Problem 1 →

Q-1-
Ans - The one downside of Information Gain is that it favours the predictor ~~values~~ variables with a large number of values. Now if we split the data using these highly branching predictors, then this data might be split into subsets with low Entropy values. Hence, the disadvantages of these splits are that our model might overfit the data or the number of nodes in the tree might increase significantly.

To remedy this Information Gain Ratio introduces a normalizing term called SplitInfo or Intrinsic Information to reduce the bias of Information Gain. ~~Split Info is given by~~

$$\boxed{\text{Split Info}(F) = -\sum \frac{|D_j|}{|D|} \cdot \log_2\left(\frac{|D_j|}{|D|}\right)}$$

&

$$\boxed{\text{Information Gain Ratio} = \frac{\text{Information Gain}}{\text{Split Info}}}$$

→ ~~We choose We split~~;

→ Whichever predictor variable has the highest Gain Ratio, ~~we choose it~~, we choose that predictor for splitting.

Q-2-

Ans - Entropy (Decision) $= -\sum p(I) \log_2 p(I)$ [where I ∈ output classes]

$= -p(Yes) \log_2 P(Yes) - p(No) \log_2 p(No)$ ~~where I ∈ {Yes, No} output classes~~

$= -\left[\frac{9}{14} \log_2 (9/14) + \frac{5}{14} \log_2 (5/14)\right]$

$= \underline{0.940}$

Gain ~~Ratio~~ (F) = Entropy (Decision)

~~$\sum p(\text{Decision} | \text{predictor } F)$ Entropy~~

$- \sum p(\text{Decision}|F) \cdot \text{Entropy}(\text{Decision}|F)$

$$\text{Gain Ratio}(F) = \frac{\text{Gain}(F)}{\text{SplitInfo}(F)}$$

$$\text{SplitInfo}(F) = -\sum \frac{|D_j|}{|D|} \cdot \log_2 \left(\frac{|D_j|}{|D|}\right)$$

Homework →

Gain (D, Homework) = Entropy (Decision) − [p(Decision | Homework = Much) · Entropy (Decision | Homework = Much) + p(Decision | Homework = Normal) ~~Entropy~~ · Entropy (Decision | Homework = Normal) + p(Decision | Homework = None) · Entropy (Decision | Homework = None)]

$$\text{Entropy}(\text{Decision} \mid \text{Homework} = \text{Much}) = -\left[\frac{2}{5}\log_2\left(\frac{2}{5}\right) + \frac{3}{5}\log_2\left(\frac{3}{5}\right)\right]$$
$$= -[0.5288 + 0.4422]$$
$$= 0.971$$

$$\text{Entropy}(\text{Decision} \mid \text{Homework} = \text{Normal}) = -\left[\frac{4}{4}\log_2\left(\frac{4}{4}\right) + 0\right] = 0$$

$$\text{Entropy}(\text{Decision} \mid \text{Homework} = \text{None}) = -\left[\frac{3}{5}\log_2\left(\frac{3}{5}\right) + \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right]$$
$$= 0.971$$

$$\text{Gain}(\text{Decision}, \text{Homework}) = 0.940 - \left[\frac{5}{14} \cdot 0.971 + \frac{5}{14} \cdot 0.971 + \frac{4}{14} \cdot 0\right]$$
$$= 0.940 - \left(\frac{5}{7} \cdot 0.971\right)$$
$$= \underline{0.246428571 4}$$

$$\text{SplitInfo}(\text{Decision}, \text{Homework}) = -\left[\frac{5}{14}\log_2\left(\frac{5}{14}\right) + \frac{4}{14}\log_2\left(\frac{4}{14}\right) + \frac{5}{14}\log_2\left(\frac{5}{14}\right)\right]$$
$$= -\left[\frac{5}{7}\log_2\left(\frac{5}{14}\right) + \frac{2}{7}\log_2\left(\frac{4}{14}\right)\right]$$
$$= -[-1.061 - 0.5164]$$
$$= 1.5774$$

$$\boxed{\text{Gain Ratio}(\text{Homework}) = \frac{0.2464285714}{1.5774} = \underline{\underline{0.1562245286}}}$$

