09-09-24.

22AIE301 Probabilistic Reasoning

Credits : 3.

Continuous Evaluation Pattern

* 2 Quizzes (2×10 = 20)

* 30 marks Assignment.
  - 15 → Assgn. 1 (Paper study) (Viva)
  - 15 → Assgn. 2 (Exam/Quiz)

* Mid term → 20 marks.

* End sem → 30 marks.

Sources.

* Christopher M. Bishop
  "Pattern Recognition & ML"

Main Topics

* Sampling Methods.
* Probabilistic Graphic Models (PGM).

- Spam classification
  A type of prob. distribution.

- Total no. of possibilities or parameters = 2^{n+1}

- learning parameter of a distribution. (Task 1)

  ⇓

  We have some data but we don't know from which underlying prob. dist. comes from, then we have to learn those parameters (θ). MLE is a method used for it. (Max. Likelihood Estimation).

- Inference (Task 2)

  ⇓

  Once we learn those parameters, we need to know on how to use them for classification purpose / required task.

In task 1, we model P_{true} by P_θ, where P_{true} is the actual data.

In task 2,

$$P_{\hat{\theta}}(y=1 \mid \underset{x_1 \dots x_n}{\text{"doc"}})$$

which is the inference.

"Joint distribution of y & x";

$$P(y \& x) = P(y \mid \hat{x}) \cdot p(x).$$

* Marginal inference :
From joint distribution (of y & x ?) we find ~~to~~ the inference / we find the marginal of some variable

* Arg max $P(x_1 \dots x_n \mid y=1)$
$x_1 \dots x_n$ <MAP inference>

for e.g. Suppose there are a certain no: of words what is the prob. that it is spam (OR) Given a doc. is spam what the max. no. of words in it.

- Modeling / Representation (Task 3) :

The one in which we efficiently model / rep. the data (joint data).

$$p(x_1 \dots x_n, y) = p(x_1 \dots x_n \mid y) \cdot p(y)$$

Assume :

$$= p(x_1 \mid y) \cdot p(x_2 \mid y) \dots p(x_n \mid y) \cdot p(y)$$

$x_1 \dots x_n$ are independent given y.
(Naive Bayes Classifier)

$O(n)$

We are doing this modeling inorder to make the problems more linear or simple so that it is more tractable.

The Graphical repre. helps in forming a proper inference & then it will in turn help in creating algorithms that can be useful in performing several tasks.

In "marginal inference",
$P(y, x_1, x_2, \dots x_n)$
$y \rightarrow$ cost of house
$x \rightarrow$ features such as no: of bedrooms, floors, etc.
& the above is a joint distribution. Suppose x_1 says about the bedrooms. & we are asked some qs. on x_1. ∴ From the joint dist., we are only answering about x_1 which comes under marginal distribution / inference.

* If the no: of parameters (2^{n+1}) is very large compared to the training examples / data then there occurs the problem of "poor generalisation" & "computational issues."

⇓

* To overcome this we make a "conditional independence" assumption.

* Naive Bayes Assumption

→ In task 3, we look into:
- Graphical models &
- Conditional independence.

Bayesian Networks.

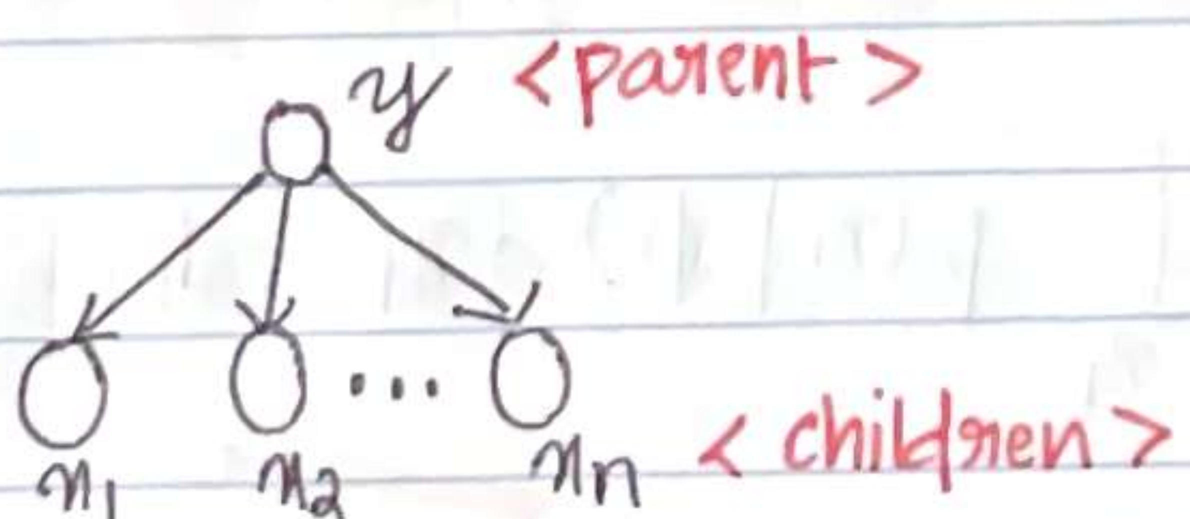
Represent $P(\bar{x}, y)$ by directed graphs.

