# Homework 1: CS 6347 Stats for AI and ML

Thennannamalai Malligarjunan – txm230003

## Problem 1: Separation and Independence (30 pts)

1. Consider the following joint probability distribution, p. Find a directed graph G such that G is a perfect I-map for p. Is G the only directed graph with this property?

| $A$ | $B$ | $C$ | $p(A,B,C)$ |
|---|---|---|---|
| 0 | 0 | 0 | 1/4 |
| 0 | 0 | 1 | 1/4 |
| 0 | 1 | 0 | 1/24 |
| 0 | 1 | 1 | 1/8 |
| 1 | 0 | 0 | 1/8 |
| 1 | 0 | 1 | 1/8 |
| 1 | 1 | 0 | 1/48 |
| 1 | 1 | 1 | 1/16 |

**Solution:**

To find a Bayesian network that represents this joint probability distribution, we first need to find the independence relationships implied in this distribution. We need to consider both conditional and unconditional independence.

We can verify unconditional independence using,

$$P(A \cap B) = P(A).P(B)$$

and conditional independence using,

$$P(A \cap B \mid C) = P(A|C).P(B|C)$$

This is to be done for all possible combinations of variables.

Since we are only given the joint probability distribution, the first obvious step is to find the marginal distributions over each combination of variables.

Starting with P(A),

$$P(A) = P(A,0,0) + P(A,0,1) + P(A,1,0) + P(A,1,1)$$

$$P(A=0) = \frac{1}{4} + \frac{1}{4} + \frac{1}{24} + \frac{1}{8} = \frac{6+6+1+3}{24} = \frac{16}{24} = \frac{2}{3}$$

$$P(A=1)=\frac{1}{8}+\frac{1}{8}+\frac{1}{48}+\frac{1}{16}=\frac{6+6+1+3}{48}=\frac{16}{48}=\frac{1}{3}$$

Next P(B),

$$P(B)=P(0,B,0)+P(0,B,1)+P(1,B,0)+P(1,B,1)$$

$$P(B=0)=\frac{1}{4}+\frac{1}{4}+\frac{1}{8}+\frac{1}{8}=\frac{6}{8}=\frac{3}{4}$$

$$P(B=1)=\frac{1}{24}+\frac{1}{8}+\frac{1}{48}+\frac{1}{16}=\frac{12}{48}=\frac{1}{4}$$

Next P(C),

$$P(C)=P(0,0,C)+P(0,1,C)+P(1,0,C)+P(1,1,C)$$

$$P(C=0)=\frac{1}{4}+\frac{1}{24}+\frac{1}{8}+\frac{1}{48}=\frac{12+2+6+1}{48}=\frac{21}{48}=\frac{7}{16}$$

$$P(C=1)=\frac{1}{4}+\frac{1}{8}+\frac{1}{8}+\frac{1}{16}=\frac{4+2+2+1}{16}=\frac{9}{16}$$

Lets, first check whether unconditional independence relations hold.

$$P(A,B)=P(A).P(B)$$

$$P(B,C)=P(B).P(C)$$

$$P(A,C)=P(A).P(C)$$

P(A,B):

P(0,0) = P(0,0,1) + P(0,0,0) = 1/4 + 1/4 = 1/2

P(A=0).P(B=0) = 2/3 * 3/4 = 1/2

P(0,1) = P(0,1,0) + P(0,1,1) = 1/24 + 1/8 = 4/24 = 1/6

P(A=0).P(B=1) = 2/3 * 1/4 = 1/6

P(1,0) = P(1,0,0) + P(1,0,1) = 1/8 + 1/8 = 1/4

P(A=1).P(B=0) = 1/3 * 3/4 = 1/4

P(1,1) = P(1,1,1) + P(1,1,0) = 1/16 + 1/48 = 1/12

P(A=1).P(B=1) = 1/3 * 1/4 = 1/12

Thus, A and B are independent.

P(B,C):

P(0,0) = P(1,0,0) + P(0,0,0) = 1/8 + 1/4 = 3/8

P(B=0).P(C=0) = 3/4 * 7/16

Thus, B and C are NOT independent.

