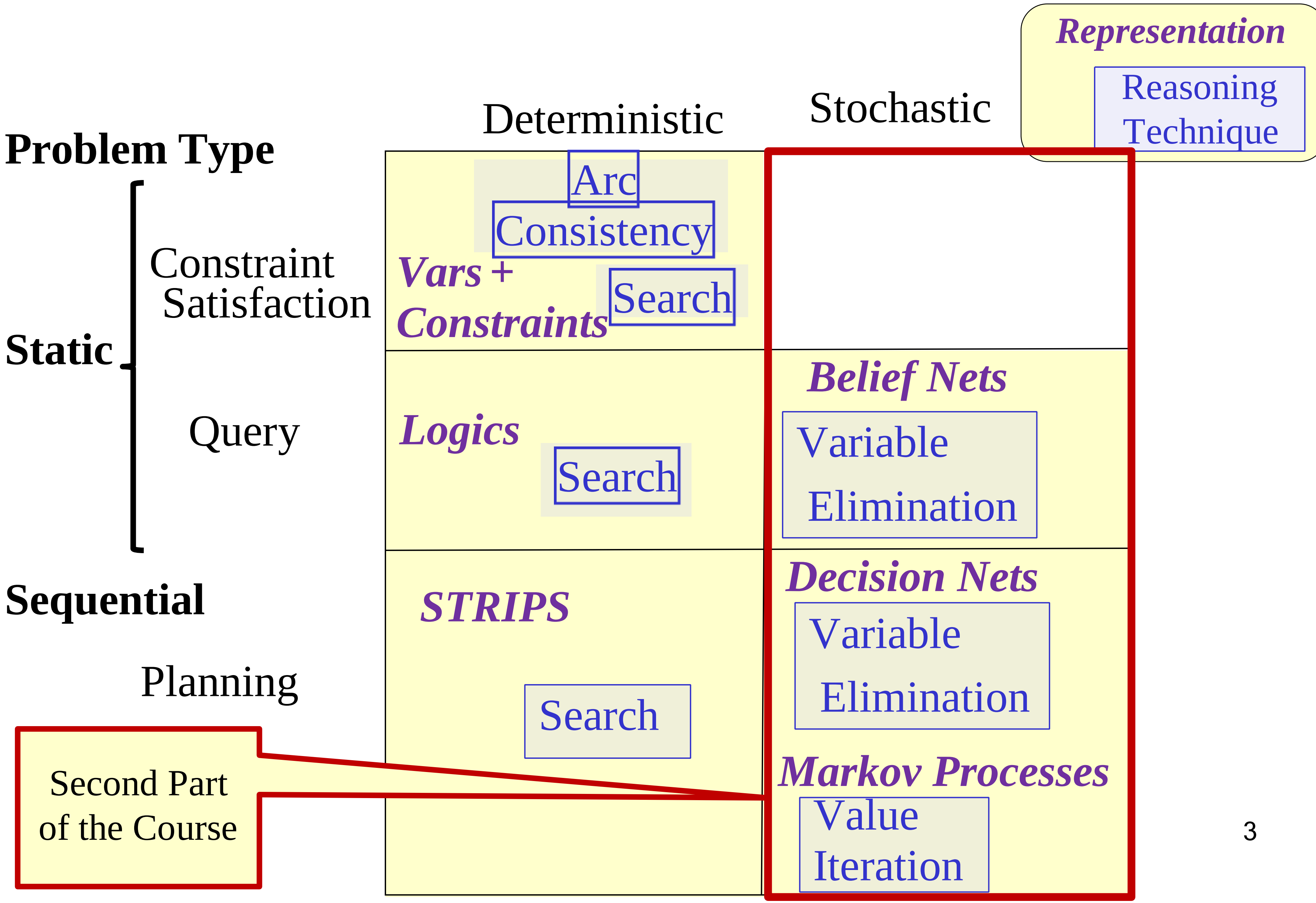
Lecture 18

Marginalization, Conditioning

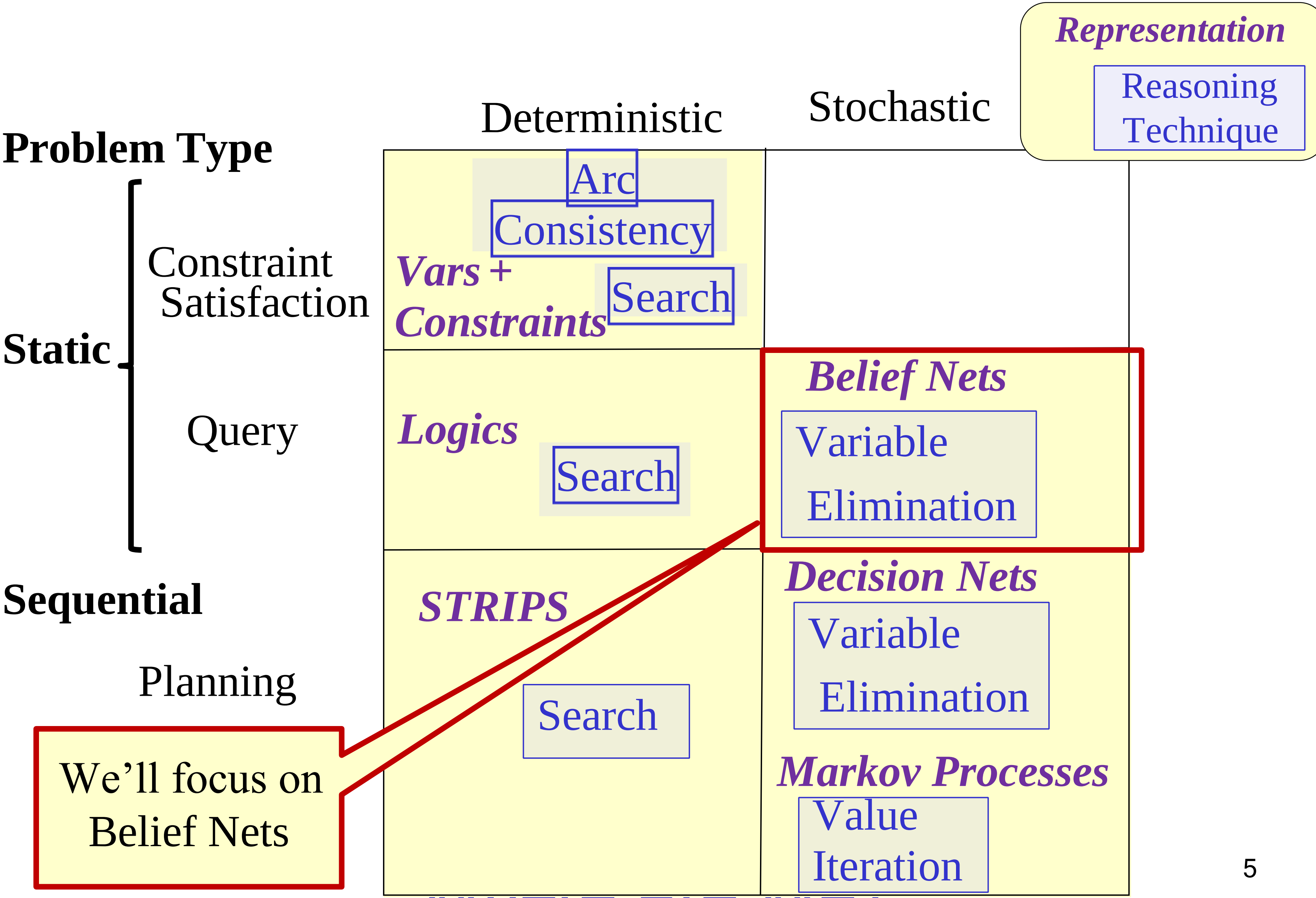
Lecture Overview

- Recap Lecture 17
- Joint Probability Distribution, Marginalization
- Conditioning

- Inference by Enumeration
- Bayes Rule, Chain Rule (time permitting)


Representation
Reasoning Technique
Deterministic
Stochastic
Problem Type
Static
Constraint Satisfaction
Query
Sequential
Planning
Vars + Constraints
Arc Consistency
Search
Logics
Search
STRIPS
Search
Belief Nets
Variable Elimination
Decision Nets
Variable Elimination
Markov Processes
Value Iteration
Second Part of the Course

Environment


Representation
Reasoning Technique
Deterministic
Stochastic
Problem Type
Static
Constraint Satisfaction
Query
Sequential
Planning
Vars + Constraints
Arc Consistency
Search
Logics
Search
STRIPS
Search
Belief Nets
Variable Elimination
Decision Nets
Variable Elimination
Markov Processes
Value Iteration
We'll focus on Belief Nets

Where Are We?

Probability as a measure of uncertainty/ignorance

- Probability measures an agent's degree of belief in truth of Environment

Probability as a measure of uncertainty/ignorance

- Probability measures an agent's degree of belief in truth of propositions (Boolean statements) about states of the world
- It does not measure how true a proposition is
- Propositions are true or false. We simply may not know exactly which.
- Belief in a proposition fcan be measured in terms of a number between 0 and 1
- this is the probability of f
- E.g. P("roll of fair die came out as a 6") = 1/6 ≈ 16.7% = 0.167
- Using probabilities between 0 and 1 is purely a convention.

Probability as a measure of uncertainty/ignorance

- Probability measures an agent's degree of belief in truth of
  - P(f) = 0 means that f is believed to be
  - Definitely false: the probability of f being true is zero.
  - Likewise, P(f) = 1 means fis believed to be definitely true

propositions about states of the world

- It does not measure how true a proposition is
- Propositions are true or false. We simply may not know exactly which.
- Example:
- I roll a fair dice. What is 'the' (my) probability that the result is a '6'?

Probability as a measure of uncertainty/ignorance

- Probability measures an agent's degree of belief in truth of

7

  propositions about states of the world

  - It does not measure how true a proposition is
  - Propositions are true or false. We simply may not know exactly which.

  - Example:
  - I roll a fair dice. What is ‘the’ (my) probability that the result is a ‘6’?

    ✓It is 1/6 ≈ 16.7%.