Traffic →

Gain (Decision, Traffic) = Entropy (Decision) − [p (Decision | Traffic = Busy) · Entropy (Decision | Traffic = Busy) + p (Decision | Traffic = Ok) · Entropy (Decision | Traffic = Ok) + p (Decision | Traffic = Chill) · Entropy (Decision | Traffic = Chill)]

$$\text{Entropy (Decision | Traffic = Busy)} = -\left[\frac{2}{4}\log_2\left(\frac{2}{4}\right) + \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right]$$
$$= 1$$

$$\text{Entropy (Decision | Traffic = Ok)} = -\left[\frac{4}{6}\log_2\left(\frac{4}{6}\right) + \frac{2}{6}\log_2\left(\frac{2}{6}\right)\right]$$
$$= 0.9183333$$

$$\text{Entropy (Decision | Traffic = Chill)} = -\left[\frac{3}{4}\log_2\left(\frac{3}{4}\right) + \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right]$$
$$= -[-0.31125 - 0.5]$$
$$= 0.81125$$

$$\text{Gain (Decision, Traffic)} = 0.940 - \left[\frac{4}{14}\cdot 1 + \frac{6}{14}\cdot 0.9183333 + \frac{4}{14}\cdot 0.81125\right]$$

$$= \underline{0.028928\,5857}$$

$$\text{SplitInfo (Decision, Traffic)} = -\left[\frac{4}{14}\log_2\left(\frac{4}{14}\right) + \frac{6}{14}\log_2\left(\frac{6}{14}\right) + \frac{4}{14}\log_2\left(\frac{4}{14}\right)\right] = 1.556685254$$

$$\boxed{\text{GainRatio (Traffic)} = \frac{0.0289285857}{1.556685714 3} = 0.018583446 5}$$

Hunger →

$$\text{Gain (Decision, Hunger)} = \text{Entropy (Decision)} -$$

$$[\,p(\text{Decision} \mid \text{Hunger} = \text{A little}) \cdot \text{Entropy (Decision} \mid \text{Hunger} = \text{A little}) \cdot$$

$$+ p(\text{Decision} \mid \text{Hunger} = \text{A lot}) \cdot \text{Entropy (Decision} \mid \text{Hunger} = \text{A lot})\,]$$

$$\text{Entropy (Decision} \mid \text{Hunger} = \text{A little}) = -\left[\frac{3}{7}\log_2\left(\frac{3}{7}\right) + \frac{4}{7}\log_2\left(\frac{4}{7}\right)\right]$$

$$= 0.985257142 9$$

$$\text{Entropy (Decision} \mid \text{Hunger} = \text{A lot}) = -\left[\frac{6}{7}\log_2\left(\frac{6}{7}\right) + \frac{1}{7}\log_2\left(\frac{1}{7}\right)\right]$$

$$= 0.591685714 3$$

$$\text{Gain (Decision, Hunger)} = 0.940 - \left[\frac{7}{14} \cdot 0.985257142 9 + \frac{7}{14} \cdot (0.591685714 3)\right]$$

$$= \underline{0.151528571 4}$$

$$\text{SplitInfo (Decision, Hunger)} = -\left[\frac{7}{14}\log_2\left(\frac{7}{14}\right) + \frac{7}{14}\log_2\left(\frac{7}{14}\right)\right]$$

$$= 1$$

$$\boxed{\text{GainRatio (Hunger)} = \underline{\underline{0.151528571 4}}}$$

Lauren →

$$\text{Gain(Decision, Lauren)} = \text{Entropy(Decision)} - [p(\text{Decision}|\text{Lauren}=\text{Available}) \cdot \text{Entropy(Decision}|\text{Lauren}=\text{Available}) + p(\text{Decision}|\text{Lauren}=\text{Not Available}) \cdot \text{Entropy(Decision}|\text{Lauren}=\text{Not Available})]$$

$$\text{Entropy(Decision}|\text{Lauren}=\text{Available}) = -\left[\frac{6}{8}\log_2\left(\frac{6}{8}\right) + \frac{2}{8}\log_2\left(\frac{2}{8}\right)\right] = 0.81125$$

$$\text{Entropy(Decision}|\text{Lauren}=\text{Not Available}) = -\left[\frac{3}{6}\log_2\left(\frac{3}{6}\right) + \frac{3}{6}\log_2\left(\frac{3}{6}\right)\right] = 1$$

$$\text{Gain(Decision, Lauren)} = 0.940 - \left[\frac{8}{14}(0.81125) + \frac{6}{14}(1)\right] = 0.047857142\underline{9}$$

$$\text{SplitInfo(Decision, Lauren)} = -\left[\frac{8}{14}\log_2\left(\frac{8}{14}\right) + \frac{6}{14}\log_2\left(\frac{6}{14}\right)\right] = 0.9852571429$$

$$\boxed{\text{GainRatio(Lauren)} = \frac{0.0478571429}{0.9852571429} = 0.0485732514}$$

As we can see that Homework feature has the highest gain ratio, hence Homework feature will be our first split in the decision tree.