P(A,C):

P(0,0) = P(0,0,0) + P(0,1,0) = 1/4 + 1/24 = 7/24

P(A=0).P(C=0) = 2/3 * 7/16 = 7/24

P(0,1) = P(0,0,1) + P(0,1,1) = 1/4 + 1/8 = 3/8

P(A=0).P(C=1) = 2/3 * 9/16 = 3/8

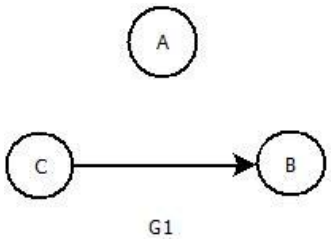
P(1,0) = P(1,0,0) + P(1,1,0) = 1/8 + 1/48 = 7/48

P(A=1).P(C=0) = 1/3 * 7/16 = 7/48

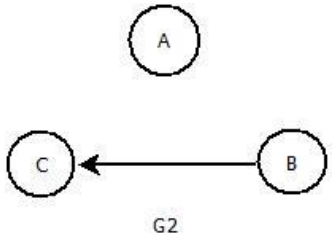
P(1,1) = P(1,0,1) + P(1,1,1) = 1/8 + 1/16 = 3/16

P(A=1).P(C=1) = 1/3 * 9/16 = 3/16

Thus, A and C are independent.
The DAGs that can represent these relationships are shown in the figure. Since there is only one edge in both graphs, there is no need to check the conditional independence relationships. (Even if checked, they would still be represented with these graphs)



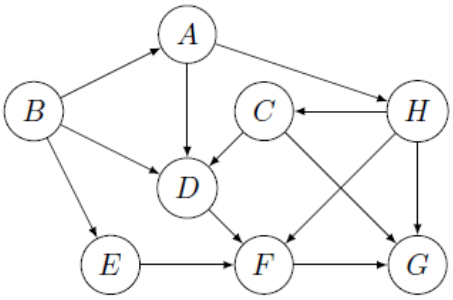
G1



G2

[Discussed with Rohan Vishal Rachamadugu]

2. Consider the following Bayesian network.




Use the above Bayesian network to answer the following questions. You must explain your answer for full credit.

(a) List all of the local Markov independence relations implied by this graph.

(b) Is G d-separated from A given C,H?

(c) Is B d-separated from H given A?

(d) Is H d-separated from E given A,G?

(e) Is E d-separated from A given B, F,H?

(f) Is A d-separated from D given C,B?

**Solution**:

Considering that A is the random variable of node A, B is for node B and so on.

a) We can find the local Markov independence relations for all nodes in the graph by taking their non-descendants and conditioning them on their parents.
   The local Markov independence relations are:
   A ⊥ E | B
   B ⊥ {} | {}
   C ⊥ A,B,E | H
   D ⊥ E,H | A,B,C
   E ⊥ A,C,D,H | B
   F ⊥ A,B,C | E,D,H
   G ⊥ A,B,D,E | F,C,H
   H ⊥ B,E | A

b) G is NOT d-separated from A given C & H.
   Example of one path where there is no node to block the path : A -> D -> F -> G

c) YES. B is d-separated from H given A.
   The paths are

1. B -> A -> H [Blocked at A]
2. B -> D <- C <- H [Blocked at D because of convergent sequence]
3. B -> E -> F -> G <- H [Blocked at G because of convergent sequence]
4. B -> E -> F <- H [Blocked at F because of convergent sequence]
5. B -> D <- A -> H [Blocked at both A and D]
6. B -> D -> F -> G <- H [Blocked at G because of convergent sequence]
7. B -> A -> D <- C <- H [Blocked at A and D]

d) H is NOT d-separated from E given A and G
Example path where H is not blocked from E : H -> G <- F <-E [since HGF is convergent sequence and G is given]

e) E is NOT d-separated from A given B,F and H
Example path where E is not blocked from A : E -> F <- D <- A [since EFD is convergent sequence and F is given]

f) A is NOT d-separated from D given C and B. This is because there is an edge connecting A and D.

