Bayesian Networks

Lecture 6:
Bayesian Networks

ECE/CS 498 DS
Professor Ravi K. Iyer
Department of Electrical and Computer Engineering
University of Illinois

Announcements

- HW 1 released today, due **Mon Feb 17th @ 11:59 PM** on Compass2G
  - Topic: basic Pandas review with Python
  - To be done individually
- **In class activity 2 this Wed, Feb 12th**
  - Will cover example related to Bayesian Networks
- TA office hours have been moved to **ECEB 3015**
  - Same time: MW 4-5 PM
- Discuss section this week (2/14) will be additional practice with Bayesian Networks

Naïve Bayes Classifier: Recap

- Driving Question: Given a **previously unseen** data point $\boldsymbol{x} = (x_1, x_2, \dots, x_n)$, which class C_k does the data point belong to?
- A Naïve Bayes classifier can be derived from three core concepts:
  - Chain rule

$$p(\boldsymbol{x}|C_k) = p(x_1, x_2, \dots, x_n|C_k)$$
$$= p(x_1|x_2, \dots, x_n, C_k)p(x_2|x_3, \dots, x_n, C_k) \dots p(x_{n-1}|x_n, C_k)p(x_n|C_k)$$

  - Class Conditional Independence

$$p(\boldsymbol{x}|C_k) = p(x_1|x_2, \dots, x_n, C_k)p(x_2|x_3, \dots, x_n, C_k) \dots p(x_{n-1}|x_n, C_k)p(x_n|C_k)$$
$$= \prod_{i=1}^{n} p(x_i|C_k)$$

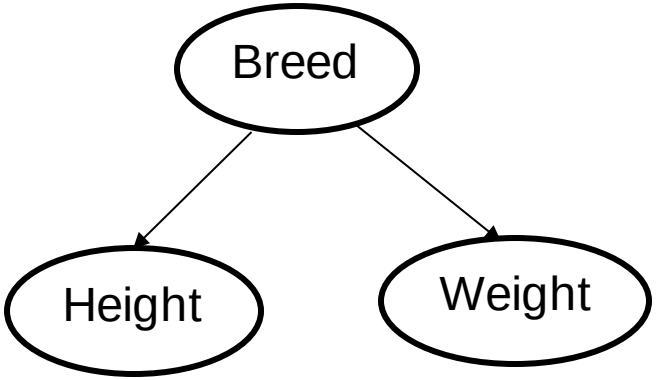
  - Maximum a posteriori (MAP) rule

$$\text{If } Z = p(\boldsymbol{x}),\ C^* = \underset{k\in\{1,..,K\}}{\text{argmax}}\ p(C_k|\boldsymbol{x}) = \underset{k\in\{1,..,K\}}{\text{argmax}} \frac{1}{z} p(C_k) \prod_{i=1}^{n} p(x_i|C_k)$$
$$= \underset{k\in\{1,\dots,K\}}{\text{argmax}}\ p(C_k) \prod_{i=1}^{n} p(x_i|C_k)$$

Naïve Bayes: Dog Breed Example

- Two classes (breeds): C_1= German Shepherd (GS), and C_2 = Dalmatian (D)
- Two features
  - height ∈ {tall, short}
  - weight ∈ {heavy, light}
- Conditional Probability Tables (CPTs) – $P(height|breed)$ and $P(weight|breed)$

German Shepherd



Dalmatian

$$P(height, weight|breed) = P(height|breed) * P(weight|breed)$$

Naïve Bayes Case Study – MESCC: Background

- Metastatic cancer patients have a primary tumor (breast, lung, kidney, etc.) that spreads to other parts of the body
- Sometimes, this tumor can spread to the spine, where it will compress the spinal cord and cause (i) immense pain and (ii) difficulty with movement
- This condition is called Metastatic Epidural Spinal Cord Compression (MESCC)
- Patients with MESCC have the option to get decompressive surgery, which will remove the spinal tumor (but not the primary tumor)
- Surgical treatment is only recommended for those who are expected to live at least 3 months after surgery

- Motivating question: ***What is the most probable outcome for 3-month postoperative survival in a patient with MESCC given their preoperative/baseline features?***

Naïve Bayes Case Study – MESCC: Variables

- Motivating question: ***What is the most probable outcome for 3-month postoperative survival in a patient with MESCC given their preoperative/baseline features?***

- Define two classes/outcomes:
  - $\boldsymbol{S}$ = survives at least 3 months after surgery
  - $\boldsymbol{D}$ = dies within 3 months after surgery

- Let us consider the following baseline features for prediction:
  - **Other-Meta ($\boldsymbol{O}$)** $\in \{\boldsymbol{yes}, \boldsymbol{no}\}$ – does the patient have any other metastases outside of the spine?
  - **Comorbidities ($\boldsymbol{C}$)** $\in \{\boldsymbol{mild}, \boldsymbol{moderate}, \boldsymbol{severe}\}$ – how severe are the other comorbidities that the patient has?
  - **Radiation ($\boldsymbol{R}$)** $\in \{\boldsymbol{yes}, \boldsymbol{no}\}$ – has the patient had any historical radiation therapy to the MESCC lesion?
  - **SF36-Phys ($\boldsymbol{F}$)** $\in \{\boldsymbol{healthy}, \boldsymbol{poor}\}$ – how good is the patient's physical health score from the SF36-v2 health questionnaire?

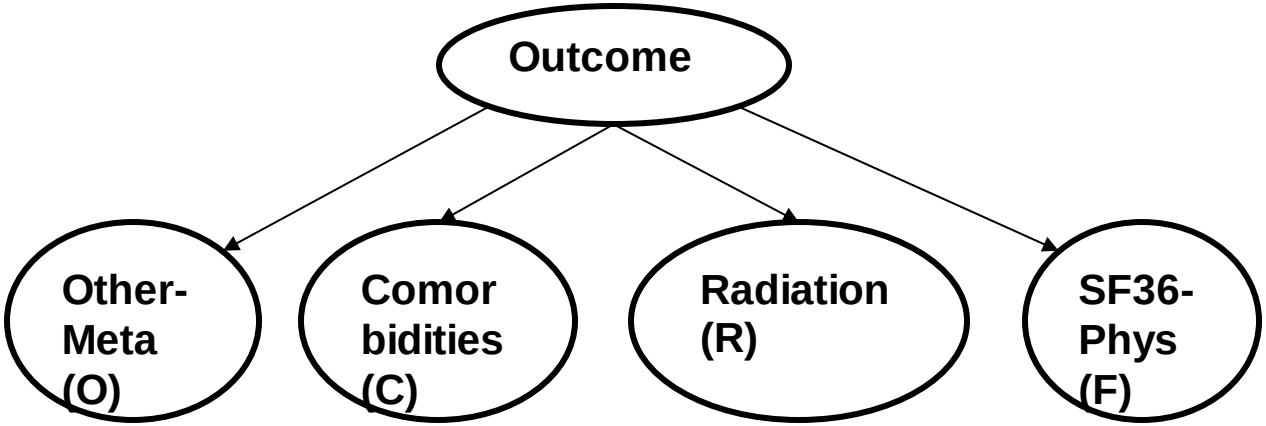
Naïve Bayes Case Study – MESCC: Setup

Naïve Bayes Model – Posterior Probabilities

- We need to calculate the posterior probabilities for survival and death given the evidence, and then use the higher one to decide an outcome
- For example, one such calculation might be

