Johns Hopkins University
EN. 601.475/675 Machine Learning

EM & Graphical Models

Xuan Zhang

Nov. 18. 2022

I. EM

- EM
- GMM
- EM for a GMM
- K-means as a special case of EM

II. Graphical Models

- Terminology
- D-separation

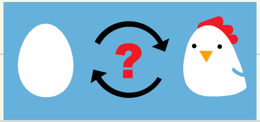
I. EM

The EM Algorithm

X : observed data

Z : hidden variable

θ : model parameters



e.g. $Z \sim Ber(\theta)$

$Z = \theta^T X$

	k-means	GMM
Z	cluster assignments	responsibility
θ	μ	μ, Σ, π

E step:

fix θ, <u>return</u> Z or $q(Z)$

M step:

fix Z, <u>find</u> θ that maximizes likelihood

The EM Algorithm

Goal:

$$\max \sum_{n=1}^{N} \log p(x_n | \theta)$$ i.e. maximize the log likelihood of the observed data

find a θ

Alternatively:

$$\max \sum_{n=1}^{N} \log \sum_{z_n} p(x_n, z_n | \theta)$$ i.e. maximize the complete data log likelihood

Decomposition of Complete data log likelihood

Let's introduce a set of arbitrary distributions $q_n(z_n)$ over each hidden variable z_n.

$$\log \sum_{z_n} P(x_n, z_n | \theta)$$

$$= \underbrace{\boxed{\sum_{z_n} q_n(z_n) \log \frac{P(x_n, z_n | \theta)}{q_n(z_n)}}}_{\mathcal{L}(q, \theta)} - \underbrace{\boxed{\sum_{z_n} q_n(z_n) \log \frac{P(z_n | x_n, \theta)}{q_n(z_n)}}}_{-KL(q \| P)}$$

$$= \sum_{z_n} q_n(z_n) \left[\log \frac{P(x_n, z_n | \theta)}{q_n(z_n)} - \log \frac{P(z_n | x_n, \theta)}{q_n(z_n)} \right]$$

$$= \sum_{z_n} q_n(z_n) \log P(x_n | \theta) = \log P(x_n | \theta)$$

* KL Divergence: $D_{KL}(q \| p) = \sum q \log \frac{q}{p}$

Decomposition of Complete data log likelihood

$$\log p(X_n, \theta) = \log \sum_{Z_n} p(X_n, Z_n \mid \theta)$$

$$= \underbrace{\boxed{\sum_{Z_n} q_n(Z_n) \log \frac{P(X_n, Z_n \mid \theta)}{q_n(Z_n)}}}_{\mathcal{L}(q,\theta)} + \underbrace{\boxed{-\sum_{Z_n} q_n(Z_n) \log \frac{P(Z_n \mid X_n, \theta)}{q_n(Z_n)}}}_{KL(q \| P)}$$

$$\mathcal{L}(q,\theta) = \sum_{Z_n} q_n(Z_n) \log \frac{P(X_n, Z_n \mid \theta)}{q_n(Z_n)} \leq \log \sum_{Z_n} q_n(Z_n) \frac{P(X_n, Z_n \mid \theta)}{q_n(Z_n)} = \log P(X_n \mid \theta)$$

equal when $KL(q \| p) = 0$, i.e. $q_n(Z_n) = P(Z_n \mid X_n, \theta)$

* Jensen's Inequality: For a concave function f, $\sum_i \lambda_i f(X_i) \leq f(\sum_i \lambda_i X_i)$

Understand EM with Decomposition

$$\log p(X_n, \theta) = \log \sum_{Z_n} p(X_n, Z_n \mid \theta)$$

$$= \underbrace{\boxed{\sum_{Z_n} q_n(Z_n) \log \frac{p(X_n, Z_n \mid \theta)}{q_n(Z_n)}}}_{\mathcal{L}(q,\theta)} + \underbrace{\boxed{-\sum_{Z_n} q_n(Z_n) \log \frac{p(Z_n \mid X_n, \theta)}{q_n(Z_n)}}}_{KL(q \| p)}$$

E step:

determines

fix θ, then $q_n(Z_n) = p(Z_n \mid X_n, \theta)$, $KL = 0$, $\mathcal{L} \uparrow$

(Recall: fix θ, <u>return</u> Z or $q(Z)$)

M step:

fix $q_n(Z_n)$, find $\underset{\theta}{\operatorname{argmax}}\, \mathcal{L}(q, \theta)$, $KL \uparrow$, $\mathcal{L} \uparrow$

(Recall: fix Z, <u>find</u> θ that maximizes likelihood)

Another way of Decomposition

$$\sum_n \log p(X_n, \theta) = \sum_n \log \sum_{z_n} p(X_n, z_n \mid \theta)$$

$$= \sum_n \log \left[\sum_{z_n} q_n(z_n) \frac{p(X_n, z_n \mid \theta)}{q_n(z_n)} \right]$$

$$\geq \sum_n \sum_{z_n} q_n(z_n) \log \frac{p(X_n, z_n \mid \theta)}{q_n(z_n)} = \sum_n \sum_{z_n} q_n(z_n) \log p(X_n, z_n \mid \theta) + \sum_n \sum_{z_n} q_n(z_n) \log \frac{1}{q_n(z_n)}$$

$$\underbrace{\boxed{\sum_n \log \sum_{z_n} p(X_n, z_n \mid \theta)}}_{\text{complete data log likelihood}} \geq \underbrace{\underbrace{\boxed{\sum_n \mathbb{E}_{q_n} \log p(X_n, z_n \mid \theta)}}_{\text{expected complete data log likelihood}} + \sum_n H(q_n)}_{\text{ELBO (evidence lower bound)}}$$

Another way of Decomposition

$$\underbrace{\boxed{\sum_n \log \sum_{z_n} p(x_n, z_n \mid \theta)}}_{\text{complete data log likelihood}} \geq \underbrace{\underbrace{\boxed{\sum_n \mathbb{E}_{q_n} \log p(x_n, z_n \mid \theta)}}_{\text{expected complete data log likelihood}} + \sum_n H(q_n)}_{= \sum_n \log p(x_n \mid \theta) - KL(q_n(z_n) \,\|\, p(z_n \mid x_n, \theta))}$$

* see Book 8.7.2.2 for derivation

E step:

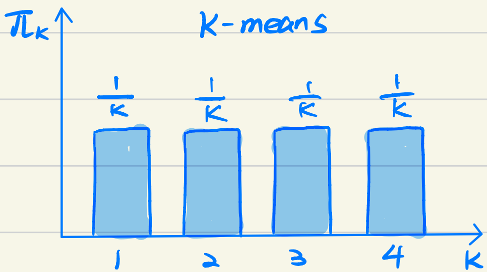
fix θ, $q_n(z_n) = \underbrace{p(z_n \mid x_n, \theta)}_{\text{the posterior}}$, $KL = 0$

