Lecture scribe for Machine Learning(CS60050), Spring 2020-21

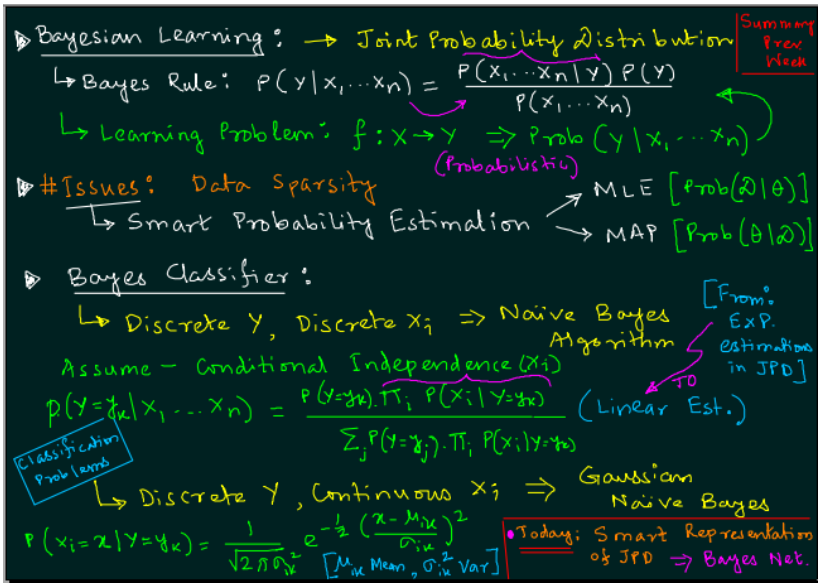
Instructor: Prof. Aritra Hazra

Topic: Bayesian Networks

Date: 27-1-2021

Prepared By: Somarpita Dutta(20CD92R02)

Board 1 (refer slide time 0:09)



Recapitulation of the previous week's agendas

1. Probabilistic notion of joint probability distribution

-It is computationally very expensive because for n attributes, there will be 2^n entries in Joint Probability Distribution table (JPDT)

- **Data sparsity**: This problem arises due to lack of data to fill up all the 2^n entries of JPDT

- In reality, we need to deal with lakhs of attributes. In that case, the number of entries in JPDT will be huge and it will be tough to get all the data. The problem of data sparsity will be manifold.

2. **Solution to the above problem**: We can be smart in 2 ways-

- **Maximum Likelihood Estimation (MLE) and Maximum Aposteriori Estimation (MAP)** : We may not get all the probabilities but we have to estimate it smartly.

- **Bayesian networks** (This will be discussed in the lecture today)

3. **Bayes rule and Naive Bayes classifier**

- Bayes' rule, being used directly with all the attributes gives us an exponential number of estimates

- We assume the notion of **conditional independence inherent in Naive Bayes classifier** where each attribute is assumed to be conditionally independent of the other and this reduces the estimate of probability drastically to a linear range of values.

- After assuming conditional independence, we can pre-estimate the probability using MLE and when a new attribute comes, we can classify it based on that easily.

- In certain cases, the attributes possess continuous values for eg. classification problem dealing with whether one has brain tumor or not from x-ray scan image.

- In such cases, we approximate the continuous probability distribution using Gaussian based approximations. The algorithm to be used for Naive Bayes classifier is almost the same with the difference that we need to learn the value of mu and Sigma instead of learning each and every x_i given y_k.

4. **Theory of conditional independence**

- The theory of conditional independence is a strong assumption and it does not hold always.

- Naive Bayes finds use in applications like spam filtering.