  - I now look at the dice. What is ‘the’ (my) probability now?

    ✓My probability is now

    ✓Your probability (you have not looked at the dice)

propositions about states of the world

- It does not measure how true a proposition is
- Propositions are true or false. We simply may not know exactly which.

- Example:
- I roll a fair dice. What is ‘the’ (my) probability that the result is a ‘6’?

✓It is 1/6 ≈ 16.7%.

- I now look at the dice. What is ‘the’ (my) probability now?

✓My probability is now either 1 or 0, depending on what I observed.

Probability as a measure of uncertainty/ignorance

- Probability measures an agent's degree of belief in truth of
    - ✓Your probability hasn't changed: 1/6 ≈ 16.7%
  - What if I tell some of you the result is even?
    - ✓Their probability

Probability as a measure of uncertainty/ignorance

- Probability measures an agent's degree of belief in truth of propositions about states of the world
- It does not measure how true a proposition is
- Propositions are true or false. We simply may not know exactly which.
- Example:
- I roll a fair dice. What is 'the' (my) probability that the result is a '6'?

  ✓It is 1/6 ≈ 16.7%.
- I now look at the dice. What is 'the' (my) probability now?
  ✓My probability is now either 1 or 0, depending on what I observed.
  ✓Your probability hasn't changed: 1/6 ≈ 16.7%

- What if I tell some of you the result is even?
  ✓Their probability increases to 1/3 ≈ 33.3%, if they believe me
- Different agents can have different degrees of belief in (probabilities for) a proposition, based on the evidence they have.

Lecture Overview

- Recap Lecture 17
- Joint Probability Distribution, Marginalization
- Conditioning

-
- Inference by Enumeration
- Bayes Rule, Chain Rule (time permitting)

Probability Theory and Random Variables

Probability Theory

- system of logical axioms and formal operations for sound reasoning under uncertainty
- Basic element: random variable X
- X is a variable like the ones we have seen in CSP/Planning/Logic
- but the agent can be uncertain about the value of X
- As usual, the domain of a random variable X, written dom(X), is the set of values Xcan take
- Types of variables

- Boolean: e.g., Cancer (does the patient have cancer or not?)
- Categorical: e.g., Cancer Type could be one of {breast Cancer, lung Cancer, skin Melanomas}
- Numeric: e.g., Temperature (integer or real)
- We will focus on Boolean and categorical variables

•

Random Variables (cont')

A tuple of random variables $<X_1 ,...., X_n>$ is a joint random variable with domain..