Q-3-
Ans - ~~As~~ ~~Hom~~ root node of our decision tree will be Homework, thus we will again make the table based on different values of Homework ⟹

(Homework = Much)

Homework	Traffic	Hunger	Lauren	Go Out ?
Much	Busy	A little	Available	No
Much	Busy	A little	Not Available	No
Much	Ok	A little	Available	No
Much	chill	A lot	Available	Yes
Much	Ok	A lot	Not Available	Yes

From this table we can see that Going out is directly correlated with Hunger & if Homework is much.

(Homework = Normal)

Homework	Traffic	Hunger	Lauren	Go Out ?
Normal	Busy	A little	Available	Yes
Normal	chill	A lot	Not Available	Yes
Normal	Ok	A little	Not Available	Yes
Normal	Busy	A lot	Available	Yes

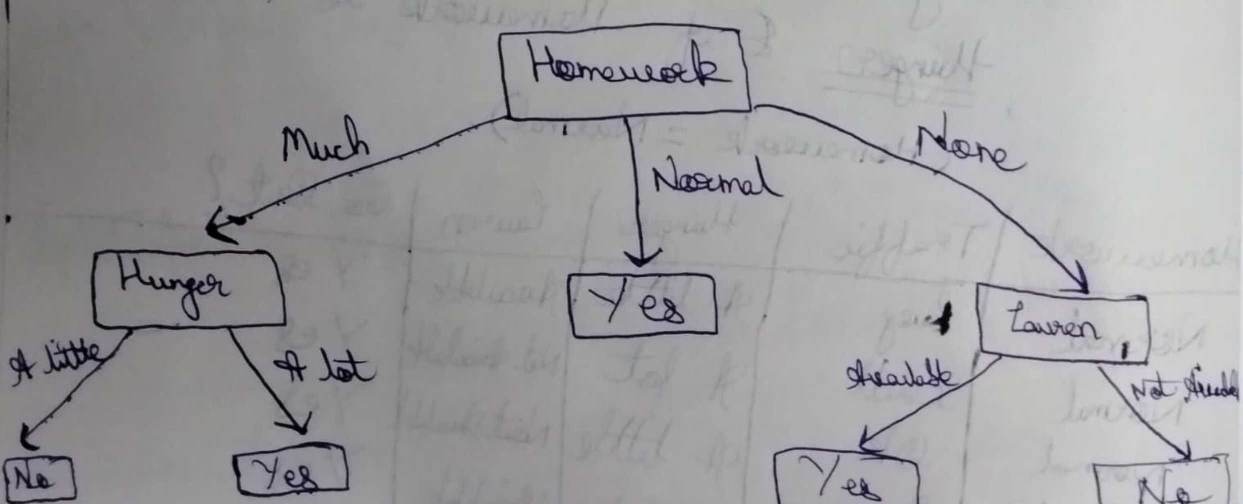
From this table we can see that if Homework is Normal then yo will always go out

(Homework = None)

Homework	Traffic	Hunger	Lauren	Go Out?
None	Ok	A little	Available	Yes
None	Chill	A lot	Available	Yes
None	Chill	A lot	Not Available	No
None	Ok	A lot	Available	Yes
None	Ok	A little	Not Available	No

From this table we can see that the decision to go out is directly correlated to availability of Lauren if there is no Homework

Hence from above three tables, we can construct our decision tree as →



Q-4-

Ans- Since, I have a normal amount of homework, therefore according to the decision tree, I will go out.

