COMP474/6741

Uncertainty and Reasoning Under Uncertainty

Poole: Chap.6

Some slides are part of my Research Work.

Today

- Uncertainty and knowledge
- Reasoning with uncertainty
- Probabilistic Theory
- Bayes' Rule
- Bayesian reasoning
- Bayesian networks

Uncertainty and knowledge (1/2)

- All knowledge representation formalism and problem solving mechanisms that we have seen until now are based on the following assumptions:
  - All facts can be evaluated to true or false
  - All the facts needed to solve the problem are available
  - The decision mechanism, if applied, always obtains a conclusion that is true
- This only happens in a perfect world

Well our world
is not that perfect !!!

Motivation

- How do we represent and reason about non-factual knowledge?
  - It *might* rain tonight
  - If you have red spots on your face, you *might* have the measles
  - This e-mail is *most likely* spam
  - I can't read this character, but it *looks* like a "B"
  - These 2 pictures are *very likely* of the same person
  - …

Uncertainty and knowledge (2/2)

- In practice we make decision without knowing all the facts and with incomplete/heuristic decision mechanisms
- Incomplete knowledge
  - It is impossible to include all the facts that represent a problem
  - Not all the decision mechanisms to solve problems are known
- Uncertain/Imprecise Knowledge
  - There is not absolute confidence on the veracity of the facts
  - Decision mechanisms are heuristic, the conclusion is not always true
  - Decision mechanisms are used on uncertain data, the conclusion is also uncertain

Today

- Uncertainty and knowledge
- Reasoning with uncertainty
- Probabilistic Theory
- Bayes' Rule
- Bayesian reasoning
- Bayesian networks

Reasoning with uncertainty

AI has developed different formalisms that allow to reason under uncertainty and incomplete knowledge, we will talk about two of them:

- Probabilistic models (Bayesian reasoning/networks)
- Possibilistic models (**Fuzzy logic**)

Next class

Today

- Uncertainty and knowledge
- Reasoning with uncertainty
- Probabilistic Theory
- Bayes' Rule
- Bayesian reasoning
- Bayesian networks

Probabilistic Models

- Probabilistic models are based on probability theory
- Probabilities are used to model our belief on the probability distribution of the values of a fact
- Each fact has associated a probability distribution (the model) that is used for reasoning
- The probability of a fact could be modified by our belief on the values of other related facts

Probabilistic Decisions- example

Will you take your umbrella tomorrow?

It never rains in Florida (Yes: 10%)

The weather forecast is cloudy skies

May be I will, may be I won't (Yes: 50 %)

Today the streets are wet

Better if I take my umbrella (Yes: 95 %)

Probability Theory

- The basic element of probability theory is the **random variable**
- A random variable has a domain of values, we have **boolean**, **discrete** o **continuous** random variables
- A **logic proposition** is defined as a formula in propositional or predicate calculus
- A logic proposition has associated a random variable that represents the degree of belief on its values
- A random variable has associated a **probability distribution**
- We will only talk about discrete random variables
- The union of random variables can be described by the **joint probability distribution**

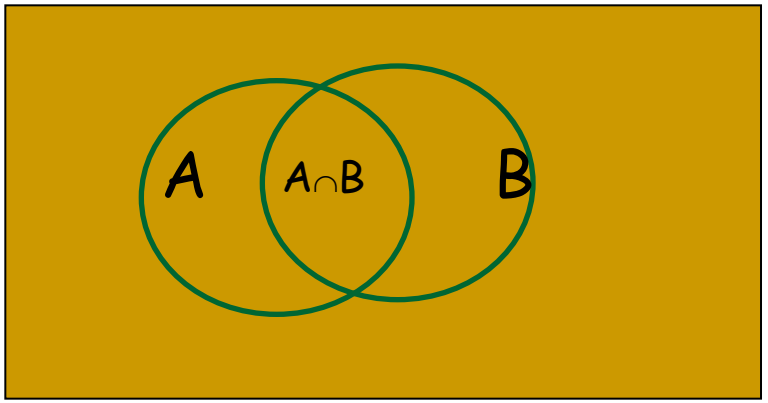
Probability Theory

- Let:
  - A random variable X (eg. throwing a die)
  - $A_1,\ldots,A_n$ be events in $X = \{A_1,\ldots,A_n\}$
- $P(A)$ gives the probability that A will take place
- P is a probability function:
  - $0 \leq P(A) \leq 1$
  - $P(A) = 0 \Rightarrow$ the event A will never take place
  - $P(A) = 1 \Rightarrow$ the event A must take place
  - $\sum_i P(A_i) = 1 \Rightarrow$ one of the events A_i will take place
  - $P(A) + P(\neg A) = 1$

