Graphical Models

INSTRUCTOR: HONGJIE CHEN

JUNE 13TH 2022

Graphical Models

- **Goal**:
  - Express sets of conditional independence assumptions via a graph structure
  - A Graph structure with associated parameters define joint probability distribution over set of variables/nodes

Recall Conditional Independence

- X is conditionally independent of Y given Z, if the probability distribution governing X is independent of the value of Y given the value of Z
  - $\forall i, j, k, P(X = x_i \mid Y = y_j, Z = z_k) = P(X = x_i \mid Z = z_k)$
  - Or equivalently $P(X \mid YZ) = P(X \mid Z)$

- P(Thunder|Rain, Lightning)= P(Thunder|Lighting)
  - If there is lighting, the probability of thunder is indepedent to the probability of rain, or they are conditionally indepedent.

Marginal Independence

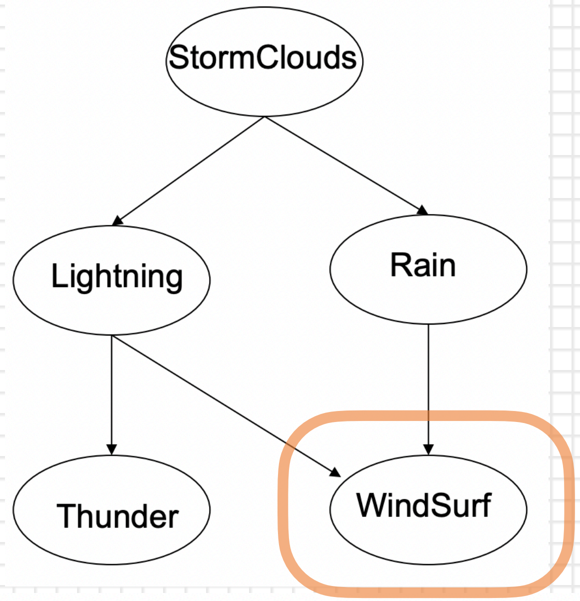
- X is marginally indepedent of Y if
  - $\forall i, j, P(X = x_i \,|\, Y = y_j) = P(X = x_i)$
  - Or equivalently $\forall i, j, P(Y = y_i \,|\, X = x_j) = P(Y = y_i)$

VIRGINIA TECH

Bayesian Network

- A Bayes network is a Directed Acyclic Graph (DAG) defining a joint probability distribution over a set of random variables
- Each node denotes a random variable
- Each edge denote a dependency of the edge receiver on the edge sender
- A conditional probability distribution (CPD) is associated with each node N, defining P(N | Parents(N))
- The joint distribution over all variables is defined as

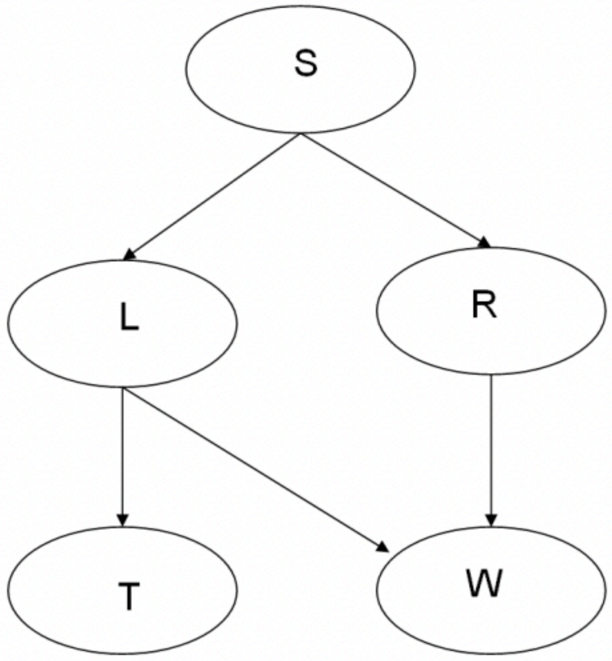
$$P(X_1, \ldots, X_n) = \prod_i P(X_i \mid Pa(X_i))$$



