Bayesian AI

07 January 2021 10:42

Objectives:

- Bayesian Probability
- Inference using joint probability distributions
- introduction to Bayesian Belief Networks
  - entailed independence relations
- Inference in Bayesian Belief Networks
  - exact inference
- Making decisions with outcome probabilities
  - decision trees
  - influence diagrams

Reading: AIFCA 8.1-8.4 & 9.1-9.3, AIAMA: Ch. 13 & 14

Probability Recap

A problem with FOL is that agents never have access to the whole truth about their environment. In almost every case, even in simple worlds, there will be important questions to which the agent cannot find a categorical answer. The agent has to act under *uncertainty*.

- At best, our agent can only provide a **degree of belief**
- The tool we use for dealing with degrees of belief is **probability theory**.
- This provides a way of summarising the uncertainty that comes from our laziness and ignorance.
- Degree of truth, as opposed to degree of belief, is the subject of fuzzy logic, which is covered later.

Just as entailment status can change when more sentences are added to the KB, probabilities can change when more evidence is acquired.

All probability statements must therefore indicate the evidence with respect to which the probability is being assessed. As the agent receives new percepts its probability assessments are updated to reflect the new evidence.

- Before the evidence is obtained, we talk about **prior** or **unconditional probability**
- After evidence is obtained, we talk about **posterior** or **conditional probability**

Uncertainty and Rationality

To make choices between options, an agent must first have preferences between possible outcomes of various plans.

We will be using **utility theory** to represent and reason with preferences. The term utility is used in the sense of "*the quality of being useful*".
Preferences as expressed by utilities, are combined with probabilities in the general theory of rational decisions called decision theory.

- **Utility theory -** every state has a degree of usefulness to an agent, and the agent will prefer states with higher utility.
- **Decision theory = probability theory + utility theory**

An agent is rational if and only if it choose the action that yields the highest expected utility, average over all the possible outcomes of the action - This is called the principle of **Maximum Expected Utility** (MEU).

Probabilities and utilities are therefore combined in the evaluation of an action by weighting the utility of a possible outcome by the probability that it occurs.

Prior Probability

We use the notation **P(A)** for the *unconditional* or *prior probability* that the condition A is true.

For example if Cavity denotes the probability that a patient has a cavity, then P(Cavity) = 0.1 means that in the absence of any other information, the agent will assign a probability of 0.1 to the event that the patient has a cavity.

Remember, P(A) can only be used when there is no other information.

As soon as some more information, **B**, is known, we have to reason with the conditional probability of **A given B**, instead of **P(A)**.

The proposition that is the subject of a probability statement can be represented by a proposition symbol, as in the **P(A)** example. Propositions can also include equalities involving random variables.

For example, if we are concerned about the random variable Weather,

- P(Weather = Sunny) = 0.7
- P(Weather = Rain) = 0.2
- P(Weather = Cloud) = 0.08
- P(Weather = Snow) = 0.02

Each random variable X has a domain of possible values, **<x1, x2, ... , xn>** that it can take on.

We can view proposition symbols as random variables as well, if we assume that they have a domain **<true, false>.**

Thus, **P(Cavity)** can be viewed as **P(Cavity == True)** and **P(!Cavity) = P(Cavity == False)**

If we want to talk about the probabilities for a random variable, we can do so with P(Weather) which denotes a vector of values for the probabilities of each state of weather. Given the preceding values, for example, we would write:

- **P(Weather) = <0.7, 0.2, 0.08, 0.02>**

This statement defines a probability distribution for the random variable Weather

You can look at the probability of many random variables at once with **P(Weather, Cavity)**, which

creates a 4x2 table of probabilities, containing all the combinations of the random variables.

Conditional Probability

Once the agent has obtained some evidence concerning the previously unknown propositions making up the domain, prior probabilities are no longer applicable. Instead, we use conditional or posterior probabilities, with the notation **P(A|B)**. This is read as "*The probability of A given that all we know is B*"

e.g. **P(Cavity|Toothache) = 0.8** reads "*The probability of a cavity given that we know the patient has a toothache = 0.8*"

As soon as we know C, we can't compute **P(A|B)**, we must instead compute **P(A|B /\ C)**

We can think of the prior probability as just a conditional probability that looks like **P(A |)**, where the probability is conditioned on no evidence.