$Dom(X_1) \times Dom(X_2)... \times Dom(X_n)...$ (cross product)

- A proposition is a Boolean formula (i.e., true or false) made from assignments of values to (some of) the variables in the joint

Example:

Given the joint random variable <Cavity, Weather>, with

Dom (Cavity) = {T,F}
Dom (Weather) = {sunny, cloudy},
possible propositions are

Possible Worlds

Random Variables (cont')

A tuple of random variables <X_1 ,…., X_n> is a joint random variable with domain..

$Dom(X_1) \times Dom(X_2)\ldots \times Dom(X_n)\ldots$ (cross product)

- A proposition is a Boolean formula (i.e., true or false) made from assignments of values to (some of) the variables in the joint

Example:

Given the joint random variable <Cavity, Weather>, with

Dom (Cavity) = {T,F}

- Dom (Weather) = {sunny, cloudy},
  possible propositions are

$$\boldsymbol{CCCCCCCCCCCC = TT\ , WWWWCCCCWWWWWW = ccccccccccCC}$$

$$\boldsymbol{CCCCCCCCCCCCC = FF}$$

- A possible world specifies an assignment to each random variable

- E.g., if we model only two Boolean variables Cavityand Toothache, then there are 4 distinct possible worlds:

w1: Cavity = T ∧Toothache = T w2: Cavity = T ∧ Toothache = F w3; Cavity = F ∧ Toothache = T w4: Cavity = T ∧ Toothache = T possible worlds are mutually exclusive and exhaustive

Cavity	*Toothache*
T	T
T	F
F	T
F	F

Possible Worlds

- $w \models f$ means that proposition f is true in world w
- A probability measure $\mu(w)$ over possible worlds w is a nonnegative real number such that

Why does this make sense?

- $\mu(w)$ sums to 1 over all possible worlds w

Possible Worlds

- A possible world specifies an assignment to each random variable
- E.g., if we model only two Boolean variables Cavityand Toothache, then there are 4 distinct possible worlds:

Cavity	*Toothache*
T	T
T	F
F	T
F	F

w1: Cavity = T $\wedge$ Toothache = T w2: Cavity = T $\wedge$ Toothache = F w3; Cavity = F $\wedge$ Toothache = T w4: Cavity = T $\wedge$ Toothache = T possible worlds are mutually exclusive and exhaustive

- w $\models$ f means that proposition fis true in world w
- A probability measure μ(w) over possible worlds w is a nonnegative real number such that

Possible Worlds

Because for sure we are in one of these worlds

- $\mu(w)$ sums to 1 over all possible worlds w

- A possible world specifies an assignment to each random variable

- E.g., if we model only two Boolean variables Cavityand Toothache, then there are 4 distinct possible worlds:

- $w \models f$ means that proposition fis true in world w
- A probability measure $\mu(w)$ over possible worlds w is a nonnegative real number such that
  - $\mu(w)$ sums to 1 over all possible worlds w

- The probability of proposition f is defined by:

$$P(f)=\Sigma_{w \models f}\mu(w).$$ i.e.

sum of the probabilities of the worlds w in which f is true[18]

w1: Cavity = T ∧ Toothache = T w2: Cavity = T ∧ Toothache = F w3; Cavity = F ∧ Toothache = T w4: Cavity = T ∧ Toothache = T possible worlds are mutually exclusive and exhaustive

Cavity	*Toothache*
T	T
T	F
F	T
F	F

Example

Example: weather in Vancouver •

two Boolean variable:

- Weather with domain {sunny, cloudy}
- Temperature, with domain {hot, mild, cold}

Possible Worlds

- There are 6 possible worlds:
- What's the probability of it being cloudy and cold?

Weather	*Temperature*	$\mu(w)$
sunny	hot	0.10
sunny	mild	0.20
sunny	cold	0.10
cloudy	hot	0.05
cloudy	mild	0.35
cloudy	cold	?

A. 0.1 B. 0.2 C. 0.3 D. 1 E. Not enough info

Example

Example: weather in Vancouver •

two Boolean variable:

- Weather with domain {sunny, cloudy}
- Temperature, with domain {hot, mild, cold}

Weather	*Temperature*	*$\mu(w)$*
sunny	hot	0.10
sunny	mild	0.20
sunny	cold	0.10
cloudy	hot	0.05
cloudy	mild	0.35

- There are 6 possible worlds:

cloudy	cold	?

- What's the probability of it being cloudy and cold?

0.10 + 0.20 + 0.10 + 0.05 + 0.35 = 0.8

It is 0.2: the probability has to sum to 1 over all possible worlds

One more example

- What's the probability of it being cloudy or cold?

A. 1

B. 0.6 C. 0.3 D. 0.7

Weather	*Temperature*	*µ(w)*
sunny	hot	0.10
sunny	mild	0.20
sunny	cold	0.10
cloudy	hot	0.05
cloudy	mild	0.35
cloudy	cold	0.20

- Remember
  - The probability of proposition f is defined by: $P(f)=\Sigma_{w \models f} \mu(w)$
  - sum of the probabilities of the worlds w in which f is true

One more example

- What's the probability of it being cloudy or cold?
- $\mu(w3) + \mu(w4) + \mu(w5) + \mu(w6) =$

0.7

	Weather	*Temperature*	$\mu(w)$
w1	sunny	hot	0.10
w2	sunny	mild	0.20
w3	sunny	cold	0.10
w4	cloudy	hot	0.05
w5	cloudy	mild	0.35
w6	cloudy	cold	0.20

- Remember
  - The probability of proposition f is defined by: $P(f)=\Sigma_{w \models f}\, \mu(w)$
  - sum of the probabilities of the worlds w in which *f* is true

Probability Distributions

Consider the case where possible worlds are simply assignments to one random variable.