Parents	P(W|Pa)	P(¬W|Pa)
L, R	0	1.0
L, ¬R	0	1.0
¬L, R	0.2	0.8
¬L, ¬R	0.9	0.1

Conditional Independence in Bayesian Network

- Each node is conditionally independent of its non-descendents, given only its immediate parents
- How to represent $P(S, L, R, T, W)$

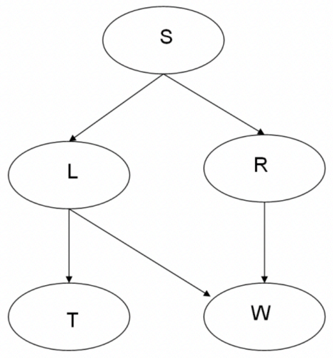


Represent $P(S, L, R, T, W)$

- Chain rule of probability

$$P(S, L, R, T, W) = P(S)P(L \mid S)P(R \mid S, L)P(T \mid S, L, R)P(W \mid S, L, R, T)$$

- With $P(X_1, \ldots, X_n) = \prod_i P(X_i \mid Pa(X_i))$

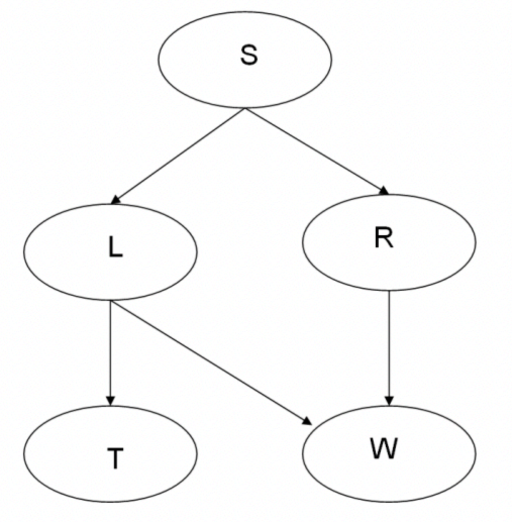


$$P(S, L, R, T, W) = P(S)P(L \mid S)P(R \mid S)P(T \mid L)P(W \mid L, R)$$

VIRGINIA TECH

Parameter Reduction

- How many parameters are needed?
- Without Bayesian network, $P(S, L, R, T, W)$, each random variable is boolean,
- With the Bayesian network?
  - Count the number of rows of each conditional probability table, and sum them up



Bayes Network Construction Algorithm

- Choose an ordering over variables, e.g. $X_1, X_2, \ldots, X_n$
- For i=1 to n
  - Add X_i to the network
  - Select parents $Pa(X_i)$ as minimal subset of $X_1, X_2, \ldots, X_{i-1}$ such that $P(X_i \mid Pa(X_i)) = P(X_i \mid X_1, \ldots, X_{i-1})$
- This assures

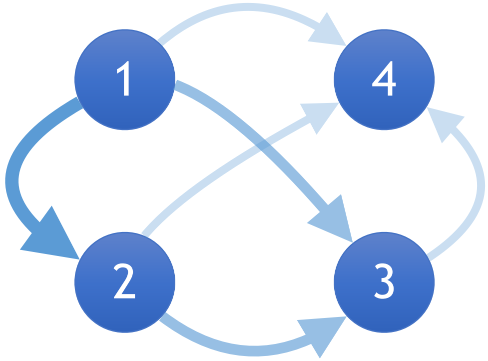
$$P(X_1, \ldots, X_n) = \prod_i P(X_i \mid X_1, \ldots, X_{i-1}) = \prod_i P(X_i \mid Pa(X_i))$$

Bayes Network with a Full Distribution

- What is the Bayes Network for $X_1, X_2 \ldots, X_n$ with no assumed conditional independece?
- X_1, X_2, X_3, X_4
- $P(X_1, \ldots, X_4) = \prod_i P(X_i \mid X_1, \ldots, X_{i-1}) = \prod_i P(X_i \mid Pa(X_i))$

Bayes Network with a Full Distribution

- What is the Bayes Network for $X_1, X_2 \ldots, X_n$ with no assumed conditional independedce?
- X_1, X_2, X_3, X_4
- $$P(X_1, \ldots, X_4) = \prod_i P(X_i | X_1, \ldots, X_{i-1}) = \prod_i P(X_i | Pa(X_i))$$
- Number of parameters 15