3. For an undirected graph $G = (V,E)$, let $X$, $Y$, $Z_1$, $Z_2 \subseteq V$ such that all sets except for possibly $Z_1$ and $Z_2$ are mutually disjoint. Argue that if $X$ is graph separated from $Y$ given $Z_1 \cap Z_2$, then $X$ is graph separated from $Y$ given $Z_1$ and $X$ is graph separated from $Y$ given $Z_2$. Is the converse also true?

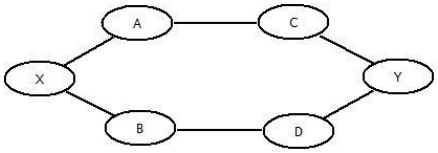
**Answer:**

1. Given that X is graph separated from Y given Z1 ⋂ Z2:
   Let S be the set of vertices from Z1 ⋂ Z2. This means that if we remove set of vertices S from the graph, then it becomes disconnected and X and Y are in different components.
2. X is graph separated from Y given Z1:
   We know that S ⊆ Z1. So removing vertices in Z1 will also remove vertices in S. Which means that X and Y will be in different components similar to case 1. So X is graph separated from Y given Z1.
3. X is graph separated from Y given Z2:
   The same logic from case 2 applies here since S is still a subset of Z2. Thus, X and Y are graph separated given Z2.

4. if X is graph separated from Y given $Z_1$ and X is graph separated from Y given $Z_2$ then X is graph separated from Y given $Z_1 \cap Z_2$:

This statement is not true. Suppose there is a vertex in Z1 that is essential for graph separating X and Y. If this vertex is not in Z2 then conditioning on $Z_1 \cap Z_2$ does not graph separate X and Y.

Example:




Z1 = {A,B,C}

Z2 = {C,D}

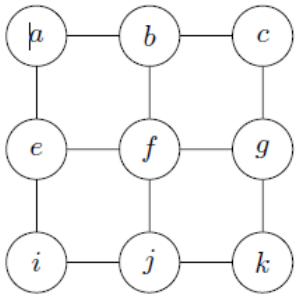
Z1 ∩ Z2 = {C}

As you can see, conditioning on either Z1 or Z2 makes X and Y separated but conditioning on Z1 ∩ Z2 does not.

---

## Problem 2: Treewidth (10 pts)

1. Compute the treewidth of the following 3 × 3 grid graph. Provide an optimal elimination ordering.




**Solution:**

We can see that eliminating any node from this graph will result in its neighbors getting connected to each other. Thus, in this graph eliminating a node will form a clique whose size is equal to the degree of the eliminated node. The optimal strategy is to eliminate

nodes so that the clique formed has minimal size. So, we select vertices based on their degree in ascending order. This is the min-degree heuristic.
One optimal elimination order is as follows,
a, i, k, c, b, j, e, g, f.

The treewidth of the graph is 3.

2. If the grid is enlarged to $n \times n$ nodes, how does the treewidth scale with $n$ (an asymptotic result is sufficient here)?

**Solution:**

The optimal elimination strategy is using the min-degree heuristic. Suppose we have an n x n grid. The vertices that have the minimum degree are at the four corners of the grid and their degrees are 2. After they are eliminated, their neighbors will have gained an edge between them. The eliminations after this will be done symmetrically. By continuing to eliminate the nodes with the minimum degree, at some point, there will be a fully connected component of the graph of size n+1. This would be the largest clique possible in the graph (by repeated optimal elimination). Elimination of a node in this clique would result in another clique of size n. So treewidth of the graph is n.
Hence for n x n grids, the treewidth scales by $\theta(n)$.

3. Explain why the treewidth of a graph is always at least as large as the size of a maximal clique in the graph minus one.

**Answer:**

Claim: For any graph G, the treewidth is always at least as large as the size of a maximal clique in G minus one.