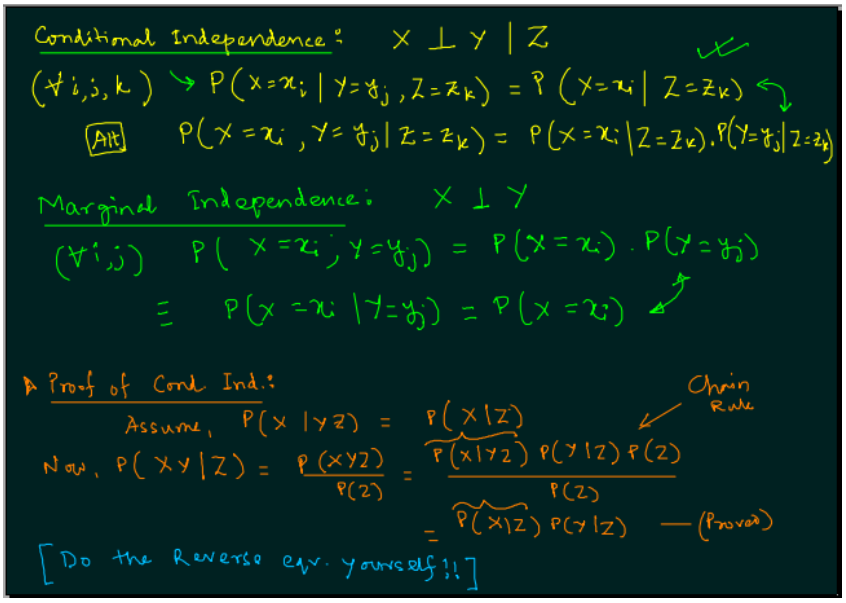
5. **Challenge is not over yet**

- It will be great if we could relax the notion of conditional independence a bit.

- Data sparsity will be obvious when we have to handle large number of attributes. So, a much structured representation of JPDT is desirable.

- These 2 aspects will be handled in the lecture today.

Board 2 (refer slide time 6:03)



2 definitions are applicable in this regard-

1. **Conditional independence: X is independent of Y given Z**

If X can take X_1, X_2, X_3,...X_k values, Y can take Y_1, Y_2,... values, similarly Z can take Z_1, Z_2,... values.

for all i,j,k $P(X=x_i|Y=y_j, Z=z_k)=P(X=x_i|Z=z_k)$

This implies that for all i,j,k, probability of $X=x_i$ given $Y=y_j$ and $Z=z_k$ is equal to probability of $X=x_i$ given $Z=z_k$.

Alternate representation:

$P(X=x_i, Y=y_j|Z=z_k)=P(X=x_i|Z=z_k)P(Y=y_j|Z=z_k)$

This implies that probability of $X=x_i$, $Y=y_j$ given $Z=z_k$ is equal to $X=x_i$ given $Z=z_k$ multiplied by probability of $Y=y_j$ given $Z=z_k$.

These 2 definitions are equivalent.

2. **Marginal independence**

X is marginally independent of Y if for all i,j, probability of $X=x_i$, $Y=y_j$ is equal to prob of $X=x_i$, multiplied by probability of $Y=y_j$.

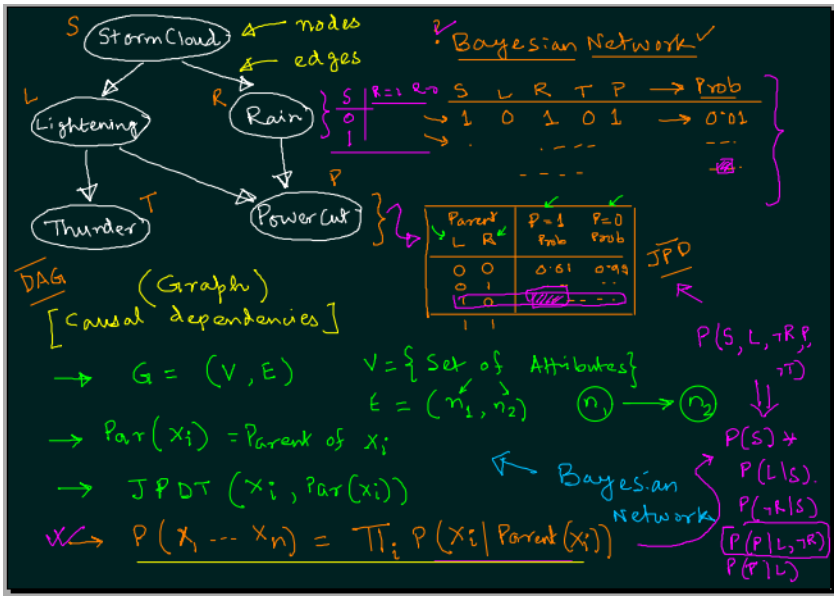
for all i,j $P(X=x_i, Y=y_j)=P(X=x_i)P(Y=y_j)$