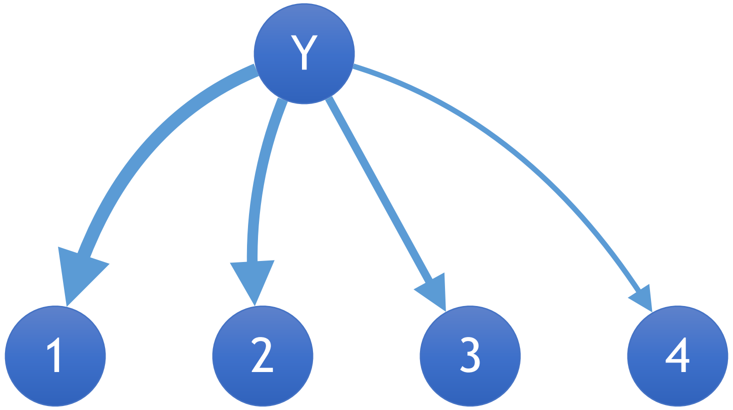


Bayes Network for Naïve Bayes

- $P(Y \mid X_1, \ldots, X_4) \propto P(Y)P(X_1 \mid Y)P(X_2 \mid Y)P(X_3 \mid Y)P(X_4 \mid Y)$

Bayes Network for Naïve Bayes

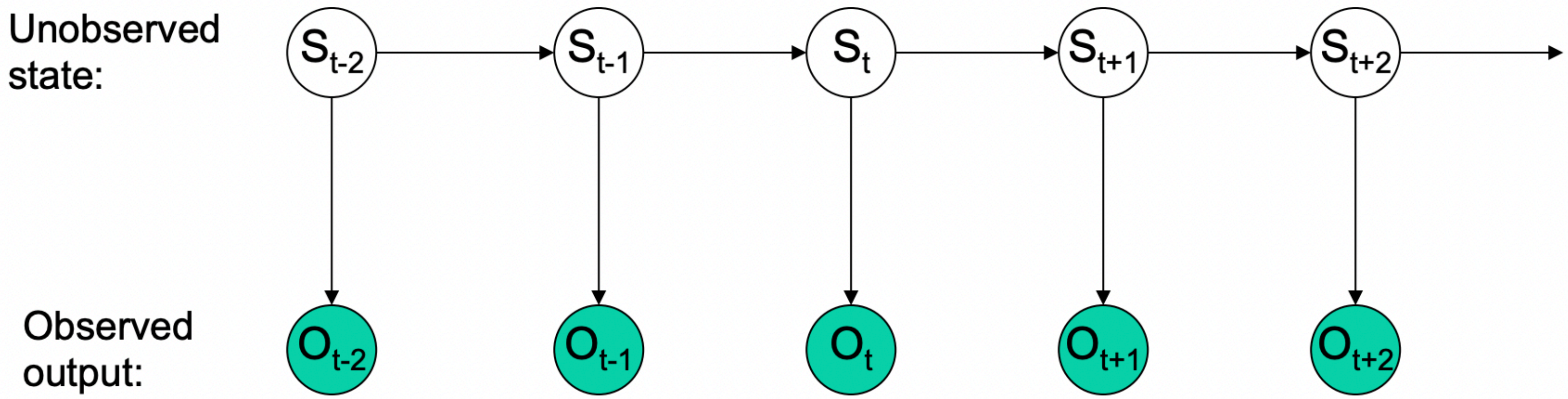
- $P(Y \mid X_1, \ldots, X_4) \propto P(Y)P(X_1 \mid Y)P(X_2 \mid Y)P(X_3 \mid Y)P(X_4 \mid Y)$



- Assumption help reduce parameters

Hidden Markov Model

- Assume the future is condtionally independent of the past, given the present.