Definition (probability distribution)
A probability distribution P on a random variable X is a function $dom(X) \rightarrow [0,1]$ such that

$$x \rightarrow P(X=x)$$

- When dom(X) is infinite we need a probability density function
- We will focus on the finite case

Probability Distributions

Consider the case where possible worlds are simply assignments to one random variable.

Definition (probability distribution)
A probability distribution P on a random variable X is a function dom(X) → [0,1] such that x → P(X=x)

Example: X represents a female adult's hight in Canada with domain {short, normal, tall} – based on some definition of these terms

short → P(hight = short) = 0.2

normal → P(hight = normal) = 0.5

tall → P(hight = tall) = 0.3

Joint Probability Distribution (JPD)

- Joint probability distribution over random variables X_1, …, X_n:
- a probability distribution over the joint random variable <X_1, …, X_n> with domain $dom(X_1) \times \ldots \times dom(X_n)$ (the Cartesian product)
- Think of a joint distribution over nvariables as the table of the corresponding possible worlds

- There is a column (dimension) for each variable, and one for the probability
- Each row corresponds to an assignment

$X_1 = x_1, \ldots, X_n = x_n$ and its probability $P(X_1 = x_1, \ldots, X_n = x_n)$

- We can also write $P(X_1 = x_1 \wedge \ldots \wedge X_n = x_n)$
- The sum of probabilities across the whole table is 1.

{Weather, Temperature} example from before

Weather	*Temperature*	*μ(w)*
sunny	hot	0.10
sunny	mild	0.20
sunny	cold	0.10
cloudy	hot	0.05
cloudy	mild	0.35
cloudy	cold	0.20

But why do we care about all this?

Because an agent can use JPDs to answer queries in a stochastic environment

Query Answering in a Stochastic Domain

- Given
- Prior joint probability (JPD) distribution (JPD) on set of variables X
- Observations of specific values e for a subset of (X) evidence variables E (subset of X)

- We want to compute
- JPD of query variables Y (a subset of X) given evidence e

To do this, we need to work through a few more definitions and operations

Marginalization

- Given the joint distribution, we can compute distributions over subsets of the variables through marginalization:

$\boxed{P(X=x) = \Sigma_{z\in dom(Z)} P(X=x, Z = z)}$ Marginalization over Z

We also write this as $P(X) = \Sigma_{z\in dom(Z)} P(X, Z = z)$.

- Simply an application of the definition of probability measure!

- Remember?
  - The probability of proposition f is defined by: $P(f) = \Sigma_{w \models f} \mu(w)$
  - sum of the probabilities of the worlds w in which f is true

Marginalization

- Given the joint distribution, we can compute distributions over subsets of the variables through marginalization:

$$\boxed{P(X=x) = \Sigma_{z\in dom(Z)}\ P(X=x, Z = z)}$$

Marginalization over Z

- We also write this as $P(X) = \Sigma_{z\in dom(Z)}\ P(X, Z = z)$.

- This corresponds to summing out a dimension in the table..

Weather	*Temperature*	*µ(w)*
sunny	hot	0.10
sunny	mild	0.20
sunny	cold	0.10
cloudy	hot	0.05
cloudy	mild	0.35
cloudy	cold	0.20

Temperature	*µ(w)*
hot	?
mild	?
cold	?

Marginalizationover Weather

Marginalization

- Given the joint distribution, we can compute distributions

Probabilities in new table still sum to 1 over subsets of the variables through marginalization:

$$\boxed{P(X{=}x) = \Sigma_{z \in dom(Z)}\, P(X{=}x, Z = z)}$$

Marginalization over Z

- We also write this as $P(X) = \Sigma_{z \in dom(Z)}\, P(X, Z = z)$.

- This corresponds to summing out a dimension in the table.

Marginalizatio

Weather	*Temperature*	*µ(w)*
sunny	hot	0.10
sunny	mild	0.20
sunny	cold	0.10
cloudy	hot	0.05
cloudy	mild	0.35
cloudy	cold	0.20

How do we compute P(T =hot)?

Temperature	*µ(w)*
hot	??
mild	
cold	

Marginalization

- Given the joint distribution, we can compute distributions over subsets of the variables through marginalization:

$$P(X=x) = \Sigma_{z\in dom(Z)}\ P(X=x, Z = z)$$

Marginalization over Z

- We also write this as $P(X) = \Sigma_{z\in dom(Z)}\ P(X, Z = z)$.

- This corresponds to summing out a dimension in the table.

Weather	*Temperature*	*µ(w)*
sunny	hot	0.10
sunny	mild	0.20
sunny	cold	0.10
cloudy	hot	0.05

Marginalization

cloudy	mild	0.35
cloudy	cold	0.20

mild cold

Temperature	*µ(w)* hot ??

P(Temperature=hot) =

P(Weather=sunny, Temperature = hot)

+ P(Weather=cloudy, Temperature = hot) =

over subsets of the variables through marginalization:

$$P(X=x) = \Sigma_{z\in dom(Z)}\ P(X=x,\ Z = z)$$

Marginalization over Z

- We also write this as $P(X) = \Sigma_{z\in dom(Z)}\ P(X, Z = z)$.

- This corresponds to summing out a dimension in the table.

Marginalization

- Given the joint distribution, we can compute distributions

Weather	*Temperature*	*μ(w)*
sunny	hot	0.10
sunny	mild	0.20
sunny	cold	0.10
cloudy	hot	0.05
cloudy	mild	0.35
cloudy	cold	0.20

Temperature *μ(w)* hot 0.15

mild cold

P(Temperature=hot) =
P(Weather=sunny, Temperature = hot)
+ P(Weather=cloudy, Temperature = hot) = 0.10 + 0.05 = 0.15

- Given the joint distribution, we can compute distributions over smaller sets of variables through marginalization:

Marginalization

$$P(X{=}x) = \Sigma_{z \in dom(Z)}\ P(X{=}x, Z = z)$$

Marginalization over Z

- We also write this as $P(X) = \Sigma_{z \in dom(Z)}\ P(X, Z = z)$.

