CSE 422: Artificial Intelligence

Bayesian Nets - I

Swakkhar Shatabda

BRAC University

December 2, 2024

Contents

Random Variables

Random Variable is a mapping/function of the outcome of an experiment to a real value. Examples:

1. Price of a stock
2. Maximum Temperature of last week
3. Number of employees in the company
4. Gender of a sales person/customer
5. Types of companies
6. Amount of money in your bank account

Types of Random Variables

- Each of the experiments in the previous slide have got outcomes.
- These outcomes are of two types:
  - countable
  - infinite.
- The values that a random variable can take are said to be in a set called Domain of the random variable.
- Random variables are of two types:
  - discrete
  - continuous.

Random variables are generally denoted by a capital Letter and its values by small letters.

Discrete Random Variables

Discrete variables are those that have a countable number of values with in any range/duration.

Example:

- Type of asset,
- Seniority level of an employee,
- Number of days, etc.

Continuous Random Variables

Continuous variables on the other hand can take infinite number of values in a range.

Example:

- Fraction of time/year,
- Interest rate,
- Amount of money, etc.

Distribution

An ensemble of numbers in called a distribution if they sum to 1 and they are within the range $[0, 1]$.

- $0, 0, 0, 0, 1, 0, 0, 0$
- $0.2, 0.2, 0.3, 0.3$
- $0.2, 0.2, 0, 0.1$
- $1, 0, -1, 1$
- $10, -9, 0$

If we normalize (divide them by their sum) any ensemble of positive integers we get a distribution that is proportional to the original ensemble.

Probability Distribution

For any random variable if you associate a distribution with all the values it can take, we can call it a probability distribution.
For example: Say there is a random variable called X which denotes the number of courses taken by Dr. Swakkhar in a year in MSCSE program. It is either 0,1,2 or 3. This is definitely a case of discrete random variable. Now lets associate a distribution $0, 0, 1, 0$ to this values. This distribution, we can call it a probability distribution of X.
$P(X = 0) = 0$
$P(X = 1) = 0$
$P(X = 2) = 1$
$P(X = 3) = 0$

Axioms of Probability

1. $\sum_x P(X = x) = 1$
2. $\forall_x 0 \leq P(X = x) \leq 1$

An Example

Consider the following data:

Case	Gender	Height (inches)	Wage ($)
1	female	64	30,000
2	female	64	30,000
3	female	64	40,000
4	female	64	40,000
5	female	68	30,000
6	female	68	40,000
7	male	64	40,000
8	male	64	50,000
9	male	68	40,000
10	male	68	50,000
11	male	70	40,000
12	male	70	50,000

Here, we can define 3 random variables. Lets try to associate probability distributions with them. Lets find $P(G), P(H)$ and $P(W)$. The three variables have respectively 2,3 and 3 possible values. Thus distribution table of these variables will have that many number of rows summing to 1. We can use the relative frequency property to estimate the distributions for each of these variables.

An Example

Case	Gender	Height (inches)	Wage ($)
1	female	64	30,000
2	female	64	30,000
3	female	64	40,000
4	female	64	40,000
5	female	68	30,000
6	female	68	40,000
7	male	64	40,000
8	male	64	50,000
9	male	68	40,000
10	male	68	50,000
11	male	70	40,000
12	male	70	50,000

G	$P(G)$
m	$\frac{6}{12} = 0.5$
f	$\frac{6}{12} = 0.5$

H	$P(H)$
64	$\frac{6}{12} = 0.5$
68	$\frac{4}{12} = 0.334$
70	$\frac{2}{12} = 0.166$

W	$P(W)$
$30k$	$\frac{3}{12} = 0.25$
$40k$	$\frac{6}{12} = 0.5$
$50k$	$\frac{3}{12} = 0.25$

Joint Probability

- Joint probability means that both event can happen together, denoted as $P(X, Y)$.
- Here we can use the relative frequency approach once again to find join probability distribution for any combination of the variables.
- For the previous example, lets consider two variables Gender and Wage.
- We are trying to find $P(G, W)$.
- These variables have 2 and 3 possible values for each. That is why they will yield $2 \times 3 = 6$ combinations.
- Thus the joint probability distribution will have six values summing to 1.