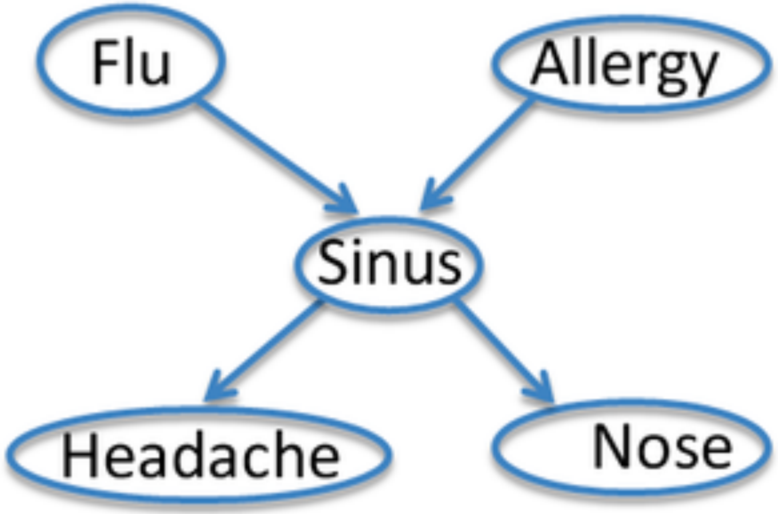


$$P(S_{t-2}, O_{t-2}, S_{t-1}, \ldots, O_{t+2}) =$$

Inference in Bayes Networks

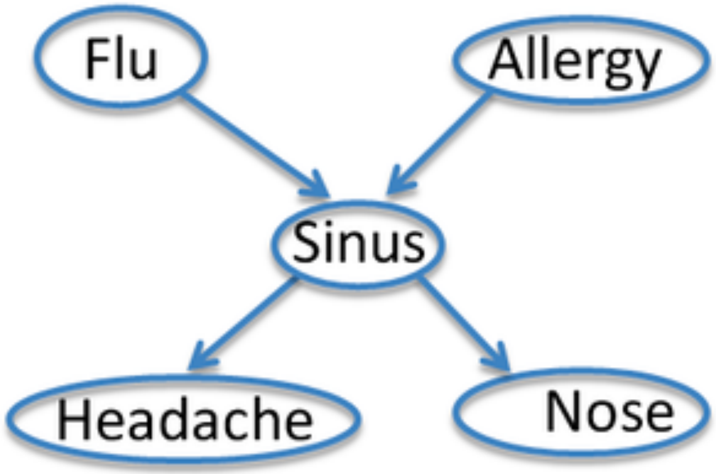
- In general, intractable (NP-complete) , Don't know $P(Data)$
- For certain cases, tractable
  - Assign probablity to fully observed set of variables
  - Or if only one variable is unobserved
  - Or for singly connected graphs (i.e. no undirected loops)
    - Belief propagation
- Monte Carlo methods
  - Generate samples and count up the result
  - Calculate π

Bayes Network Example



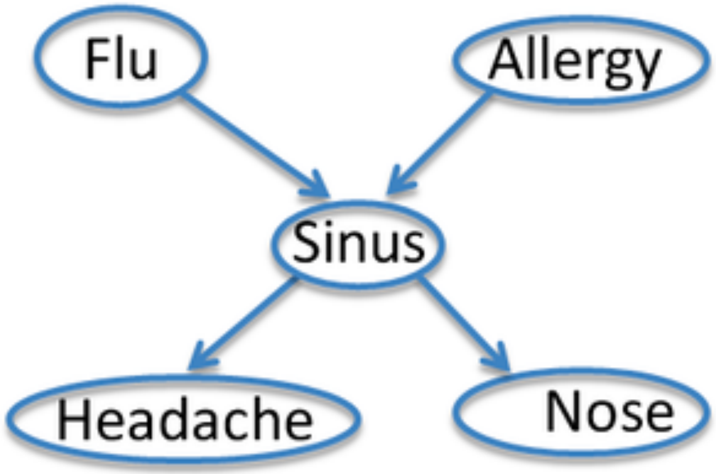
Assign joint probability

- Suppose we want to calculate joint probability of $P(F = f, A = a, S = s, H = h, N = n)$
- f, a, s, h, n are actual values.
- Let's use a shorthand representation $P(f, a, s, h, n)$