You can marginalize over any of the variables

e.g., Marginalization over Temperature

Weather	*Temperature*	*µ(w)*
sunny	hot	0.10
sunny	mild	0.20
sunny	cold	0.10
cloudy	hot	0.05
cloudy	mild	0.35
cloudy	cold	0.20

Weather	*µ(w)*
sunny	??
cloudy	

Marginalization

- We also marginalize over more than one variable at once

$$\boxed{P(X=x) =} \sum_{z1 \in dom(Z1), \dots, zn \in dom(Zn)} \boxed{P(X=x, Z_1} \boxed{= z_1} \boxed{, \dots, Z_n} \boxed{= z_n)}$$

- E.g. go from P(Wind, Weather, Temperature) to P (Weather)

Marginalization

Wind	*Weather*	*Temperature*	*µ(w)*
yes	sunny	hot	0.04
yes	sunny	mild	0.09
yes	sunny	cold	0.07
yes	cloudy	hot	0.01
yes	cloudy	mild	0.10
yes	cloudy	cold	0.12
no	sunny	hot	0.06
no	sunny	mild	0.11
no	sunny	cold	0.03
no	cloudy	hot	0.04
no	cloudy	mild	0.25
no	cloudy	cold	0.08

i. e., Marginalizationover Temperature and Wind

Weather	*µ(w)*
sunny	
cloudy	

Still simply an application of the definition of probability measure

The probability of proposition f is $P(f)=\Sigma_{w \models f} \mu(w)$: sum of the probabilities of the worlds w in which f is true

- We also marginalize over more than one variable at once

$$\boxed{P(X=x) =} \Sigma_{z1\in dom(Z1),\ldots,zn\in dom(Zn)} \boxed{P(X=x, Z}_1 \boxed{= z}_1 \boxed{, \ldots, Z}_n \boxed{= zn)}$$

Marginalization

- E.g. go from P(Wind, Weather, Temperature) to P (Weather)

Marginalization

Wind	*Weather*	*Temperature*	*µ(w)*
yes	sunny	hot	0.04
yes	sunny	mild	0.09
yes	sunny	cold	0.07
yes	cloudy	hot	0.01
yes	cloudy	mild	0.10
yes	cloudy	cold	0.12
no	sunny	hot	0.06
no	sunny	mild	0.11
no	sunny	cold	0.03
no	cloudy	hot	0.04
no	cloudy	mild	0.25
no	cloudy	cold	0.08

i. e., Marginalizationover Temperature and Wind

Weather	*µ(w)*
sunny	???
cloudy	

- We can also marginalize over **more than one variable at once**

$$P(X=x) = \sum_{z_1 \in dom(Z_1), \ldots, z_n \in dom(Z_n)} P(X=x, Z_1 = z_1, \ldots, Z_n = z_n)$$

Marginalization

Wind	*Weather*	*Temperature*	*µ(w)*
yes	sunny	hot	0.04
yes	sunny	mild	0.09
yes	sunny	cold	0.07
yes	cloudy	hot	0.01
yes	cloudy	mild	0.10
yes	cloudy	cold	0.12
no	sunny	hot	0.06
no	sunny	mild	0.11
no	sunny	cold	0.03
no	cloudy	hot	0.04
no	cloudy	mild	0.25
no	cloudy	cold	0.08

i.e., Marginalization
over Temperature and Wind

Weather	*µ(w)*
sunny	0.40
cloudy	

Marginalization

- We can also get marginals for more than one variable

$$P(X{=}x, Y{=}y) = \Sigma_{z_1 \in dom(Z_1), \ldots, z_n \in dom(Z_n)} P(X{=}x, Y{=}y, Z_1 = z_1, \ldots, Z_n = z_n)$$

Wind	*Weather*	*Temperature*	*µ(w)*
yes	sunny	hot	0.04
yes	sunny	mild	0.09
yes	sunny	cold	0.07
yes	cloudy	hot	0.01
yes	cloudy	mild	0.10
yes	cloudy	cold	0.12
no	sunny	hot	0.06
no	sunny	mild	0.11
no	sunny	cold	0.03
no	cloudy	hot	0.04
no	cloudy	mild	0.25
no	cloudy	cold	0.08

Weather	*Temperature*	*µ(w)*
sunny	hot	0.10
sunny	mild	
sunny	cold	
cloudy	hot	
cloudy	mild	
cloudy	cold	

Still simply an application of the definition of probability measure

The probability of proposition f is $P(f) = \Sigma_{w \models f} \mu(w)$: sum of the probabilities of the worlds w in which f is true

Lecture Overview

- Recap Lecture 16
- Joint Probability Distribution, Marginalization
- Conditioning
- Inference by Enumeration
- Bayes Rule, Chain Rule (time permitting)

Conditioning

- Are we done with reasoning under uncertainty? What can happen?
- Remember from last class

  I roll a fair dice. What is ‘the’ (my) probability that the result is a ‘6’?

  - It is 1/6 ≈ 16.7%.

  - I now look at the dice. What is ‘the’ (my) probability now?

- My probability is now either 1 or 0, depending on what I observed.

  - Your probability hasn’t changed: 1/6 ≈ 16.7%

  - What if I tell some of you the result is even?

  - Their probability increases to 1/3 ≈ 33.3%, if they believe me

- Different agents can have different degrees of belief in (probabilities for) a proposition, based on the evidence they have.

Conditioning

Conditioning: revise beliefs based on new observations

- Build a probabilistic model (the joint probability distribution, JPD)
  - ✓Take into account all background information
  - ✓Called the prior probability distribution
  - ✓Denote the prior probability for hypothesis h as P(h)
- Observe new information about the world
  - ✓Call all information we received subsequently the evidence e
- Integrate the two sources of information
  - ✓to compute the conditional probability P(h|e)

✓This is also called the posterior probability of h given e.