Intuition:
Let G be a graph of n vertices. Let C be a maximal clique in the graph G of size k.
Every node in C is thus connected to k-1 nodes.
We know that eliminating a node from the graph will add an edge between all pairs of neighbors of the eliminated node.
So, after eliminating a node in clique C we will obtain another clique C′ of size k-1.
Thus, the treewidth of C, tw(C) = k-1.

Explanation:

Further, let G be a graph with treewidth tw(G).

By definition, a tree decomposition of G is a tree T such that each node in T corresponds to a subset (bag) of vertices in G, satisfying the following conditions:

- Every vertex in G is in at least one bag.
- For every edge in G, there exists a bag containing both vertices of the edge.

Let C be a maximal clique in G, and let v_1, v_2, ..., v_k be the vertices in C, where k is the size of the clique. Here, C is a subgraph of G.

Since C is a clique, all pairs of vertices in C are connected by edges in G.

Consider the bags in the tree decomposition. For each vertex v_i in C, there exists at least one bag in the tree decomposition containing v_i. Let's denote these bags as B_1, B_2, ..., B_k.

Since each vertex v_i is in at least one bag, we have bags B_1, B_2, ..., B_k corresponding to the vertices in C.

The size of the largest bag in the tree decomposition is at least max(|B_1|, |B_2|, ..., |B_k|).

Without loss of generality, assume that |B_1| is the largest among |B_1|, |B_2|, ..., |B_k|.

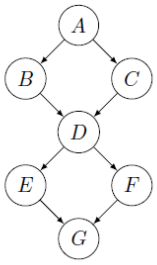
Therefore, $|B_1| \geq k$.

The treewidth (tw(G)) is defined as the size of the largest bag minus one, so $tw(G) \geq |B_1| - 1$.

Combining the inequalities, we get $tw(G) \geq |B_1| - 1 \geq k - 1$.

Since k is the size of the maximal clique C, we have $tw(G) \geq k - 1$.

## Problem 3: Marginal Bayes Nets (15 pts)




A Bayesian network is a minimal I-map for a distribution if removing any edges from the network will cause it to no longer be an I-map for the distribution. Consider the Bayesian network given above. Construct a new Bayesian network over all of the nodes except node D that is a minimal I-map for the marginal distribution over the remaining variables (A,B,C,E, F,G) for any probability distribution that factorizes over the given Bayesian network.
Be sure that your model contains all of the dependencies from the original network.

**Solution**:

Let the original graph be G. We need to construct a new network G' that is a minimal I-map for the marginal distribution over variables A, B, C, E, F and G. This means that G' must have all the dependencies from the G without adding new independence relationships.

Definition of minimal I-map: if removing an edge will no longer cause it to be an I-map for p.

Intuition: If you add an edge, you could potentially reduce the number of elements in the independent set whereas removing an edge can increase the number of elements in the independent set. If the independent set of the graph has elements not in the independent set of the distribution *p*, then the graph is no longer an I-map of *p*.

Thus, for $I(G') \subseteq I(G)$ we may only add edges in G' after removing the node D. But we also need G' to be a minimal I-map. So new edges if added should not introduce redundant dependencies.

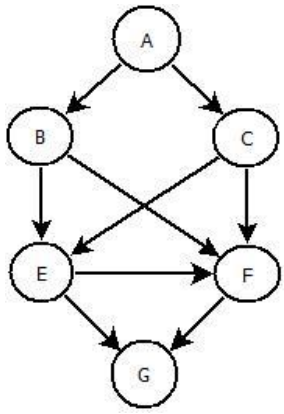
We first list out the independence relations over all nodes in G using d-separation. Since we are removing D from the graph, we only consider the independence relations not involving D.

1. $A \perp E \mid B,C$

2. $A \perp F \mid B,C$
3. $A \perp G \mid B,C$
4. $A \perp G \mid E,F$
5. $B \perp C \mid A$
6. $B \perp G \mid E,F$
7. $C \perp B \mid A$
8. $C \perp G \mid E,F$
9. $E \perp A \mid B,C$
10. $F \perp A \mid B,C$
11. $G \perp B \mid E,F$
12. $G \perp C \mid E,F$
13. $G \perp A \mid E,F$
14. $G \perp A \mid B,C$

Another thing to note is that if D is not given, then the pairs (B, E), (B, F), (C, E), (C, F) and (E, F) are not d-separated. Hence, we add edges connecting these pairs in the new network G'.