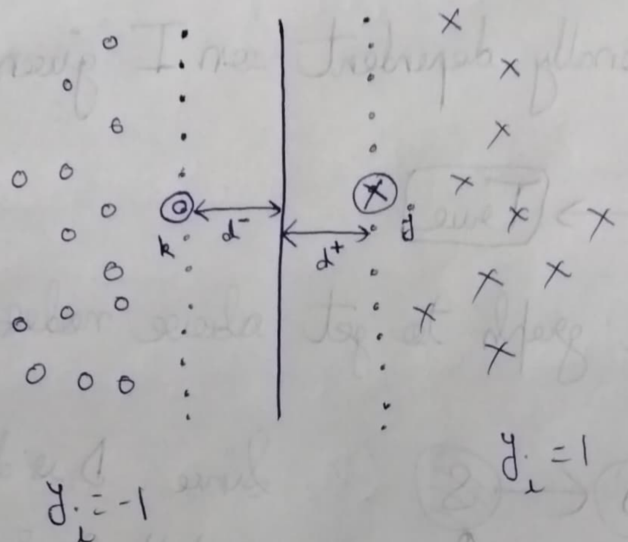
Problem - 2 $\rightarrow$

Linearly Separable data

(1)

Q-1-

Ans-



For class $y_i = 1 \Rightarrow$

$$w^T x_i + b \geq 1$$

ie

$$w^T x_j + b = 1 \Rightarrow b = 1 - w^T x_j$$

Similarly

For class $y_i = -1 \Rightarrow$

$$w^T x_i + b \leq -1$$

ie

$$w^T x_k + b = -1 \Rightarrow b = -1 - w^T x_k$$

$$\therefore \; d^+ - d^- = \frac{w^T x_i}{\|w\|_2} - \frac{w^T x_k}{\|w\|_2} = \frac{1-b}{\|w\|_2} - \frac{(-1-b)}{\|w\|_2}$$

$$\Rightarrow d^+ - d^- = \frac{2}{\|w\|_2}$$

∴ This SVM problem can be defined as ⇒

$$\max_{w,b} \frac{2}{\|w\|_2} \quad \text{such that } y_i(w^T x_i + b) \geq 1$$
$$\text{for } i \in \{1, \ldots, N\}$$

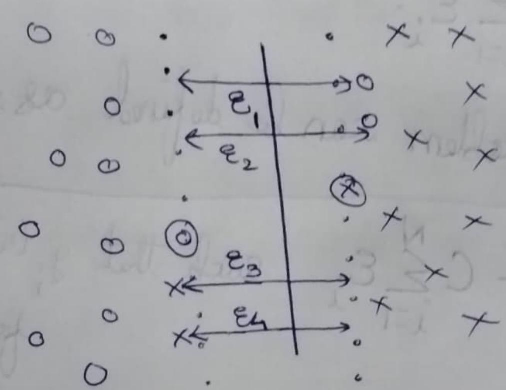
denotes no. of data points

Or

$$\min_{w,b} \frac{1}{2}\|w\|_2^2 \quad \text{such that } y_i(w^T x_i + b) \geq 1$$
$$\text{for } i \in \{1, \ldots, N\}$$

Q-2-

Ans -



Now if our data is not linearly separable, then we might have to deal with slack variables. The slack variables are used to minimize the misclassifiers along with the objective function defined in the above problem $\frac{1}{2}\|w\|_2^2$ (See the above figure)

In the case data points are correctly classified,

$$y_i(w^T x_i + b) \geq 1$$

In the case data is misclassified ;

$$\cancel{y_i} \quad y_i(w^T x_i + b) < 1$$

The slack variable that we will be using can be defined as $\Rightarrow$

$\cancel{\varepsilon}$ $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N$ for misclassifiers

(for $i \in 1, \ldots, N$) [See the above figure]

$\therefore$ Now our objective function becomes $\Rightarrow$

$$\frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{N} \varepsilon_i$$

$\therefore$ This SVM problem can be defined as $\Rightarrow$

$$\boxed{\min_{w, b, \varepsilon_i} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{N} \varepsilon_i \quad \text{such that } y_i(w^T x_i + b) \geq 1 - \varepsilon_i \text{ for } i \in \{1, \ldots, N\} \quad \varepsilon_i \geq 0 \text{ for } i \in \{1, \ldots, N\}}$$

Q-3-

Ans- SVM doesn't support multiclass classification directly. But we can still perform multiclass classification by breaking it down into smaller binary classification ~~subpro~~ subproblems. There are many approaches like One vs One, One vs All and Directed Acyclic graph ~~approach~~.