Another representation:

Probability of $X=x_i$ given $Y=y_j$ is equal to probability of $X=x_i$

$P(X=x_i|Y=y_j)=P(X=x_i)$

Board 3 (refer slide time 9:21)



An example to illustrate the aforementioned ideas is given below-

- Suppose you have stormclouds, lightning, rain, thunder and power cut(assuming, in India).

- **In Bayesian networks, we try to assume that whatever be the outcome, it is independent of the attributes**
- **But, usually, this is not so**. For eg, stormcloud may be reason for Rain or lightning. Rain may be a reason for power cut or lightning may be a reason for power cut. Lightning is definitely a reason for thunder but rain is not always the reason for thunder.
- **The attributes are represented as rows and relationships as edges**.
- A graph can be produced out of these causal dependencies.
- **This graph is DAG(Directed Acyclic Graph)**.
- It is directed because a particular attribute is responsible for it to be directed to some other attributes.
- It is acyclic because it is not possible that stormcloud causes rain as well rain also causes stormcloud and vice versa or there could not be any cycle in the graph.
- Such a network is referred to as the Bayesian network.
- There are certain causal dependencies among the attributes, for eg, rain can only be caused by stormcloud, power cut can only happen if there is rain as well as lightning.
- In the JPDT being constructed, every entry is equipped with a probability value for eg there is stormcloud, rain , power cut which is assigned probability 0.01. JPDT is complete and it has to specify all possibilities. But, **in Bayesian networks, every attribute has relationship with only its parents.**
- Hence, **every attribute will have its local table of joint probability distribution instead of a global table accounting all attributes**.
- In a local table, the parents of a node will be listed and for each possible value of the child attribute, a separate probability will be assigned.
- Now **the local JPDT is attached with every node**.
- **Bayesian network scores over the global JPDT by reducing the number of entries in the table because it involves only the parent attributes of a particular attribute** eg. in the local table for power cut, there are only lightning and rain, who are the parents of power cut. Hence, local JPDT for power cut contains 4 entries.

Formal definition of Bayesian networks-

It is a graph containing sets of vertices and edges.

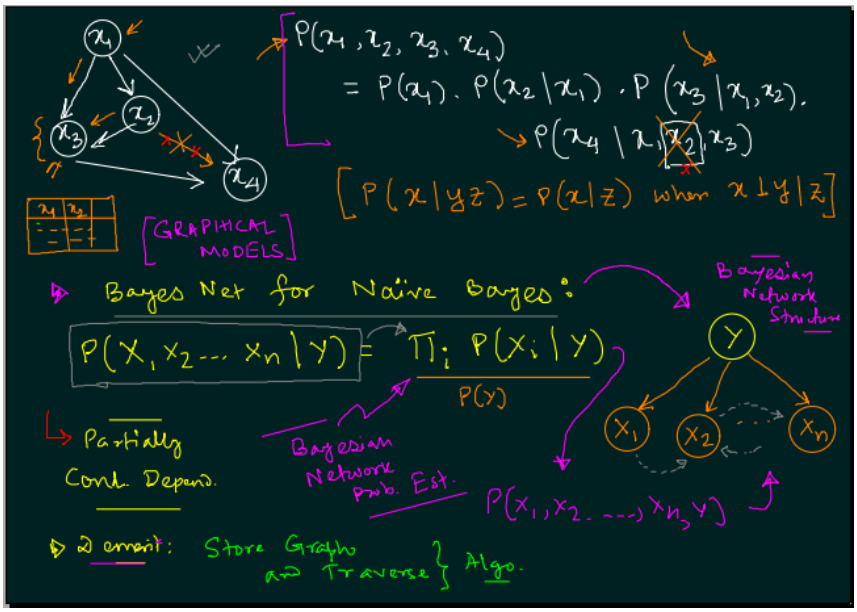
It is a directed graph.