Joint Probability

Case	Gender	Height (inches)	Wage ($)
1	female	64	30,000
2	female	64	30,000
3	female	64	40,000
4	female	64	40,000
5	female	68	30,000
6	female	68	40,000
7	male	64	40,000
8	male	64	50,000
9	male	68	40,000
10	male	68	50,000
11	male	70	40,000
12	male	70	50,000

G	W	$P(G, W)$
m	30k	$\frac{0}{12} = 0.0$
m	40k	$\frac{3}{12} = 0.25$
m	50k	$\frac{3}{12} = 0.25$
f	30k	$\frac{3}{12} = 0.25$
f	40k	$\frac{3}{12} = 0.25$
f	50k	$\frac{0}{12} = 0.0$

Similar table / distribution is possible from three variables. In the case of $P(G, H, W)$, the number of rows will be 18. For n binary or two values variables number of rows are 2^n.

Joint Probability

G	H	W	$P(G,H,W)$
m	64	30k	$\frac{0}{12}$
m	64	40k	$\frac{1}{12}$
m	64	50k	$\frac{1}{12}$
m	68	30k	$\frac{0}{12}$
m	68	40k	$\frac{1}{12}$
m	68	50k	$\frac{1}{12}$
m	70	30k	$\frac{0}{12}$
m	70	40k	$\frac{1}{12}$
m	70	50k	$\frac{1}{12}$
f	64	30k	$\frac{2}{12}$
f	64	40k	$\frac{2}{12}$
f	64	50k	$\frac{0}{12}$
f	68	30k	$\frac{1}{12}$
f	68	40k	$\frac{1}{12}$
f	68	50k	$\frac{0}{12}$
f	70	30k	$\frac{0}{12}$
f	70	40k	$\frac{0}{12}$
f	70	50k	$\frac{0}{12}$

Observations

Data Sampling

Look at the 18 combinations here. Due to only 12 samples, several of the values are 0. Thus we can assume for such tables with large number of variables we often need a lots of samples if we are to calculate or estimate joint probability distribution using relative frequency approach.

Space and Computation

Another trouble is the exponential number of rows. It makes the storage of such tables into the memory a challenging issue. There are solution to both of the problems. But first lets get an idea on the strength of this distribution.

Conditional Probability

- Conditional probability of a random variable is the probability when the outcome of the other variable is known.
- For example, suppose I is a variable denoting the increase of price of stock for a certain company share and B denotes its directors buying shares from the market.
- Lets assume both are binary variables.
- Now, conditional probability $P(I|B = yes)$ or $P(I|B = no)$ denotes two conditional probability distributions, one considering the directions bought company shares and the other the opposite.
- Knowing the value of one variable outcome might change / not change the probability distribution of the other variable.

Conditional Probability

Lets try to calculate the probability distribution of W given the other variable G. There are definitely two cases: male and female.

W	$P(W \mid G = male)$
30k	$\frac{0}{6} = 0$
40k	$\frac{3}{6} = 0.5$
50k	$\frac{3}{6} = 0.5$

W	$P(W \mid G = female)$
30k	$\frac{3}{6} = 0.5$
40k	$\frac{3}{6} = 0.5$
50k	$\frac{0}{6} = 0$

How did we find it? We followed the same relative frequency approach to find this values with the condition that now we look at only those samples that satisfies the given condition.

Independence

- Sometimes two random variables are independent of each other.
- Suppose, there are two variables: X and Y.
- We say X is independent of Y if any knowledge of X does not change the probability distribution of Y.
- Lets assume initially the probability distribution was $P(Y)$ and after we gained the knowledge of X, the conditional probability is now $P(Y|X)$.
- Now if this knowledge have no effect, i.e., $P(Y|X) = p(X)$ we say $Y \perp X$ i.e. Y is independent of X and vice-versa.

Independence Example

Lets examine one example. Suppose there is a roll of dice.
$X = \{\text{the outcome is at least 3}\}$
$Y = \{\text{the outcome is an odd number}\}$
here, $P(Y = yes) = 0.5, P(Y = no) = 0.5$. When X is known, i.e. the number is at least 3,
$P(Y = yes|X = yes) = 0.5, P(Y = no|X = yes) = 0.5$
and also if the number is not at least 3,
$P(Y = yes|X = no) = 0.5, P(Y = no|X = no) = 0.5$
Note that the conditional probability of Y remains same. Thus we can conclude, $X \perp Y$.

In the previous example, compare G, W and we see that $P(W) \neq P(W|G)$, that is why $G \not\perp W$.