Joint Probability

- The intersection $A_1 \cap \ldots \cap A_n$ is an event that takes place
  - if all the events $A_1,\ldots,A_n$ take place

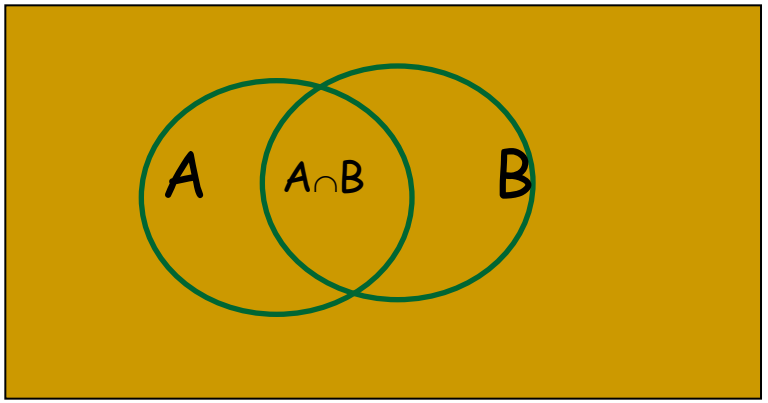
- Joint probability of A and B:
  - $P(A \cap B) = P(A,B)$



Sum Rule

- The union $A_1 \cup \ldots \cup A_n$ is an event that takes place
  - if at least one of the events $A_1,\ldots,A_n$ takes place

- Sum rule:
  - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$



Conditional Probability

The **a priori probability** of a proposition ($P(a)$) is defined as the degree of belief on the proposition if we have no other information

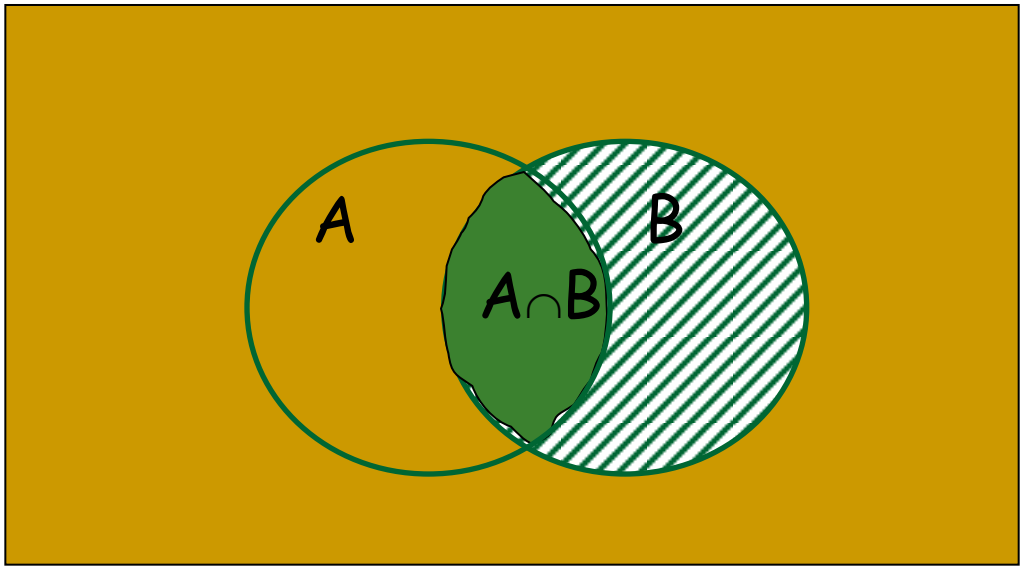
- Prior (or unconditional) probability
  - Probability of an event before any evidence is obtained
  - P(A) = 0.1 *P(rain today) = 0.1*
  - i.e. Your belief about A given that you have no evidence

The **a posteriori or conditional probability** ($P(a|b)$) is defined as the degree of belief on a propositions after observing other proposition associated to it

- Posterior (or conditional) probability
  - Probability of an event given that you know that B is true (B = some evidence)
  - P(A|B) = 0.8 *P(rain today| cloudy) = 0.8*
  - i.e. Your belief about A given that you know B

Conditional Probability (con't)

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A, B)}{P(B)}$$



Chain Rule

$$P(A \mid B) = \frac{P(A,B)}{P(B)} \text{ so } P(A,B) = P(A \mid B) \times P(B) \qquad P(B \mid A) = \frac{P(A,B)}{P(A)} \text{ so } P(A,B) = P(B \mid A) \times P(A)$$

- With 3 events, the probability that A, B and C occur is:
  - The probability that A occurs
  - Times, the probability that B occurs, assuming that A occurred
  - Times, the probability that C occurs, assuming that A and B have occurred

- With multiple events, we can generalize to the Chain rule:

$$P(A_1, A_2, A_3, A_4, \ldots, A_n)$$
$$= P\left(\cap A_i\right)$$
$$= P(A_1) \times P(A_2 \mid A_1) \times P(A_3 \mid A_1, A_2) \times \ldots \times P(A_n \mid A_1, A_2, A_3, \ldots, A_{n-1})$$

So what?