Vertices are the set of attributes.

Edges represent the set of causal dependencies (directed) parent of a node Xi is/are all the node(s) from where there is incoming edge to Xi.

There is a JPDT for each of the nodes (attributes) Xi and its parents.

Board 4 (refer slide time 19:15)



Graphical representation of dependencies

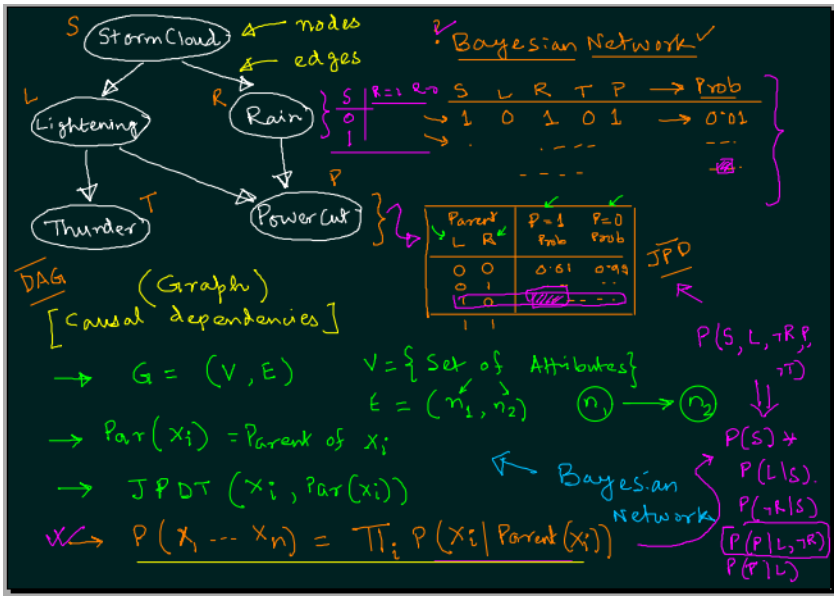
- There is a simple graph containing nodes X_1, X_2, X_3, X_4 and dependencies between attributes are shown by directed edges in the graph.
- Probability of X_1, X_2, X_3, X_4 or. **$P(X_1, X_2, X_3, X_4)=P(X_1).P(X_2|X_1).P(X_3|X_1,X_2).P(X_4|X_1, X_2 X_3)$.**
    - **X_1 is standalone** (not dependent on any event). Hence, probability of X_1 can be expressed without any dependence on any other attribute ie. $P(X_1)$
    - **There is a directed edge from X_1 to X_2 which shows that X_2 has a dependency on X_1**. Hence, to estimate the probability of X_2, we have to consider the dependence of X_2 on X_1 ie. $P(X_2|X_1)$.
    - **Probability of X_3 is dependent on X_1 and X_2 values**. So, we can write $P(X_3|X_1, X_2)$.
    - **Probability of X4 is dependent on X_1, X_2 and X_3**. Hence, we write $P(X_4|X_1, X_2, X_3)$.
    - This follows from the chain rule and conditional independence factors.
- **Removal of edge from X_2 to X_4 implies that X_2 and X_3 are not conditionally dependent any more.**
- Rule: If there are 2 descendants of the same parent, if they are descended on different branches and if they do not share any dependency edge, then the 2 descendants are conditionally independent.