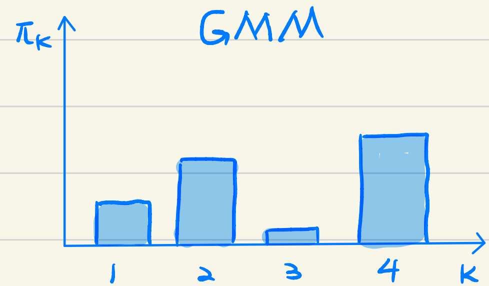
M step:

fix $q_n(z_n)$, $H(q_n)$ stays constant, $\underset{\theta}{\operatorname{argmax}} \sum_n \mathbb{E}_{q_n} \log p(x_n, z_n \mid \theta)$

Gausian Mixture Models (GMM)

π_k : cluster weights , $0 \le \pi_k \le 1$, $\sum_{k=1}^{K} \pi_k = 1$





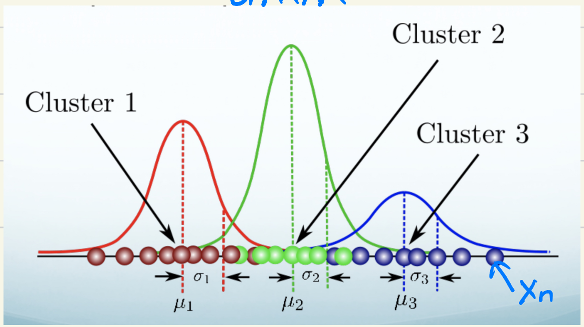
μ_k, Σ_k : probability distribution of points in cluster k

kmeas

$\Sigma_k = I$

$\mu_k = \frac{\sum_{n:z_{nk}=1} x_n}{N_k}$

GMM



GMM

π_k : cluster weights , $0 \le \pi_k \le 1$, $\sum_{k=1}^{K} \pi_k = 1$

μ_k , Σ_k : $\mathcal{N}(x_n | \mu_k, \Sigma_k)$

z_n : a selector, $[0, 0, \cdots, 1, \cdots]$, $P(z_{nk} = 1) = \pi_k$

$P(x_n | z_{nk} = 1) = \mathcal{N}(x_n | \mu_k, \Sigma_k)$ $\qquad P(X | Z) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}(x_n | \mu_k, \Sigma_k)^{z_{nk}}$

r_{nk} : responsibility of cluster k for generating example n
determined by x_n, π_k, μ_k, Σ_k

GMM

$$\pi_k = \underbrace{P(z_{nk}=1)}_{\text{prior}} \qquad \underbrace{P(x_n \mid z_{nk}=1)}_{\text{likelihood}} = \mathcal{N}(x_n \mid \mu_k, \Sigma_k)$$

$$r_{nk} = \underbrace{P(z_{nk}=1 \mid x_n)}_{\text{posterior}}$$

$$= \frac{P(z_{nk}=1)\,P(x_n \mid z_{nk}=1)}{P(x_n)} = \frac{P(z_{nk}=1)\,P(x_n \mid z_{nk}=1)}{\sum_{j=1}^{k} P(z_{nj}=1)\,P(x_n \mid z_{nj}=1)}$$

$$= \frac{\pi_k \, \mathcal{N}(x_n \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{k} \pi_j \, \mathcal{N}(x_n \mid \mu_j, \Sigma_j)}$$

EM for a GMM

Goal:

$$\max \sum_n \log \sum_k p(x_n, z_{nk} | \theta)$$

$$= \max \sum_n \log \left\{ \sum_k \pi_k \mathcal{N}(x_n | \mu_k, \Sigma_k) \right\}$$

E step:

fix π, μ, Σ, get $r_{nk} = \dfrac{\pi_k \mathcal{N}(x_n | \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n | \mu_j, \Sigma_j)}$

M step:

fix r_{nk}, maximization (take derivative, set it to 0)

get updated π, μ, Σ

Another way of Decomposition

$$\boxed{\sum_n \log \sum_{z_n} p(x_n, z_n \mid \theta)} \geq \boxed{\sum_n \mathbb{E}_{q_n} \log p(x_n, z_n \mid \theta)} + \sum_n H(q_n)$$

complete data log likelihood

expected complete data log likelihood

E step:

fix θ, $q_n(z_n) = \underline{p(z_n \mid x_n, \theta)}$

the posterior, r_{nk}!

M step:

fix $q_n(z_n)$, $H(q_n)$ stays constant, $\arg\max_\theta \sum_n \mathbb{E}_{q_n} \log p(x_n, z_n \mid \theta)$

EM for a GMM

$$0,1$$

$$E_{q_n}(z_{nk}) = \sum_n \sum_k q_n(z_{nk})\, z_{nk}$$

E step: set $q_n(z_{nk}) = p(z_{nk} \mid x_n, \theta)$

$$= r_{nk}$$

$$E_{q_n}(z_{nk}) = r_{nk}$$

E step:

fix π, μ, Σ, get $r_{nk} = \dfrac{\pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n \mid \mu_j, \Sigma_j)}$

write out expected complete data log likelihood:

$$E_{q_n}\left[\sum_n \log p(x_n, z_n \mid \theta)\right] = E_{q_n}\left[\sum_n \log p(z_n \mid \pi) + \sum_n \log p(x_n \mid z_n, \mu, \Sigma)\right]$$

$$= E_{q_n}\left[\sum_n \log\left(\prod_k \pi_k^{z_{nk}}\right) + \sum_n \log\left(\prod_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k)^{z_{nk}}\right)\right]$$

$$= \sum_n \sum_k E_{q_n}[z_{nk}] \log \pi_k + \sum_n \sum_k E_{q_n}[z_{nk}] \log \mathcal{N}(x_n \mid \mu_k, \Sigma_k)$$

$$= \sum_n \sum_k r_{nk} \log \pi_k + \sum_n \sum_k r_{nk} \log \mathcal{N}(x_n \mid \mu_k, \Sigma_k)$$

EM for a GMM

E step:

expected complete data log likelihood:

$$\mathbb{E}_{q_n}\left[\sum_n \log p(x_n, z_n \mid \theta)\right] = \sum_n \sum_k r_{nk} \log \pi_k + \sum_n \sum_k r_{nk} \log \mathcal{N}(x_n \mid \mu_k, \Sigma_k)$$