$$P(x_1 \ldots x_n | y) = P(x_1|y) \cdot P(x_2|y) \ldots P(x_n|y)$$

Conditional Probability

$$P(\bar{x}, y) = P(\bar{x}|y) \cdot P(y)$$
$$= P(y) \cdot P(x_1|y) \ldots P(x_n|y)$$

Conditional Independence.



Arrow goes from parent to child.

Suppose, $x_1 \longrightarrow x_2$

then,

$$P(x_1, x_2) = P(x_1) \cdot P(x_2 | x_1)$$

Likewise, $x_1 \xrightarrow{\text{no edge}} x_2$

$$P(x_1, x_2) = P(x_1) \cdot P(x_2)$$

* Suppose x_1 & x_2 are discrete value that take one of K values. i.e. if we take binary then $K = 2$. but we don't exactly care what the K values are in specific.

We can get the x_1 value using, 1-of-K-representation OR one-hot. which means if we assign a value to x_1 at a particular position then all others are zero. & is mathematically represented as (for any discrete valued random variable):-

$$P(x_1) = P_1^{x_1^{(1)}} \cdot P_2^{x_1^{(2)}} \cdot \ldots \cdot P_K^{x_1^{(K)}}$$

* Independent parameter ⇒ $K-1$.

∴ for $x_1 \longrightarrow x_2$

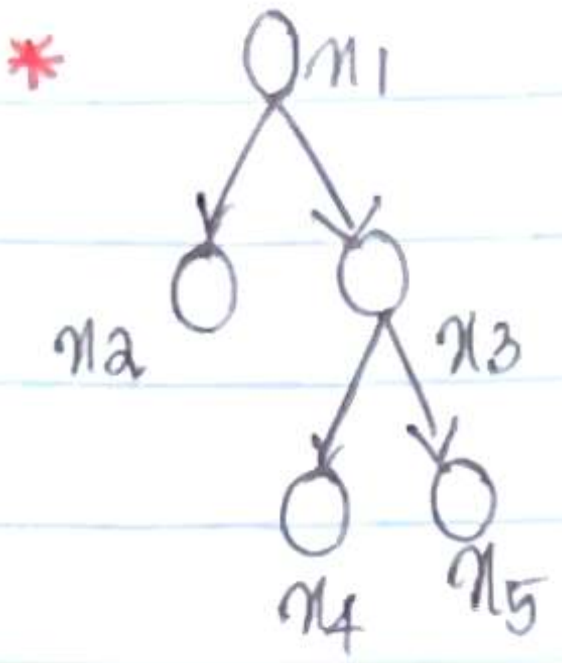
its $(K-1) + K \cdot (K-1)$

$= K^2 - 1$.

& for x_1 x_2, its $(K-1) + (K-1)$

$= 2(K-1)$

Summary ⇒ Making more conditional independence assumptions reduces no: of model parameters, but the class of distributions ~~representations~~ represented becomes restricted.

* 

$P(x_1, x_2 \dots x_5)$
$= P(x_1) \cdot P(x_2 | x_1) \cdot P(x_3 | x_1) \cdot P(x_4 | x_3) \cdot P(x_5 | x_3)$

* Generally,
$$P(x_1, \dots x_k) = \prod_{i=1}^{k} P(x_i \mid \underbrace{pa(x_i)}_{\text{set of parents}})$$

* The probability distribution should not be negative & the integration should be equal to one.

$\rightarrow p(\bar{x}) \geq 0.$

$\rightarrow \int p(\bar{x})\, d\bar{x} = 1.$

※ Q. When will $\prod_{i=1}^{k} P(x_i \mid pa(x_i))$ integrate to 1? Only if this happens then we can say it is a valid assumption.

Assumption A: Assume, (without any loss of generality) that all the directed edges go from lower-indexed to higher-indexed nodes.

Then,

It integrates to 1 if each factor integrates to one.

$$\int_{x_1, x_2, \dots x_k} \prod_{i=1}^{k} P(x_i \mid pa(x_i))\, dx_1\, dx_2 \dots dx_k$$

$\Downarrow$

need to prove this is equal to one.