- Similarly, if X_2 and X_4 are descendants of X_1 , both descend from different branch and also do not share any dependency edge. Hence, X_2 and X_4 are conditionally independent (after edge removal).
- By definition of conditional independence,

  P(X|YZ)=P(X|Z) when X is conditionally independent of Y given Z.

- In this case X_4 is conditionally independent of X_2.
- This implies X_2 vanishes from the term $P(X_4|X_1,X_2,X_3)$ (since there is no conditional dependency between X_4 and X_2.
- Each node has a local JPDT. Hence, the probability of each attribute can be computed from the partial JPDT and the conditional independence assumption.

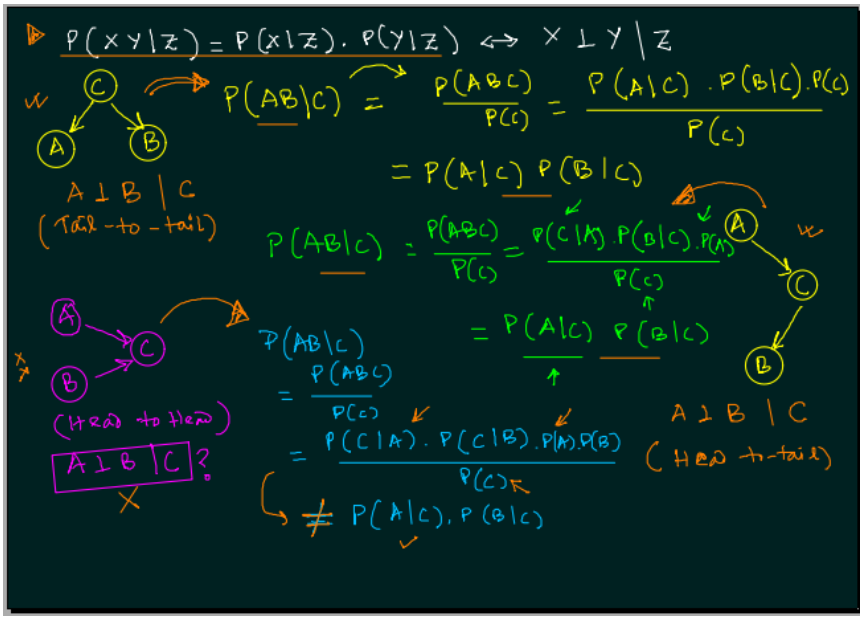
Let us refer Board 3 once again,



- **The fundamental rule of joint probability distribution of attributes X_1, X_2, X_3,...,X_n is given as $P(X_1, X_2,...,X_n)=\sum_i P(X_i|$ parent$(X_i))$**
- Each $P(X_i|$ parent$(X_i))$ can be obtained from the partial JPDT associated with that node.
- Hence**, the problem of data sparsity has been reduced by the smart use of causal dependency, local JPDT and notions of probability.**
- Following the above notion,
  $P(S,L,\neg R,P,T)=P(S)*P(L|S)*P(\neg R|S)*P(P|L,\neg R)*P(P|L)$
- We basically need to traverse the graph using any graph traversal algorithm and JPDT will pick up the value.

- **Bayes network for Naive Bayes**
    - As per Naive Bayes'rule, $P(X1,X2,....,Xn|Y)=\sum_i P(Xi|Y)/P(Y)$
    - The Bayes network corresponding to this rule will contain Y as the parent of each X_i ie. there will be edges from Y to each X_i. So, we have to calculate $P(X_i|Y)$ from the local JPDT of each X_i and then divide the whole expression by probability of Y(Y is not dependent on any attribute) which will be obtained from the local JPDT of Y.
    - But **it overlooks the dependencies between X_1, X_2, X_3,.....,X_n which counts for a major drawback of Naive Bayes classifier.**
    - **Bayes network provides a much more compact representation of the dependencies between attributes.**
    - If the graph changes, the probability notions of attributes (as well as the dependencies among attributes) changes.
    - Bayes network is used for assessing causal relationships for eg. medical diagnosis of patients.
    - **Demerits : Bayes network has a representational structure and substantial memory required to store the graph and we also need graph traversal algorithm for traversing the graph.**
    - Bayes network is a much older concept as it is related to AI.
    - The colloquial term for Bayes network was Graphical models.