M step:

set derivative to 0 w.r.t. π_k, μ_k, Σ_k

$$\pi_k = \frac{1}{N} \sum_n r_{nk} = \frac{r_k}{N}$$

r_k: weighted # points in cluster k $(r_n = 1)$

$$\mu_k = \frac{\sum_n r_{nk} x_n}{r_k}$$

weighted average

$$\Sigma_k = \frac{\sum_n r_{nk} x_n x_n^T}{r_k} - \mu_k \mu_k^T$$

also weighted

K-means as a special case of EM

	GMM	K-means
parameters	π, μ, Σ	μ $(\pi_k = \frac{1}{K}, \Sigma_k = I)$
E-step	$r_{nk} = \frac{\pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n \mid \mu_j, \Sigma_j)}$	$z_{nk^*} = 1$, where $k^* = \arg\max_k r_{nk}$ $(r_{nk} = \mathbb{I}(z_{nk^*} = 1))$
M-step	$\pi_k = \frac{1}{N} \sum_n r_{nk} = \frac{r_k}{N}$ $\mu_k = \frac{\sum_n r_{nk} x_n}{r_k}$ $\Sigma_k = \frac{\sum_n r_{nk} x_n x_n^T}{r_k} - \mu_k \mu_k^T$	$\mu_k = \frac{\sum_{n: z_{nk}=1} x_n}{N_k}$

II. Graphical Models

Terminology

PGM: probablistic graphical models

DAG: directed acyclic graph

Bayesian networks (Bayes nets), belief networks: PGMs based on DAGs

topological ordering: In DAG, nodes are ordered s.t. parents come before children

ordered Markov property: assumption that a node is conditionally independent (CI) of all predecessors given its parents

Markov chain: $X_1 \to X_2 \to X_3 \cdots \to X_T$

$$P(x_{1:T}) = p(x_1) \prod_{t=2}^{T} p(x_t \mid x_{1:t-1})$$ first-order Markov chain

CPT: conditional probability table, 2d table representing conditional probability distribution (CPD)

$X \to Y$

	Y=0	Y=1
X=0	0.05	0.95
X=1	0.2	0.8

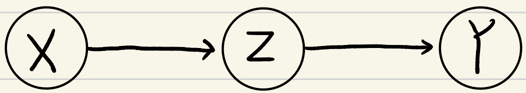
P(Y|X)

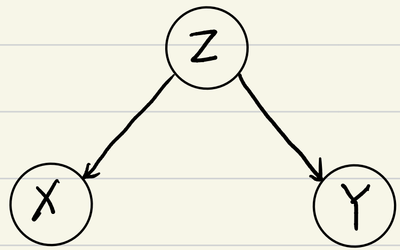
D-separation

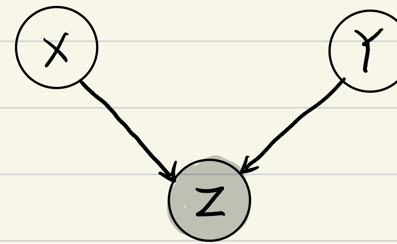
$X_A \perp_G X_B \mid X_C$ A is conditionally independent of B given C in graph G.
A is d-separated from B given C

We "shade" a node to indicate it's observed.

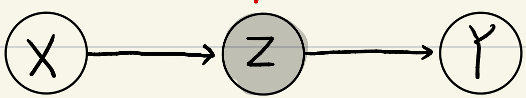
X, Y are not d-separated

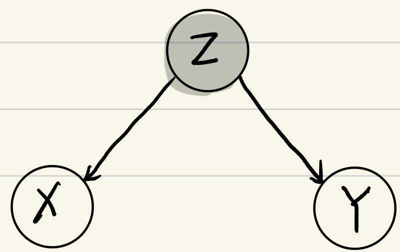


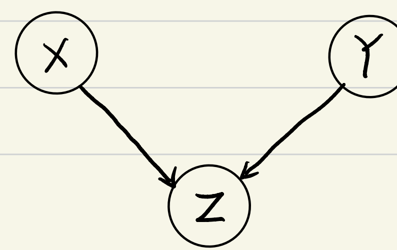




X, Y are d-separated







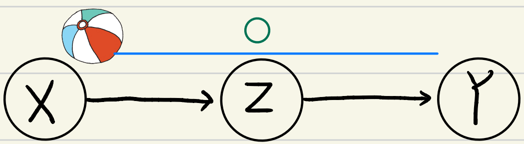
D- separation

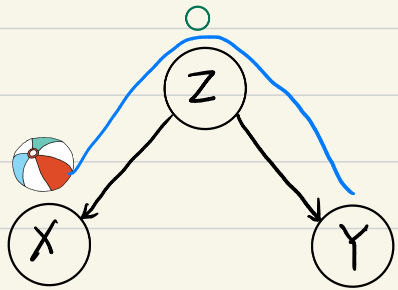
Can X pass a ball to Y ?

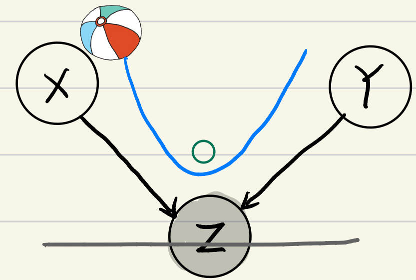
- yes : not d-separated
- no : d-separated

(Balls travel along edges and can travel opposite to edge directions.)

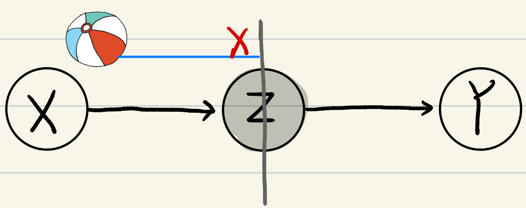
X, Y are not d-separated

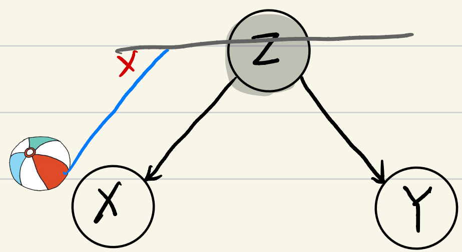


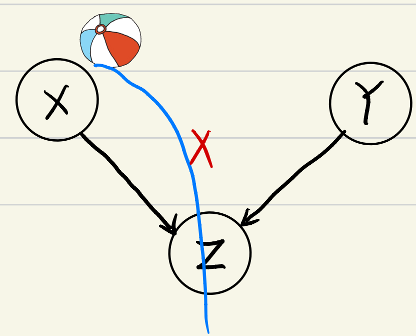




X, Y are d-separated

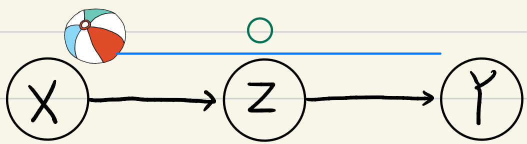




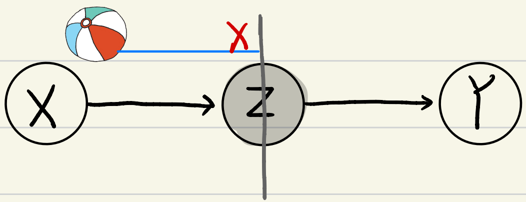


Can X pass a ball to Y ?

The ball travels along edges and can travel opposite to edge directions.