Doing individually w.r.t. x_k;

$$\int_{x_1 \dots x_{k-1}} \underbrace{\prod_{i=1}^{k-1} P(x_i \mid pa(x_i))}_{\text{this doesn't depend on } k.} \left[\int \underbrace{p(x_k \mid pa(x_k))\, dx_k}_{} \right] dx_1\, dx_2 \dots dx_{k-1}$$

We know this integrates to 1 & to apply that we need to ensure that the other parts are not involved with x_k.

$Pa(x_i)$ only involves nodes with index $\leq k-1$.

$\therefore$,

We are left with,

$$\int_{x_1, \dots x_{k-1}} \prod_{i=1}^{k-1} p(x_i \mid pa(x_i))\, dx_i \dots dx_{k-1}.$$

The same arguments keep repeating, until we end up with:

$$\int_{x_1} p(x_1 \mid \phi)\, dx_1 = 1.$$

Q. What does the Assumption A mean in terms of the graph?

Claim $\rightarrow$ If parents have lower index than the child, then the directed graph is "acyclic"!!

→ As long as the graph is a directed graph that is acyclic then the probability is valid.

→ Assumption A ⇒ No cycles.
(implies)

⇒ There is a cycle ⇒ Assump. A invalidated
< a contrapositive stmnt. >

Summary 2
∴ The Bayesian n/w is always a DAG.
& factorisation represented by a DAG is a valid probability distribution

From the above we can further say that, as long as the factors when integrated gives '1', making it valid then their products will also be valid.

e.g. $\underset{x_1}{O} \longrightarrow \underset{x_2}{O}$

$$\sum_{x_2} p(x_2 | x_1) = 1$$

&

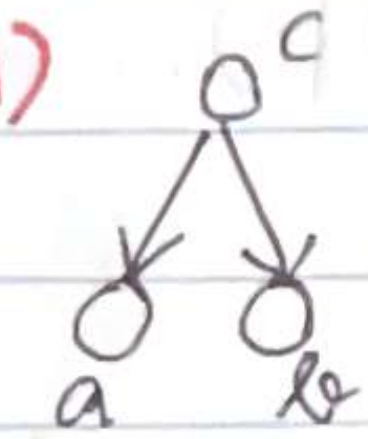
$$\sum_{x_1} p(x_1) = 1 \ ;$$

then $\sum_{x_1 x_2} p(x_1) \cdot p(x_2|x_1) = 1.$

This whole thing/assump. is an example of CI.

Goal ⇒ Find the set of CI (Cond. Independence) properties satisfied by a given graph.

In order to find the goal let's look into some basic graphs;

i) 

$$^{*}p(a|b, c) = p(a|c)$$

$a \perp\!\!\!\perp b \mid c$

means a is independent of b condition c.

Coz a doesn't depend upon b. Thus when c is observed, a doesn't depend upon b, showing CI. This equation basically shows/defines CI.

Similarly,

As we saw earlier

$O_{x_1} \longrightarrow O_{x_2}$.

$$P(x_1, x_2) = P(x_1) \cdot P(x_2)$$

< Here it is conditioned on ϕ >.

Thus,

Same

$$^{*}p(a, b | c) = p(a|c) \cdot p(b|c)$$

OR

$$P_c(a, b) := p(a, b | c)$$

OR

$$P_c(a, b) = P_c(a|b) \ P_c(b)$$
$$= P_c(a) \ P_c(b).$$

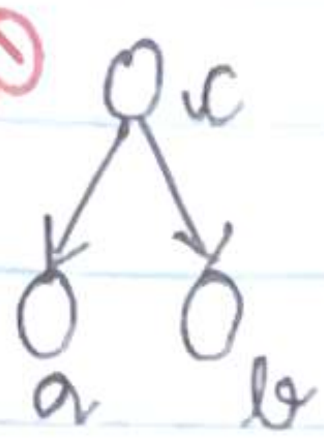
< Just use one for defining CI, everything is basically same >.

< The * ones are the main definitions used >.

< The above points are generalised version, not base exactly on the graph >

lets come back to the graph,

graph①



$$p(a,b,c) = p(c)\cdot p(a|c)\cdot p(b|c). \quad -①$$

$$p(a,b|\phi) \stackrel{?}{=} p(a,b) \stackrel{?}{=} p(a|\phi)\cdot p(b|\phi) \stackrel{?}{=} p(a)\cdot p(b). \quad -②$$

From ①, how do we get ② ?

Doing marginal inference,

$$\sum_c p(a,b,c) = p(a,b).$$

⇓

Here since we need to eliminate c we are taking it like this. Basically we are marginalizing w.r.t. c.

$$p(a,b) = \sum_c p(a,b,c)$$

$$= \sum_c p(a|c)\cdot p(b|c)$$

$$\neq p(a)\cdot p(b)$$

∴ For this graph,

$$p(a,b|\phi) \stackrel{?}{\neq} p(a|\phi)\cdot p(b|\phi)$$

∴ In general, is not unconditionally independent.