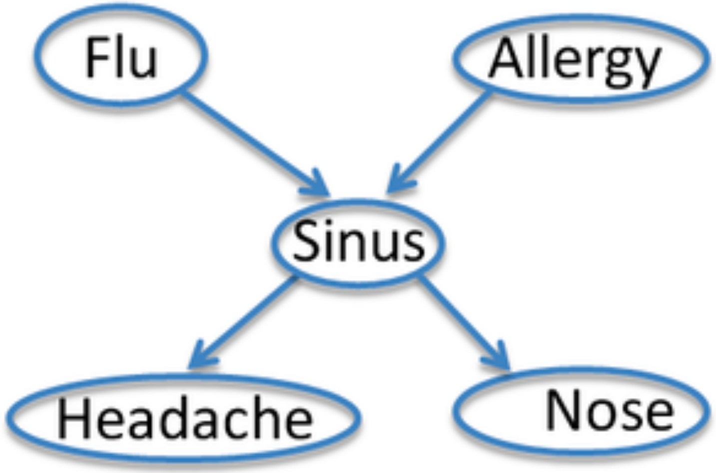
Calculate $P(f, a, s, h, n)$

- $P(f, a, s, h, n) = P(f)P(a)P(s \mid fa)P(h \mid s)P(n \mid s)$
- Inference is linear to number of random variables.



Calculate Marginal Probability

- For example, calculate $P(N = n)$

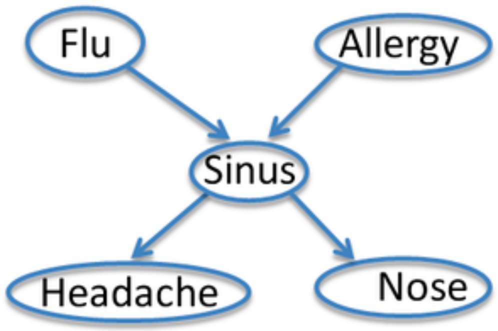


Calculate $P(N = n)$

- $P(N=n) = \sum_{s} P(N=n \mid S=s) P(S=s)$
- Now we have to calcualte $P(S = s)$, and go all the way up

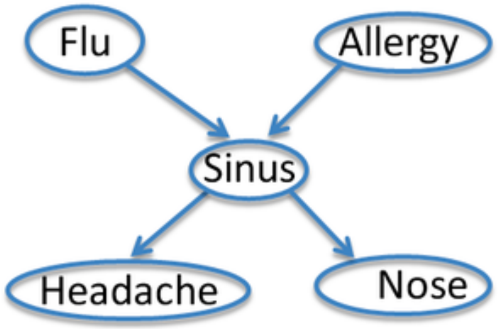
$$P(N=n) = \sum_{f,a,h,s} P(f,a,h,s,n)$$

$$= \sum_{f,a,h,s} P(f)P(a)P(s \mid fa)p(n \mid s)$$



- Exponential growth: computationally expensive

Monte Carlo



- To generate random samples is easy

- Assume a $P(F = 1) = \theta$, draw a value r uniformly randomly from $[0,1]$, if $r < \theta$ then let $F = 1$
- Also draw for other random variables.

- Then we count the fraction of samples where $N = n$

VIRGINIA TECH.

Learning Bayes Network

- Case 1: When graph is *known*, data are *fully observed*
- Case 2: When graph is *known*, data are *partly known*

Learning Bayes Network with Fully Observed Data

- For example

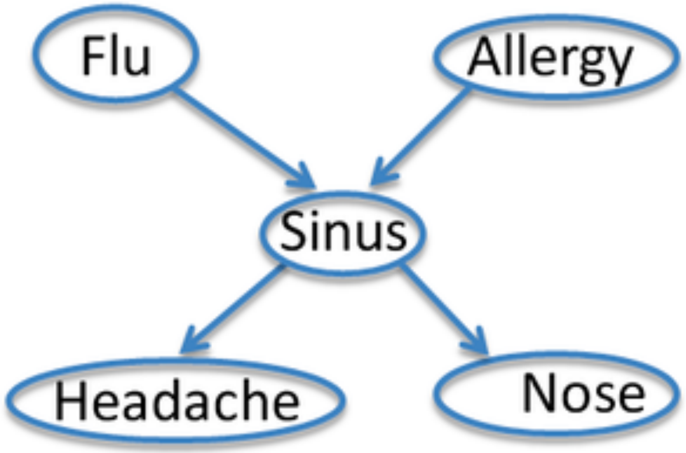
$$\theta_{s|ij} = P(S = 1 \mid F = i, A = j)$$