I. 

can — not d-separated

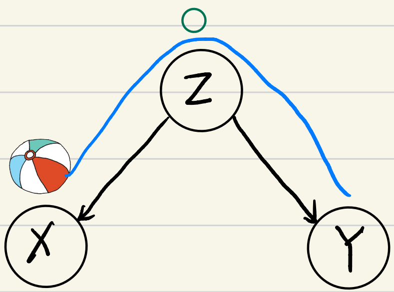
II. 

can't — d-separated

Can X pass a ball to Y ?

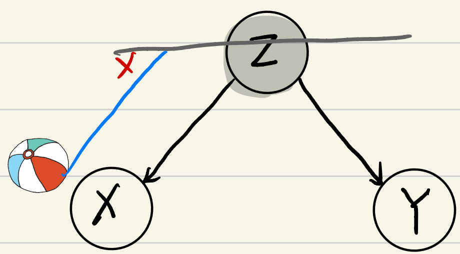
The ball travels along edges and can travel opposite to edge directions.

I.



can not d-separated

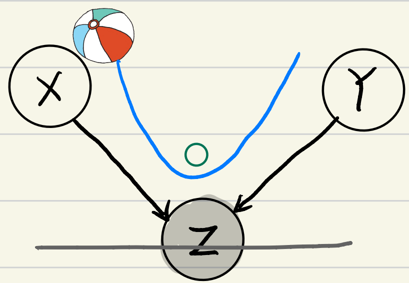
II.



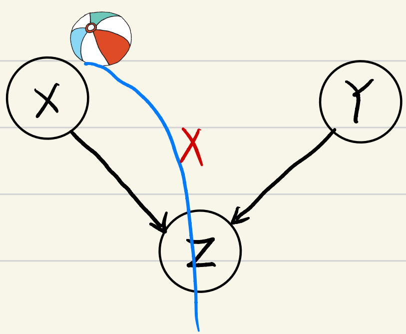
can't d-separated

Can X pass a ball to Y ?

The ball travels along edges and can travel opposite to edge directions.

I. 

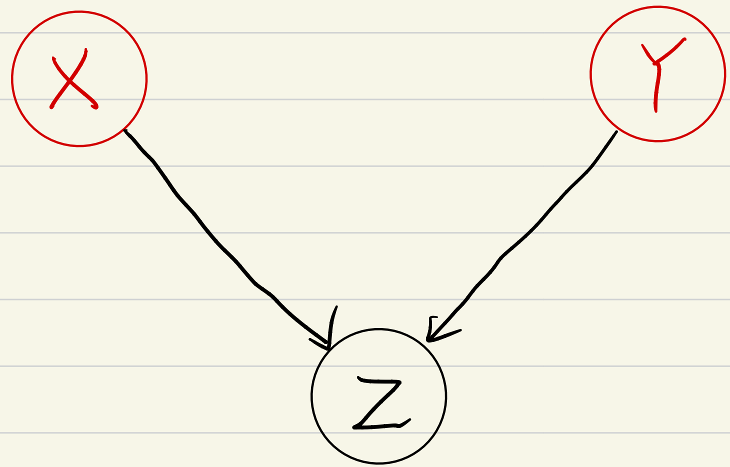
can not d-separated

II. 

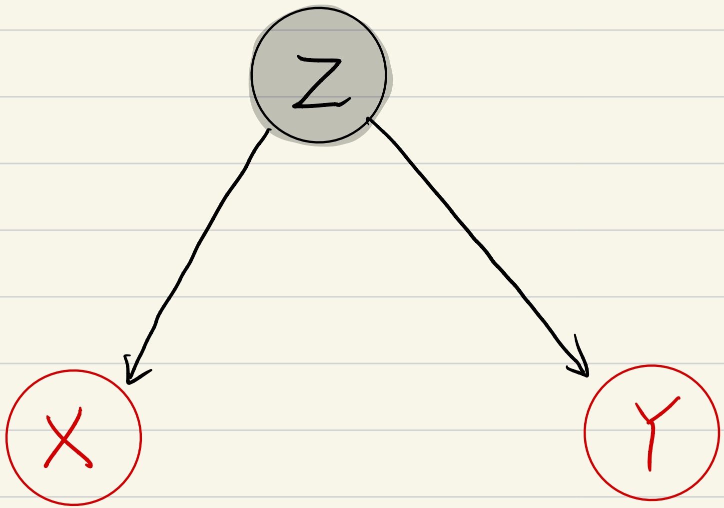
can't d-separated

Are X and Y d-separated or not?

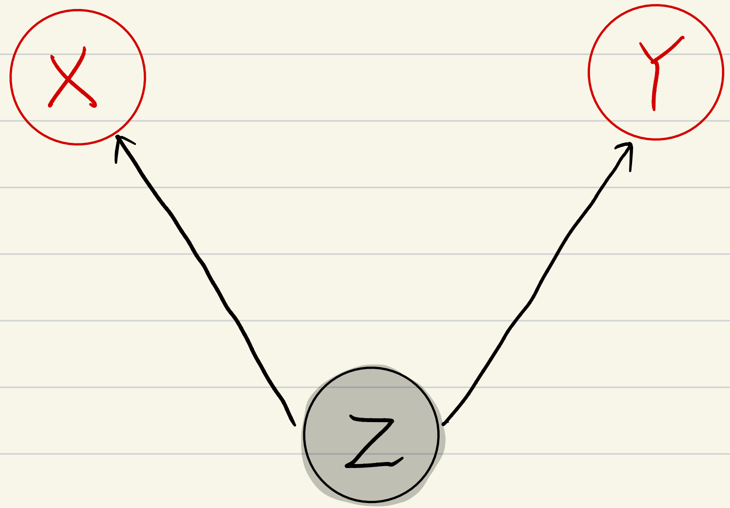
1. d-separated ?



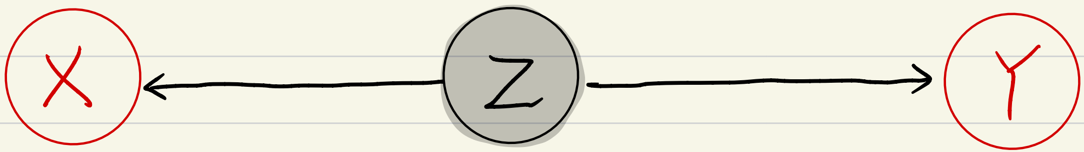
2. d-separated?



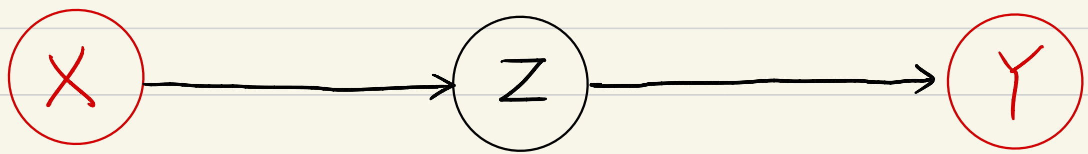
3. d-separated ?