$$p(\boldsymbol{S}|O = yes, C = moderate, R = yes, F = poor)$$

$$= \frac{p(O = yes, C = moderate, R = yes, F = poor|\boldsymbol{S})\, p(\boldsymbol{S})}{p(O = yes, C = moderate, R = yes, F = poor)}$$



Naïve Bayes Case Study – MESCC: Assumptions

- **Class-conditional Independence Assumption**
  - Given the class, the features are independent
  - **This greatly simplifies calculations for the likelihood**. The posterior becomes

$$p(\boldsymbol{S}|O = yes, C = moderate, R = yes, F = poor)$$

$$= \frac{p(O = yes, C = moderate, R = yes, F = poor|\boldsymbol{S})\, p(\boldsymbol{S})}{p(O = yes, C = moderate, R = yes, F = poor)}$$

$$\propto p(O = yes, C = moderate, R = yes, F = poor|\boldsymbol{S})\, p(\boldsymbol{S})$$

$$= p(O = yes|\boldsymbol{S}) * p(C = moderate|\boldsymbol{S}) * p(R = yes, |\boldsymbol{S}) * p(F = poor|\boldsymbol{S}) * p(\boldsymbol{S})$$

- **Conditional Probability Tables (CPTs)**
  - Contain information about relevant conditional probabilities needed for calculations
  - In this case, it will contain data on $P(O|outcome), P(C|outcome)$, $P(R|outcome)$, and $P(F|outcome)$

Naïve Bayes Case Study – MESCC: Extensions

- With the model, we can also ask other inference questions
- Ex. 1: How likely is it that the patient has received radiation therapy to their MESCC lesion given that they don't survive 3 months after surgery?
  - $P(R = yes|\boldsymbol{D})$: We can get this probability from the CPT
- Ex. 2: How likely is it that the patient has severe comorbidities given that they have extraspinal metastases?
  - Note that comorbidities and other_meta are *conditionally independent* given the class ($C \perp O \mid outcome$), but not *generally independent* ($C \not\perp O$)

$$P(C = severe|O = yes) = \frac{P(C = severe, O = yes)}{P(O = yes)}$$

$$= \frac{P(C = severe, O = yes|\boldsymbol{D})P(\boldsymbol{D}) + P(C = severe, O = yes|\boldsymbol{S})P(\boldsymbol{S})}{P(O = yes)}$$

$$= \frac{P(C = severe|\boldsymbol{D})P(O = yes|\boldsymbol{D})P(\boldsymbol{D}) + P(C = severe|\boldsymbol{S})P(O = yes|\boldsymbol{S})P(\boldsymbol{S})}{P(O = yes)}$$

  - We can get each of these probabilities from the CPT

Joint Distributions

- Inference questions like the ones we answered can be computed from the *joint probability distribution* over the variables

$$P(O, C, R, F, outcome)$$

- Generally, the number of parameters required to specify an arbitrary joint distribution is (*# unique combinations of variable values*) -1
  - In our example, O, R, F and $outcome$ each take on 2 values, while C takes on 3. So we need $(2 * 3 * 2 * 2 * 2)$ - 1 = 47 parameters
  - ***Where does the "-1" come from?***

- The advantage of Naïve Bayes is that, if the class conditional independence assumption is true, then the joint distribution can be specified with much fewer parameters

$$P(O, C, R, F, outcome) = P(O|outcome)P(C|outcome)P(R|outcome)P(F|outcome)P(outcome)$$

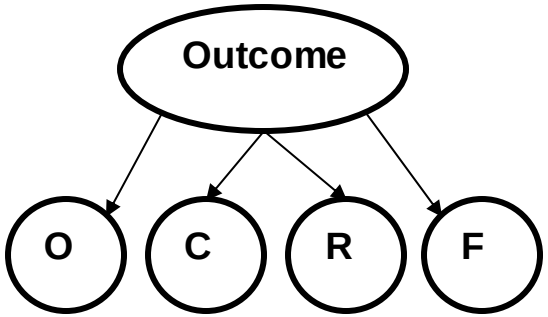
  - **In this case, we only need 11 parameters instead of 47 parameters!**

$P(O = yes|\boldsymbol{S})$, $P(O = yes|\boldsymbol{D})$, $P(C = mild|\boldsymbol{S})$, $P(C = mild|\boldsymbol{D})$, $P(C = moderate|\boldsymbol{S})$, $P(C = moderate|\boldsymbol{D})$, $P(R = yes|\boldsymbol{S})$, $P(R = yes|\boldsymbol{D})$, $P(F = yes|\boldsymbol{S})$, $P(F = yes|\boldsymbol{D})$, $P(\boldsymbol{S})$

** This set of parameters is complete, but not unique*

Joint Distributions

- The structure of the Naïve Bayes graph helps determine how to factorize the joint distribution
- In general, the structure of the graph encodes the **conditional independence assumptions** about a distribution