Now,

$$p(a,b|c) \stackrel{?}{=} p(a|c)\,p(b|c).$$

Now does this work ?

$$\therefore,\ p(a,b|c) = \frac{p(a,b,c)}{p(c)}$$

Basic conditional probability.

$$= \frac{p(c)\,p(a|c)\,p(b|c)}{p(c)}$$

Got from ①

$$= \underline{p(a|c)\,p(b|c)}.$$

Thus, we can say

$a \perp\!\!\!\perp b|c$ & $a \not\perp\!\!\!\perp b\,|\,\phi$.

We can also say that, the path of a & b is $\underbrace{\text{tail-to-tail}}$ w.r.t. node c.

saying it here coz these wordings are commonly used in graphical representations

We will also say like, when c is observed the path from a to b is blocked.

⇑

All this in relation to this graph. ①.

OMIT

ii) Graph ② $a \rightarrow c \rightarrow b$

Here, c is head-to-tail w.r.t path from a to b.

$$p(a,b,c) = p(a) \cdot p(c|a) \cdot p(b|c)$$

$$p(a,b|\phi) \stackrel{?}{=} p(a) \cdot p(b).$$

$$p(a,b|\phi) = \sum_c p(a,b,c)$$

$$= p(a) \sum_c p(b|c) p(c|a)$$

Taken out as it doesn't depend on c.

$$= p(a) \cdot p(b|a).$$
$$\neq p(a) \cdot p(b).$$

Thus, $a \not\perp\!\!\!\perp b \mid \phi$.
<Not unconditionally independent>.

Now,
$$p(a,b|c) \stackrel{?}{=} p(a|c)\, p(b|c).$$

$$p(a,b|c) = \frac{p(a,b,c)}{p(c)}$$

$$= \frac{p(a)\, p(c|a)\, p(b|c)}{p(c)}$$

$$< \frac{p(a)\, p(c|a)}{p(c)} = p(a|c) >$$

$$= p(a|c)\, p(b|c)$$

Thus, $a \perp\!\!\!\perp b \mid c$.

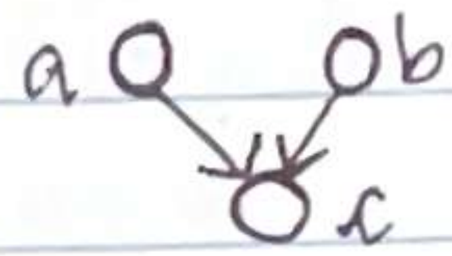
$$< \frac{p(a) \cdot p(c|a)}{p(c)} = \frac{p(a)}{p(c)} \cdot \frac{p(c \cap a)}{p(a)}$$

$$= \frac{p(c \cap a)}{p(c)}$$

$$= p(a|c)$$

This $\cap$ is basically the ',' we need before indicating the joint distribution $\cap$ is used in case for sets. >

iii) Graph ③



$$~~p(a|b)~~\; p(a,b|\phi) \stackrel{?}{=} p(a) \cdot p(b)$$
$$p(a,b|c) \stackrel{?}{=} p(a|c)\, p(b|c).$$

$$p(a,b|\phi) = \sum_c p(a,b,c).$$

$$p(a,b,c) = p(a) \cdot p(b) \cdot p(c|a,b)$$

(continuing)

$$p(a,b|\phi) = p(a) \cdot p(b) \underbrace{\sum_c p(c|a,b)}_{\text{This will be equal to 1.}}$$

$$= p(a) \cdot p(b).$$

Thus, $a \perp\!\!\!\perp b \mid \phi$. i.e. they are unconditionally independent.

$$p(a,b|c) = \frac{p(a,b,c)}{p(c)} = \frac{p(a) \cdot p(b) \cdot p(c|a,b)}{p(c)}$$
$$\neq p(a|c) \cdot p(b|c).$$

Thus, $a \not\perp\!\!\!\perp b \mid c$.

Question ⇒ B (state of battery);
F (state of fuel tank);
G (state of fuel guage).

$B, F, G = 0$ or 1
$0 \downarrow$ Bad, $1 \downarrow$ Good <In general>

B: 0 Battery drained. 1 " charged.

F: 0 Tank empty. 1 " Full.

G: 0 Indicator poor. 1 " good.

B, F are independent.

$P(B=1) = 0.9$

$P(F=1) = 0.9$

$P(G=1 | B=1, F=1) = 0.8$

$P(G=1 | B=1, F=0) = 0.2$

$P(G=1 | B=0, F=1) = 0.2$

$P(G=1 | B=0, F=0) = 0.1$

Answer :-

i) Find $P(F=0 | G=0)$.

