Foundations of Machine Learning: Week 1: Bayesian Networks

Professor Christopher Yau
christopher.yau@manchester.ac.uk

October 21, 2020

Foundations of Machine Learning: Week 1: Bayesian Networks
Professor Christopher Yau
christopher.yau@manchester.ac.uk
October 21, 2020

Bayesian Networks

A **Bayesian network** is a *probabilistic graphical model* that represents a set of variables and their conditional dependencies via a *directed acyclic graph* (DAG).

1. Collections of random variables and joint probability distributions
2. Probabilistic Graphical Models:
   2.1 Directed acyclic graphs
   2.2 Marginalisation and conditional probabilities on DAGs
3. Problems involving a Bayesian Network

Bayesian Networks

A **Bayesian network** is a *probabilistic graphical model* that represents a set of variables and their conditional dependencies via a *directed acyclic graph* (DAG).

1. Collections of random variables and joint probability distributions
2. Probabilistic Graphical Models:
   2.1 Directed acyclic graphs
   2.2 Marginalisation and conditional probabilities on DAGs
3. Problems involving a Bayesian Network

In this lecture, I will be discussing Bayesian networks.
These are probabilistic graphical models that represent a set of variables and their conditional dependencies via directed acyclic graphs or DAGs.
Before we do this, we will first examine collections of random variables and joint probability distributions.
Then I will discuss directed acyclic graphs and how marginalisation and conditional probabilities can be computed with the use of DAGs.
We will then consider some problems involving a Bayesian Network.

Joint probability distributions

A **joint probability distribution** (or *joint distribution*) is the probability of a collection of random variables.

For example, suppose we have three random variables X, Y and Z.

The joint distribution is defined by $p(X = x \cap Y = y \cap Z = z)$.

It is the probability that the r.v. X takes on the value x *AND* r.v. Y takes on the value y *AND* r.v. Z takes on the value z.

We often simply write $p(X, Y, Z)$ or $p(x, y, z)$ for convenience.

Joint probability distributions

A **joint probability distribution** (or *joint distribution*) is the probability of a collection of random variables.
For example, suppose we have three random variables X, Y and Z.
The joint distribution is defined by $p(X = x \cap Y = y \cap Z = z)$.
It is the probability that the r.v. X takes on the value x *AND* r.v. Y takes on the value y *AND* r.v. Z takes on the value z.
We often simply write $p(X, Y, Z)$ or $p(x, y, z)$ for convenience.

A complex probability model may contain many random variables with a variety of conditional dependencies.
A **joint probability distribution** or simply the *joint distribution* is the probability of a collection of random variables.
For example, suppose we have three random variables X, Y and Z.
The joint distribution is defined by the probability of $X = x$ and $Y = y$ and $Z = z$.

Joint probability distributions

x	y	z	$p(x,y,z)$
0	0	0	p_1
0	0	1	p_2
0	1	0	p_3
0	1	1	p_4
1	0	0	p_5
1	0	1	p_6
1	1	0	p_7
1	1	1	p_8
Total			1

Example

- If each random variable X, Y and Z is discrete and can take on q, r and s different values respectively
- The joint distribution defines probabilities for each of $q \times r \times s$ combinations of the actual values that the random variables can take.
- In this case, $q = r = s = 2$,
- There are 2^3 combinations.
- If X, Y and Z are independent then $p(x,y,z) = p(x)p(y)p(z)$.

Joint probability distributions

x	y	z	$p(x,y,z)$
0	0	0	p_1
0	0	1	p_2
0	1	0	p_3
0	1	1	p_4
1	0	0	p_5
1	0	1	p_6
1	1	0	p_7
1	1	1	p_8
Total			1

Example

- If each random variable X, Y and Z is discrete and can take on q, r and s different values respectively
- The joint distribution defines probabilities for each of $q \times r \times s$ combinations of the actual values that the random variables can take.
- In this case, $q = r = s = 2$,
- There are 2^3 combinations.
- If X, Y and Z are independent then $p(x,y,z) = p(x)p(y)p(z)$.