- we can do probabilistic inference
  - i.e. infer new knowledge from observed evidence

Example 1

- Joint probability distribution:

evidence

hypothesis

	Toothache	¬Toothache
Cavity	0.04	0.06
~Cavity	0.01	0.89

$$P(H \mid E) = \frac{P(H \cap E)}{P(E)}$$

$$P(\text{cavity} \mid \text{toothache}) = \frac{P(\text{cavity} \cap \text{toothache})}{P(\text{toothache})} = \frac{0.04}{0.04 + 0.01} = 0.8$$

Getting the Probabilities

- in most applications, you just count from a set of observations

$$P(A) = \frac{count\_of\_A}{count\_of\_all\_events}$$

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{count\_of\_A\_and\_B\_together}{count\_of\_all\_B}$$

- Subjectivist position:
  - if data from observations are not available, ask an expert (ex. a doctor)

Combining Evidence

- Suppose, we know that
  - P(Cavity | Toothache) = 0.12
  - P(Cavity | Young) = 0.18
- A patient complains about Toothache and is Young…
  - what is P(Cavity | Toothache $\cap$ Young) ?

Combining Evidence

	Toothache		~Toothache	
	Young	~ Young	Young	~ Young
Cavity	0.108	0.012	0.072	0.008
~Cavity	0.016	0.064	0.144	0.576

- But how do we get the data ?
- In reality, we may have dozens, hundreds of variables
- We cannot have a table with the probability of all possible combinations of variables
  - Ex. with 16 variables, we would need 2^{16} entries

Probabilistic Inference Problems

- To compute these inferences it is required to store and search the joint probability distribution of all the propositions
- Assuming binary propositions the cost in space and time is $O(2^n)$ being n the number of propositions
- For any real problem this is impracticable
- It is needed a mechanism that allows to reduce the inference computational cost

Probabilistic independence

- Usually not all the propositions from a problem are related to each other
- They have the property of **probabilistic independence**
  - Two events A and B are independent:
    if the occurrence of one of them does not influence the occurrence of the other
    i.e. A is independent of B if P(A) = P(A|B)
- This means that some propositions do not have influence over others and their probabilities can be expressed as:

$$P(X|Y) = P(X); \quad P(Y|X) = P(Y); \quad \boxed{P(X,Y) = P(X)P(Y)}$$

- Because of this property the joint probability distribution can be expressed in a more compact way, reducing the computational complexity

Independent Events

- So we often assume independence of events
  - $P(A,B) = P(A) \times P(B)$
- In reality:
  - some variables are independent…
    - ex: living in Montreal & tossing a coin
    - P(Montreal, head) = P(Montreal) * P(head)
    - probability of 2 heads in a row: P(head, head) = 1/2 * 1/2 = 1/4
  - some variables are not independent…
    - ex: living in Montreal & wearing boots
    - P(Montreal, boots) ≠ P(Montreal) * P(boots)

Conditional Independent Events

- Two events A and B are conditionally independent given C:
  - once we *know* that C is true, then any evidence about B cannot change our belief about A
  - $P(A, B \mid C) = P(A \mid C) \times P(B \mid C)$.

Today

- Uncertainty and knowledge
- Reasoning with uncertainty
- Probabilistic Theory
- Bayes' Rule 

- Bayesian reasoning
- Bayesian networks

Bayes rule

- The product rule can been expressed as:

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

- This derives to the **Bayes rule**

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

- This rule and the probabilistic independence property is the basis of probabilistic reasoning and it will allow us to propagate the probabilities of propositions to others

So?

