Probabilistic Reasoning and Bayesian Network

Ning Xiong

Madardalen University

1

Probabilistics versus Fuzzy

- Probability theory aims at stochastic uncertainty, where the occurrences of events are uncertain
  - example: tossing a coin
- Fuzzy theory deals subjective uncertainty with vague, ambiguous concepts and rough, inexact information
  - examples : low price, tall people, young age

This lecture focuses on probabilistic reasoning to address stochastic uncertainty

2

Agenda

- Basic about probability theory
- Bayes theorem (update belief based on evidence)
- Selection of most probable hypothesis
- Bayesian networks

3

Basic Properties of Probability

1. Probabilities must lie between 0 and 1, i.e. for any occurance A, we have $0 \le P(A) \le 1$

2 . Sum of probabilities must equal one, i.e. the probabilities of a set of exhaustive and mutually exclusive events must be summed to unity

$$P(A_1)+P(A_2)+.....+P(A_n)=1$$

4

Disjunction

The probability of the occurrence of A or B is given as

$$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$

Especially when A and B are mutually exclusive, it becomes

$$P(A \vee B) = P(A) + P(B)$$

5

Production Rule

The probability of the occurrence of both A and B is given by

$$\begin{aligned} P(A \wedge B) &= P(A|B)P(B) \\ &= P(B|A)P(A) \end{aligned}$$

6

Conditional Independence

Two events A and B are conditionally independent iff

$$P(A|B)=P(A)$$

If two events A and B are conditionally indenpent, we have

$$P(A \wedge B) = P(A|B)P(B)=P(A)P(B)$$

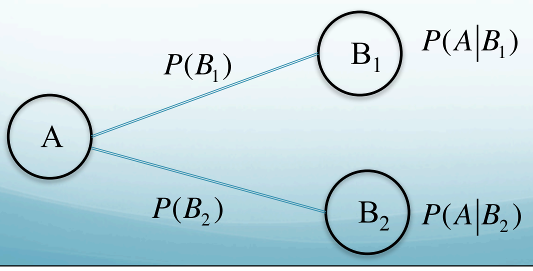
7

Total Probability of an Event

Let B_1, B_2, …., B_n be a set of exhaustive and mutually exclusive situations, the probability of occurrence of A can be calculated as:

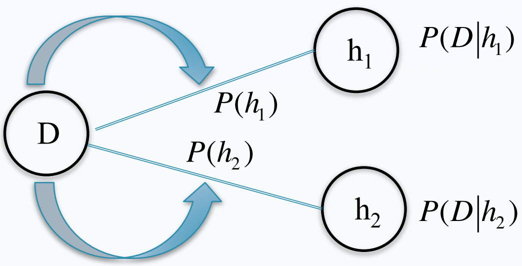
$$P(A)=\sum_{i=1}^{n} P(A|B_i)P(B_i)$$

Weighted averaging of probabilities in different situations



8

Bayes Theorem



Based on the evidence or observation D, we can revise our prior belief (probabilities on hypotheses) to the a posteriori probability P(h|D)

$$P(h \mid D) = P(h)P(D \mid h) / P(D)$$

Bayes Theorem

Let h_i be a hypothesis about something in the real world, the probability of h_i being true given the occurrence of event D is given by

$$P(h_i|D) = \frac{P(D|h_i)P(h_i)}{P(D)} = \frac{P(D|h_i)P(h_i)}{\sum_j P(D|h_j)P(h_j)}$$

P(D) : prior probability of the event D (*evidence*)

P(h) : prior probability of the hypothesis h

P(h|D) : posterior probability of the hypothesis given the event D

P(D|h) : probability of the event D given the hypothesis h, likelihood of D given h

Example (Bayes Theorem)

Consider a target detection problem in which there is a need to locate an important target. In terms of intelligence analysis, the target is possibly located in a certain area. The probability for that is initially estimated to be 60%. Then a sensor (say radar) is applied to scan that area. However, this sensor is imperfect and can occasionally give wrong alarm or miss something. The measurement properties are characterised by the **probability of false alarm** and the **probability of missed detection**, which are 4% and 2% respectively. What will be our belief after the sensor reports alarm? What will be our belief if the sensor "sees" no target?