Now, suppose X can take on q possible values, Y r values and Z s different values then the total number of possibilities for all values of X, Y and Z is $q \times r \times s$.
Therefore the joint distribution contains q times r times s probabilities each corresponding to each combination of X, Y, and Z.
In this illustrated example, $q = r = s = 2$, so the variables are binary and there are $2^3 = 8$ combinations and therefore 8 values the joint distribution can take on.
If the random variables are independent then we would simply multiply the individual probabilities to get the joint probability but this is not the case in general so how do we describe the dependencies?

Probabilistic Graphical Models

A probability model may contain many variables with a variety of conditional dependencies which are determined by the application.

For example, suppose we have three random variables X, Y and Z whose joint distribution is defined by:

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

This says:

- the distribution of X only depends on Y and Z ($p(X|Y, Z)$)
- the distribution of Y only depends on Z ($p(Y|Z)$)
- the distribution of Z is independent of X and Y ($p(Z)$)

The ability to rewrite the joint distribution as a product of different conditional probabilities is known as **factorisation**.

Probabilistic Graphical Models

A probability model may contain many variables with a variety of conditional dependencies which are determined by the application.

For example, suppose we have three random variables X, Y and Z whose joint distribution is defined by:

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

This says:

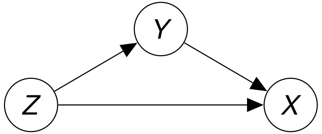
- the distribution of X only depends on Y and Z ($p(X|Y, Z)$)
- the distribution of Y only depends on Z ($p(Y|Z)$)
- the distribution of Z is independent of X and Y ($p(Z)$)

The ability to rewrite the joint distribution as a product of different conditional probabilities is known as **factorisation**.

The conditional dependencies between variables will generally be determined by the problem you are examining.
For example, suppose we have three random variables X, Y and Z whose joint distribution is defined by a product of conditional probabilities.
This says that the distribution of X only depends on Y and Z ($p(X|Y, Z)$). While the distribution of Y only depends on Z ($p(Y|Z)$) and the distribution of Z is independent of X and Y ($p(Z)$).
The ability to rewrite the joint distribution as a product of these conditional probabilities is known as factorisation.

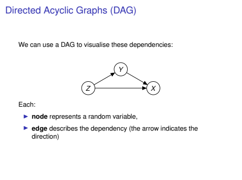
Directed Acyclic Graphs (DAG)

We can use a DAG to visualise these dependencies:



Each:

- **node** represents a random variable,
- **edge** describes the dependency (the arrow indicates the direction)



It can be convenient to visual these relationships using a directed acyclic graph.
This is the DAG corresponding the the factorisation we saw previously.
Each node represents one of our random variables X, Y and Z.
And each edge describes the dependency and the arrow indicates the direction.
So here, we have arrows from Y and Z to X because the probability of X depends on Y and Z.
While Z has no arrows going to it because its probability doesn't depend on anything else.

Try It Yourself

Before we move on, take a moment, to see if you can draw the DAG for this probability model:

$$p(A, B, C, D) = p(D|B, A)p(C|A)p(A|B)p(B)$$

PAUSE VIDEO

Try It Yourself

Before we move on, take a moment, to see if you can draw the DAG for this probability model:

$p(A, B, C, D) = p(D|B, A)p(C|A)p(A|B)p(B)$

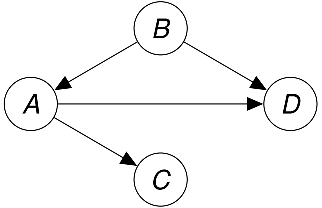
PAUSE VIDEO

Before we move on, take a moment, to see if you can draw the DAG for this probability model.
[PAUSE]

Solution

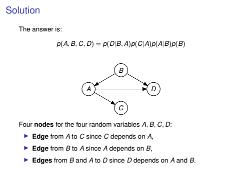
The answer is:

$$p(A, B, C, D) = p(D|B, A)p(C|A)p(A|B)p(B)$$



Four **nodes** for the four random variables A, B, C, D:

- **Edge** from A to C since C depends on A,
- **Edge** from B to A since A depends on B,
- **Edges** from B and A to D since D depends on A and B.