- We typically want to know: P(Hypothesis | Evidence)
  - ex: P(Disease | Symptoms)… P(meningitis | spots)
  - ex: P(Cause | Side Effect)… P(misaligned brakesb| squeaking wheels)
- But P(Hypothesis| Evidence) is hard to gather
  - ex: out of all people who have red spots… how many have meningitis?
- However P(Evidence | Hypothesis) is easier to gather
  - ex: out of all people who have the meningitis … how many have red spots?
- So

$$P(\text{Hypothesis} \mid \text{Evidence}) = \frac{P(\text{Evidence} \mid \text{Hypothesis}) \times P(\text{Hypothesis})}{P(\text{Evidence})}$$

with Y= H, and X=E in $P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

Combining Evidence

- With Bayes rule and independence assumption:

$$P(\text{Cavity} \mid \text{Toothache} \cap \text{Young})$$

$$= \frac{P(\text{Toothache} \cap \text{Young} \mid \text{Cavity}) \times P(\text{Cavity})}{P(\text{Toothache} \cap \text{Young})}$$

$$= \frac{P(\text{Toothache} \cap \text{Young} \mid \text{Cavity}) \times P(\text{Cavity})}{P(\text{Toothache}) \times P(\text{Young})}$$

E_1:toothache
E_2:young
H: Cavity

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

With conditional independence assumption:

- i.e. Toothache & Young are independent given the absence or presence of cavity

$$P(\text{Toothache} \cap \text{Young} \mid \text{Cavity}) = P(\text{Toothache} \mid \text{Cavity}) \times P(\text{Young} \mid \text{Cavity})$$

Now we have decomposed the joint probability distribution into much smaller pieces…

$$P(\text{Cavity} \mid \text{Toothache} \cap \text{Young})$$

$$= \frac{P(\text{Toothache} \mid \text{Cavity}) \times P(\text{Toothache} \mid \text{Young}) P(\text{Cavity})}{P(\text{Toothache}) \times P(\text{Young})}$$

Bayes rule + independence

- Assuming that we can exhaustively estimate all the probabilities that are related to the values of the variable Y the Bayes rules can be rewritten as:

$$P(Y|X) = \alpha P(X|Y)P(Y)$$

- Assuming conditional independence between two variables we have that:

$$P(X, Y|Z) = P(X|Z)P(Y|Z)$$

- so:

$$P(Z|X, Y) = \alpha P(X, Y|Z)P(Z) = \alpha P(X|Z)P(Y|Z)P(Z)$$

- This means that we can decompose the computation of joint probabilities on independent computations

Example 2

E: spots
H: meningitis

P(spots | meningitis) = 0.4
P(meningitis) = 0.00003
P(spots) = 0.05

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

$$P(\text{meningitis} \mid \text{spots}) = \frac{P(\text{spots} \mid \text{meningitis}) \times P(\text{meningitis})}{P(\text{spots})}$$

$$= \frac{0.4 \times 0.00003}{0.05} = 0.00024$$

- If you have spots… you are more likely to have meningitis

Example 3

A patient takes a lab test and the result comes back positive for cancer. The test returns a correct positive result in only 98% of the cases in which the cancer is actually present, and a correct negative result in only 97% of the cases in which the cancer is not present. Furthermore, only 0.08% of the entire population have this cancer.

1. What is the probability that this patient has cancer?
2. What is the probability that he does not have cancer?
3. What is the diagnosis?

Example 3

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

P(cancer) = 0.08%
P(pos | cancer) = 98%
P(neg | not cancer) = 97%

P(not cancer) = 99.92%
P(neg | cancer) = 2%
P(pos | not cancer) = 3%

$$P(\text{cancer} \mid \text{pos}) = \frac{P(\text{pos} \mid \text{cancer})\, P(\text{cancer})}{P(\text{pos})} = \frac{0.98 \times 0.0008}{P(\text{pos})}$$

$$P(\text{pos}) = P(\text{pos} \mid \text{cancer}) \times P(\text{cancer}) + P(\text{pos} \mid \text{not cancer}) \times P(\text{not cancer})$$

$$P(\text{pos}) = 0.98 \times 0.0008 + 0.03 \times 0.9992 = 0.03076$$

E: positive
H: cancer

$$P(\text{cancer} \mid \text{pos}) = \frac{0.98 \times 0.0008}{0.03076} = 0.0255$$

Diagnosis: he does not have cancer

E: positive
H: not cancer

$$P(\text{not cancer} \mid \text{pos}) = \frac{P(\text{pos} \mid \text{not cancer}) \times P(\text{not cancer})}{P(\text{pos})} = \frac{0.03 \times 0.99992}{0.03076} = 0.9745$$

Today

- Uncertainty and knowledge
- Reasoning with uncertainty
- Probabilistic Theory
- Bayes' Rule
- Bayesian reasoning
- Bayesian networks