From above,

$$P(F=0) = 0.1.$$

$$P(F=0 | G=0) = \frac{P(G=0 | F=0)\, P(F=0)}{P(G=0)}$$

$$P(G=0 | F=0) = \frac{P(G=0, F=0)}{P(F=0)}$$

$$= \sum_{b=0,1} P(G=0 | F=0, B=b)\, P(B=b) \quad (1)$$

Just trying to add B into it coz that is the form we have in qs.

$$P(G=0 | F=0, B=b) \times P(B=b)$$

$$= \frac{P(G=0, F=0, B=b)}{P(F=0, B=b)} \times P(B=b)$$

$$= \frac{P(G=0, F=0, B=b)}{P(F=0)\, P(B=b)} \times P(B=b)$$

given F & B are independent.

$$= \frac{P(G=0, F=0, B=b)}{P(F=0)}$$

Now we are gon doing a sum over b & thus due to marginalisation we will get;

$$= \sum_b \left[\frac{P(G=0, F=0, B=b)}{P(F=0)} \right]$$

$$= \frac{P(G=0, F=0)}{P(F=0)}.$$

The above was done to show that incorporating B doesn't cause any problems.

$\therefore$ (1) $P(G=0 | F=0) =$

$P(G=0 | F=0, B=0) \cdot P(B=0) +$

$P(G=0 | F=0, B=1) \cdot P(B=1)$

$$= 0.9 \times 0.1 + 0.8 \times 0.9$$

$$= 0.09 + 0.72$$

$$= \underline{0.81}.$$

Similarly now we need to find,

$P(G=0)$.

$$P(G=0 \mid F=f, B=b) \cdot P(B=b) \cdot P(F=f)$$

$$= \frac{P(G=0, F=f, B=b)}{P(F=f, B=b)} \cdot P(B=b) \cdot P(F=f)$$

$$= \frac{P(G=0, F=f, B=b)}{\cancel{P(F=f)} \cdot \cancel{P(B=b)}} \cdot \cancel{P(B=b)} \cdot \cancel{P(F=f)}$$

Marginalizing over both f & b.

$$= \Sigma_b \Sigma_f \, P(G=0, B=b, F=f).$$

$$= \underline{\underline{P(G=0)}}.$$

$$\Rightarrow 0.2 \times 0.9 \times 0.9 + 0.8 \times 0.9 \times 0.1 + 0.8 \times 0.1 \times 0.9 + 0.9 \times 0.1 \times 0.1$$

Applying to all the given prob.

$$= 0.162 + 0.072 + 0.072 + 0.009$$

$$= \underline{0.315}.$$

$\therefore$

$$P(F=0 \mid G=0) = \frac{P(G=0 \mid F=0) \; P(F=0)}{P(G=0)}$$

$$= \frac{0.81 \times 0.1}{0.315}$$

$$= \underline{\underline{0.257}}$$

< "> 0.1" >

ii) "$P(F=0 \mid G=0, B=0)$" we give battery is drained.

Ans: ~~P(B~~ $P_B(F=0 \mid G=0)$

$$= \frac{P_B(G=0 \mid F=0) \cdot P_B(F=0)}{P_B(G=0)}$$

$\therefore P_B(G=0 \mid F=0)$

$$= P(G=0 \mid F=0)(F=0)(B=0) + P(G=0 \mid F=1)(F=1)(B=0)$$

$$= 0.9 \times 0.1 \times 0.1 + 0.8 \times 0.9 \times 0.1$$

$$= \underline{\underline{0.081}}.$$

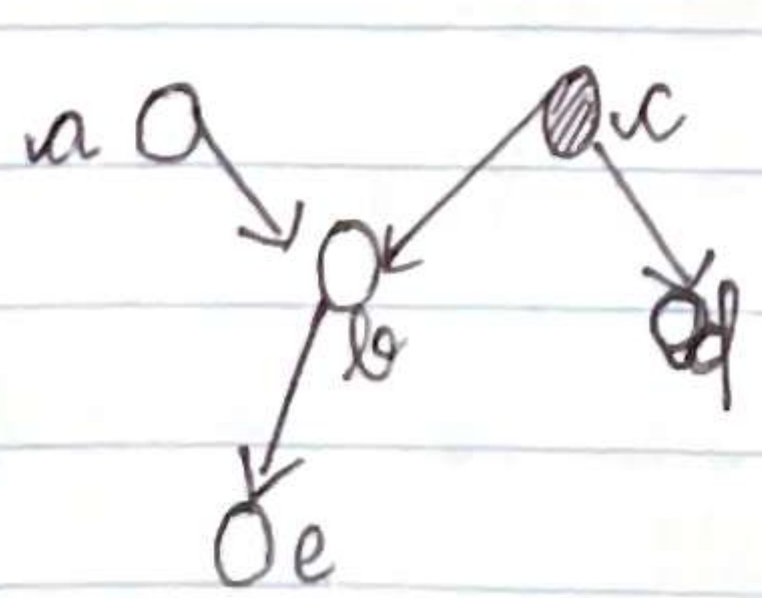
* "When will we have $A \perp\!\!\!\perp B \mid C$?"