Your answer should look like the following.
We have four nodes to represent each of the four random variables.
The random variable D depends on A and B so we draw arrows from B and A to D.
C depends on A so we draw an arrow from A to C.
A depends on B so we draw an arrow from B to A.
The variable B does not depend on anything so there are no arrows into it.

Marginalisation on graphs

The total or marginal probability of any one variable:

$$P(D) = \sum_{A,B,C} p(A, B, C, D)$$

requires us to "average" out the effect of all the other variables.

Marginalisation on graphs

The total or marginal probability of any one variable:

$$P(D) = \sum_{A,B,C} p(A, B, C, D)$$

requires us to "average" out the effect of all the other variables.

Using DAGs can help us when doing computations.
Suppose we are interested in computing the total or marginal probability of any one variable.
This requires us to "average" the effect of all other variables, in this case, to get the total probability of D, we need to sum over the effect of A, B and C.

Solution

Lets do this directly:

$$p(D) = \sum_A \sum_B \sum_C p(A, B, C, D),$$
$$= \sum_A \sum_B \sum_C p(D|B, A)p(C|A)p(A|B)p(B),$$

Solution

Lets do this directly:

$$p(D) = \sum_A \sum_B \sum_C p(A, B, C, D),$$
$$= \sum_A \sum_B \sum_C p(D|B, A)p(C|A)p(A|B)p(B),$$

Lets solve this directly.
First we use the factorised form of the joint distribution.

Solution

Lets do this directly:

$$
\begin{aligned}
p(D) &= \sum_A \sum_B \sum_C p(A, B, C, D), \\
&= \sum_A \sum_B \sum_C p(D|B, A) p(C|A) p(A|B) p(B), \\
&= \sum_A \sum_B p(D|B, A) \underbrace{\left(\color{red}{\sum_C p(C|A)} \right)}_{\color{red}{=1}} p(A|B) p(B),
\end{aligned}
$$

Solution
Lets do this directly:
$p(D) = \sum_A \sum_B \sum_C p(A, B, C, D), = \sum_A \sum_B \sum_C p(D|B, A) p(C|A) p(A|B) p(B), = \sum_A \sum_B p(D|B, A) \underbrace{\left(\sum_C p(C|A) \right)}_{=1} p(A|B) p(B),$

Now we exchange the ordering of the summations and move the sum over C inside to the only term involving C which is the conditional probability of C given A.
This term is summed overall possible values of C but if we do this we know the answer must be one so the entire sum over C can be removed.

Solution

Lets do this directly:

$$
\begin{aligned}
p(D) &= \sum_A \sum_B \sum_C p(A, B, C, D), \\
&= \sum_A \sum_B \sum_C p(D|B, A)p(C|A)p(A|B)p(B), \\
&= \sum_A \sum_B p(D|B, A) \underbrace{\left(\sum_C p(C|A) \right)}_{=1} p(A|B)p(B), \\
&= \sum_A \sum_B p(D|B, A)p(A|B)p(B).
\end{aligned}
$$

Solution

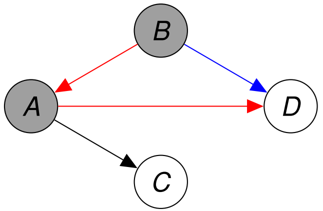
Lets do this directly:

$$
\begin{aligned}
p(D) &= \sum_A \sum_B \sum_C p(A, B, C, D), \\
&= \sum_A \sum_B \sum_C p(D|B, A)p(C|A)p(A|B)p(B), \\
&= \sum_A \sum_B p(D|B, A) \underbrace{\left(\sum_C p(C|A) \right)}_{=1} p(A|B)p(B), \\
&= \sum_A \sum_B p(D|B, A)p(A|B)p(B).
\end{aligned}
$$

This leaves us only with the summation and terms involving the variables A and B.

Solution

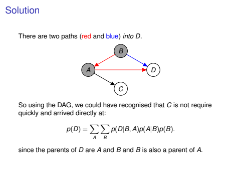
There are two paths (red and blue) *into* D.