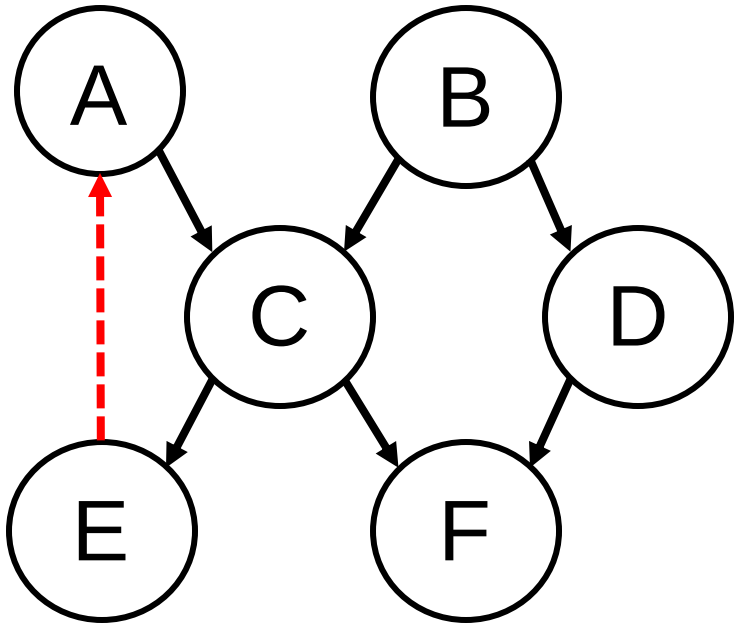


$$P(O, C, R, F, outcome)$$
$$= P(O|C, R, F, outcome)P(C|R, F, outcome)P(R|F, outcome)P(F|outcome)P(outcome)$$
$$= P(O|outcome)P(C|outcome)P(R|outcome)P(F|outcome)P(outcome)$$

- By assuming class conditional independence between features in Naïve Bayes, we have a much simpler graph
- When this assumption is no longer valid, we need another model
- We can handle more realistic scenarios while maintaining tractable calculations (with fewer parameters) using a Bayesian Network

Basic Graph Review

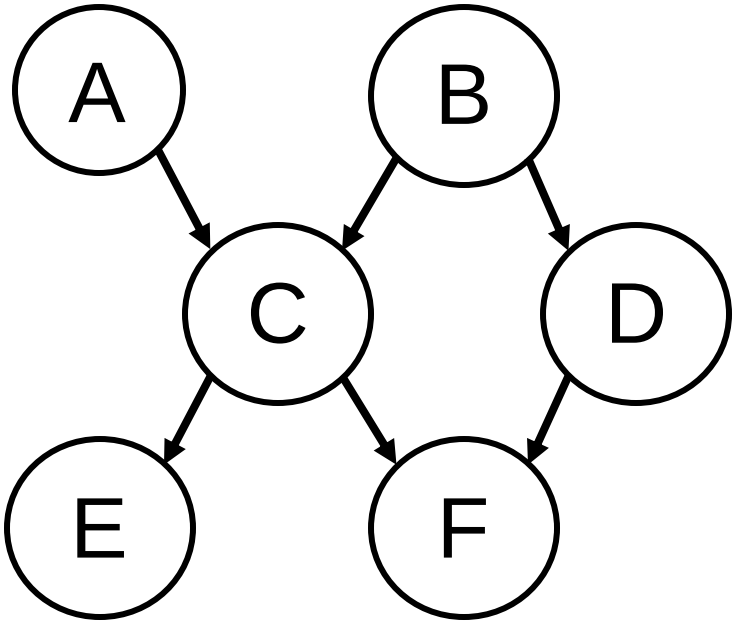
- A **directed** graph $G = (V, E)$ is defined by (i) V, the set of nodes/vertices and (ii) E, the set of directed edges
  - In the below example, $V = \{A, B, C, D, E, F\}$ and $E = \{(A, C), (B, C), (B, D), (C, E), (C, F), (D, F)\}$



- A **trail** is a sequence of vertices/edges in a graph where no edge is repeated
  - e.g. $A \rightarrow C \rightarrow F$ is a trail
  - If the graph is directed, then edges must be traversed in the defined direction
- G is **acyclic** if there aren't any cycles present in it
  - i.e. there are no non-empty trails with only the first and last vertex V_i repeated
  - If there was an edge connecting nodes E and A in the left graph, we would have a cycle: $A \rightarrow C \rightarrow E \rightarrow A$

Basic Graph Review

- With a directed acyclic graph (DAG) $G = (V, E)$, we can specify relationships between nodes



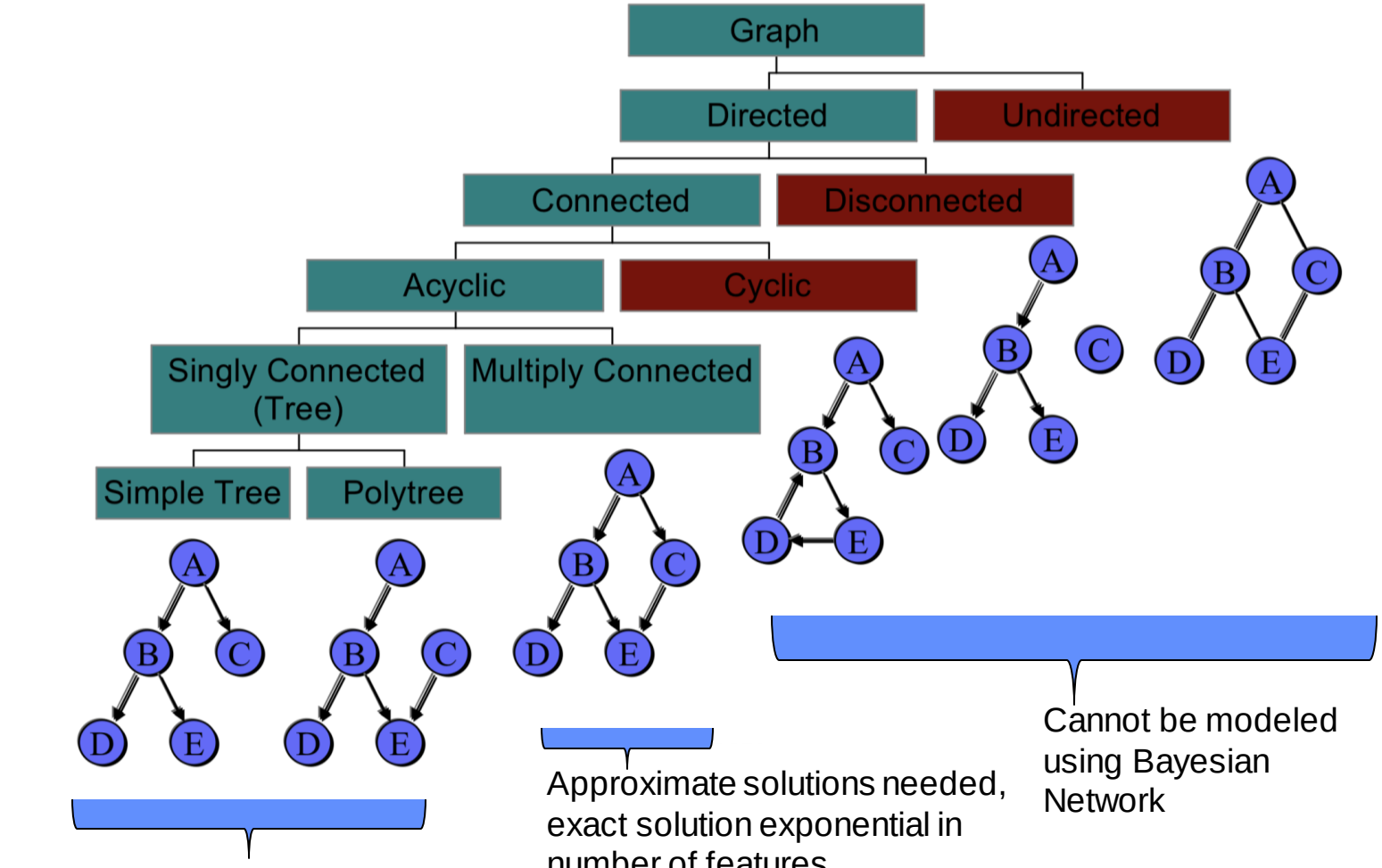
- Whenever there is a trail $V_1 \rightarrow V_2 \rightarrow V_3$,
  - V_1 is a **parent** of V_2 and V_2 is a **child** of V_1
    - Similarly, V_2 is a parent of V_3 and V_3 is a child of V_2
  - V_2 and V_3 are **descendants** of V_1
  - V_1 and V_2 are **non-descendants** of V_3

