Probabilistic Reasoning and Bayesian Belief Networks:-

Probability of an Event

Consider an experiment that may have different outcomes. We are interested to know what is the probability of a particular set of outcomes.

Let sample space S be the set of all possible outcomes
Let Event A be any subset of S

Definition 1: probability(A) = (number of outcomes in A)/ (total number of outcomes) P(A) = |A| / |S|
i.e. the probability of A is equal to the number of outcomes of interest divided by the number of all possible outcomes.

P(A) is called prior (unconditional) probability of A
P(~A) is the probability event A not to take place.

Example 1: the probability to pick a spade card out of a deck of 52 cards is 13/52 = ¼ The probability to pick an Ace out of a deck of 52 cards is 4/52 = 1/13

Probability Axioms:

(1) 0 ≤ P(A) ≤ 1

(2) P(A) = 1 – P(~A)

(3) P(A v B) = P(A) + P(B) – P(A & B)

P(A v B) means the probability of either A or B or both to be true
P(A&B) means the probability of both A and B to be true.

Example 2: P(~A) – The probability to pick a card that is not a spade out of a deck of 52 cards is 1 – 1/4 = 3/4
Example 3: P(A v B) – The probability to pick a card that is either a spade or an Ace is 1/4 + 1/13 - 1/4 *1/13 = 16/52 = 4/13
Another way to obtain the same result: There are 13 spade cards and 3 additional Ace cards in the set of desired outcomes. The total number of cards is 52, thus the probability is 16/52.

Example 4: P(A&B) – The probability to pick the spade Ace is 1/52

2. Random Variables and Probability Distributions

To handle more conveniently the outcomes, we can treat them as values of so called ***random variables***.
For example “spade” is one possible value of the variable Suit, “clubs” is another possible value. In the card example, all values of the variable Suit are equally probable. This is not always so however. We may be interested in the probabilities of each separate value.

The set of the probabilities of each value is called *probability distribution* of the random variable.

Let X be a random variable with a domain $<x_1, x_2, \ldots, x_n>$
The probability distribution of X is denoted by **P**(X) = $<P(X = x_1), P(X=x_2), \ldots, P(X=x_n)>$ Note that $P(X = x_1) + P(X = x_2) + \ldots + P(X = x_n) = 1$

Example 5: Let Weather be a random variable with values <sunny, cloudy, rainy, snowy>
Assume that records for some town show that in a year 100 days are rainy, 50 days are snowy, 120 days are cloudy (but without snow or rain) and 95 days are sunny.

i.e. P(Weather = sunny) = 95/365 = 0.26
P(Weather = cloudy) = 120/365 = 0.33
P(Weather = rainy) = 100/365 = 0.27
P(Weather = snowy) = 50/365 = 0.14

Thus **P**(Weather) = <0.26, 0.33, 0.27, 0.14> is the probability distribution of the random variable Weather.

3. Joint Distributions

The following example is used to illustrate conditional probabilities and joint distributions

Example 6: Consider a sample S of of 1000 individuals age 25 – 30. Assume that 600 individuals come from high-income families, 570 of those with high income have college education and 100 individuals with low income have college education.

The following table illustrates the example:

		C College ed.	~C Not college ed.	
H	High income	570	30	600
~H	Low income	100	300	400
		670	330	1000

Let H be the subset of S of individuals coming from high-income families, |H| = 600
Let C be the subset of S of individuals that have college education, |C| = 670

The prior probabilities of H, ~H, C and ~C are:

P(H) = 600 / 1000 = 0.6 (60%)   P(~H) = 400 / 1000 = 0.4 (40%)

P(C) = 670 / 1000 = 0.67 (67%)   P(~C) = 330 / 1000 = 0.33 (33%)

We can compute also P(H&C), P(H & ~C), P(~H & C), P(~H & ~C)

P(H&S) = |H &C| / |S| = 570/1000 = 0.57 (57%) - the probability of a randomly selected individual in S to be of high-income family and to have college education.

P(H & ~C) = |H& ~C| / |S| = 30/1000 = 0.03 (3%) - the probability of a randomly selected individual in S to be of high-income family and not to have college education.

P(~H & C) = |~H& C| / |S| = 100/1000 = 0.1 (10%) - the probability of a randomly selected individual in S to be of low-income family and to have college education.

P(~H & ~C) = |~H& ~C| / |S| = 300/1000 = 0.3(30%) - the probability of a randomly selected individual in S to be of low-income family and not to have college education.