- In One vs One we breakdown our multiclass classification problem into various binary classification subproblems. For the final prediction we use the concept of majority voting.

- In One vs All approach we train N SVMs

  SVM (1) → for class 1

  SVM (2) → for class 2

  ⋮

  SVM (N) → for class N

Now to predict the output of new inputs predict output using all the SVMs & then just identify which model puts the prediction farthest into the positive region.

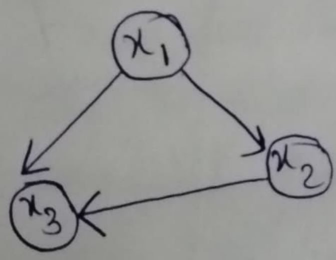
- In Directed Acyclic Graph approach we try to first group the classes based on some logical grouping & then train SVMs. Thus, at the end we might need to train less number of SVMs and this approach also reduces the diversity from the majority class.

Problem 3 →

Q-1-

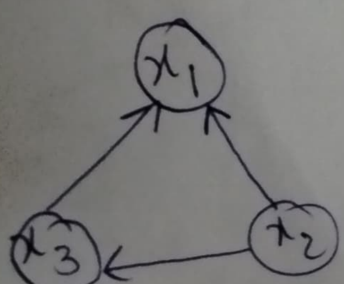
Ans- (1) We are given N random variables $x_1, x_2, \ldots, x_N$. And now we want to graph a Bayesian Network showing the joint distribution of all N variables.

Let us consider joint distribution of ~~p(a,b,c)~~ x_1, x_2, x_3 random variables.



Then $p(x_1,x_2,x_3) = p(x_1)\cdot p(x_2|x_1)\, p(x_3|x_1,x_2)$

Similarly if we change the ordering, & build the following graph ⇒



then $p(x_1,x_2,x_3) = p(x_2)\cdot p(x_3|x_2)\cdot p(x_1|x_3,x_2)$

Thus, if we list out all the possibilities, then we will see that for 3 random variables we can make total $3! = 6$ distinct graphs.

Hence, for N random variables, we can ~~say that~~ make $N!$ distinct graphs.

(2) For a single discrete variable x, having M possible states, the probability distribution $p(x|\sigma)$ is given by:

$$p(x|\sigma) = \prod_{m=1}^{M} \sigma_m^{x_m}$$

Similarly for discrete random variables x_1, x_2; we have two states $\rightarrow s_1$ & s_2

$$\therefore p(x_1, x_2|\sigma) = \prod_{a=1}^{S_1} \prod_{b=1}^{S_2} \sigma_{ab}^{x_{1a} x_{2b}}$$

& we can see that total number of parameters here are $= (s_1 \times s_2 - 1)$

Thus, after generalizing this we can say that if we have N random variables $x_1, x_2, x_3, \ldots, x_N$ & they have N states $s_1, s_2, s_3, \ldots, s_N$ then total number of parameters in the joint probability distribution will be :- $[(s_1 \times s_2 \times s_3 \cdots \times s_n) - 1]$

This shows that we will eventually end up with a graph that grows exponentially and ~~thus it would be very~~ so the task to calculate joint probability in this case will become very complex.

Now if we consider all our random variables to be conditionally independent then,

for two random variables x_1, x_2

joint probability $p(x_1, x_2 | \sigma)$ becomes →

$$p(x_1, x_2 | \sigma) = \prod_{a=1}^{s_1} \sigma_{1a}^{x_{1a}} \prod_{b=1}^{s_2} \sigma_{2b}^{x_{2b}}$$

& no of parameters in this joint probability will be $(s_1 + s_2 - 1)$

∴ After Generalizing this we can say that N random variables with N states will have ⇒ $(s_1 + s_2 + \cdots + s_N - 1)$ parameters

& we will get a graph which grows linearly instead of exponentially.

Hence, if we maintain conditional independence of all N random variables, then it will be less complex to calculate joint probability $p(x_1, x_2, \ldots, x_N)$, as the graph will grow linearly & not exponentially.