- For example, in the provided graph,
  - A and B are **parents** of C
  - F is a **child** of C and D
  - C, D, F and E are the **descendants** of B
  - A is a **non-descendant** of B

Bayesian Network

- Definition: A Bayesian Network (BN) structure $\mathcal{G}$ is a directed acyclic graph whose nodes represent random variables $X_1, \ldots, X_n$.
- Examples of Bayesian Networks:
  - Naïve Bayes
  - Dynamic Bayesian networks
  - Decision graphs
  - Hidden Markov Models

Valid Bayesian Networks



Bayesian Networks: A Burglary Example

Mr. Holmes received a phone call at work from his neighbor notifying him that he heard a burglar alarm sound from the direction of his home. As he is preparing to rush home, Mr. Holmes recalls that recently the alarm had been triggered by an earthquake. Driving home, he hears a radio newscast reporting an earthquake 200 miles away.

Question: Mr. Holmes is at work. Neighbor John calls to say Mr. Holmes' alarm is ringing, but neighbor Mary doesn't call. Sometimes the alarm set off by minor earthquakes. Is there a burglar?

- Can the question be represented in terms of probability?

$$\mathrm{P}(\mathrm{Burglary} = \mathrm{T} | \mathrm{Alarm} = \mathrm{T}, \mathrm{Earthquake} = \mathrm{T}, \mathrm{John} = \mathrm{T}, \mathrm{Mary} = \mathrm{F})$$

- Need the joint distribution $P(Burglary, Alarm, Earthquake, John, Mary)$ to evaluate the above probability
  - This requires (2*2*2*2*2) -1 = 31 parameters. Can we do better?

Credits: Judea Pearl, 1986

Solution formulation

- Using a **graph** specific to this problem, we can reduce the number of parameters needed
- As in the NB case, there are two components needed to formulate a solution

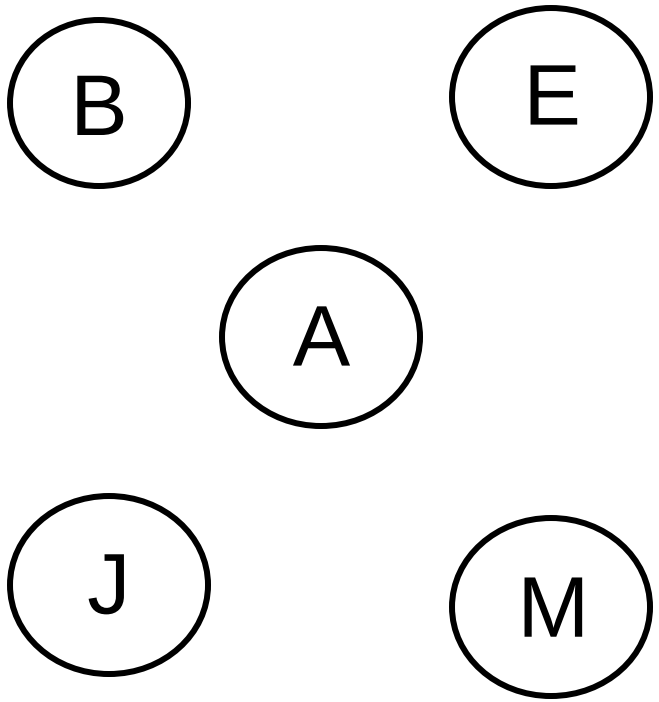
1. Graph representation: This will inform us how to factorize the joint distribution. The graph should be **acyclic** and have
   - Nodes representing variables/features
   - Directed edges representing causal influences of variables on each other
2. CPTs: These will be extracted from the training data or provided by some oracle

Burglary Network Example

Define the variables that completely describe the problem. They become the **nodes** of the graph.

Following are the variables involved in the burglary example:

- Burglary (**B**) – whether there was a burglary or not
- Earthquake (**E**) – whether there was an earthquake or not
- Alarm (**A**) – did the alarm sound or not
- John (**J**) – did the neighbor John call or not
- Mary (**M**) – did the neighbor Mary call or not



Burglary Network Example

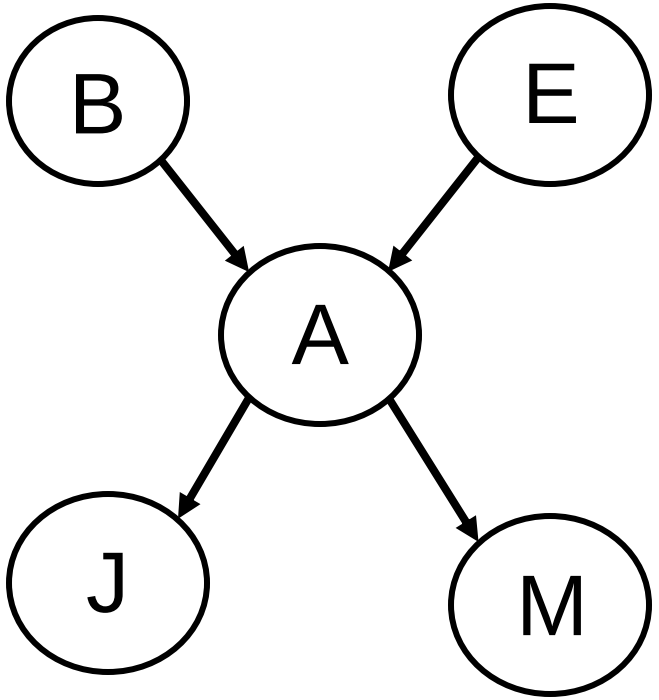
Define the causal relationships between features. These will be encoded as **directed edges** in the graph. Note that for Bayesian Networks, the graph must be **acyclic**.