Example

- Prior probability for having a disease (typically small)
- Evidence: a test for the disease comes out positive
  - ✓But diagnostic tests have false positives
- Posterior probability: integrate prior and evidence

Example for conditioning

- You have a prior for the joint distribution of weather and temperature

Possible world	*Weather*	*Temperature*	*μ(w)*
w_1	sunny	hot	0.10
w_2	sunny	mild	0.20
w_3	sunny	cold	0.10
w_4	cloudy	hot	0.05
w_5	cloudy	mild	0.35
w_6	cloudy	cold	0.20

T	*P(T|W=sunny)*
hot	0.10/0.40=0.25
mild	
cold	

- Now, you look outside and see that it's sunny
- You are now certain that you're in one of worlds w_1, w_2, or w_3
  - To get the conditional probability P(T|W=sunny)
    - renormalize $\mu(w_1)$, $\mu(w_2)$, $\mu(w_3)$ to sum to 1
  - $\mu(w_1) + \mu(w_2) + \mu(w_3) = 0.10+0.20+0.10=0.40$

Example for conditioning

- You have a prior for the joint distribution of weather and temperature

Possible world	*Weather*	*Temperature*	*µ(w)*
w_1	sunny	hot	0.10
w_2	sunny	mild	0.20
w_3	sunny	cold	0.10
w_4	cloudy	hot	0.05
w_5	cloudy	mild	0.35
w_6	cloudy	cold	0.20

T	*P(T|W=sunny)*
hot	0.10/0.40=0.25
mild	??
cold	

- Now, you look outside and see that it's sunny
- You are nowcertain that you're in one of worlds w_1, w_2, or w_3

- To get the conditional probability P(T|W=sunny)
  - renormalize $\mu(w_1)$, $\mu(w_2)$, $\mu(w_3)$ to sum to 1
- $\mu(w_1) + \mu(w_2) + \mu(w_3) = 0.10+0.20+0.10=0.40$

Example for conditioning

- You have a prior for the joint distribution of weather and

Possible Weather Temperature µ(w) world

w_1 sunny hot 0.10 w_2 sunny mild 0.20 w_3 sunny

cold 0.10 w_4 cloudy hot 0.05 w_5 cloudy mild 0.35

w_6 cloudy cold 0.20

T	*P(T|W=sunny)*
hot	0.10/0.40=0.25
mild	0.20/0.40=0.50
cold	0.10/0.40=0.25

- Now, you look outside and see that it's sunny

- You are nowcertain that you're in one of worlds w_1, w_2, or w_3

- To get the conditional probability P(T|W=sunny)
  - renormalize $\mu(w_1)$, $\mu(w_2)$, $\mu(w_3)$ to sum to 1
- $\mu(w_1) + \mu(w_2) + \mu(w_3) = 0.10+0.20+0.10=0.40$

Conditional Probability

Definition (conditional probability)

The conditional probability of proposition h given evidence e is

$$P(h \mid e) = \frac{P(h \wedge e)}{P(e)}$$

- P(e): Sum of probability for all worlds in which e is true
- P(h∧e): Sum of probability for all worlds in which both h and e are true

-

Recap: Conditional probability

Definition (conditional probability)

The conditional probability of formula h given evidence e is

$$P(h|e) = \frac{P(h \wedge e)}{P(e)}$$

E.g. $P(T = hot|W = sunny) = \frac{P(T=hot \wedge W=sunny)}{P(W=sunny)}$

Possible world	Weather	Temperature	*µ(w)*
w_1	sunny	hot	0.10
w_2	sunny	mild	0.20
w_3	sunny	cold	0.10
~~w_4~~	~~cloudy~~	~~hot~~	~~0.05~~
~~w_5~~	~~cloudy~~	~~mild~~	~~0.35~~
~~w_6~~	~~cloudy~~	~~cold~~	~~0.20~~

T	*P(T|W=sunny)*
hot	0.10/0.40=0.25
mild	0.20/0.40=0.50
cold	0.10/0.40=0.25

Example for conditioning

- Note how the belief over the possible values of T changed given the new evidence

T	*P(*
hot	0.
mild	0.
cold	0.

T	*P(T|W=sunny)*
hot	0.10/0.40=0.25
mild	0.20/0.40=0.50
cold	0.10/0.40=0.25

How do we get this distribution from the original joint distribution P(W, T)?

Marginalization

- By Marginalizing over weather!

Weather	*Temperature*	*μ(w)*

sunny	hot	0.10
sunny	mild	0.20
sunny	cold	0.10
cloudy	hot	0.05
cloudy	mild	0.35
cloudy	cold	0.20

Temperature *μ(w)* hot	0.15

mild

cold

P(Temperature=hot) =
P(Weather=sunny, Temperature = hot)
+ P(Weather=cloudy, Temperature = hot) =
0.10 + 0.05 = 0.15

Conditional Probability among Random Variables

P(X | Y) = P(X , Y) / P(Y)

It expresses the conditional probability of each possible value for X given each possible value for Y

P(X | Y) = P(Temperature | Weather) = P(Temperature $\wedge$ Weather) / P(Weather)

	T = hot	
W = sunny	P(hot|sunny)	
W = cloudy	P(hot|cloudy)	

Crucial that you can answer this question. Think about it at home and let me know if you have questions next time

Which of the following is true?

A. The probabilities in each row should sum to 1

B. The probabilities in each column should sum to 1

C. Both of the above

D. None of the above

Lecture Overview

- Recap Lecture 16
- Joint Probability Distribution, Marginalization
- Conditioning
- Inference by Enumeration
- Bayes Rule, Chain Rule (time permitting)

Query Answering in a Stochastic Domain

Great, we can compute arbitrary probabilities now!