Bayes' Reasoning

- Out of n hypothesis…
  - If we want to find the most probable H_i given the evidence E
- So we chose the H_i with the largest $P(H_i|E)$

$$H_{NB} = \underset{H_i}{\text{argmax}}\ P(H_i \mid E) = \underset{H_i}{\text{argmax}}\ \frac{P(E \mid H_i) \times P(H_i)}{P(E)}$$

- P(E)
  - is the same for all possible Hs (and is hard to gather anyways)
  - so we can drop it
- So Bayesian reasoning:

$$H_{NB} = \underset{H_i}{\text{argmax}}\ P(E \mid H_i) \times P(H_i)$$

Example 4

- Predict the weather tomorow based on tonight's sunset...
  - H_1: *weather is nice tomorrow* $P(H_1) = 0.2$
  - H_2: *weather is bad tomorrow* $P(H_2) = 0.5$
  - H_3: *weather is mixed tomorrow* $P(H_3) = 0.3$

  - E_1: today, there's a *beautiful sunset*
  - E_2: today, there's a *average sunset*
  - E_3: today, there's *no sunset*

$P(E_x \mid H_i)$	E_1	E_2	E_3
H_1	0.7	0.2	0.1
H_2	0.3	0.3	0.4
H_3	0.4	0.4	0.2

Example 4

- Observation: average sunset (E_2)
- Question: how will be the weather tomorrow?
  - $P(H_i \mid E_2)$?
  - predict the weather that minimizes your error !
  - select H_i such that $P(H_i \mid E_2)$ is the greatest

$$P(H_i \mid E_2) = \frac{P(H_i) \times P(E_2 \mid H_i)}{P(E_2)}$$

$$P(E_2) = P(H_1) \times P(E_2 \mid H_1) + P(H_2) \times P(E_2 \mid H_2) + P(H_3) \times P(E_2 \mid H_3)$$
$$= .2x.2 + .5x.3 + .3x.4 = .04 + .15 + .12 = 0.31$$

Example 4

$$P(H_1 \mid E_2) = \frac{P(H_1) \times P(E_2 \mid H_1)}{P(E_2)} = \frac{.2 \times .2}{.31} = .129$$

$$P(H_2 \mid E_2) = \frac{P(H_2) \times P(E_2 \mid H_2)}{P(E_2)} = \frac{.5 \times .3}{.31} = .484$$

$$P(H_3 \mid E_2) = \frac{P(H_3) \times P(E_2 \mid H_3)}{P(E_2)} = \frac{.3 \times .4}{.31} = .387$$

$$H_{NB} = \underset{H_i}{\operatorname{argmax}} \; P(E \mid H_i) \times P(H_i)$$

$\Rightarrow H_2$ is the most likely hypothesis, given the evidence

$P(H_2 \mid E_2)$ is the highest

Tomorrow the weather will be bad

Representing the Evidence

- The evidence is typically represented by many attributes/features
  - beautiful sunset? clouds? temperature? summer?, …
- so often represented as a feature/attribute vector

Evidence	sunset	clouds	temp.	summer	weather tomorrow
e1	beautiful	no	high	yes	*Nice*

$e1 = \langle a_1, \ldots, a_n \rangle$

e1 = <sunset:beautiful, clouds:no, temp:high, summer:yes>

Example 5

evidence

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

Example 5

- Goal: Given a new instance $X=\langle a_1,\ldots, a_n\rangle$, classify as Yes/No

$$\arg\max_{H_i} P(H_i)xP(X \mid H_i) = \arg\max_{H_i} P(H_i)xP(< a_1,\ldots, a_n > \mid H_i)$$

- Naïve Bayes: Assumes that the attributes/features are conditionally independent

$$\arg\max_{H_i} P(H_i) \times P(< a_1,\ldots, a_n > \mid H_i)$$

$$= \arg\max_{H_i} P(H_i) \times P(a_1 \mid H_i) x\, P(a_2 \mid H_i) x \ldots x\, P(a_n \mid H_i)$$

$$= \arg\max_{H_i} P(H_i) \times \prod_{j=1}^{n} P(a_j \mid H_i)$$

Example 5

- Goal: Given a new instance X=<a_1,…, a_n>, classify as Yes/No

$$\underset{H_i}{\operatorname{argmax}}\, P(H_i)P(X \mid H_i) = \underset{H_i}{\operatorname{argmax}}\, P(H_i)\prod_{j=1}^{n} P(a_j \mid H_i)$$

- 1st estimate the probabilities from the training examples:
  - For each target value (hypothesis) H_i

    estimate $P(H_i)$
  - For each attribute value a_j of each instance (evidence)

    estimate $P(a_j \mid H_i)$

Example 5

1. TRAIN:
   - compute the probabilities from the training set

$$P(\text{PlayTennis} = \text{yes}) = 9/14 = 0.64$$
$$P(\text{PlayTennis} = \text{no}) = 5/14 = 0.36$$