Board 5 (refer slide time 40:05)



A study of tail-to-tail, head-to-tail and head-to-head structures

Revisiting the definition of conditional independence-

If P(XY|Z)=P(X|Z)P(Y|Z), we say X is conditionally independent of Y given Z.

First graph(tail-to-tail structure)-

- Is it possible to say that A and B are conditionally independent given C?
- Applying Bayesian rule,
- P(AB|C)=P(ABC)/P(C)
- P(ABC)=P(A|C)P(B|C)P(C) (as per Bayesian network causal dependencies)
- Now **P(ABC)/P(C)=P(A|C)P(B|C)P(C)/P(C)=P(A|C)P(B|C)**
- This implies A and B are conditionally independent of each other.
- Hence, P(AB|C)=P(A|C)P(B|C)
- **In this particular network structure A and B are conditionally independent of each other.**
- **This network structure is referred to as tail-to-tail structure (A and B are 2 tails).**

Second graph (Head-to-tail structure)

- Are A and B conditionally independent of each other given C?
- P(AB|C)=P(ABC)/P(C) (as per Bayes' rule)
- As per the network structure, P(ABC)=P(A)P(C|A)P(B|C)
- **P(ABC)/P(C)=P(A)P(C|A)P(B|C)/P(C)=P(A|C)P(B|C)**
- Hence, it is proved that alike tail-to-tail structure, in head-to-tail structure also, A and B are conditionally independent of each other given C.

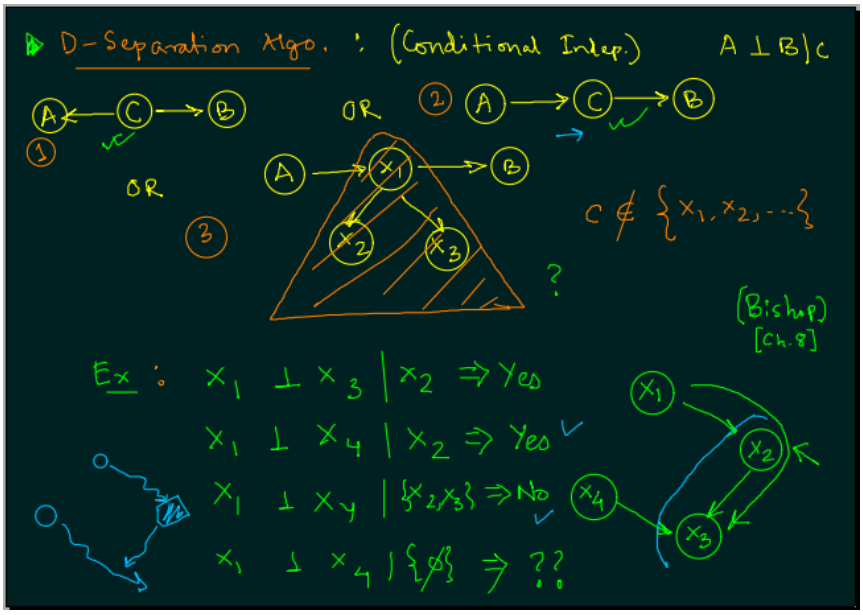
Third Graph (Head-to-head structure)

- As per Bayes' rule, P(AB|C)=P(ABC)/P(C)
- From Bayes' network, we observe, P(ABC)=P(A)P(B)P(C|A)P(C|B)
- **P(ABC)/P(C)=P(A)P(B)P(C|A)P(C|B)/P(C)**

=P(AC)P(B)P(C|B)/P(C)
=P(A|C)P(B)P(C|B)

- The above is not equal to P(A|C)P(B|C)
- Hence for this structure A and B are not conditionally independent of each other given C.
- Also, **it is observed that A and B cannot be conditionally independent of each other if both are the causes of C.**

Board 6 (refer slide time 52:05)



D separation algorithm

- This algorithm is used to determine conditional independence.
- A is said to be conditionally independent of B given C if the node C sits between the attributes A and B.
- In all 3 cases, we can say A is conditionally independent of B given C.