- Given
- Prior joint probability (JPD) distribution on set of variables X
- Observations of specific values e for a subset of (X) evidence variables E (subset of X)

- We want to compute
- JPD of query variables Y (a subset of X) given evidence e (posterior joint distribution)

- Step 1: Condition to get distribution $P(X|e)$
- Step 2: Marginalize to get distribution $P(Y|e)$

Inference by Enumeration: example

- Given P(X) as JPD below, and evidence e : “Wind=yes” • What is the probability that it is hot? I.e., P(Temperature=hot | Wind=yes)
- Step 1: condition to get distribution P(X|e)

Windy W	*Cloudy C*	*Temperature T*	*P(W, C, T)*
yes	no	hot	0.04
yes	no	mild	0.09
yes	no	cold	0.07
yes	yes	hot	0.01
yes	yes	mild	0.10
yes	yes	cold	0.12
no	no	hot	0.06
no	no	mild	0.11
no	no	cold	0.03

Inference by Enumeration: example

no	yes	hot	0.04
no	yes	mild	0.25
no	yes	cold	0.08

- Given P(X) as JPD below, and evidence e : “Wind=yes” • What is the probability that it is hot? I.e., P(Temperature=hot | Wind=yes)
- Step 1: condition to get distribution P(X|e)

$$P(C = c \wedge T = t | W = yes) = \frac{P(C = c \wedge T = t \wedge W = yes)}{P(W = yes)}$$

Inference by Enumeration: example

P(X|Windy e) W	Cloudy C	Temperature T	P(W, C, T)
yes	no	hot	0.04
yes	no	mild	0.09
yes	no	cold	0.07
yes	yes	hot	0.01
yes	yes	mild	0.10
yes	yes	cold	0.12
no	no	hot	0.06
no	no	mild	0.11
		cold	0.03
no	no		
no	yes	hot	0.04

Cloudy C	Temperature T	P(C, T| W=yes)
no	hot	
no	mild	
no	cold	
yes	hot	
yes	mild	
yes	cold	

Inference by Enumeration: example

no	yes	mild	0.25
no	yes	cold	0.08

Inference by Enumeration: example

- Given P(X) as JPD below, and evidence e : “Wind=yes” • What is the probability that it is hot? I.e., P(Temperature=hot | Wind=yes)
- Step 1: condition to get distribution P(X|e)

Windy *W*	*Cloudy* *C*	*Temperature* *T*	*P(W, C, T)*
yes	no	hot	0.04
yes	no	mild	0.09
yes	no	cold	0.07
yes	yes	hot	0.01
yes	yes	mild	0.10
yes	yes	cold	0.12
~~no~~	~~no~~	~~hot~~	~~0.06~~
~~no~~	~~no~~	~~mild~~	~~0.11~~
~~no~~	~~no~~	~~cold~~	~~0.03~~
~~no~~	~~yes~~	~~hot~~	~~0.04~~
~~no~~	~~yes~~	~~mild~~	~~0.25~~
~~no~~	~~yes~~	~~cold~~	~~0.08~~

$$P(C = c \wedge T = t | W = yes) = \frac{P(C = c \wedge T = t \wedge W = yes)}{P(W = yes)}$$

Inference by Enumeration: example

Cloudy C	*Temperature T*		*P(C W=y*
	hot	0.04/0.43 ≅ 0.10	
	mild	0.09/0.43 ≅ 0.21	
	cold	0.07/0.43 ≅ 0.16	
	hot	0.01/0.43 ≅ 0.02	
	mild	0.10/0.43 ≅ 0.23	
	cold	0.12/0.43 ≅ 0.28	

- Given P(X) as JPD below, and evidence e : “Wind=yes”

Inference by Enumeration: example

- What is the probability that it is hot? I.e., P(Temperature=hot | Wind=yes)
- Step 2: marginalize to get distribution P(Y|e)

Cloudy C	*Temperature T*	*P(C, T| W=yes)*
sunny	hot	0.10
sunny	mild	0.21
sunny	cold	0.16
cloudy	hot	0.02
cloudy	mild	0.23
cloudy	cold	0.28

Temperature T	*P(T| W=yes)*
hot	0.10+0.02 = 0.12
mild	0.21+0.23 = 0.44
cold	0.16+0.28 = 0.44

Problems of Inference by Enumeration

- If we have n variables, and d is the size of the largest domain
- What is the space complexity to store the joint distribution?
- We need to store the probability for each possible world • There are possible worlds, so the space complexity is
- How do we find the numbers for entries?
- Time complexity
- In the worse case, need to sum over all entries in the JPD
- We have some of our basic tools, but to gain computational efficiency we need to do more

- We will exploit (conditional) independence between variables
- But first, we will look at a neat application of conditioning

Learning Goals For Probability so Far

- Given a JPD
- Marginalize over specific variables
- Compute distributions over any subset of the variables
- Apply the formula to compute conditional probability P(h|e)
- Use inference by enumeration • to compute joint posterior probability distributions over any subset of variables given evidence

Marginalization and conditioning are crucial

They are core to reasoning under uncertainty

Be sure you understand them and be able to use them!

Lecture Overview

- Recap Lecture 16
- Joint Probability Distribution, Marginalization
- Conditioning
- Inference by Enumeration
- Bayes Rule, Chain Rule (time permitting)

Using conditional probability

- Often you have causal knowledge (from cause to evidence):
- For example
  - ✓P(symptom | disease)
  - ✓P(light is off | status of switches and switch positions)
  - ✓P(alarm | fire)
- In general: P(evidence e | hypothesis h)

- ... and you want to do evidential reasoning (from evidence to cause):
- For example
  - ✓P(disease | symptom)
  - ✓P(status of switches | light is off and switch positions)
  - ✓P(fire | alarm)
- In general: P(hypothesis h | evidence e)

Bayes Rule

- By definition, we know that :

$$P(h \mid e) = \frac{P(h \wedge e)}{P(e)} \qquad P(e \mid h) = \frac{P(e \wedge h)}{P(h)}$$

- We can rearrange terms to write

$$P(h \wedge e) = P(h \mid e) \times P(e) \qquad (1)$$

$$P(e \wedge h) = P(e \mid h) \times P(h) \qquad (2)$$

- But

$$P(h \wedge e) = P(e \wedge h) \qquad (3)$$

- From (1) (2) and (3) we can derive

- On average, the alarm rings once a year
  - $P(alarm) = ?$
- If there is a fire, the alarm will almost always ring
- On average, we have a fire every 10 years
- The fire alarm rings. What is the probability there is a fire?

Bayes Rule

$$P(h|e) = \frac{P(e|h)P(h)}{P(e)} \qquad (3)$$

Example for Bayes rule

- On average, the alarm rings once a year
  - $P(alarm) = 1/365$
- If there is a fire, the alarm will almost always ring
  - $P(alarm|fire) = 0.999$
- On average, we have a fire every 10 years
  - $P(fire) = 1/3650$
- The fire alarm rings. What is the probability there is a fire?
  - Take a few minutes to do the math!