Q-2-

Ans- (1) G ⇒ Going out right now

H ⇒ contracting Hakie plague

D ⇒ Drowsy

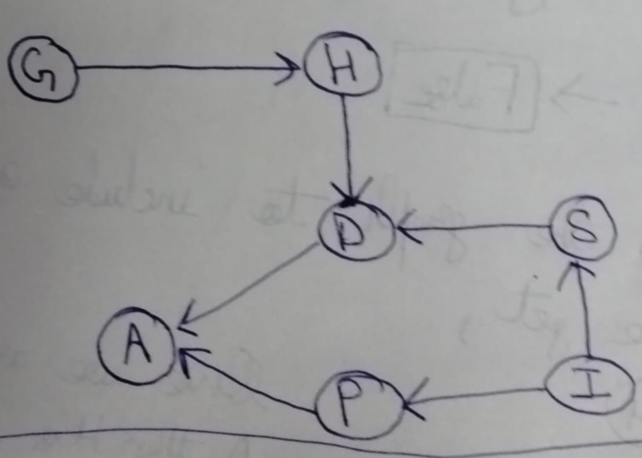
S ⇒ Lack of Sleep.

I ⇒ Insomnia

P ⇒ Pay Attention

A ⇒ Unable to finish the assignment.

Bayesian Network ⇒



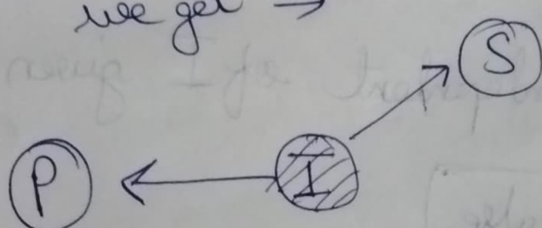
$$\boxed{p(G,H,D,S,I,P,A) = p(G)\cdot p(H|G)\cdot p(D|H,S)\cdot p(S|I)\cdot p(I)\cdot p(P|I)\cdot p(A|P,D)}$$

Nodes in the S's Markov Blanket ⇒

I, D & H as ~~we~~ if we know the values of these nodes, then we can determine the value of S.

② S $\perp\!\!\!\perp$ P | I → $\boxed{\text{True}}$

Simplifying the graph to get above nodes, we get ⇒

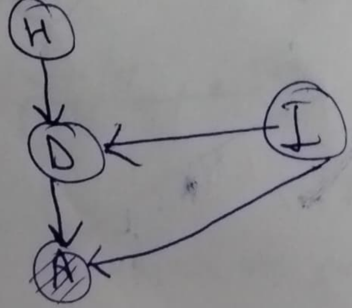


Since, I is dependent on both S & P ~~and we~~ ~~know the value of I~~ thus even if we know the value of I still P & S are independent

∴ S is conditionally independent of P given I

③ H $\perp\!\!\!\perp$ I | A → $\boxed{\text{False}}$

Simplifying the graph to get above nodes, we get ⇒



Since A is dependent on D & I, & D is dependent on both I & H

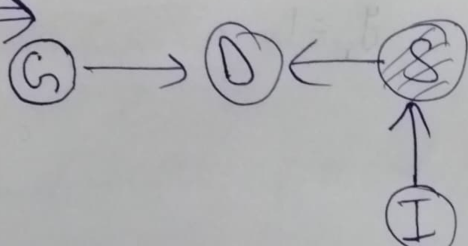
Hence, H is also dependent on I, even when A is given

∴ H is conditionally dependent on I given A.

④ $G \perp\!\!\!\perp I \mid S \rightarrow \boxed{\text{True}}$

Simplifying the graph to get above nodes, we get

⇒



Since, D is dependent on both G & S and S is dependent on I

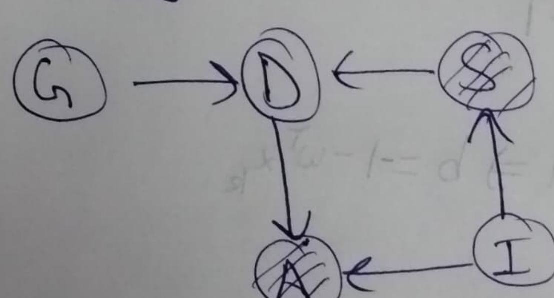
Hence, G is independent of I even when S is given

∴ G is conditionally independent on I given S.

⑤ $G \perp\!\!\!\perp I \mid S, A \rightarrow \boxed{\text{False}}$

Simplifying the graph to get above nodes, we get

⇒



Since, A is dependent on both I & D and D is dependent on both S & G and S is dependent on I

Hence, G is dependent on I even when both S & A are known.

∴ G is dependent on I given S and A.