The new network G' is shown in the following figure. This is a minimal I-map for the marginal distribution over the remaining variables (A,B,C,E, F,G) for any probability distribution that factorizes over the given Bayesian network.




[Discussed with Rohan Vishal Rachamadugu]

## Problem 4: Vertex Covers (15 pts)

Consider an undirected graph G = ($V_G$,$E_G$). A vertex cover is a subset S ⊆ $V_G$ such that every edge in $E_G$ is incident to at least one vertex in S.

1. Explain how to construct a Markov random field to represent the uniform probability distribution, p, over valid vertex covers of G.

**Answer:**

Let $X_1, X_2, \dots, X_n$ be n binary random variables each corresponding to the n vertices in G.

$$i \in V, X_i = \begin{cases} 1 & if\ i \in S \\ 0 & Otherwise \end{cases}$$

The random variables are said to form a Markov Random Field over G if it factorizes as follows,

$$P(X_1, X_2 \dots, X_n) = \frac{1}{Z} \prod_{C \in C(G)} \psi_C(X_C)$$

Let the potential functions be defined over all edges in graph G since each edge in general is a clique.

$$\psi_{i,j}(X_i, X_j) = \begin{cases} 1 & if\ X_i + X_j \geq 1 \\ 0 & otherwise \end{cases}$$

So the distribution can be expressed as,

$$P(X_1, X_2 \dots, X_n) = \frac{1}{Z} \prod_{(i,j) \in E_G} \psi_{i,j}(X_i, X_j)$$

And Z = the size of set containing all valid vertex covers of G.

[Discussed with Rohan Vishal Rachamadugu]

2. Explain why your construction always yields a valid probability distribution.

**Answer**:

The vertex covers of G and the potential functions are closely related in this construction. This is because we have defined the potential functions over all edges in the graph.

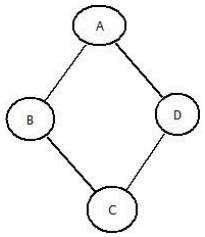
Now if we want to find the probability that a set of vertices represented by the binary vector $X_v$ of order n forms a vertex cover of G, we just select the potential functions[in a sense the

edges] corresponding to the vertices by setting it to 1. Setting the potential function to 0 means that the edge is not covered by the vertices in the set which satisfies the definition of vertex covers.
The product of these potential functions must be 1 if the set of vertices is a vertex cover so that $P(X_v) = \frac{1}{z}$ and 0 if the set of vertices is not a vertex cover.
This corresponds to the joint probability distribution we wish to represent.

3. Let G be the graph consisting of a single cycle on four nodes. What is the partition function (normalizing constant) of the MRF for this choice of G?

**Answer:**

Let G = ({A,B,C,D},{(A,B), (B,C), (C,D), (D,A)})
Valid vertex covers are {{A,C}, {B,D}, {A,B,C}, {A,D,C}, {B,A,D}, {B,C,D}, {A,B,C,D}}




The normalizing constant z = 7 in this case.
Since the size of the valid vertex cover set is 7 and the probability must be distributed uniformly over this space.

4. Explain how to construct an MRF to represent a probability distribution over vertex covers such that for all vertex covers S, p(S) $\propto$ exp(|S|).

**Answer:**

Now the potential function needs to be modified to include the exp() so that P(S) is proportional to exp(|S|).

Let the potential functions be defined over all edges in graph G .

Now,

$$\psi_{i,j}(X_i, X_j) = \begin{cases} e^1 & if\ X_i + X_j > 1 \\ e^{1/2} & if\ X_i + X_j = 1 \\ 0 & if\ X_i + X_j < 1 \end{cases}$$

The distribution can be expressed as,

$$P(X_1, X_2 \dots, X_n) = \frac{1}{Z} \prod_{(i,j) \in E_G} \psi_{i,j}(X_i, X_j)$$

---