We can also use the **P** notation with conditional probabilities. **P(X|Y)** is a 2D table giving the values of **P(X = xi | Y = yj)** for each possible **i,j**. Conditional probabilities can be defined in terms of unconditional probabilities

P(A|B) = P(A/\B) / P(B) holds whenever **P(B) > 0**, and can be written as **P(A/\B) = P(B|A)P(B)**

P(A/\B) = **P(A|B)P(B)** = **P(B|A)P(A)**

We can also extend our P notation to handle equations like these, providing conciseness.

e.g. **P(X, Y) = P(X|Y)P(Y)** which denotes a set of equations relating to the corresponding individual entries in the tables (not a matrix multiplication of the tables)

Thus, one of the equations might be

P(X = x1 /\ Y = y2) = P(X = x1 | Y = y2) P(Y = y2)

In general, if we are interested in the probability of a proposition **A**, and we have accumulated evidence **B**, then the quantity we must calculate is **P(A|B)**. Sometimes we will not have this conditional probability, available directly in the KB, and we must resort to probabilistic inference.

- Let α and β be two propositions such that $P(\beta) \neq 0$. Then the conditional probability of α given β, denoted by $p(\alpha|\beta)$, is defined as:
- $p(\alpha|\beta) = p(\alpha \wedge \beta)/p(\beta)$
- This is also known as the **posterior probability** of α being true when β (described as the *evidence*) is true.
  - $P(\alpha)$ is also known as the **prior probability** and is equivalent to $P(\alpha|\ true)$
  - Reflects our background knowledge about the chance of α being true
- Exercise: Given that the card drawn from the top of the pack is from a black suit, what is the probability of the card being a King or Queen?
- Note that conditional probability is **not** a measure of causality

- Not all random variables or events affect the probability of each other.
- If we have two random variables, X and Y then:
  - If $p(X|Y) = p(X)$ we say that X and Y are independent of each other, as Y occurring has not affected the chance of X occurring.
  - $p(X|Y) = p(X \wedge Y)/p(Y) = p(X) \implies p(X \wedge Y) = p(X) * p(Y)$
  - This is also known as **unconditional independence**
- X and Y are **conditionally independent** given random variable Z if:
  - $p(X \wedge Y|Z) = p(X|Z) * p(Y|Z)$
  - $p(X|Y \wedge Z) = p(X|Z)$
  - $p(Y|X \wedge Z) = p(Y|Z)$
- Assuming independence is a useful tool, as it means we do not need a list of exhaustive conditional probabilities, which can often be infeasible to compute.

Probability Theorems

- Total Probability
  - Given a set of disjoint events A_i that partition the Sample Space: $p(\Omega) = \sum_i p(A_i)$
  - Also: $p(B) = \sum_i p(B \wedge A_i) = \sum_i p(B|A_i)p(A_i)$
- Product Rule
  - $p(A \wedge B) = p(B|A)p(A)$
- Chain Rule (Generalization of the product rule)
  - If we rearrange the definition of conditional probability, we find that $P(\alpha \wedge \beta) = P(\alpha|\beta) * P(\beta)$
  - This means any conjunction of propositions and can be expressed as a product of conditional probabilities
    - $P(a_1 \wedge a_2 \wedge ... \wedge a_i) = P(a_1) * P(a_1|a_2) * P(a_3|a_1 \wedge a_2) * ... * P(a_i|a_1 \wedge ... \wedge a_{i-1})$

Joint Probability Distribution

A joint completely specifies an agent's probability assignments to all propositions in the domain (both simple and complex)

A probabilistic model of a domain consists of a set of random variables that can take on particular values with certain probabilities. Let **X1 ... Xn** be the variables. An atomic event is an assignment of particular values to all the variables, in other words a complete specification of the state of the domain.
The joint probability distribution **P(X1 ... X2)** assigns probabilities to all possible atomic events. Recall that **P(Xi)** is a one dimensional vector of probabilities for the possible values of the variable **Xi.**

Then the joint is an n-dimension table with a value in every cell, giving the probability of that particular state occurring.

	toothache	!toothache

cavity	0.04	0.06
!cavity	0.01	0.89

- Any conjunction of atomic events is necessarily **false**.
- Since they are collectively exhaustive, their disjunction is necessarily **true**.

Adding across a row or columns gives the unconditional probability of a variable, e.g.

P(Cavity) = 0.06 + 0.04 = 0.1

Thus the probability that the patient has a cavity given the evidence that they have a toothache

P(Cavity|Toothache) = P(Cavity /\ Toothache) / P(Toothache) = 0.04/(0.04 + 0.01) = 0.8

Computing the joint is useful, but expensive. Bayes' rule gets us around this cost.