For the burglary example:

- A burglar can set the alarm off $(B \rightarrow A)$
- An earthquake can set the alarm off $(E \rightarrow A)$
- The alarm can cause Mary to call $(A \rightarrow M)$
- The alarm can cause John to call $(A \rightarrow J)$

From the constructed graph, we can identify node relationships. For example,

- Burglary is a **parent** of Alarm
- Mary is a **child** of Alarm
- John and Alarm are **descendants** of Earthquake
- Burglary is a **non-descendant** of Mary



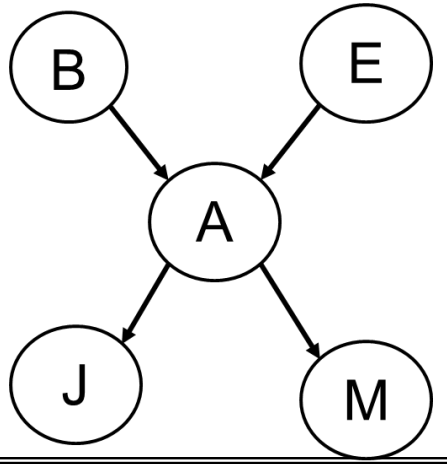
Factorizing the Joint distribution

- How did we go from a joint to its factorized distribution in Naïve Bayes?
  1. Apply chain rule to expand joint probability
  2. Simplify using conditional independences specified by the graph structure
- The above steps are followed for a BN as well!

1. We begin by using chain rule to specify the joint probability:

$$P(J, M, A, E, B) = P(J|M, A, E, B)P(M|A, E, B)P(A|E, B)P(E|B)P(B)$$

- **Note:** In order to simplify calculations, the conditional probabilities are written in such a way that the parents of a node are the ones conditioned on.
  - For example: $P(A|E, B)$ is preferred over $P(B|E, A)$

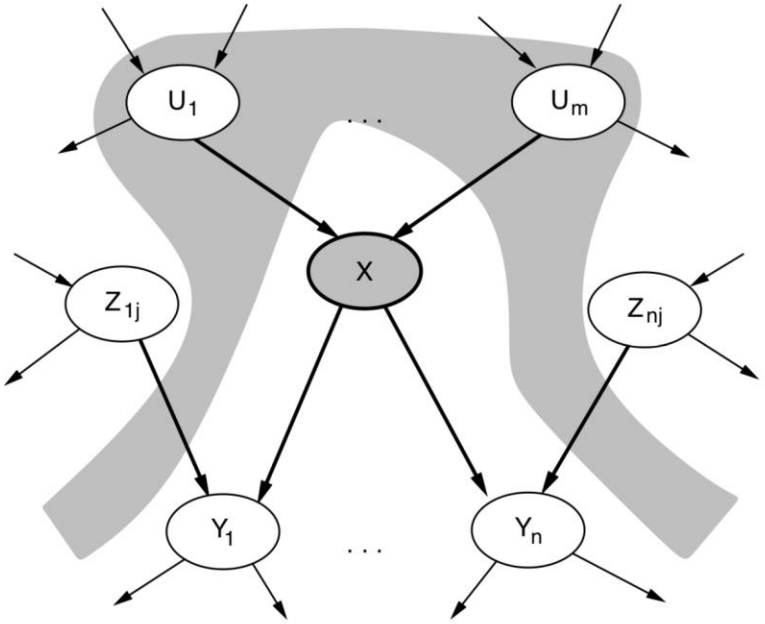


Factorizing the Joint distribution

2. Now we use the **conditional independences** from the graph structure to factorize the joint distribution

- Conditional independence was easy in NB, but for a general BN it is more involved – we will use local semantics.

- Local semantics: Each node is conditionally independent of its non-descendants given its parents

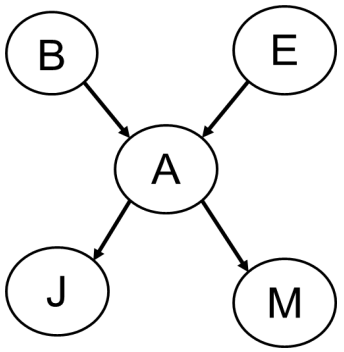


Application of Local Semantics

$$P(J, M, A, E, B) = P(J|M, A, E, B)P(M|A, E, B)P(A|E, B)P(E|B)P(B)$$

- P(J|M,A,E,B) = P(J|A)
  - M, B, E are non-descendants
  - A is the parent
- P(M|A,E,B) = P(M|A)
  - E, B are non-descendants
  - A is the parent

…



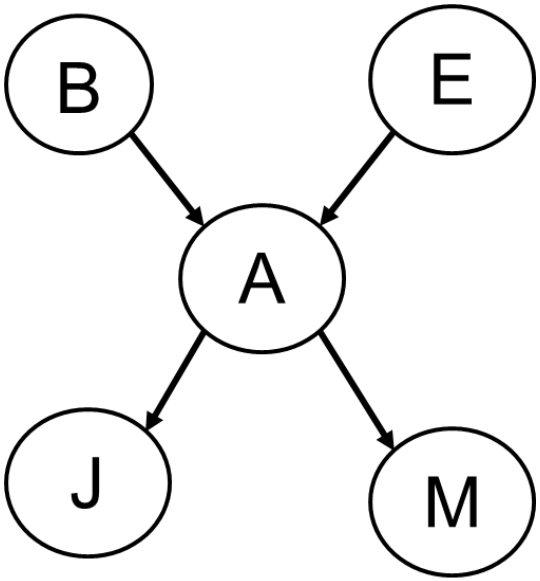
After applying local semantics, we simplify the joint distribution as follows:

$$P(J, M, A, E, B) = P(J|A)P(M|A)P(A|E, B)P(E)P(B)$$

Burglary Example: CPTs

P(B=T) = 0.001

P(E=T) = 0.002



B	E	P (A= T|B,E)	P (A= F|B,E)
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