Thus we come to the following table:

		C	~C	
		College ed.	Not college ed.	
H	High income	0.57	0.03	0.6
~H	Low income	0.10	0.30	0.4
		0.67	0.33	1

Here we will treat C and H as random variables with values "yes" and "no". The values in the table represent the ***joint distribution*** of C and H, forexample
P(C = yes, H = yes) = 0.57

Formally, joint distribution is defined as follows:

Definition 2: Let X1, X2, .., Xn be a set of random variables each with a range of specific values.

P(X1,X2,…,Xn) is called **joint distribution** of the variables X1, X2, …, Xn and it is defined by a n-dimensional table, where each cell corresponds to one particular assignment of values to the variables X1, X2, …, Xn

Each cell in the table corresponds to an **atomic event** – described by a particular assignment of values to the variables.

Since the atomic events are mutually exclusive, their conjunction is necessarily false.

Since they are collectively exhaustive, the disjunction is necessarily true.

So by axioms (2) and (3) the sum of all entries in the table is 1

Given a joint distribution table we can compute prior probabilities:

P(H) = P(H & C) + P(H& ~C) = 0.57 + 0.03 = 0.6

Given a joint distribution table we can compute conditional probabilities, discussed in the next section.

4. Conditional Probabilities

We may ask: what is the probability of an individual in S to have a college education given that he/she comes from a high income family?

In this case we consider only those individuals that come from high income families. Their number is 600. The number of individuals with college edication within the group of high-family income is 570. Thus the probability to have college education given high-income family is 570/600 = 0.95.

This type of probability is called ***conditional probability***

The probability of event B given event A is denoted as P(B|A), read “P of B given A”

In our example, $P(C|H) = \dfrac{|C \& H|}{|H|}$

We will represent P(C|H) by P(C&H) and P(H)

$$P(C|H) = \frac{|C \& H|}{|H|} = \frac{\dfrac{|C \& H|}{|S|}}{\dfrac{|H|}{|S|}} = \frac{P(C\&H)}{P(H)}$$

Therefore

P(C|H) = P(C&H) / P(H)

Definition 3: The conditional probability of an event B to occur given that event A has occurred is

P(B|A) = P(B&A) / P(A)

P(B|A) is known also as ***posterior probability*** of B

P(B & A) is an element of the joint distribution of the random variables A and B.

In our example, P(C&H) = P(C = yes, H = yes). Thus given the joint distribution **P** (H, C), we can compute the prior probability P(H), P(~H), P(C), P(~C) and then the conditional probability P(C|H), P(C|~H), P(H|C), P(H|~C) .

Independent events

Some events are not related, for example each outcome in a sequence of coin flips is independent on the previous outcome.

Definition 4: Two events **A and B are independent** if $P(A|B) = P(A)$, and $P(B|A) = P(B)$.

Theorem: A and B are independent if and only if $P(A \& B) = P(A)*P(B)$

The proof follows directly from Definition 3 and Definition 4.

Another definition: X and Y are conditionally independent iff $P(X|Y \& Z) = P(X|Z)$

Bayes' Theorem:-

From Definition 3 we have

$$P(A\&B) = P(A|B)*P(B)$$
$$P(B\&A) = P(B|A)*P(A)$$

However, $P(A\&B) = P(B\&A)$
Therefore

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

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

This is the Bayes' formula for conditional probabilities, known also as Bayes' theorem

More than 2 variables

Bayes' theorem can represent conditional probability for more than two variables:

$$P(X|Y1\&Y2 \& \ldots \& Yn) = P(Y1 \& Y2 \& \ldots \& Yn \mid X) * P(X) / P(Y1 \& Y2 \& \ldots \& Yn)$$

Think of X as being a hypothesis, and Y1, Y2, …, Yn as being n pieces of evidence for the hypothesis. When Y1, Y2, …, Yn are independent on each other, the formula takes the form:

$$P(X|Y1\&Y2 \& \ldots \& Yn) = \frac{P(Y1|X)*P(Y2|X)*\ldots*P(Yn \mid X) * P(X)}{P(Y1)*P(Y2)*\ldots*P(Yn)}$$

In case of several related events, the Bayes' formula is used in the following form:

$$P(X1 \& X2 \& \ldots \& Xn) = P(X1) * P(X2|X1) * P(X3 \mid X2 \& X1) \ldots P(Xn \mid Xn-1 \& \ldots X1)$$

Normalization

Consider the probability of malaria given headache

$$P(M|H) = P(H | M)*P(M) / P(H)$$

It may be more difficult to compute P(H) than P(H|M) and P(H | ~M).
We can represent P(H) trough P(H|M) and P(H | ~M).