Bayes' Rule

Recall the two forms of the product rule:
- **P(A /\ B) = P(A|B)P(B)**
- **P(A /\ B) = P(B|A)P(A)**

Equating the RHS we get

$$\mathbf{P(B|A)} = \frac{P(\boldsymbol{A}|\boldsymbol{B})P(\boldsymbol{B})}{\mathbf{P(A)}}$$

Bayes' Rule needs 3 terms - a conditional probability and 2 unconditional probabilities - just to compute one conditional probability

Example

- A doctor knows that meningitis causes a stiff neck, 50% of the time
- The doctor knows some unconditional facts, the prior probability of a patient having meningitis is 1/50,000
- The prior probability of a patient having a stiff neck is 1/20

Let S = stiff neck
Let M = meningitis

P(S|M) = 0.5
P(M) = 1/50000
P(S) = 1/20

P(M|S) = P(S|M) * P(M) / P(S) = (0.5 * 1/50000) / (1/20) = 0.0002

This is very useful to know, since if there is suddenly an outbreak of meningitis, we can simply run the equation with new numbers and work out the new relationship between the two probabilities.

Normalisation

The goal of a normalising constant is to reduce any probability function t oa probability density function with a total probability of one.

For example:

P(M|S) = P(S|M)P(M) / P(S)
P(W|S) = P(S|W)P(W) / P(S)

if we divide the top by the bottom we get P(S|M)P(M) / P(S|W)P(W) - the relative likelihood of whiplash W compared to meningitis M.

In some cases, relative likelihood is sufficient for decision making, but when the two possibilities yield radically different utilities for various treatment actions, one needs exact values in ordre to make rational decisions.

Combining Evidence

Given many variables, we might need an exponential number of probability values to complete a task. At this point we may as well go back to the joint.

in many domains, we can simplify the application of Bayes' rule so that it requires fewer probabilities in order to compute a result.

The first step is to take a slightly different view of the process of incorporating multiple pieces of evidence. The process of Bayesian updating incorporates evidence on piece at a time, modifying the previously held belief in the unknown variable.

In Bayesian updating, when each new piece of evidence is observed the belief in the unknown variable is multiplied by a factor that depends on the new evidence.

Working out this multiplication factor depends not just on the new evidence, but also on the evidence already obtained.

The key observation here is that of conditional independence

P(Catch|Cavity /\ Toothache) == P(Catch|Cavity)

The probability of the probe catching does not depend on the presence of a toothache. Thus we can simplify our expression.

Conditional Independence

A useful way to limit the amount of information required is to assume that each variable only directly depends on a dew other variables. This uses assumptions of conditional independence. Not

only does it reduce how many numbers are required to specify a model, but also the independence structure may be exploited for efficient reasoning.

Random variable X is conditionally independent of random variable Y given a set of random variables **Zs** if
P(X|Y), Zs = P(X|Zs)

whenever the probabilities are well defined. This means that for all x in the domain of X, for all y in the domain of Y and for all z in the domain of z, if **P(Y = y /\ | Zs = s) > 0**, then

P(X = x | Y = y /\ Zs = z)
is equal to
P(X = x | Zs = z)

In other words, given a value of each variable in Zs, knowing Y's value does not affect the belief in the value of X

Example

*consider the probabilistic model of students and exams. it is reasonable to assume that the random variable Intelligence is independent of **Works_hard,** given no other observations. If you find a student that works hard, it does not tell you anything about their level of intelligence.*

*the answers to the exams, **Answers**, would depend on whether the student is intelligent and works hard. Thus, given **Answers**, intelligent would be dependent on **Works_hard;** if you found someone had insightful answers, and did not work hard, your belief that they are intelligent would go up.*