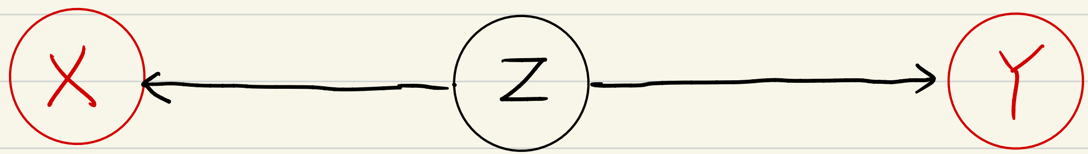
4. d-separated?



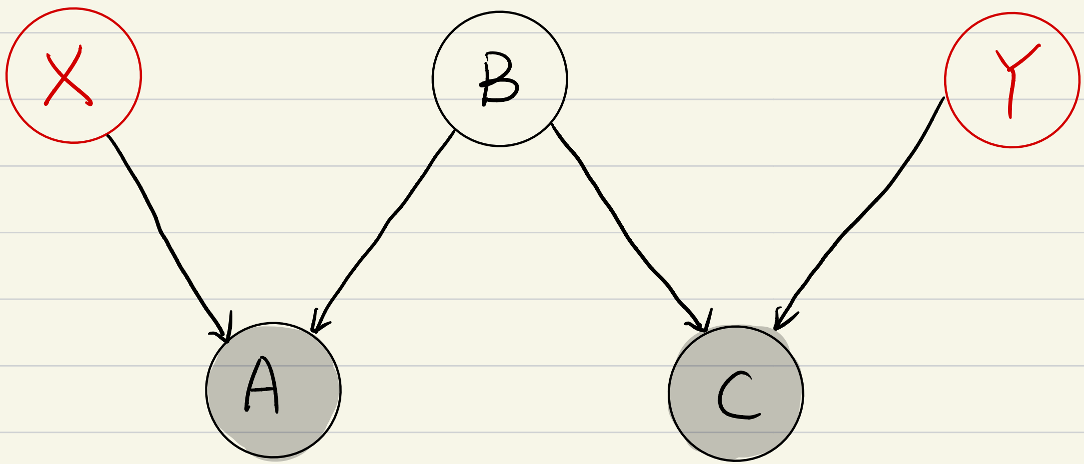
5. d-separated ?



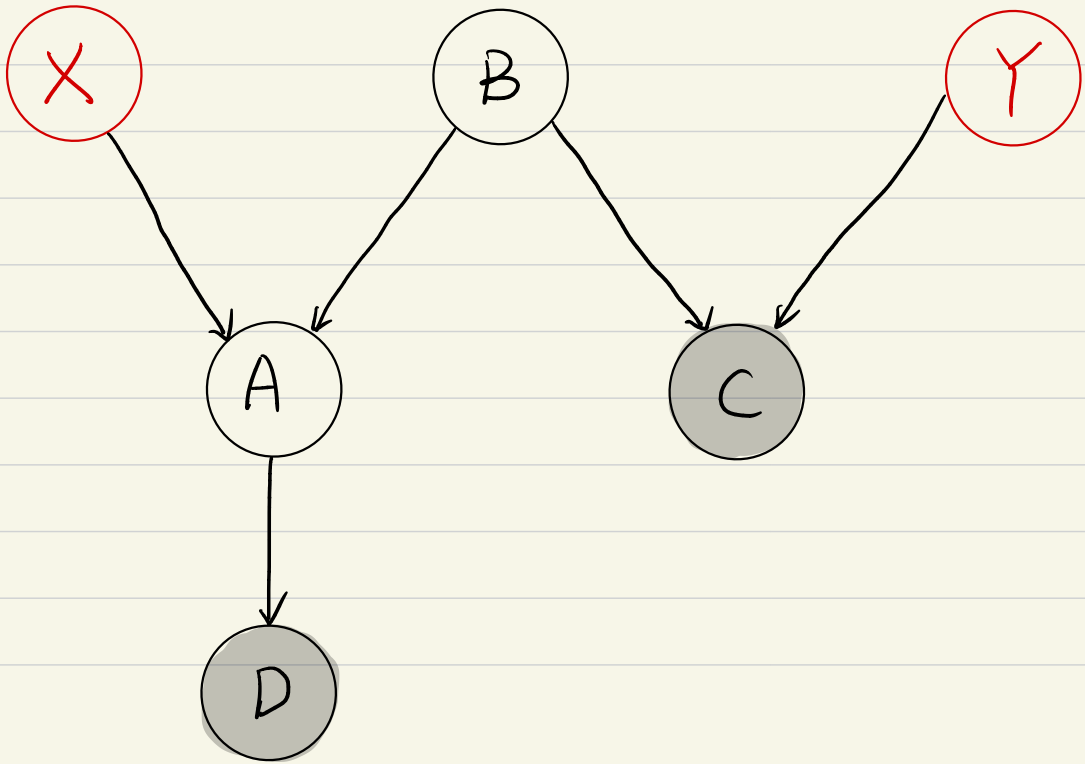
6. d-separated?



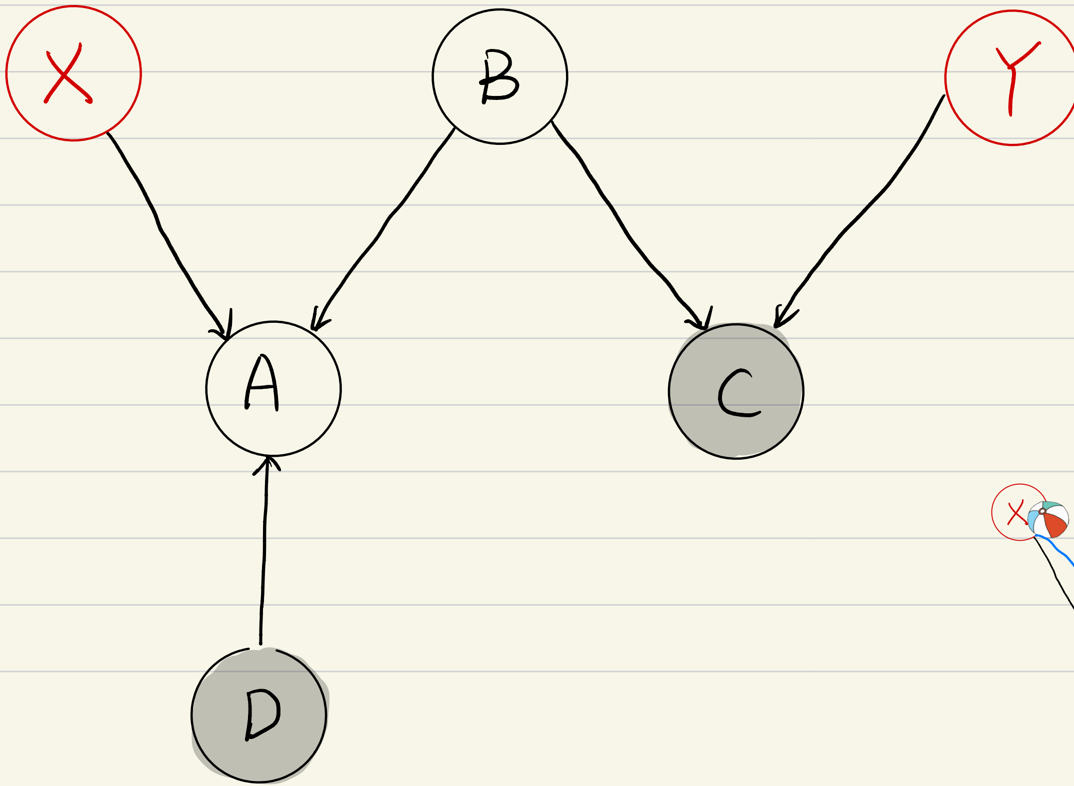
7. d-separated?