Independence Example - II

$P(T)$

T	P
hot	0.5
cold	0.5

$P_1(T, W)$

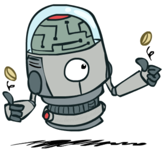
T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$P_2(T, W)$

T	W	P
hot	sun	0.3
hot	rain	0.2
cold	sun	0.3
cold	rain	0.2

$P(W)$

W	P
sun	0.6
rain	0.4

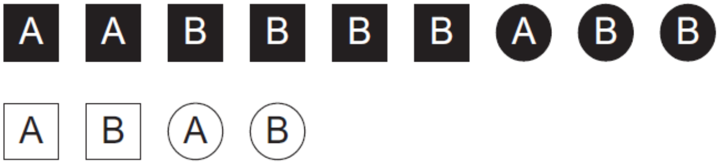


Conditional Independence

Let us explain the idea of conditional independence using three random variables. $X \perp Y|Z$ if $P(X|Z) = P(X|Y,Z)$.

Here note that, if Z is known, then the knowledge of Y does not change the conditional distribution of $X|Z$ which is already conditioned by Z.

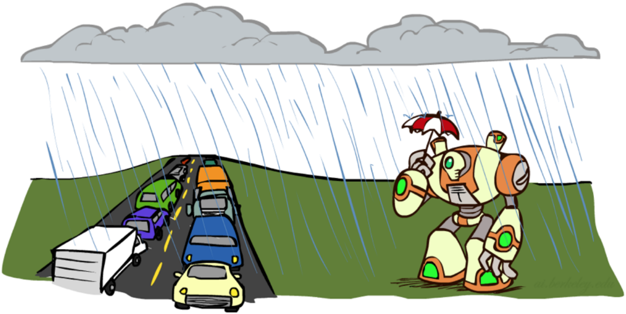
Conditional Independence



Lets define three variables, A (alphabet = A), B (color = black), C (shape = circle).
Here $P(A) = \frac{5}{13}$
$P(A|C = square) = \frac{3}{8}$
we can conclude $A \not\perp C$
But, $P(A|B = yes) = \frac{3}{9} = \frac{1}{3}$
Now if we add the knowledge of the shape,
$P(A|B = yes, C = square) = \frac{2}{6} = \frac{1}{3}$
$P(A|B = yes, C = circle) = \frac{1}{3}$
the conditional probability distribution does not change. Here we conclude, $A \perp C|B$.
That it though they are not independent, they are conditionally independent given B is known. This will lead us to one of the most important structures in graphical models.

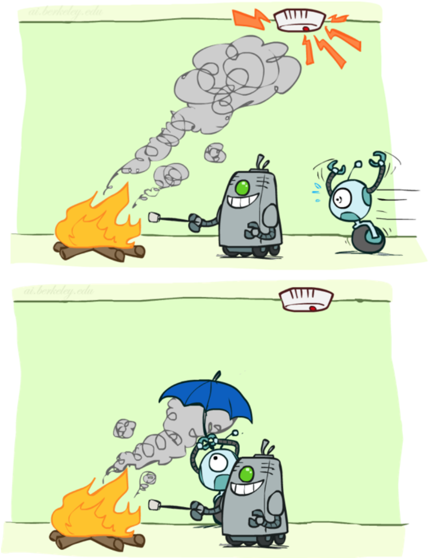
Example

- Rain
- Traffic
- Umbrella



Example

- Fire
- Smoke
- Alarm



Enumeration from Joint Distribution

From the joint distribution of any combination of variables it is possible to enumerate and aggregate the probabilities to find the marginal or conditional probabilities. Later, we will extend this idea for **Bayesian Enumeration** technique.

Marginal Probability

$$P(X) = \sum_{y} P(X, y)$$

$$P(X) = \sum_{y} \sum_{z} P(X, y, z)$$

$$P(X, Y) = \sum_{z} P(X, Y, z)$$

This formula indicates that marginal probability of X is a sum of all joint probabilities considering all combinations of Y. Lets verify that with one example.