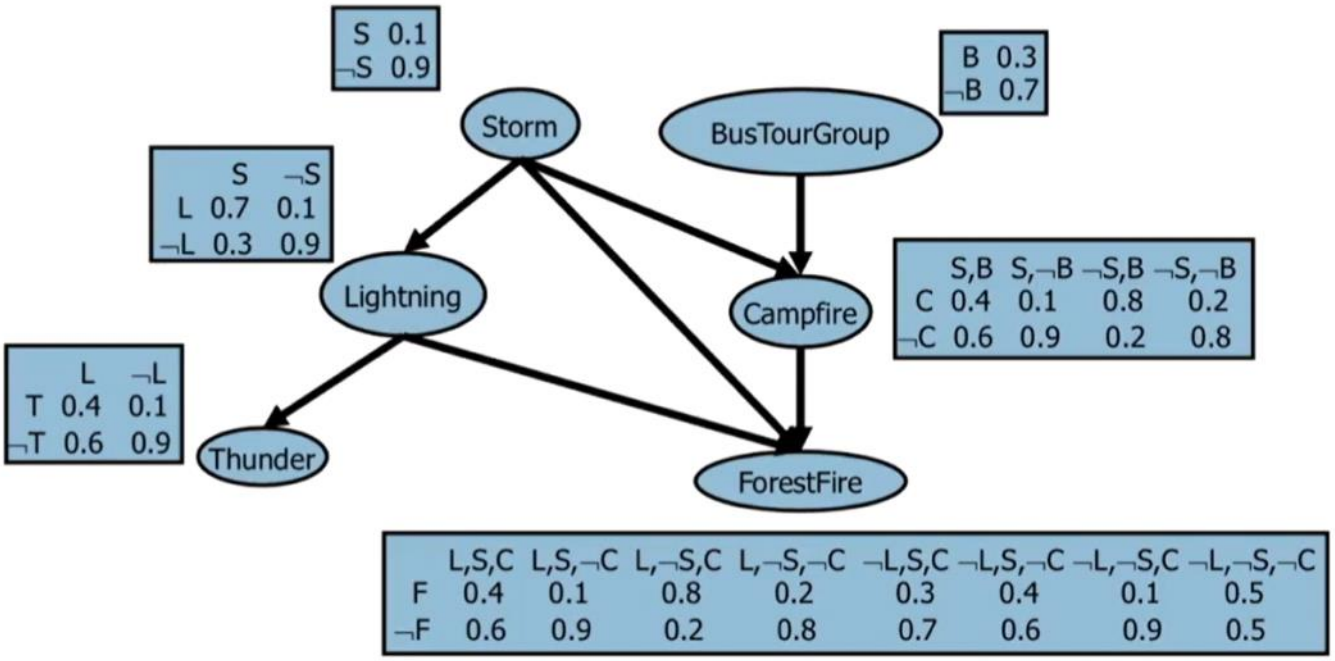
*The grade on the exam, **Grade**, should depend on the student's answers, not on the intelligence or whether they work hard. Thus **Grade** would be independent of **Intelligence** given **Answers**. However, if the answers were not observed, **Intelligence** will affect **Grade** because highly intelligent students would be expected to have different answers than less intelligent students. Thus **Grade** is dependent on I**ntelligence** given no observations.*

Conditional independence is a useful assumption that is often natural to assess and can be exploited in inference. It is very rare that we should have a table of probabilities of worlds and assess independence numerically.

Summary

- Probabilities represent an inability to reach a definite decision regarding truth
- Basic probability statements include prior probabilities and conditional probabilities.
- The axioms of probability specify constraints on reasonable assignments of probabilities to axioms. An agent violating the axioms will behave irrationally and can be manipulated.
- The joint probability distribution specifies the probability of each complete assignment of values to random variables. It is usually far too large to create or use.
- Bayes' rule allows unknown probabilities to be computed from known, stable ones.
- In the general case, combining many pieces of evidence may require assessing a large number of conditional probabilities, as in the joint probability distribution
- Conditional independence brought about by direct causal relationships in the domain allow Bayesian updating to work effectively even with multiple pieces of evidence.

- The computation, from observed evidence, of posterior probabilities for query propositions
  - The full joint probability distribution is the Knowledge Base
- The General Inference Procedure is as follows:
  - Let X be the query variable
  - Let E be the evidence variables and e be the observed values for them
  - Let Y be the unobserved variables
  - We need to calculate $p(X|e)$
  - $p(X|e) = \alpha p(X|e) = \alpha \sum_y p(X, e, y)$
  - α is the normalization constant, $1/p(e)$
  - $p(e) = \sum_{x,y} p(x, e, y)$


S 0.1
¬S 0.9
Storm
BusTourGroup
B 0.3
¬B 0.7
S ¬S
L 0.7 0.1
¬L 0.3 0.9
Lightning
Campfire
S,B S,¬B ¬S,B ¬S,¬B
C 0.4 0.1 0.8 0.2
¬C 0.6 0.9 0.2 0.8
L ¬L
T 0.4 0.1
¬T 0.6 0.9
Thunder
ForestFire
L,S,C L,S,¬C L,¬S,C L,¬S,¬C ¬L,S,C ¬L,S,¬C ¬L,¬S,C ¬L,¬S,¬C
F 0.4 0.1 0.8 0.2 0.3 0.4 0.1 0.5
¬F 0.6 0.9 0.2 0.8 0.7 0.6 0.9 0.5