- MLE

***K* data points**

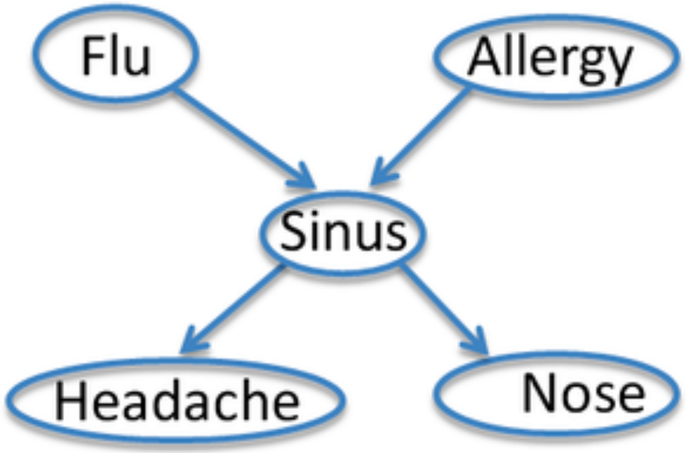
$$\theta_{s|ij} = \frac{\sum_{k=1}^{K} \delta(f_k = i, a_k = j, s_k = 1)}{\sum_{k=1}^{K} \delta(f_k = i, a_k = j)}$$

The fraction of rows under the given condition('s rows)



MLE of $\theta_{s|ij}$ from Fully Observed Data

- MLE
  - $\theta \leftarrow \text{argmax}_\theta \log P(D \mid \theta)$
- Calculation



$$P(D \mid \theta) = \prod_{k=1}^{K} P(f_k, a_k, s_k, h_k, n_k)$$

$$= \prod_{k=1}^{K} P(f_k)P(a_k)P(s_k \mid f_k a_k)P(h_k \mid s_k)P(n_k \mid s_k)$$

$$\log P(D \mid \theta) = \sum_{k=1}^{K} \log P(f_k) + \log P(a_k) + \log P(s_k \mid f_k a_k) + \log P(h_k \mid s_k) + \log P(n_k \mid s_k)$$

$$\frac{\partial \log P(D \mid \theta)}{\partial \theta_{s|ij}} = \sum_{k=1}^{K} \frac{\partial \log P(s_k \mid f_k a_k)}{\partial \theta_{s|ij}}$$

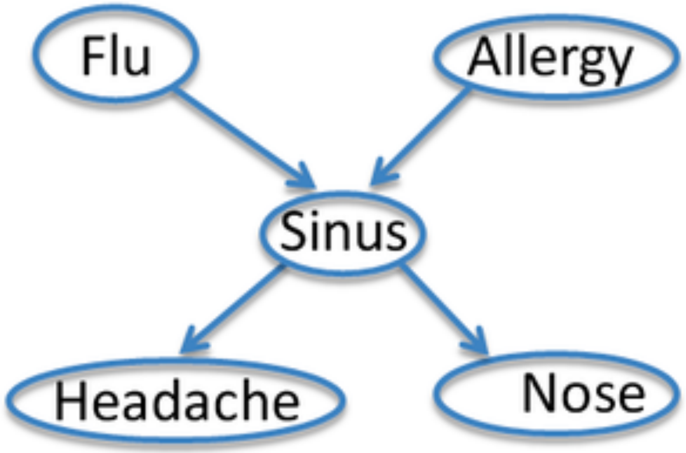
$$\hat{\theta}_{s|ij} = \frac{\sum_{k=1}^{K} \delta(f_k = i, a_k = j, s_k = 1)}{\sum_{k=1}^{K} \delta(f_k = i, a_k = j)}$$

$$\theta_{s|ij} = P(S = 1 \mid F = i, A = j)$$

VIRGINIA TECH

MLE of $\theta_{s|ij}$ from Partially Observed Data

- If data are partially observed
  - For example, S is not observed
- $\theta \leftarrow \text{argmax}_\theta \log P(D|\theta)$
  - Let X be all observed variables
  - Let Z be all unobserved variables
  - $\theta \leftarrow \text{argmax}_\theta \log P(X, Z|\theta)$
- Can’t calculate since Z is unknown, solution?
- Expectation Maximation



VIRGINIA TECH.