A	P (J = T|A)	P (J = F|A)
T	0.90	0.10
F	0.05	0.95

A	P (M = T|A)	P (M = F|A)
T	0.70	0.30
F	0.01	0.99

Inference Tasks

Simplifying the expression used in inference

$$P(B|A,E,J,M) = \frac{P(B,A,E,J,M)}{P(A,E,J,M)} = \frac{P(J|A)P(M|A)P(A|E,B)P(E)P(B)}{P(J|A)P(M|A)P(A|E)P(E)}$$

$$= \frac{P(A|E,B)P(B)}{P(A|E)} = \frac{P(A|E,B)P(B)}{\sum_B P(A|E,B)P(B)}$$

Substituting values from the CPT to evaluate the exact value

$$P(B = T|A = T, E = T,\ J = T,\ M = F)$$

$$= \frac{P(A = T|E = T, B = T)P(B = T)}{P(A = T|E = T, B = T)P(B = T) + P(A = T|E = T, B = F)P(B = F)}$$

$$= \frac{0.95 * 0.001}{0.95 * 0.001 + 0.29 * 0.999} = \frac{0.00095}{0.29066} = 0.00327$$

Inference Tasks

- Now that we have the joint probability and the CPTs, what other inference questions can we answer?

- Simple queries: Computer posterior marginal
  - E.g., P(Alarm = False| Earthquake = False, Burglary = True). Here we are marginalizing over John and Mary

- Conjunctive queries:
  - P(Alarm, Mary | Earthquake=True) =
    P(Alarm | Earthquake = True) x P(Mary| Alarm , Earthquake =T)

- Sensitivity analysis: Which probability values are most critical?

- Explanation: How good is the current alarm system?
  - Probability of false positive for the current alarm system P(A=T|E=F, B=F) = 0.001

Another inference example

Calculate P(John=True, Mary=True, Alarm = True, Burglary=False, Earthquake = True).

From BN, we get:

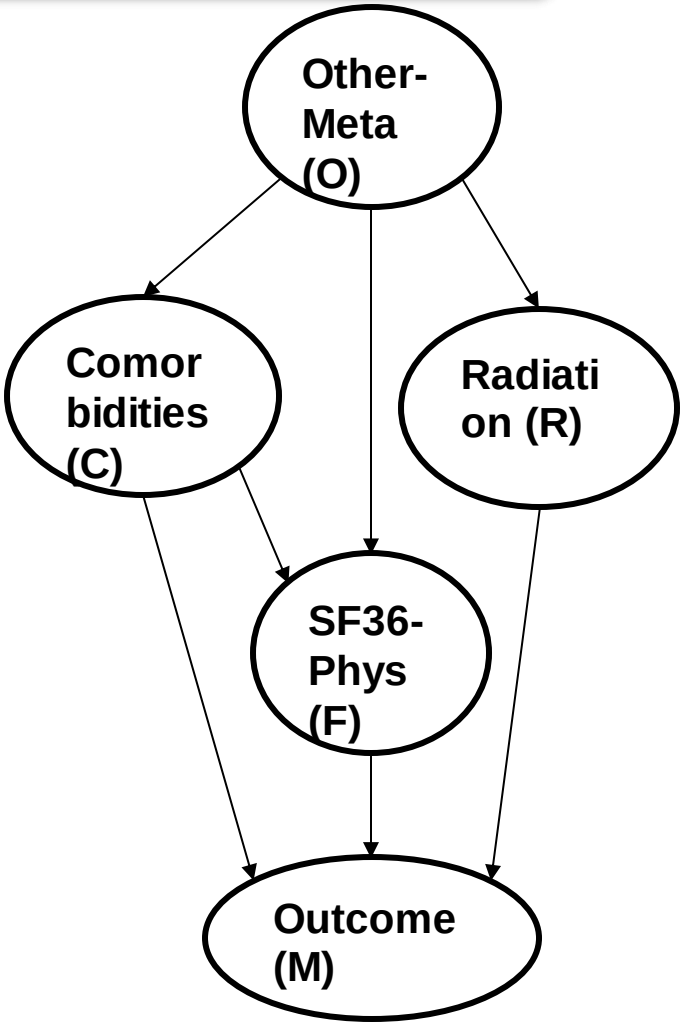
$$P(J, M, A, E, B) = P(J|A)P(M|A)P(A|E, B)P(E)P(B)$$

Substituting values from the CPT, we get:

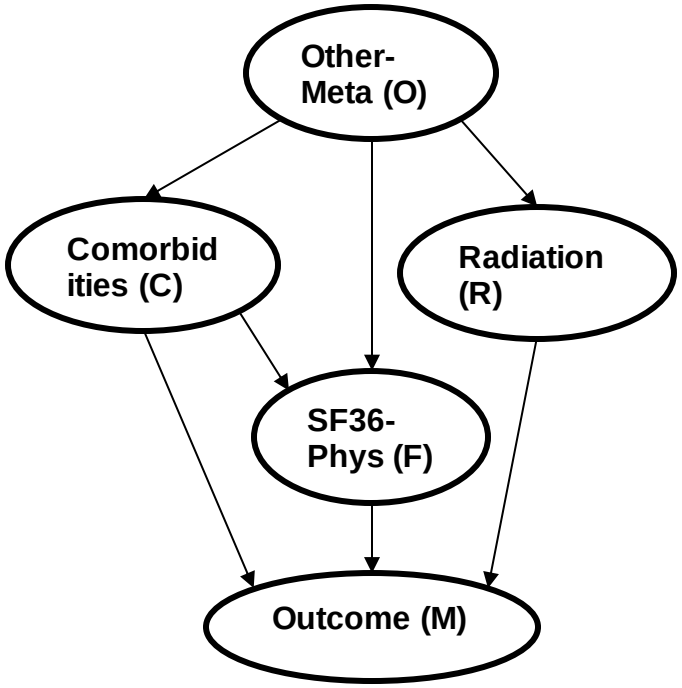
$$P(J = T, M = T, A = T, B = F, E = F)$$
$$= P(J = T|A = T)P(M = T|A = T)P(A = T|B = F, E = F)P(B = F)P(E = F)$$
$$= 0.9 * 0.7 * 0.001 * 0.999 * 0.998$$
$$= 0.00063$$