Example for Bayes rule

Product Rule

- By definition, we know that :

$$P(f_2 \mid f_1) = \frac{P(f_2 \wedge f_1)}{P(f_1)}$$

- We can rewrite this to

$$P(f_2 \wedge f_1) = P(f_2 \mid f_1) \times P(f_1)$$

Theorem (Product Rule)

$$P(f_n \wedge \cdots \wedge f_{i+1} \wedge f_i \wedge \cdots \wedge f_1) = P(f_n \wedge \cdots \wedge f_{i+1} \mid f_i \wedge \cdots \wedge f_1) \times P(f_i \wedge \cdots \wedge f_1)$$

- In general

Chain Rule

- We know

$$P(f_2 \wedge f_1) = P(f_2|f_1) \times P(f_1)$$

- In general:

$$
\begin{aligned}
&P(f_n \wedge f_{n-1} \wedge \cdots \wedge f_1) \\
&\quad = P(f_n|f_{n-1} \wedge \cdots \wedge f_1) \times P(f_{n-1} \wedge \cdots \wedge f_1) \\
&\quad = P(f_n|f_{n-1} \wedge \cdots \wedge f_1) \times P(f_{n-1}|f_{n-2} \wedge \cdots \wedge f_1) \\
&\qquad\qquad \times P(f_{n-2} \wedge \cdots \wedge f_1) \\
&\quad = \ldots \\
&\quad = \prod_{i=1}^{n} P(f_i|f_{i-1} \wedge \cdots \wedge f_1)
\end{aligned}
$$

Chain Rule

- We know

$$P(f_2 \wedge f_1) = P(f_2 | f_1) \times P(f_1)$$

- In general:

$$
\begin{aligned}
&P(f_n \wedge f_{n-1} \wedge \cdots \wedge f_1) \\
&\quad = P(f_n | f_{n-1} \wedge \cdots \wedge f_1) \times P(f_{n-1} \wedge \cdots \wedge f_1) \\
&\quad = P(f_n | f_{n-1} \wedge \cdots \wedge f_1) \times P(f_{n-1} | f_{n-2} \wedge \cdots \wedge f_1) \\
&\qquad\qquad \times P(f_{n-2} \wedge \cdots \wedge f_1) \\
&\quad = \ldots \\
&\quad = \prod_{i=1}^{n} P(f_i | f_{i-1} \wedge \cdots \wedge f_1)
\end{aligned}
$$

Chain Rule

- We know

$$P(f_2 \wedge f_1) = P(f_2|f_1) \times P(f_1)$$

- In general:

$$\begin{aligned}
&P(f_n \wedge f_{n-1} \wedge \cdots \wedge f_1) \\
&\quad = P(f_n|f_{n-1} \wedge \cdots \wedge f_1) \times P(f_{n-1} \wedge \cdots \wedge f_1) \\
&\quad = P(f_n|f_{n-1} \wedge \cdots \wedge f_1) \times P(f_{n-1}|f_{n-2} \wedge \cdots \wedge f_1) \\
&\qquad\qquad \times P(f_{n-2} \wedge \cdots \wedge f_1) \\
&\quad = \ldots \\
&\quad = \prod_{i=1}^{n} P(f_i|f_{i-1} \wedge \cdots \wedge f_1)
\end{aligned}$$

Chain Rule

- We know

$$P(f_2 \wedge f_1) = P(f_2 | f_1) \times P(f_1)$$

- In general:

$$
\begin{aligned}
&P(f_n \wedge f_{n-1} \wedge \cdots \wedge f_1) \\
&\quad = P(f_n | f_{n-1} \wedge \cdots \wedge f_1) \times P(f_{n-1} \wedge \cdots \wedge f_1) \\
&\quad = P(f_n | f_{n-1} \wedge \cdots \wedge f_1) \times P(f_{n-1} | f_{n-2} \wedge \cdots \wedge f_1) \\
&\qquad\qquad \times P(f_{n-2} \wedge \cdots \wedge f_1) \\
&\quad = \ldots \\
&\quad = \prod_{i=1}^{n} P(f_i | f_{i-1} \wedge \cdots \wedge f_1)
\end{aligned}
$$

Chain Rule

- We know

$$P(f_2 \wedge f_1) = P(f_2|f_1) \times P(f_1)$$

- In general:

$$P(f_n \wedge f_{n-1} \wedge \cdots \wedge f_1)$$
$$= P(f_n|f_{n-1} \wedge \cdots \wedge f_1) \times P(f_{n-1} \wedge \cdots \wedge f_1)$$
$$= P(f_n|f_{n-1} \wedge \cdots \wedge f_1) \times P(f_{n-1}|f_{n-2} \wedge \cdots \wedge f_1) \times P(f_{n-2} \wedge \cdots \wedge f_1)$$
$$= \ldots$$
$$= \prod_{i=1}^{n} P(f_i|f_{i-1} \wedge \cdots \wedge f_1)$$

Chain Rule

Theorem (Chain Rule)

$$P(f_n \wedge \cdots \wedge f_1) = \prod_{i=1}^{n} P(fi|f_{i-1} \wedge \cdots \wedge f_1)$$

Bayes rule and Chain Rule

Theorem (Chain Rule)

$$P(f_n \wedge \cdots \wedge f_1) = \prod_{i=1}^{n} P(fi|f_{i-1} \wedge \cdots \wedge f_1)$$

E.g. $P(A,B,C,D) = P(A) \times P(B|A) \times P(C|A,B) \times P(D|A,B,C)$

Bayes rule and Chain Rule

Theorem (Chain Rule)

$$P(f_n \wedge \cdots \wedge f_1) = \prod_{i=1}^{n} P(fi|f_{i-1} \wedge \cdots \wedge f_1)$$

E.g. $P(A,B,C,D) = P(A) \times P(B|A) \times P(C|A,B) \times P(D|A,B,C)$

Why does the chain rule help us?

We will see how, under specific circumstances (variables independence), this rule helps gain compactness

- We can represent the JPD as a product of marginal distributions
- We can simplify some terms when the variables involved are marginally independent or conditionally independent