Example

Consider the following join probability distribution of $P(G, H)$

G	H	$P(G,H)$
m	64	$\frac{2}{12} = \frac{1}{6}$
m	68	$\frac{2}{12} = \frac{1}{6}$
m	70	$\frac{2}{12} = \frac{1}{6}$
f	64	$\frac{4}{12} = \frac{1}{3}$
f	68	$\frac{2}{12} = \frac{1}{6}$
f	70	0

G	H	$P(G,H)$
m	64	$\frac{2}{12} = \frac{1}{6}$
m	68	$\frac{2}{12} = \frac{1}{6}$
m	70	$\frac{2}{12} = \frac{1}{6}$
f	64	$\frac{4}{12} = \frac{1}{3}$
f	68	$\frac{2}{12} = \frac{1}{6}$
f	70	0

Now suppose we wish to calculate marginal of H. Here is the formula:

$P(H = 64) = P(H = 64, G = m) + P(H = 64, G = f)$

$P(H = 68) = P(H = 68, G = m) + P(H = 68, G = f)$

$\cdots$ and so on

Marginals

Now summing up the corresponding rows we should find the following distribution:

H	$P(H)$
64	$\frac{1}{6}+\frac{1}{3}=0.5$
68	$\frac{1}{6}+\frac{1}{6}=0.334$
70	$\frac{1}{6}+0=0.166$

In the similar way, any combination of variables and their joint distribution can be derived from a larger join distribution table.

Conditional Probability

In case of conditional probability, the idea is a little tricky. It follows the same procedure. For any given condition, we only consider the rows that matches with the condition and then use marginal probability law to find the distributions. However, the remaining rows might not sum to 1. That is why we need to normalize.

$$P(X|Y=y) \propto P(X, Y=y)$$

$$P(X|Y=y, Z=z) \propto P(X, Y=y, Z=z)$$

$$P(X|Y=y) \propto \sum_{z} P(X, Y=y, Z=z)$$

Conditional Probability - Enumeration

Here is an example. Lets suppose we wish to find $P(H|G = m)$ from $P(G, H)$

G	H	$P(G,H)$
m	64	$\frac{2}{12} = \frac{1}{6}$
m	68	$\frac{2}{12} = \frac{1}{6}$
m	70	$\frac{2}{12} = \frac{1}{6}$
f	64	$\frac{4}{12} = \frac{1}{3}$
f	68	$\frac{2}{12} = \frac{1}{6}$
f	70	0

H	$P(H\|G = m)$
64	$\frac{1}{6}$ after normalization $= \frac{1}{3}$
68	$\frac{1}{6}$ after normalization $= \frac{1}{3}$
70	$\frac{1}{6}$ after normalization $= \frac{1}{3}$

H	$P(H\|G = f)$
64	
68	
70	

Chain Rule

From conditional probability, we derive the following:

$$P(A|B) = \frac{P(A, B)}{P(B)} \tag{1}$$

Hence,

$$P(A, B) = P(A|B) \times P(B) \tag{2}$$

Chain Rule

This is the form of chain rule for two variables. Lets extend that to three variables, A, B, C.
$P(A, B, C) = P(X, C) = P(X|C)P(C) = P(A, B|C)P(C) = P(A|B, C)P(B|C)P(C)$

$$P(A, B, C) = P(A|B, C)P(B|C)P(C) \tag{3}$$

General Chain Rule

In the same way, chain rule is applicable to any number of variables:

$$P(X_n, X_{n-1}, \cdots, X_2, X_1) = P(X_n|X_{n-1}, \cdots, X_1) \cdots P(X_2|X_1)P(X_1) \quad (4)$$

Chain Rule and Ordering

Now the order that the list of variables are arbitrary. However, independence and conditional independence will give us freedom to choose the order.
Suppose, two variables are independent of each other, $A, B \implies A \perp B$
Then Eq. 1. becomes

$$P(A, B) = P(A)P(B) \tag{5}$$

Similar way, if $A \perp B|C$, Eq. 3 becomes:

$$P(A, B, C) = P(A|C)P(B|C)P(C) \tag{6}$$

Bayes Theorem

Now, we are ready to apply Bayes theorem! First lets get the theorem.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{7}$$

Bayes Theorem Example

lets illustrate this with one example. Consider the following data:

Stock Price	Interest Rates			
		Decline	Increase	Unit Frequency
	Decline	200	950	1150
	Increase	800	50	850
		1000	1000	2000

Here lets define two variables, S, I to denote stock price increment and interest rate increment. Now,

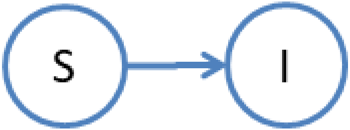
$$
\begin{aligned}
P(S|I) &= \frac{P(I|S)P(S)}{P(I)} \\
&= \frac{\frac{50}{850} \times \frac{850}{2000}}{\frac{1000}{2000}} \\
&= \frac{50}{1000} \\
&= 0.05
\end{aligned}
$$