Bayesian Network Case Study – MESCC: Construction

- Using the MESCC context from earlier, we can construct a BN that is not an NB
- Denote the survival outcome prediction we wish to make as M
- Suppose that from domain knowledge, we know
  - Presence of extraspinal metastasis can influence (i) severity of comorbidities and (ii) whether radiation therapy is performed for MESCC lesion ($O \rightarrow C$, $O \rightarrow R$)
  - Both (i) presence of extraspinal metastasis and (ii) severity of comorbidities have an influence on physical health score ($O \rightarrow F, C \rightarrow F$)
  - (i) severity of comorbidities, (ii) radiation therapy for MESCC, and (iii) physical health score all directly influence survival outcome ($C \rightarrow M, R \rightarrow M, F \rightarrow M$)



Bayesian Network Case Study – MESCC: Joint Probability



- Using this Bayesian Network, we can expand the joint probability as follows:

$$P(O,C,R,F,M)$$
$$= P(M|O,C,R,F)P(F|O,C,R)P(C|R,O)P(R|O)P(O)$$
$$= P(M|C,R,F)P(F|O,C)P(C|O)P(R|O)P(O)$$

- With this BN, we need to calculate **25 parameters**
  - (2-1)*(3*2*2)=12 parameters for $P(M|C,R,F)$
  - (2-1)*(3*2) = 6 parameters for $P(F|O,C)$
  - (3-1)*(2) = 4 parameters for $P(C|O)$
  - (2-1)*(2) = 2 parameters for $P(R|O)$
  - (2-1) = 1 parameter for $P(O)$

Bayesian Network Case Study – MESCC: Comparison

Metric	NB	BN
Joint Probability	$P(O\|M)P(C\|M)$ $P(R\|M)P(F\|M)P(M)$	$P(M\|C,R,F)P(F\|O,C)$ $P(C\|O)P(R\|O)P(O)$
#params required (original: 47)	11	25
Average Testing AUC*	0.77 ± 0.04	0.63 ± 0.09

- AUC = Area Under the Receiver Operating Characteristic Curve
  - Ranges from 0 to 1, with 1 being ideal
  - Measures probability that model reports a random positive example more highly than a random negative example[1]
  - We will learn more about this later on in the semester…

[1]: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

- NB reports higher AUC than BN
  - Sometimes, finer granularity with data representation in BN can demand a more comprehensive or well-balanced dataset
  - E.g. consider SF36-Phys (F) node. Need balanced data to fill extra data "bins" in the BN.
    - In the NB model, we calculate $P(F|M)$, which defines 2*2 = 4 "bins"

$(\{F = healthy, F = poor\} x \{M = survives, M = dies\})$

    - In the BN model, we calculate $P(F|C,O)$, which defines 2*3*2=12 "bins"

$(\{F = healthy, F = poor\} x \{C = mild, C = moderate, C = severe\} x \{O = yes, O = no\})$

Curse of Dimensionality

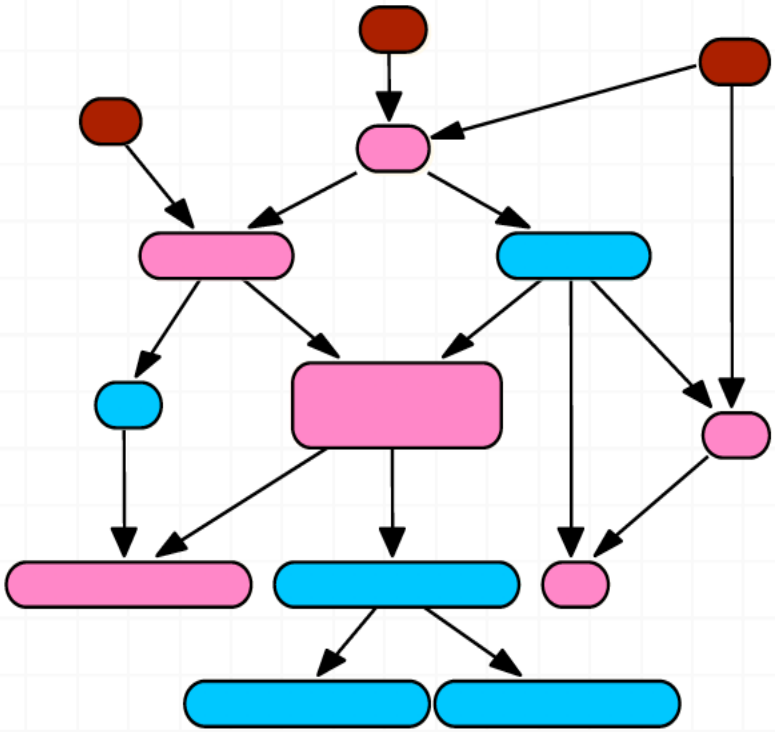
- Let Network Size = number of parameters required for joint distribution

- Network grows exponentially with number of nodes ~ 2^N
  - Each additional node doubles the size of the network!
  - A network with **100 nodes => 2^{100} -1 parameters! =>** Impractical!

- Bayesian networks can greatly reduce this complexity

Curse of Dimensionality - Example

- We are given a Bayesian network with 14 binary variables (shown on right)
  - Assume that each variable can take on value 0 or value 1
- The joint probability space requires 2^14 -1 ~**16 K** parameters, which is huge!
- Let us now begin to calculate the number of parameters required if we take advantage of the BN structure
  - There are 3 **red** nodes, each of which has no incoming edges and no parents
  - There are 5 **blue** nodes, each of which has 1 incoming edge from a parent
  - There are 6 **pink** nodes, each of which has 2 incoming edges from parents



Curse of Dimensionality - Example

- Consider a **red** node . The following table defines probabilities for all values of

(=)	(=)
	1 −

- We only need one of these two values to calculate P( ) since once we know one, the other is simply 1 minus the known one.
- With three red nodes, we have 3*1 = 3 total independent parameters from the red nodes

- Consider a **blue** node with parent node $_1$. The following table defines probabilities for all values of and $_1$.