- First consider $P = \{$ All paths from any node in A to any node in B $\}$

- We will say a path is blocked if either :-
  - i) The path involves a tail-to-tail or head-to-tail node in C which is observed.
  - ii) The path involves a head-to-head node ~~in C~~ & neither this node nor any of its descendants are observed in C.

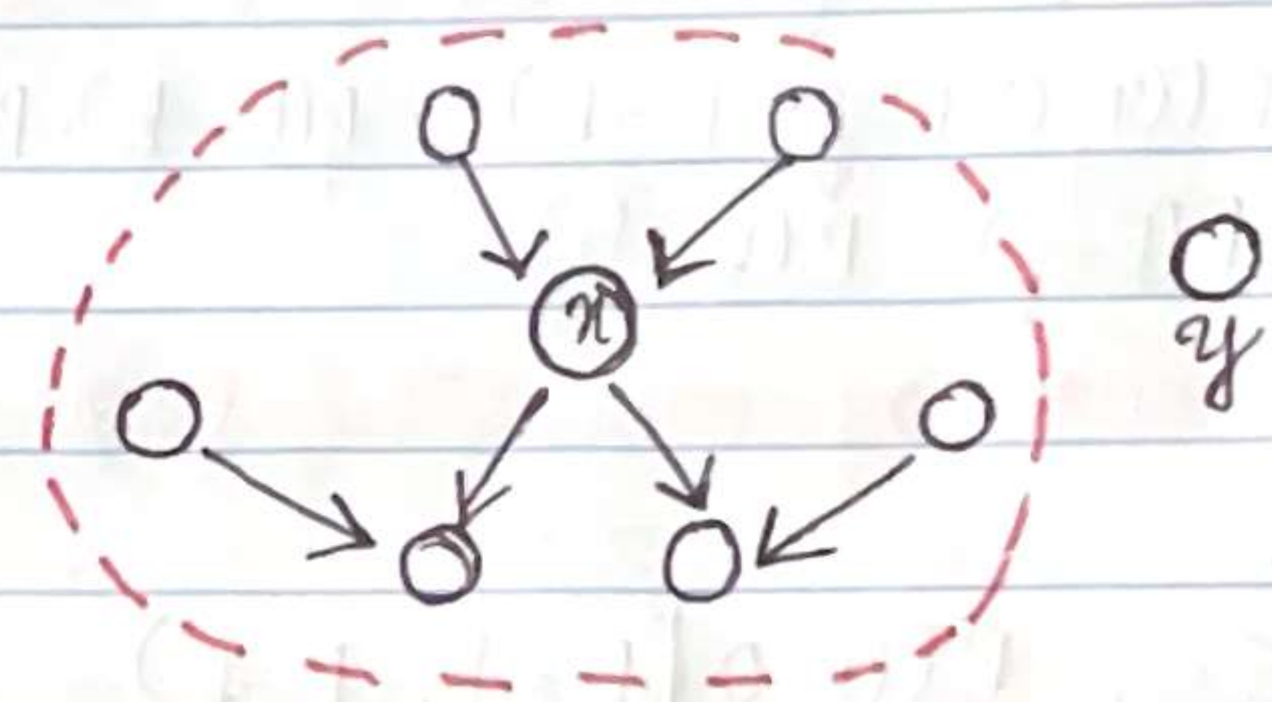
- A, B, C - group of vertices in the graph.

- So based on the above conditions we can draw the conclusion that $a \perp\!\!\!\perp d \mid c$ or $b \perp\!\!\!\perp d \mid c$, etc. for the below graph.



- We should be able to tell the conditional independencies when a graph is given & this can be done using the given 2 conditions. This method is called "d-separation."

Q. Show that $x \perp\!\!\!\perp y$ using d-sep.



$C_x = \{$ parents of x, children of x, co-parents of x $\}$.