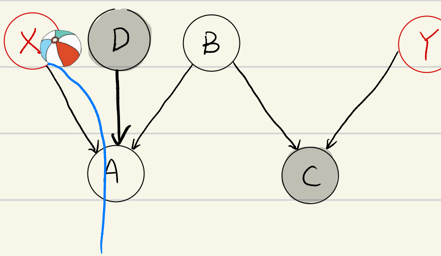
8. d-separated ?



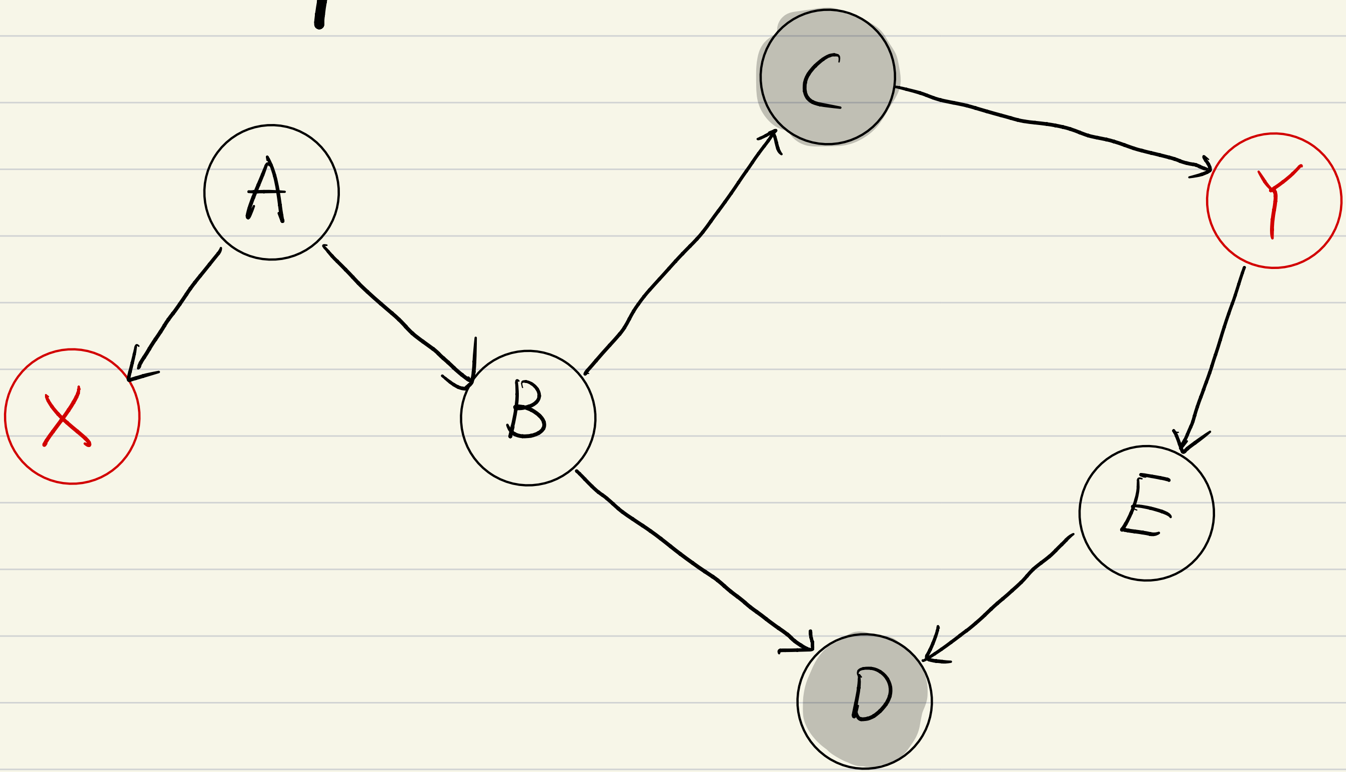
9. d-separated?