11

Example (Bayes Theorem)

P(target) = 60%   P(no target) = 40%

P_{FA} = P(alarm | no target) = 4%
P_{MD} = P(no alarm | target) = 2%

$$P(target \mid alarm) = \frac{P(alarm \mid target)P(target)}{P(alarm \mid target)P(target) + P(alarm \mid no\_target)P(no\_target)}$$

$$= \frac{0.98 \cdot 0.6}{0.98 \cdot 0.6 + 0.04 \cdot 0.4} = 0.9735$$

$$P(target \mid no\_alarm) = \frac{P(no\_alarm \mid target)P(target)}{P(no\_alarm \mid target)P(target) + P(no\_alarm \mid no\_target)P(no\_target)}$$

$$= \frac{0.02 \cdot 0.6}{0.02 \cdot 0.6 + 0.96 \cdot 0.4} = 0.0303$$

12

Hypothesis selection

- Given the observation data D we generally want decide most probable hypothesis according to posterior probability

$$P(h|D) = P(D|h)\,P(h) / P(D)$$

- *Maximum a posteriori* hypothesis h_{MAP}

$$h_{MAP} = \mathrm{argmax}_{h \in H}\, P(h|D)$$

$$= \mathrm{argmax}_{h \in H}\, P(D|h)\,P(h) / P(D)$$

$$= \mathrm{argmax}_{h \in H}\, P(D|h)\,P(h)$$

- If the prior probabilities of hypotheses are equal $P(h_i)=P(h_j)$, then we can choose the *maximum likelihood* (ML) hypothesis

$$h_{ML} = \mathrm{argmax}_{h \in H}\, P(D|h)$$

13

Hypothesis Selection Example

A lab makes tests on patients for diagnosis of a certain disease. The test returns a positive (⊕) result in 98% of the cases in which the disease is present and negative (∅) result in 97% of the cases in which the disease is not present. Furthermore, 0.8% of the entire population have the disease.

$$P(\text{disease}) = 0.008,\ P(\neg\ \text{disease})=0.992$$

$$P(\oplus|\text{disease})=0.98,\ P(\varnothing\ |\text{disease})=0.02$$

$$P(\oplus|\neg\text{disease})=0.03,\ P(\varnothing\ |\neg\text{disease})=0.97$$

Given a patient with positive result, what to diagnoze?

$$P(\oplus|\text{disease})P(\text{disease})=0.98*0.008=0.0078$$

$$P(\oplus|\neg\ \text{disease})\ P(\neg\ \text{disease})\ =0.03*0.992=0.0298$$

$$h_{MAP}= \neg\ \text{disease}$$

14

Naive Bayes Classifier

- Naïve Bayes Classifier is applied to learning tasks where each instance x is described by a set of (discrete) attribute values and **the target of x takes on a value from some finite set V (classification).** A set of training examples is provided. Now given a new instance described by a tuple of attribute values **[a_1, a_2, …, a_n],** it is required to predict the target value for the new instance.
- The Bayesian approach is to assign the most probable target value, v_{MAP}, given the attribute values, to the new instance

$$v_{MAP} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j | a_1, a_2 \ldots a_n)$$

$$\begin{aligned} v_{MAP} &= \underset{v_j \in V}{\operatorname{argmax}} \frac{P(a_1, a_2 \ldots a_n | v_j) P(v_j)}{P(a_1, a_2 \ldots a_n)} \\ &= \underset{v_j \in V}{\operatorname{argmax}} P(a_1, a_2 \ldots a_n | v_j) P(v_j) \end{aligned}$$

15

Naive Bayes Classifier

- Naive Bayes assumption: the attribute values are conditionally independent given the target value:

$$P(a_1, a_2 \ldots a_n | v_j) = \prod_i P(a_i | v_j)$$

which gives

Naive Bayes Classifier: $v_{NB} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j) \prod_i P(a_i | v_j)$

- The probabilities $P(v_j)$, $P(a_i|v_j)$ are estimated in the learning stage based on the frequencies in the training data.

16

Naive Bayes: Training Examples

Day	Outlook	Temp.	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Weak	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cold	Normal	Weak	Yes
D10	Rain	Mild	Normal	Strong	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Naive Bayes: Example

- Need to decide the activity for a new instance [Outlook=sunny, Temp=cool, Humidity=high, Wind=strong]

- Our solution is

$$VNB = \arg\max_{v_j \in \{yes,no\}} P(v_j)\Pi_i P(a_i \mid v_j)$$

$$= \arg\max_{v_j \in \{yes,no\}} P(v_j)P(sunny \mid v_j)P(cool \mid v_j)P(high \mid v_j)P(strong \mid v_j)$$

- P(yes)=9/14=0.64 P(No)=5/14=0.36
  P(strong|yes)=3/9=0.33 P(strong|no)=3/5=0.6
  ……………………………………………………………
  P(yes)P(sunny|yes)P(cool|yes)P(high|yes)P(strong|yes)=0.0053
  P(no)P(sunny|no)P(cool|no)P(high|no)P(strong|no)=0.0206