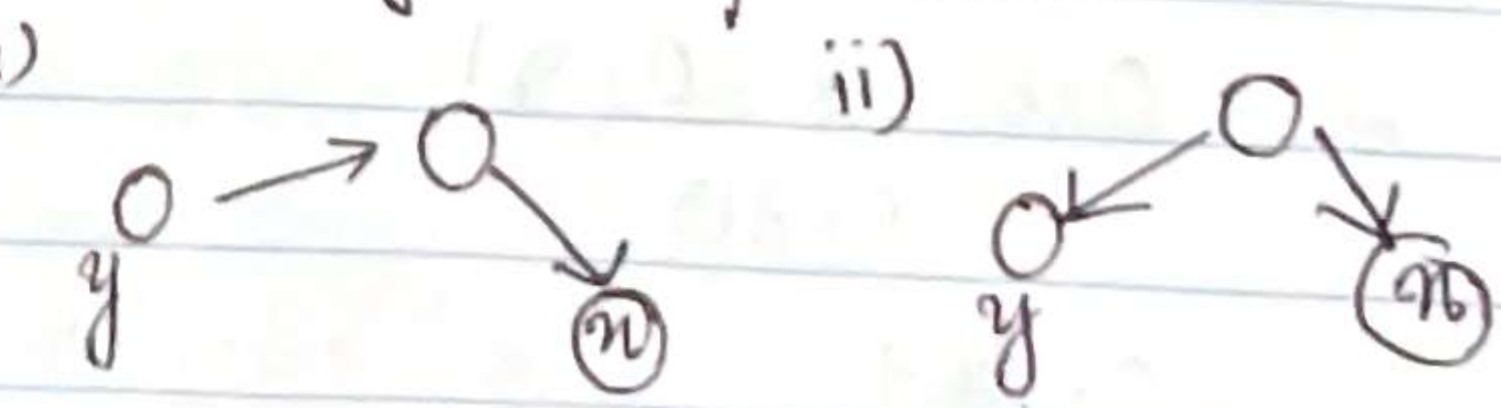
$P(x \mid$ nodes other than x in graph$)$
$= P(x \mid C_x)$

<This is like a Markov property>
< C_x is called the Markov blanket of x >

let $A = \{x\}$ & $B = \{y\}$

y can be connected x in 2 ways:- via parents :-

i) ii)



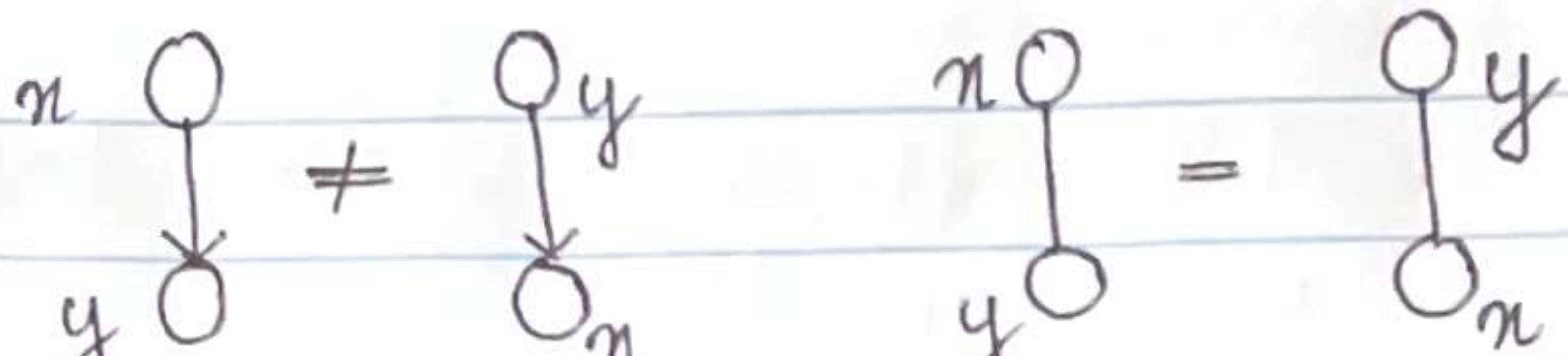
Both these ways satisfy condition 1.

$\therefore$ All paths via parents of x are blocked.

Now let's look into path via children of x.

Undirected Graphical Models (Markov n/w's)

- Correlation doesn't mean causation. But causation means / implies correlation.