} prior probabilities $P(H_i)$

$$P(\text{Out} = \text{sunny} \mid \text{PlayTennis} = \text{yes}) = 2/9 = 0.22$$
$$P(\text{Out} = \text{sunny} \mid \text{PlayTennis} = \text{no}) = 3/5 = 0.60$$
$$P(\text{Out} = \text{rain} \mid \text{PlayTennis} = \text{yes}) = 3/9 = 0.33$$
$$P(\text{Out} = \text{rain} \mid \text{PlayTennis} = \text{no}) = 2/5 = 0.4$$
$$\ldots$$
$$P(\text{Wind} = \text{strong} \mid \text{PlayTennis} = \text{yes}) = 3/9 = 0.33$$
$$P(\text{Wind} = \text{strong} \mid \text{PlayTennis} = \text{no}) = 3/5 = 0.60$$

} conditional probabilities $P(a_j \mid H_i)$

Example 6

2. TEST:

classify the new case: X=(Outlook: Sunny, Temp: Cool, Hum: High, Wind: Strong)

$$H_{NB} = \underset{H_i \in [yes,no]}{\operatorname{argmax}} \; P(H_i) \times P(X \mid H_i)$$

$$= \underset{H_i \in [yes,no]}{\operatorname{argmax}} \; P(H_i) \times \prod_j P(a_j \mid H_i)$$

$$= \underset{H_i \in [yes,no]}{\operatorname{argmax}} \; P(H_i) \times P(Outlook = sunny \mid H_i) \times P(Temp = cool \mid H_i)$$

$$\times P(Humidity = high \mid H_i) \times P(Wind = strong \mid H_i)$$

$$1 - P(yes) \times P(sunny \mid yes) \times P(cool \mid yes) \times P(high \mid yes) \times P(strong \mid yes) = 0.0053$$

$$2 - P(no) \times P(sunny \mid no) \times P(cool \mid no) \times P(high \mid no) \times P(strong \mid no) = 0.0206$$

$$\Rightarrow answer : PlayTennis(X) = no$$

Application of Bayesian Reasoning

- Categorization: P(Category | Features of Object)
  - Diagnostic systems: P(Disease | Symptoms)
  - Text classification: P(sports_news | text)
  - Character recognition: P(character | bitmap)
  - Speech recognition: P(words | acoustic signal)
  - Image processing: P(face_person | image features)
  - Spam filter: P(spam_message | words in e-mail)
  - …

Today

- Uncertainty and knowledge
- Reasoning with uncertainty
- Probabilistic Theory
- Bayes' Rule
- Bayesian reasoning
- Bayesian networks