(|)	=	=
=	$_1$	1 − $_1$
=	$_2$	1 − $_2$

- We only need one value per row to be able to calculate P(\| $_1$), which means we need 2 independent parameters per blue node
- With 5 blue nodes, we have 5*2 = 10 total independent parameters from the blue nodes

Curse of Dimensionality - Example

- Consider a pink node M with parent nodes P_1 and P_2. The following table defines probabilities for all values of M, P_1 and P_2:

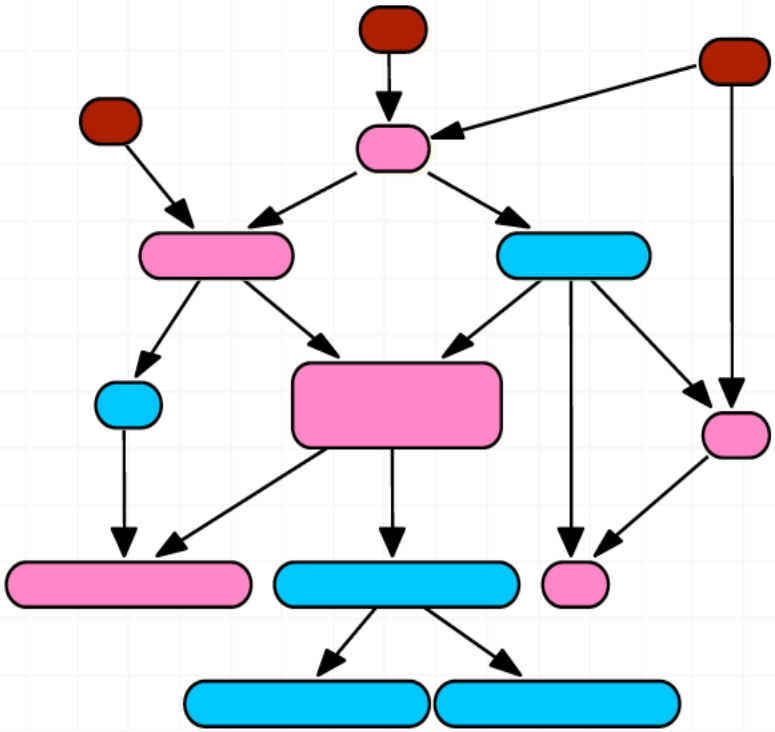
$P(M\|P_1,P_2)$	$M=0$	$M=1$
$P_1=0, P_2=0$	k_1	$1-k_1$
$P_1=0, P_2=1$	k_2	$1-k_2$
$P_1=1, P_2=0$	k_3	$1-k_3$
$P_1=1, P_2=1$	k_4	$1-k_4$

- We only need one value per row to be able to calculate $\mathrm{P}(M|P_1,P_2)$, which means we need 4 independent parameters per pink node
- With 6 blue nodes, we have 6*4 = 24 total independent parameters from the pink nodes

Curse of Dimensionality - Example

- We need ~**16 K** parameters to specify an arbitrary joint distribution

- Using the Bayesian Network structure as shown, we only need 3 + 10 + 24 = **37 parameters, which is much less than 16 K parameters!**



Bayesian Networks - Summary

- Graphical representation of dependencies among a set of random variables
  - Nodes: variables
  - Directed links to a node from its *parents*: direct probabilistic dependencies
  - Each X_i has a conditional probability distribution, $P(X_i|Parents(X_i))$ showing the effects of the parents on the node.
  - The graph is directed (DAG); hence, no cycles

- Can express dependencies, more concisely
  - Given some evidence, what are the most likely values of other variables? $\boldsymbol{argmax_X P(X|E)}$ ***- MAP explanation***
  - Given some new evidence, how does this affect the probability of some other node(s)? $\boldsymbol{P(X|E)}$***—belief propagation/updating***

Comparison

- Full joint distribution
  - Completely expressive
  - Hugely data-hungry
  - Exponential computational complexity
- Bayesian Network
- Naïve Bayes (full conditional independence)
  - Relatively concise
  - Need data ~ (#hypotheses) × (#features) × (#feature-vals)
  - Fast ~ (#features)
  - Cannot express dependencies among features or among hypotheses
  - Cannot consider possibility of multiple hypotheses co-occurring

Additional Slides

Steps for Creating Bayesian Network

- Choose a set of variables and an ordering {X1,..,Xm}
- For each variable X_i for i = 1 to m:
  - Add the variable Xi to the network
  - Set Parents(Xi) to be the minimal subset of $\{X_1, .., X_{i-1}\}$ such that Xi is conditionally independent of all the other members of $\{X_1, .., X_{i-1}\}$ given Parents(X_i)
  - Define the probability table describing $P(X_i \mid \text{Parents}(X_i))$

Methods to solve joint distribution

- Computing the conditional probabilities by enumerating all relevant entries in the joint is expensive:
  - Exponential in the number of variables!
- However, for *poly-trees* (not even undirected loops—i.e., only one connection between any pair of nodes; like our Burglary example), there are efficient linear algorithms, similar to constraint propagation
- But, for arbitrary BNs:
  - Solving for general queries in Bayes nets is NP-hard!
- Approximate methods
  - Approximate the joint distributions by drawing samples
- Exact methods
  - Factorization and variable elimination
  - Exploit special network structure (e.g., trees)
  - Transform the network structure

Maximum a posteriori probability (MAP) Decision Rule

- Maximum a-posteriori probability (MAP) decision rule declares the hypothesis which ***maximizes the posteriori probabilities***
- ***A Posteriori probability*** is a conditional probability that an observer would declare a hypothesis, given an observation *k:*

$$P(H_i|X = k)$$

- So given an observation $X = k$, the MAP decision rule chooses the hypothesis with the larger posteriori probability
- The posteriori probabilities are unknown, so we use Bayes' formula to calculate them.

Maximum a posteriori probability (MAP) Decision Rule (Cont'd)

- *A posteriori from Bayes Rule*

$$P(H_i|X=k) = \frac{P(H_i, X=k)}{\sum_{i=1}^{N} P(H_i, X=k)}$$

- Example
  - $P(MCE = C|App = F) = \frac{P(MCE=C, App=F)}{\sum_{i=1}^{N} P(H_i, App=F)}$
- So the MAP decision rule requires the computation of the ***joint probabilities*** *using* ***Chain Rule***:
- But we have:
  - $P(X = k \mid H_i)$ from the likelihood matrix
  - $P(H_i)$ are the ***prior probabilities***.

  The prior probabilities are determined independent of the experiment and prior to any observation is made.

Inference Task

- In general, any inference operation of the form *P(values of some variables | values of the other variables)* can be computed: **Probability that both John and Mary call given that there was a burglar.**

We know how to compute these sums because we know how to compute the joint as written, we still need to compute most of the entire joint table

$$P(J,M \mid B) = \frac{P(J,M,B)}{P(B)} = \frac{\sum_{\text{All entries } X \text{ that contain } J \wedge M \wedge B} P(X)}{\sum_{\text{All entries } Y \text{ that contain } B} P(Y)}$$