- 

- Here,

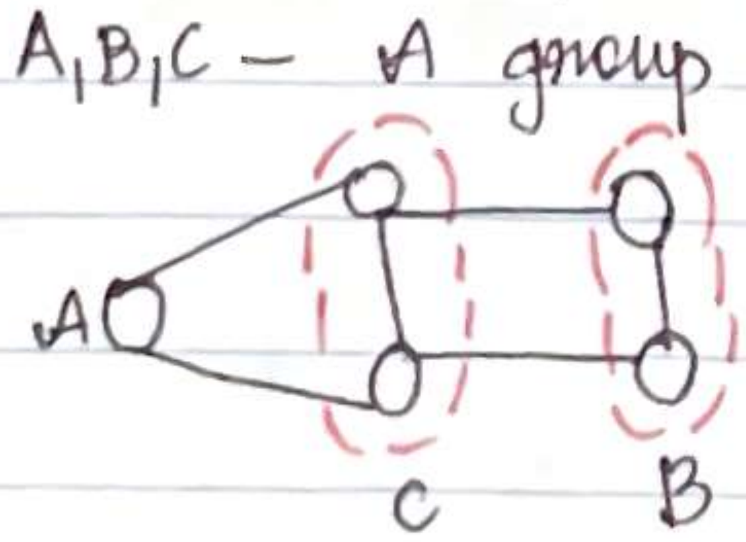
  i) Conditional independence will be discussed 1st unlike directed graphs where factorization was done 1st & then conditional independence. <d-sep & all>

  ii) Secondly we will do, factorization that corresponds to the definition of conditional independence.

- Conditional indep.?

  i) Can be defined by "graph separation"

  A, B, C – A group of nodes.



If we remove nodes in C, & there is no path b/w a node in A to any node in B, the "$A \perp\!\!\!\perp B \mid C$."

< So in the example graph, now what will be the factorization? >

< when causality doesn't matter much we use undirected graphical models.
when causality is also of importance we use directed models.
e.g. Auto-encoders >

Let,

$x \rightarrow$ set of all nodes

$x_i, x_j \rightarrow$ Nodes in graph with no direct edge b/w them.

$\therefore P(x_i x_j \mid x - \{x_i, x_j\})$ &

$A = \{x_i\}$

$B = \{x_j\}$

$C = x - \{x_i, x_j\}$.

When we remove / if we remove all the nodes in the graph except x_i & x_j, there is as given no direct edge b/w them, thus $A \perp\!\!\!\perp B \mid C$.

$x_i \perp\!\!\!\perp x_j \mid x$ / all other nodes.

OMIT

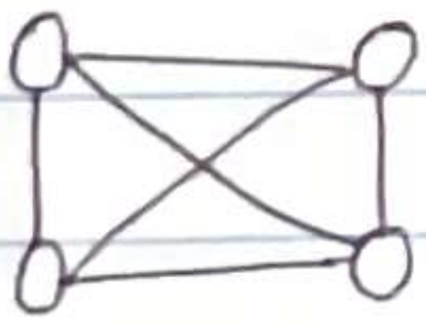
Thus,

$$P(x_i, x_j \mid x - \{x_i, x_j\})$$

$$= P(x_i \mid x - \{x_i, x_j\}) \cdot P(x_j \mid x - \{x_i, x_j\}).$$

The factorisation should be such that the nodes that doesn't have a direct edge b/w them should appear in/as 2 separate factors.

Clique → Subset of nodes where each pair is connected.



$\{x_1, x_2\}$
$\{x_1, x_2, x_3\}$
$\{x_2, x_3\}$
} All are cliques!

$\{x_2, x_3, x_4\}$ → Not a clique.

Maximal cliques → Ones with largest size. $\{x_1, x_2, x_3\}$ & $\{x_1, x_3, x_4\}$.

Let x_c be clique variables,

$$P(x_1, x_2, \dots x_D) = \prod_{\text{maximal clique graph } C} \psi(\bar{x}_c)$$

→ From the above we can understand that x_2 & x_4 needs to be 2 separate factors while writing the factorization.

Also,

$$\psi_c(\bar{x}_c) > 0.$$

Z → normalisation factor.

$$Z = \underbrace{\sum_{\bar{x}_c} \cdot \prod_c \cdot \psi_c \cdot (\bar{x}_c)}_{\text{this will be one.}}$$

∴,

$$P(x_1, x_2, \dots x_D) = \frac{1}{Z} \prod_c \psi_c(\bar{x}_c).$$

ψ_c → potential function.
Z → partition function.

$$\psi_c(\bar{x}_c) = e^{-E_c(\bar{x}_c)}$$ } +ve func.

E_c → Energy function.

If the energy func. is large, then we get a lower value as o/p.

Directed models are used in the area of generative models similarly an app. of undirected graphical modelling would be image denoising.

$x_i \rightarrow$ Pixels in a image $\in \{+1, -1\}$

- With a probability p, we flip the sign of x_i to get a noisy pixel y.

Goal $\rightarrow$ Given noisy img. y_i estimate x_i.

- When noise is low (i.e. p is small), y_i & x_i are highly correlated.

OR

When the noise is small that means the image is more similar to that of clean image & the not much flipping is required, they the clean & noisy img. values will be correlated.

- Neighbouring pixels in x are also highly correlated.

Some topics ?

* Markov blanket

* Comparison b/w directed & undirected graphical models.