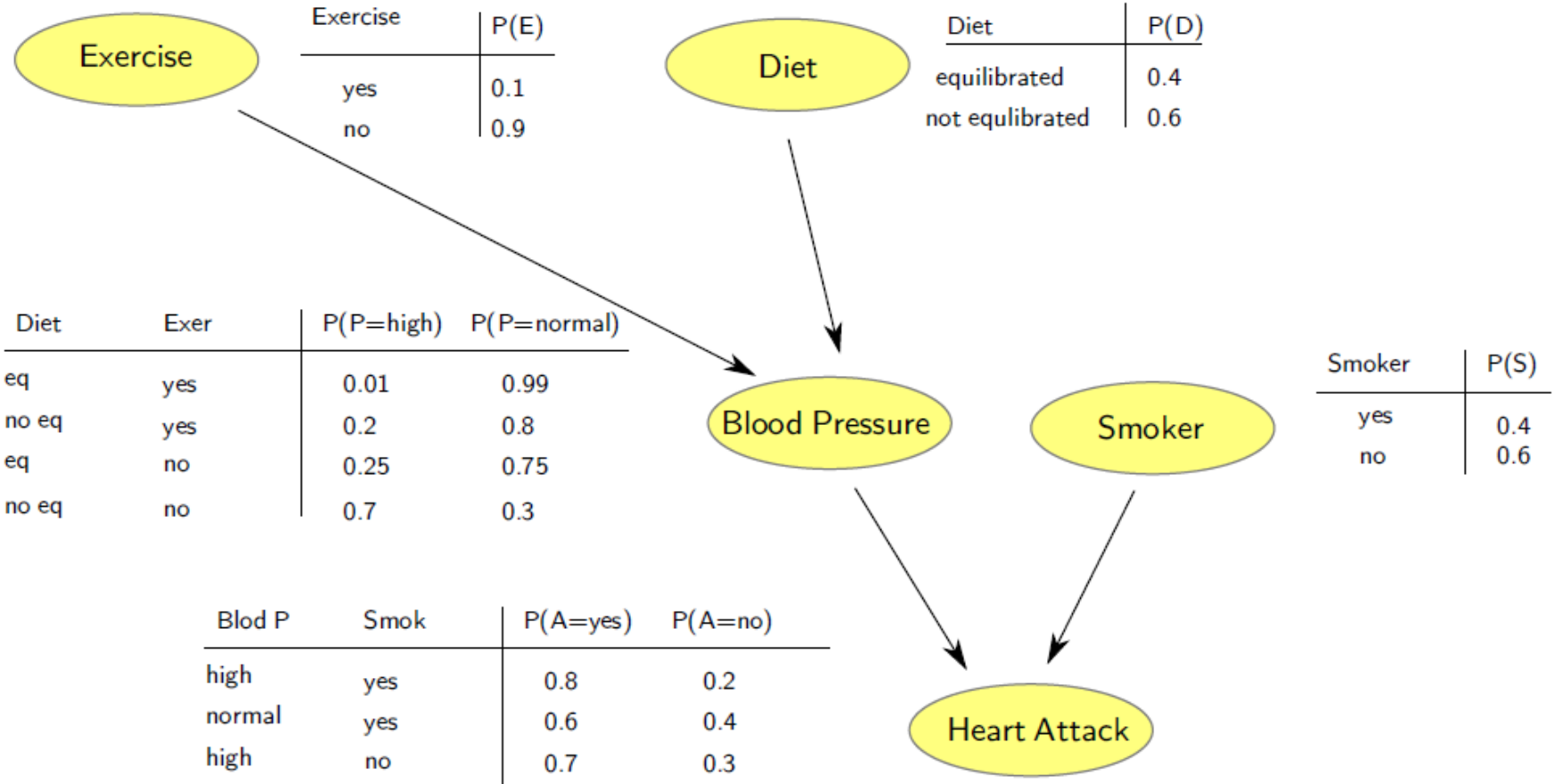
Bayesian networks

- If we know the independence relations among variables we can simplify the computation of their combination and also their representation
- **Bayesian networks** is a formalism for representing these independence relations
- A bayesian network is an **acyclic directed graph** that stores probabilistic information in its nodes that represents the influence of the ancestors of a node ($P(X_i|parent(X_i))$) on its probability distribution
- The intuitive meaning of an edge between two nodes X and Y is that the variable X has influence over Y probability
- The probabilities represented by the network describe the joint probability distribution of all the variables

Bayesian networks: example 1/2

- We want to find out the probability of having a heart attack of a person
- We know that this probability is determined by four variables: the practice of exercise, adequate diet, blood pressure and smoking
- We also know that blood pressure depends directly of exercising and diet, that are independent variables, and smoking is independent of the rest of variables
- This knowledge allows us to create a dependency network among the variables
- Our knowledge about the domain allows us to estimate the probabilities of each independent variable and their influence to the dependent variables

Bayesian networks: example 2/2



Exercise	P(E)
yes	0.1
no	0.9

Diet	P(D)
equilibrated	0.4
not equlibrated	0.6

Diet	Exer	P(P=high)	P(P=normal)
eq	yes	0.01	0.99
no eq	yes	0.2	0.8
eq	no	0.25	0.75
no eq	no	0.7	0.3

Smoker	P(S)
yes	0.4
no	0.6

Blod P	Smok	P(A=yes)	P(A=no)
high	yes	0.8	0.2
normal	yes	0.6	0.4
high	no	0.7	0.3
normal	no	0.3	0.7

Bayesian networks-Joint probability distribution

- Each node in the network has the probability distribution of the node given its parents
- This allows to factorize the joint probability distribution transforming its expression to a product of independent conditional probabilities

$$P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | parents(x_i))$$

Bayesian networks-Joint probability distribution - example

$$P(Attack = yes \wedge Pressure = high \wedge Smoker = yes \wedge Exercise = yes \wedge Diet = equil)$$
$$=$$
$$P(Attack = yes | Pressure = high, Smoker = yes)$$
$$P(Pressure = high | Exercise = yes, Diet = equil)$$
$$P(Smoker = yes)P(Exercise = yes)P(Diet = equil)$$
$$= 0{,}8 \times 0{,}01 \times 0{,}4 \times 0{,}1 \times 0{,}4$$
$$= 0{,}000128$$

Design of Bayesian networks

- The properties of bayesian networks give some ideas about how to build them. Considering that (by the product rule):

$$P(x_1, x_2, \ldots, x_n) = P(x_n|x_{n-1}, \ldots, x_1)P(x_{n-1}, \ldots, x_1)$$