So using the DAG, we could have recognised that C is not require quickly and arrived directly at:

$$p(D) = \sum_A \sum_B p(D|B, A)p(A|B)p(B).$$

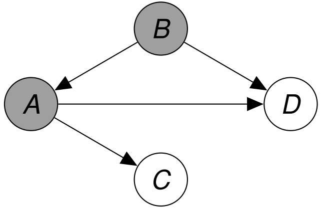
since the parents of D are A and B and B is also a parent of A.



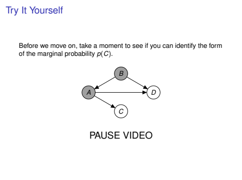
Now, if we had used the DAG, we could have arrived at this result more quickly. Notice there are two paths into D indicated here by red and blue lines. There is no path from C to D so we could have inferred quickly that C could be eliminated and that the total probability of D only depends on how much probability flows via A and B into D.

Try It Yourself

Before we move on, take a moment to see if you can identify the form of the marginal probability $p(C)$.



PAUSE VIDEO



Before we move on, see if you can identify the form of the marginal probability of C.
[PAUSE]

Solution

Using the direct approach:

$$p(C) = \sum_A \sum_B \sum_D p(A, B, C, D),$$

Solution
Using the direct approach:
$$p(C) = \sum_A \sum_B \sum_D p(A, B, C, D),$$

So, similar to before, we could try the direct approach and ignore the DAG. Once again, the total probability sums over A, B and D in the joint distribution.

Solution

Using the direct approach:

$$p(C) = \sum_A \sum_B \sum_D p(A, B, C, D),$$
$$= \sum_A \sum_B \sum_D p(D|B, A)p(C|A)p(A|B)p(B),$$

Solution

Using the direct approach:

$$p(C) = \sum_A \sum_B \sum_D p(A, B, C, D),$$
$$= \sum_A \sum_B \sum_D p(D|B, A)p(C|A)p(A|B)p(B),$$

We can factorise the joint distribution.

Solution

Using the direct approach:

$$
\begin{aligned}
p(C) &= \sum_A \sum_B \sum_D p(A, B, C, D), \\
&= \sum_A \sum_B \sum_D p(D|B, A)p(C|A)p(A|B)p(B), \\
&= \sum_A \sum_B \underbrace{\left(\sum_D p(D|B, A) \right)}_{=1} p(C|A)p(A|B)p(B), \\
&= \sum_A \sum_B p(C|A)p(A|B)p(B),
\end{aligned}
$$

Solution

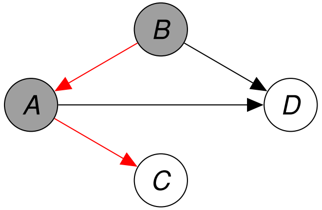
Using the direct approach:

$$
\begin{aligned}
p(C) &= \sum_A \sum_B \sum_D p(A, B, C, D), \\
&= \sum_A \sum_B \sum_D p(D|B, A)p(C|A)p(A|B)p(B), \\
&= \sum_A \sum_B \underbrace{\left(\sum_D p(D|B, A) \right)}_{=1} p(C|A)p(A|B)p(B), \\
&= \sum_A \sum_B p(C|A)p(A|B)p(B).
\end{aligned}
$$

Now we can rearrange the summations and bring the sum over D together with the only terms involving D.
Again, the summation sums over all possibilities of D which must add up to one so we can eliminate it.
This leaves us finally with the total probability of C dependent only on summing over A and B.

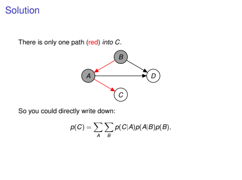
Solution

There is only one path (red) *into* C.



So you could directly write down:

$$p(C) = \sum_A \sum_B p(C|A)p(A|B)p(B),$$



Of course, using the DAG, we would have noticed that there is no path from D to C.
There is in fact only one path highlighted in red to C which flows through A from B.
So we could have written down directly that the total probability of C only depends on this path and summed over the random variables A and B.