We have:

$$P(M|H) = P(H | M)*P(M) / P(H)$$
$$P(\sim M|H) = P(H | \sim M)*P(\sim M) / P(H)$$

Adding these equations we obtain

$$P(M|H) + P(\sim M|H) = (P(H | M)*P(M) + P(H | \sim M)*P(\sim M)) / P(H)$$

For the left side we know that $P(M|H) + P(\sim M|H) = 1$

So we have

$$1 = (P(H | M)*P(M) + P(H | \sim M)*P(\sim M)) / P(H)$$

Multiply both sides by P(H):

$$P(H) = P(H | M)*P(M) + P(H | \sim M)*P(\sim M)$$

Replacing in the Byes' Theorem P(H) with the right side above, we get:

$$\mathbf{P(M|H) = \frac{P(H | M)*P(M)}{P(H | M)*P(M) + P(H | \sim M)*P(\sim M)}}$$

This process is called normalization because it resembles the normalization process for functions – multiplying a function by a chosen constan so that its values stay withun a specified range.

Relative Likelihood of two events

Given that you have a headache, is it more likely that you have flu rather than plague?

P(plague|headache) = P(headache | plague) * P(plague) / P(headache)
P(flu | headache) = P(headache | flu) * P(flu) / P(headache)

The ratio

$$\frac{P(plague|headache)}{P(flu | headache)} = \frac{P(headache | plague) * P(plague)}{P(headache | flu) * P(flu)}$$

is called relative likelihood of having plague vs having flu given headache. It can be computed without knowing P(headache).

In general, the relative likelihood of two events B and C given A is computed as follows

$$\frac{P(B \mid A)}{P(C \mid A)} = \frac{P(A \mid B) * P(B)}{P(A \mid C) * P(C)}$$

Example: The Monty Hall game

You are about to choose your winning in a game show. There are three doors behind one of which is a red Porsche and other two, goats. You will get whatever is behind the door you choose. You pick a door, say *A*. At this point the game show host opens one of the *other* two doors, which he knows to contain a goat, for example *B* and asks if you would now like to revise your choice to *C*. The question is: Should you? (Assuming you want the car and not the goat.)

Let P(PA). P(PB), and P(PC) be the probabilities of the Porsche being behind door A, door B and door C respectively. We assume that the car is randomly placed, so

P(PA) = P(PB) = P(PC) = 1/3

Let O be the event that Monty Hall opens door B.
The Monty Hall Problem can be restated as follows: is P(PA | O) = P(PC| O)

By the Bayes' Theorem we have:

$$P(PA \mid O) = \frac{P(O \mid PA) * P(PA)}{P(O)}$$

$$P(PC \mid O) = \frac{P(O \mid PC) * P(PC)}{P(O)}$$

We have to compute P(O), P(O|PA) and P(O|PC)

P(O | PA) = 1/ 2 , if the car is behind A, Monty Hall can open either B or C
P(O | PB) = 0 , if the car is behind B, Monty Hall will not open B
P(O | PC) = 1 , if the car is behind C, Monty Hall can only open door B

P(O) = P(O|PA)* P(PA) + P(O|PB) * P(PB)+ P(O|PC) * P(PC) (see section 5.2. Normalization)

P(O) = 1/3 * (1/2 + 0 + 1) = 1/2
Therefore we obtain:

P(PA | O) = (1 / 2 * 1 / 3) / (1 / 2) = 1/3
P(PC | O) = (1 * 1/3) / (1 / 2) = 2/3

So, if you switch to door C, you double your chance to win the Porsche.

Useful expressions

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

$$P(A|B) = \frac{P(A \& B)}{P(A \& B) + P(\sim A \& B)}$$

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

$$P(A | B) = \frac{P(B|A) * P(A)}{P(B|A)*P(A) + P(B|\sim A) * P(\sim A)}$$

Simple Bayesian Concept Learning

The Bayes' theorem can be used to solve the following problem:
Determine the most probable hypothesis out of *n* possible hypotheses H_1, H_2, .., H_n, given a set of evidence E. For each H_i we can compute

$$P(H_i | E) = \frac{P(E|H_i) * P(H_i)}{P(E)}$$

and take the hypothesis H_k for which $P(H_k | E)$ has the greatest value.

This is a maximization problem – we are not looking for the particular value of each $P(H_i | E)$, we are looking the hypothesis for which the posterior probability is maximum. Hence we can simplify the expression to be computed based on the following considerations:

a) The evidence is not dependent on the hypotheses, so we can remove P(E) :

$$P(H_i | E) = P(E|H_i) * P(H_i)$$

b) Assuming that all hypotheses are equally likely (same prior probability), we can remove the prior probability

$$P(H_i | E) = P(E|H_i)$$

We choose the hypothesis for which the value of $P(E|H_i)$ is highest.

$P(E|H_i)$ is known as the likelihood of the evidence E given the hypothesis H_i.