- Iterating the process we have that:

$$\begin{aligned} P(x_1, \ldots, x_n) &= P(x_n|x_{n-1}, \ldots, x_1)P(x_{n-1}|x_{n-2}, \ldots, x_1) \\ &\quad \cdots P(x_2|x_1)P(x_1) \\ &= \prod_{i=1}^{n} P(x_i|x_{i-1}, \ldots, x_1) \end{aligned}$$

- This is the **chain rule**

Design of Bayesian networks

- Given that properties we can say that if $parents(X_i) \subseteq \{X_{i-1}, \ldots, X_1\}$, then:

$$P(X_i|X_{i-1}, \ldots, X_1) = P(X_i|parents(X_i))$$

- This means that a bayesian network is a correct representation of a domain only if each node is conditional independent of its ancestors given its parents
- The parents of a variable X_i must be the variables $X_1, \ldots X_{i-1}$ that have a direct influence over X_i

Cost of the representation

- The cost of representing a joint probability distribution of n binary variables is $O(2^n)$
- Bayesian networks allow a more compact representation because of the factorization of the joint probability distribution
- Assuming that each node has no more that k parents ($k \ll n$), a node needs 2^k to represent the influence of its parents, then the space necessary is $O(n2^k)$.
- For example, with 10 variables and assuming 3 parents we have 80 instead of 1024, with 100 variables and assuming 5 parents we have 3200 instead of approximately 10^{30}

Inference in Bayesian networks

- The goal of probabilistic inference is to compute the a posteriori probability distribution of a set of variables given the observation of an event (observed values for a subset of variables)
- X is the variable we want to know its probability distribution
- $\mathbf{E}$ is the set of variables which their value is known $E_1, \ldots, E_n$
- $\mathbf{Y}$ is the set of variables not observed $Y_1, \ldots, Y_n$ (hidden variables)
- $\mathbf{X} = \{X\} \cup \mathbf{E} \cup \mathbf{Y}$ is the set of all variables
- We want to compute $P(X|\mathbf{e})$ (e = observed values for E)

Exact inference

- **Inference by enumeration**: Any conditional probability can be calculated as the sum of all possible cases from the joint probability distribution

$$P(X|\mathbf{e}) = \alpha P(X, \mathbf{e}) = \alpha \sum_{y} P(X, \mathbf{e}, \mathbf{y})$$

- The bayesian network allows to factorize the probability distribution and obtain an expression that can be evaluated in a simpler way
- For instance we can compute using the bayesian network from the example the probability of being a smoker if we have had a heart attack and do not exercise

$$P(Smoker|Attack = yes, Exercise = no)$$

Exact inference-example

The joint probability distribution of the network is:

$$P(E, D, S, P, A) = P(A|P, S)P(S)P(P|E, D)P(E)P(D)$$

We have to compute $P(S|A = yes, E = no)$, so

$$P(S|A = y, E = n) = \alpha P(S, A = y, E = n)$$
$$= \alpha \sum_{D \in \{e, \neg e\}} \sum_{P \in \{h, n\}} P(E = n, D, P, S, A = y)$$
$$= \alpha P(E = n)P(S) \sum_{D \in \{e, \neg e\}} P(D) \sum_{P \in \{h, n\}} P(P|E = n, D)P(A = y|P, S)$$

Exact inference-example

If we enumerate all the possibilities and sum them up using the joint probability distribution we have that:

$$
\begin{aligned}
& P(Smoker|Attack = yes, Exercise = no) \\
& = \alpha\langle 0{,}9 \cdot 0{,}4 \cdot (0{,}4 \cdot (0{,}25 \cdot 0{,}8 + 0{,}75 \cdot 0{,}6) + 0{,}6 \cdot (0{,}7 \cdot 0{,}8 + 0{,}3 \cdot 0{,}6)) \\
& \qquad 0{,}9 \cdot 0{,}6 \cdot (0{,}4 \cdot (0{,}25 \cdot 0{,}7 + 0{,}75 \cdot 0{,}3) + 0{,}6 \cdot (0{,}7 \cdot 0{,}7 + 0{,}3 \cdot 0{,}3)\rangle \\
& = \alpha\langle 0{,}253, 0{,}274\rangle \\
& = \langle 0{,}48, 0{,}52\rangle
\end{aligned}
$$

Today

- Uncertainty and knowledge
- Reasoning with uncertainty
- Probabilistic Theory
- Bayes' Rule
- Bayesian reasoning
- Bayesian networks