Conditional probabilities on DAGs

How do we compute $p(C|B)$:

$$p(C|B) = \sum_A \sum_D p(A, C, D|B),$$

Conditional probabilities on DAGs

How do we compute $p(C|B)$:

$p(C|B) = \sum_A \sum_D p(A, C, D|B),$

Lets look now at how we compute conditional probabilities.
The conditional probability of C given B is the joint distribution of A, C and D given B with A and B averaged out by summation.

Conditional probabilities on DAGs

How do we compute $p(C|B)$:

$$p(C|B) = \sum_A \sum_D p(A, C, D|B),$$
$$= \sum_A \sum_D \frac{p(A, B, C, D)}{p(B)},$$

Conditional probabilities on DAGs

How do we compute $p(C|B)$:

$$p(C|B) = \sum_A \sum_D p(A, C, D|B),$$
$$= \sum_A \sum_D \frac{p(A, B, C, D)}{p(B)},$$

We know from Bayes Theorem that the conditional probability of A, C, D given B can be rewritten as the joint probability involving all the variables including B divided by the total probability of B

Conditional probabilities on DAGs

How do we compute $p(C|B)$:

$$\begin{aligned} p(C|B) &= \sum_A \sum_D p(A, C, D|B), \\ &= \sum_A \sum_D \frac{p(A, B, C, D)}{p(B)}, \\ &= \frac{1}{p(B)} \sum_A \sum_D p(D|B, A)p(C|A)p(A|B)p(B). \end{aligned}$$

Conditional probabilities on DAGs

How do we compute $p(C|B)$:

$$\begin{aligned} p(C|B) &= \sum_A \sum_D p(A, C, D|B), \\ &= \sum_A \sum_D \frac{p(A, B, C, D)}{p(B)}, \\ &= \frac{1}{p(B)} \sum_A \sum_D p(D|B, A)p(C|A)p(A|B)p(B). \end{aligned}$$

We can then rewrite the joint probability in its factorised form.

Conditional probabilities on DAGs

How do we compute $p(C|B)$:

$$
\begin{aligned}
p(C|B) &= \sum_A \sum_D p(A, C, D|B), \\
&= \sum_A \sum_D \frac{p(A, B, C, D)}{p(B)}, \\
&= \frac{1}{\cancel{p(B)}} \sum_A \underbrace{\sum_D p(D|B, A)}_{=1} p(C|A)p(A|B)\cancel{p(B)}, \\
&= \sum_A p(C|A)p(A|B).
\end{aligned}
$$

However, if we look at the DAG, there is only one path from B to C via A so we could have written this directly!

Conditional probabilities on DAGs

How do we compute $p(C|B)$:

$$p(C|B) = \sum_A \sum_D p(A, C, D|B) = \sum_A \sum_D \frac{p(A, B, C, D)}{p(B)} = \frac{1}{\cancel{p(B)}} \sum_A \underbrace{\sum_D p(D|B, A)}_{=1} p(C|A)p(A|B)\cancel{p(B)} = \sum_A p(C|A)p(A|B).$$

However, if we look at the DAG, there is only one path from B to C via A so we could have written this directly!

We can eliminate the two terms involving the probability of B and recognise that the inner sum over D equates to one.

This leaves us with the conditional probability of C given B as being the conditional path from B to C via A with A summed out.

If we look at the DAG, we coul d have deduced this immediately.

Conditional probabilities on DAGs

How do we compute $p(A|D)$:

$$\begin{aligned} p(A|D) &= \sum_B \sum_C p(A, B, C|D), \\ &= \sum_B \sum_C \frac{p(A, B, C, D)}{p(D)}, \\ &= \sum_B \sum_C \frac{p(D|B, A)p(C|A)p(A|B)p(B)}{p(D)} \end{aligned}$$

Conditional probabilities on DAGs

How do we compute $p(A|D)$:

$$p(A|D) = \sum_B \sum_C p(A, B, C|D), = \sum_B \sum_C \frac{p(A, B, C, D)}{p(D)}, = \sum_B \sum_C \frac{p(D|B, A)p(C|A)p(A|B)p(B)}{p(D)}$$

Now suppose we are now interested in the conditional probability of A given D.
We can proceed directly as before by rewriting the conditional probability as the conditional joint distribution of A, B, C given D with B and C summed out.
Then, in the second line, we can apply Bayes Theorem to rewrite this as the full joint distribution divided by the total probability of D.
Which we can expand out using the factorisation of the joint distribution in line 3.