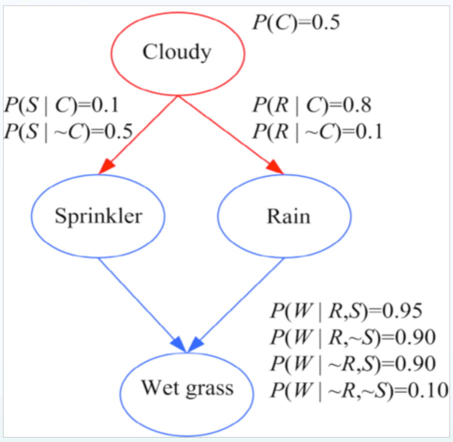
- *VNB* = not to ply tennis

Random Variable and Bayesian Networks

- A random variable is variable that takes a value at random.
- A Bayesian network is used to describe causal relationship amomg random variables

19

Bayesian Network Example



Network implies a set of conditional independence assertions:

Each node is conditionally independent of its non-descendants, given its parent values

20

Bayesian Network

A Bayesian network is a graph in which the following holds:

1. A set of random variables makes up the nodes of the graph
2. A set of directed links conects pairs of nodes, showing the causal relationship
3. Each node is associated with conditional probabilities quantifying the effects of parent nodes on it
4. The structure is represented as a directed acyclic graph (DAG)
5. Each node is conditionally independent of its non-descendants given its parent values.

21

Inference in Bayesian Networks

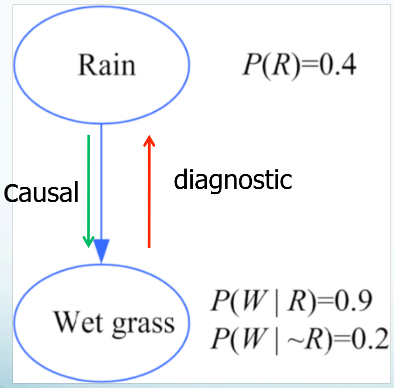
It is to infer the probabilities of values of one or more network variables given observed values of other variables.

Two basic types of inferences:

1. Diagnostic inference: from effects to cause
2. Causal inferences: from cause to effect

22

Inference with Bayesian Network



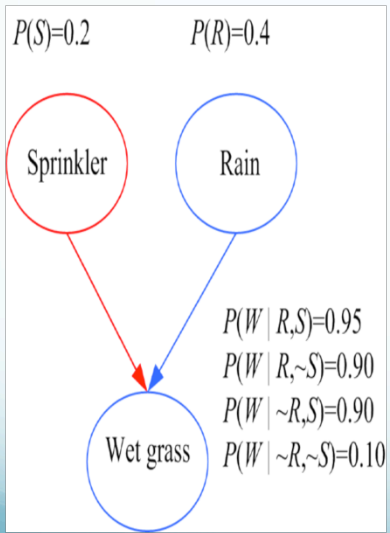
Diagnostic inference:
Knowing that the grass is wet, what is the probability for rain

$$P(R|W) = \frac{P(W|R)P(R)}{P(W|R)P(R) + P(W|notR)P(notR)}$$

$$= \frac{0.9 \cdot 0.4}{0.9 \cdot 0.4 + 0.2 \cdot 0.6} = 0.75$$

23

Inference with Bayesian Network



Causal inference: If the sprinkler is on, what is the probability that the grass is wet?

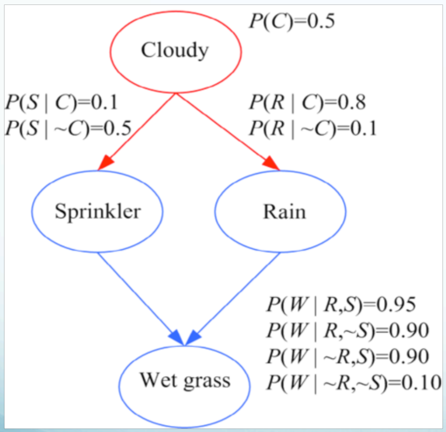
P(W|S) = P(W|R,S) P(R|S) + P(W|~R,S) P(~R|S)
= P(W|R,S) P(R) +P(W|~R,S) P(~R)
= 0.95 0.4 + 0.9 0.6 = 0.92

Diagnostic inference: If the grass is wet, what is the probability that the sprinkler is on?

P(S|W) = ?

24

Inference with Bayesian Network



Causal inference:

P(W|C) = P(W|R,S) P(R,S|C) + P(W|~R,S) P(~R,S|C) + P(W|R,~S) P(R,~S|C) + P(W|~R,~S) P(~R,~S|C)

and use the fact that P(R,S|C) = P(R|C) P(S|C)

Diagnostic: P(C|W) = ?

25

Reading Guidance

1. Read sections 6.1-- 6.2, and 6.9 in the book "Machine Learning" by Mitchell

2. Read 6.11.1 – 6.11.3 of section 6.11 in the "Machine Learning" book by Mitchell for the knowledge of Bayesian networks.

3. For inference with Bayesian network, please read the slides carefully and solve the remaining questions by yourself as exercise.

26