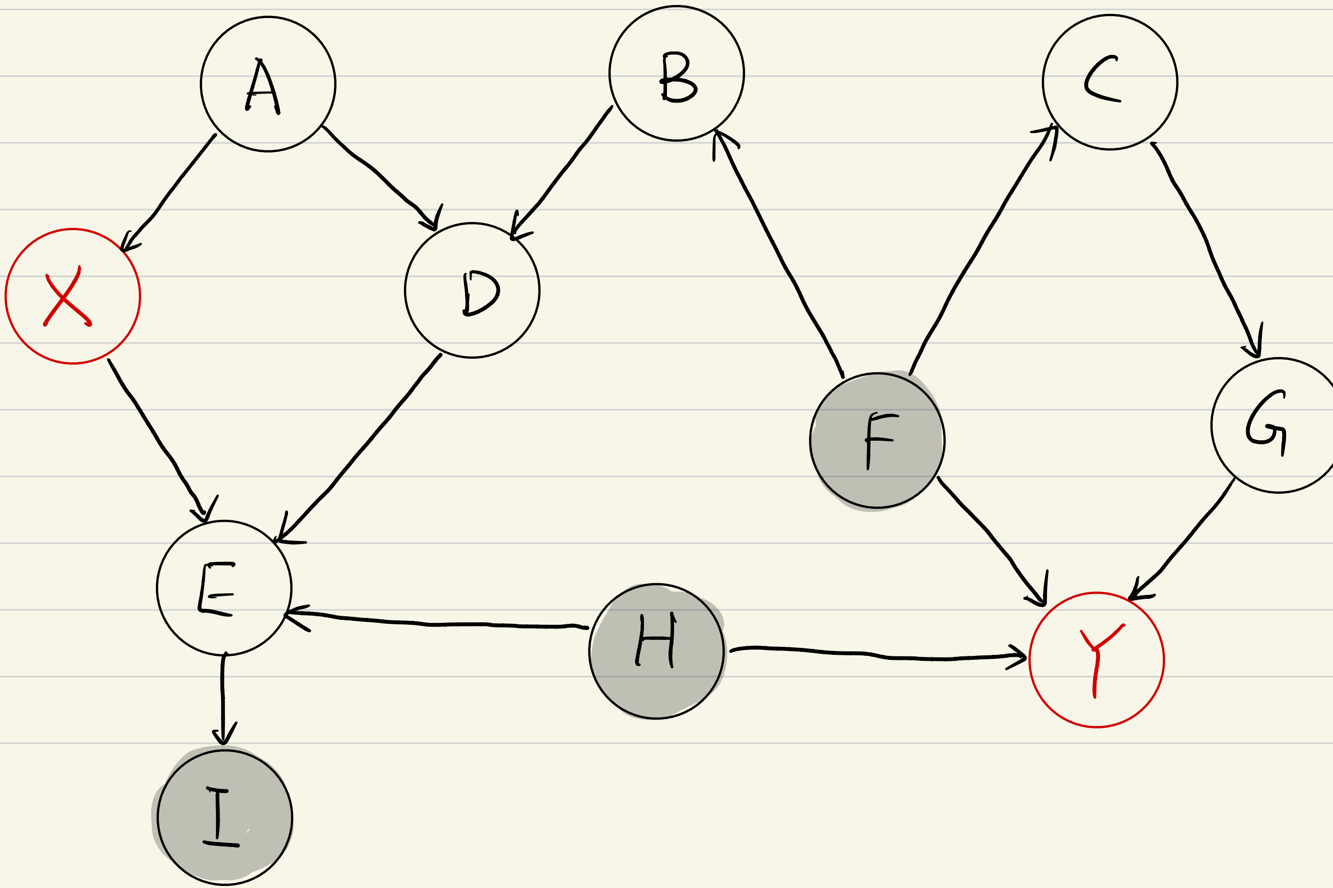
Yes!



10. d-separated ?



11. d-separated ? The last one !