Conditional probabilities on DAGs

How do we compute $p(A|D)$:

$$
\begin{aligned}
p(A|D) &= \sum_B \sum_C \frac{p(D|B,A)p(C|A)p(A|B)p(B)}{p(D)}, \\
&= \frac{1}{p(D)} \underbrace{\sum_C p(C|A)}_{=1} \underbrace{\sum_B p(D|B,A)\,\underbrace{p(A|B)p(B)}_{=p(A,B)}}_{p(D,A,B)}, \\
&= \frac{1}{p(D)} \sum_B p(D|B,A)p(A|B)p(B),
\end{aligned}
$$

Note, using the DAG we could have removed C immediately since, given D, A only depends on D via B and not C.

Conditional probabilities on DAGs

How do we compute $p(A|D)$:

$p(A|D) = \sum_B \sum_C \frac{p(D|B,A)p(C|A)p(A|B)p(B)}{p(D)}, = \frac{1}{p(D)} \underbrace{\sum_C p(C|A)}_{=1} \underbrace{\sum_B p(D|B,A)\,\underbrace{p(A|B)p(B)}_{=p(A,B)}}_{p(D,A,B)}, = \frac{1}{p(D)} \sum_B p(D|B,A)p(A|B)p(B).$

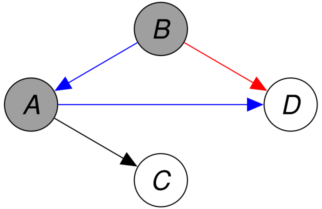
Note, using the DAG we could have removed C immediately since, given D, A only depends on D via B and not C.

Now rearranging the order of the summations we can see that we can eliminate C since the only term involving C is the conditional of C given A and summing over C gives us one.
This gives us the final expression for the desired conditional involving A, B and D only.
How do we interpret this?

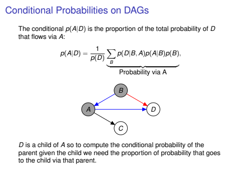
Conditional Probabilities on DAGs

The conditional $p(A|D)$ is the proportion of the total probability of D that flows via A:

$$p(A|D) = \frac{1}{p(D)} \underbrace{\sum_B p(D|B,A)p(A|B)p(B)}_{\text{Probability via A}},$$



D is a child of A so to compute the conditional probability of the parent given the child we need the proportion of probability that goes to the child via that parent.



Note that the expression involves a ratio where the denominator is the total probability of D.
We can see that the numerator sums out B so we can think of it as the contribution to the total probability of D that flowed through A.
Therefore the conditional of A given D is the proportion of the total probability of D that came via A.
It is possible to recognise this directly from the DAG by realising that A is a parent of D and not a child.
Therefore when we compute the probability of a parent given its child, we need to compute the proportion of the total probability of the child that came via that parent.

Summary

So to recap:

- Use directed acyclic graphs to represent the conditional dependencies between random variables,
- Shown how DAGs can help us to do marginalisation and conditional probability computations more quickly,
- Next, we will consider Bayesian Network examples.

Summary

So to recap:

- Use directed acyclic graphs to represent the conditional dependencies between random variables,
- Shown how DAGs can help us to do marginalisation and conditional probability computations more quickly,
- Next, we will consider Bayesian Network examples.

So to summarise, we have learnt how to use directed acyclic graphs or DAGs to represent complex conditional dependencies between random variables in probability models.

I have discussed how DAGs help us to do marginalisation and conditional probability computations more quickly by giving us visualisations of the model.

In the next lecture, we will consider a specific Bayesian Network numerical example.

END LECTURE

END LECTURE