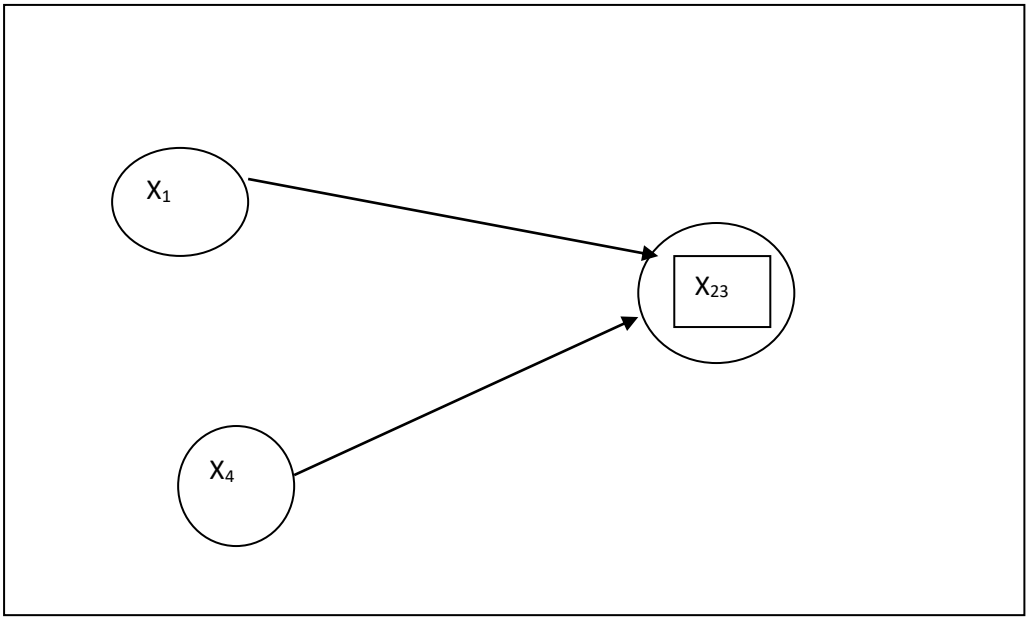
Graph structure

- **Is X_1 conditionally independent of X_3 given X_2?**
  - **Yes, because X_1, X_2, X_3 give rise to a head-to-tail structure.**
  - In head-to-tail structure head and tail are conditionally independent of each other.

- Hence X_1 and X_3 are conditionally independent of each other given X_2.

- **Is X_1 conditionally independent of X_4 given X_2?**
  - There is an edge from X_1 to X_2.
  - There is no edge from X_4 to X_1 or X_2.
  - Hence, **there is no scope of head-to-head structure arising here.**
  - **Hence, X_1 is conditionally independent of X_4 given X_2.**

- **Is X_1 conditionally independent of X_4 given X_2, X_3?**
  - Consider the nodes X_2 and X_3 being merged. Let it be X_{23}.
  - Hence there are edges from X_1 to X_{23} and X_4 to X_{23}(shown below).
  - **A head-to-head structure is created with X_1 and X_4 being 2 heads and edges directed from X_4 to X_{23} and from X_1 to X_{23}.**
  - **Hence, X_1 is not conditionally independent of X_4 given X_2, X_3.**



All in all, Bayes network helps us to handle the dependencies which was not that flexible to be handled in Naive Bayes. The demerits are that the graph structure needs to be stored.