Answers

1 – 4 : Yes

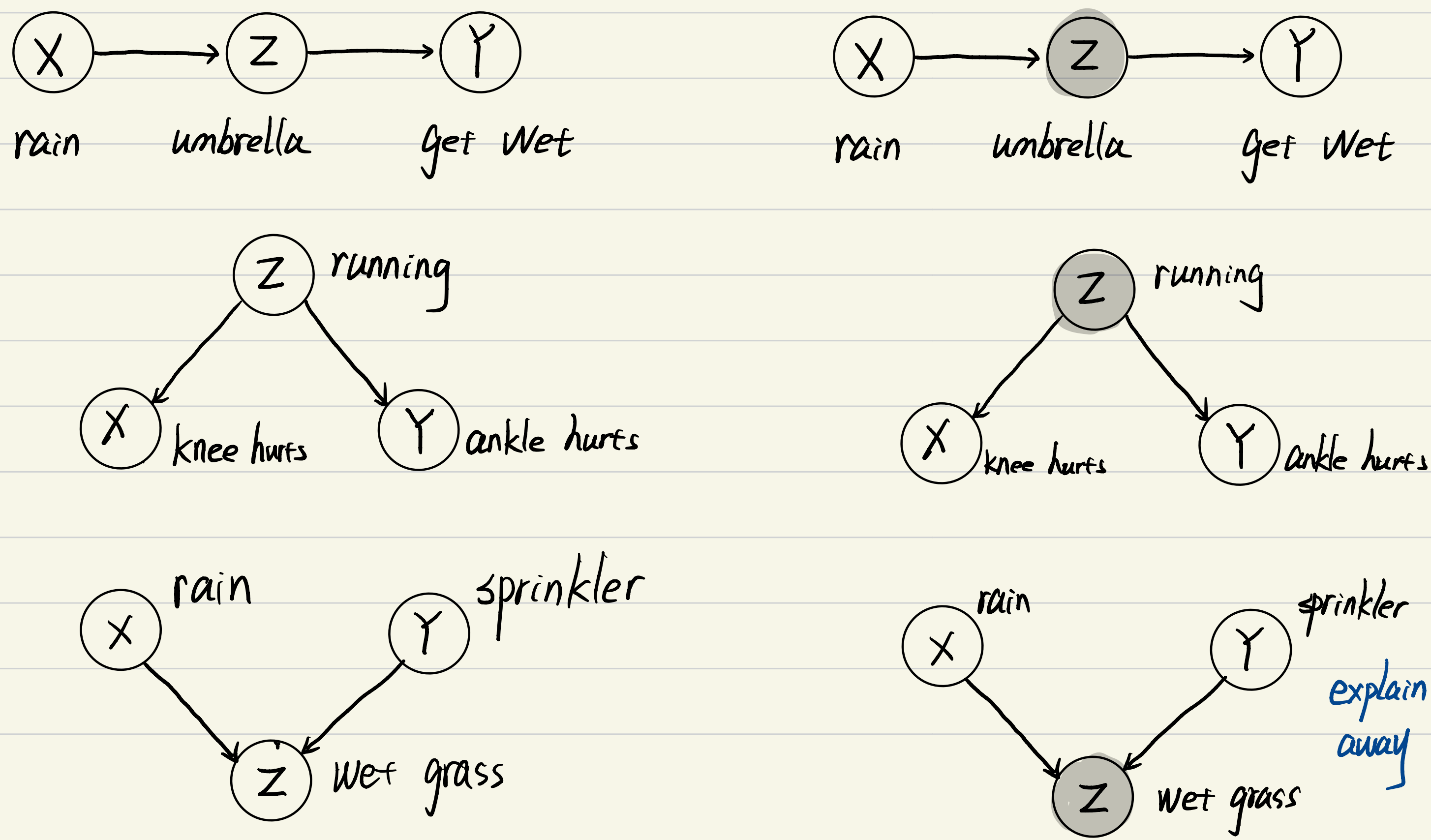
5 – 8 : No

9 : Yes

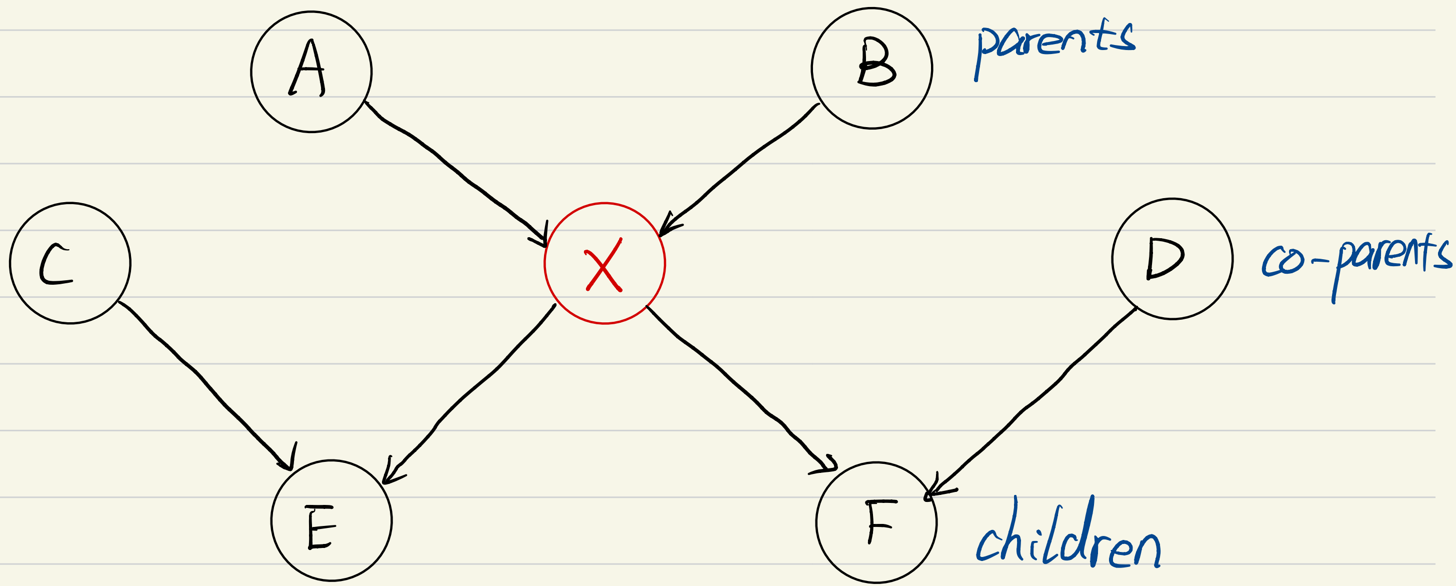
10 : No

11 : Yes

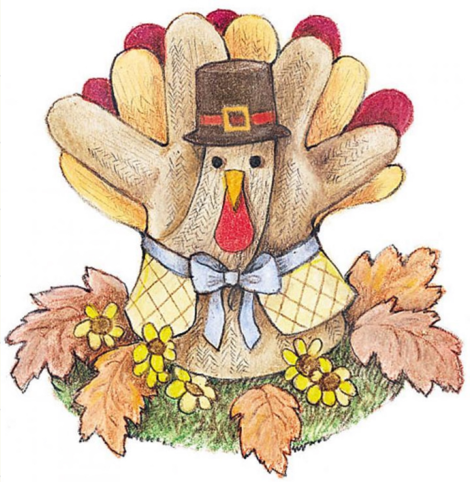
What does it mean?



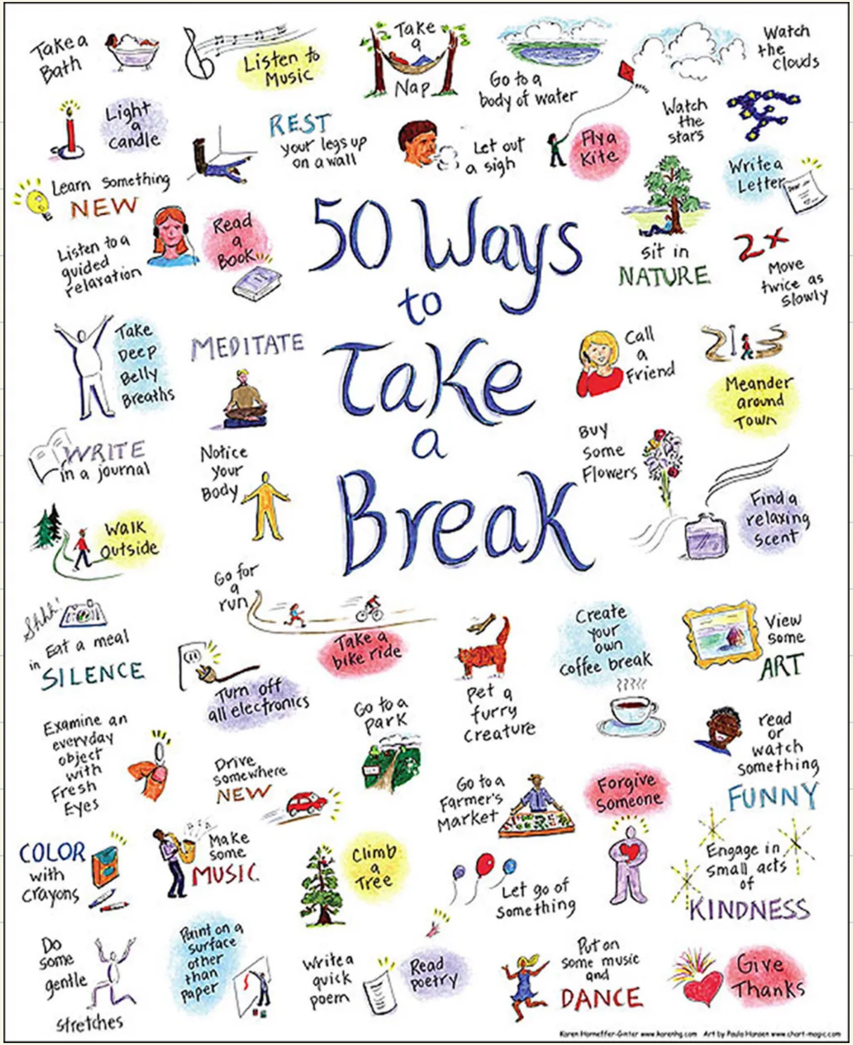
Markov Blanket



A node conditioned on its Markov blanket is independent from all other nodes in the graph.




HAPPY
thanksgiving


50 Ways to Take a Break
Take a Bath
Listen to Music
Take a Nap
Go to a body of water
Watch the clouds
Light a candle
REST your legs up on a wall
Let out a sigh
Fly a Kite
Watch the stars
Write a Letter
Learn Something NEW
Listen to a guided relaxation
Read a Book
Sit in NATURE
2x Move twice as slowly
Take Deep Belly Breaths
MEDITATE
Call a Friend
Meander around Town
WRITE in a journal
Notice your Body
Buy some Flowers
Find a relaxing scent
Walk outside
Go for a run
Shhh!
Eat a meal in SILENCE
Turn off all electronics
Take a bike ride
Pet a furry creature
Create your own coffee break
View some ART
Examine an everyday object with Fresh Eyes
Drive somewhere NEW
Go to a Park
Go to a Farmer's Market
Forgive Someone
read or watch something FUNNY
COLOR with Crayons
Make some MUSIC
Climb a Tree
Let go of Something
Engage in small acts of KINDNESS
Do some gentle stretches
Paint on a surface other than paper
Write a quick poem
Read poetry
Put on some music and DANCE
Give Thanks