Exercise

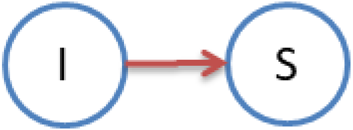
Suppose you feel there is a 0.4 probability the NASDAQ will go up at least 1% today. This is based on your knowledge that, after trading closed yesterday, excellent earnings were reported by several big companies in the technology sector, and that U.S. crude oil supplies unexpectedly increased. Furthermore, if the NASDAQ does go up at least 1% today, you feel there is a 0.1 probability that your favorite stock NTPA will go up at least 10% today. If the NASDAQ does not go up at least 1% today, you feel there is only a 0.02 probability NTPA will go up at least 10% today. You have these beliefs because you know from the past that NTPA's performance is linked to overall performance in the technology sector. You checked NTPA after the close of trading, and you noticed it went up over 10%. What is the probability that the NASDAQ went up at least 1%?

A simple network

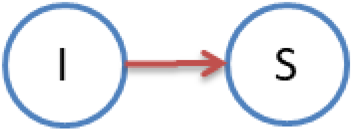
The previous problem can be expressed using a simple Bayesian net with two variables.



This is a graph with two variables denoted by the two nodes and the edge represents the relationship between them. There are four possible graphs possible with two variables. Such graphical models usually repsents the casual relationships. Here is a more plausible relationship:



A simple network

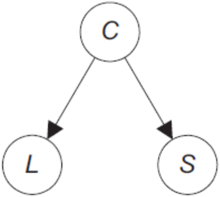


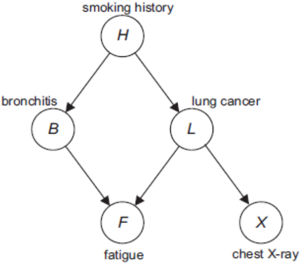
Note that for a Joint Distribution Table with these two variables would require 3 entries (actually 4, however, 3 is sufficient). Here, in this case too we need only 3. For the variable I, $P(I)$ and for S, $P(S|I)$. In the case, where both are independent of each other only 2 values are required. Now, it is a question of how we model the system or which model best fits to our observations.

Exercise: How many graphs are possible with 3 variables?

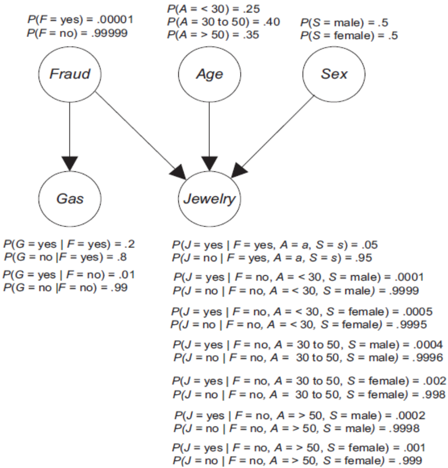
Representation of a Bayesian Net

Bayesian network is a directed acyclic graph. Arrows of the directed edges represent the casual flow of reasoning. In a graphical model, each variable is presented as a node and arrows represent the casual relationship. Generally, an arrow is placed from parent variable to the child. Thus the casual flow is always from parent to child. **Each node must be represented with a distribution conditioned by its parents**.
Try to find representations for the following Bayesian Nets.





A Fraud Detection Problem



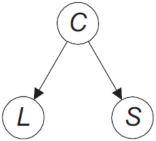
Variable	What the Variable Represents
Fraud (F)	Whether the current purchase is fraudulent
Gas (G)	Whether gas has been purchased in the past 24 hours
Jewelry (J)	Whether jewelry has been purchased in the past 24 hours
Age (A)	Age of the card holder
Sex (S)	Sex of the card holder

Markov Condition

Suppose we have a joint probability distribution P of the random variables in some set V and a DAG $\mathbb{G} = (V, E)$. We say that $(\mathbb{G}, P)$ satisfies the Markov condition if for each variable $X \in V$, X is conditionally independent of the set of all its nondescendants given the set of all its parents.

A A B B B B A B B

A B A B



Find all the nodes.
Find all the parents.
Find the descendants.
Now prove the following:

Exercise

Try the similar with the following DAG.