Introduction to Bayesian Networks

ECE/CS 498 DS U/G

Lecture 6

Ravi K. Iyer

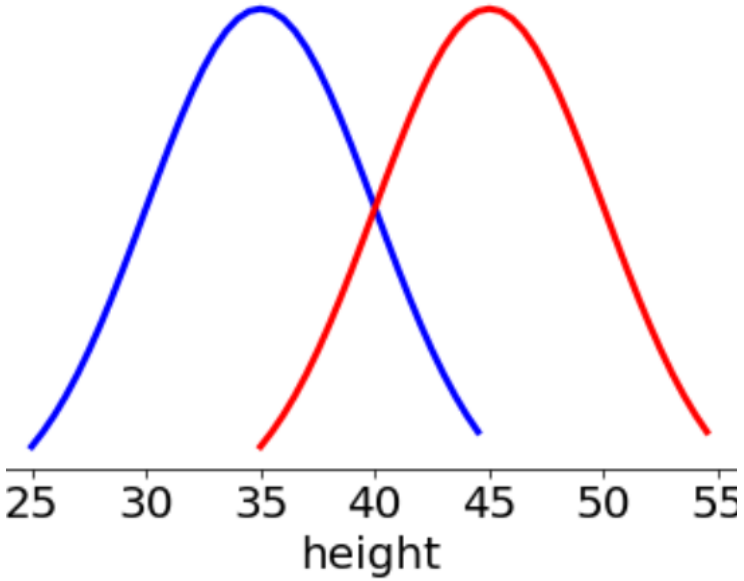
Department of Electrical and Computer Engineering

University of Illinois at Urbana Champaign

Naïve Bayes with continuous variables

- Height can be any real number; therefore we its **distribution** will be **continuous**
- With enough data, approximate the distribution of height for the two classes
  - We can also assume a parametric form for the distribution

- Following is the distribution of height based on some training data
  - German Shepherd: $\mathcal{N}(\mathbf{35}, \mathbf{25})$
  - Dalmatian: $\mathcal{N}(\mathbf{45}, \mathbf{25})$
- Priors are equal



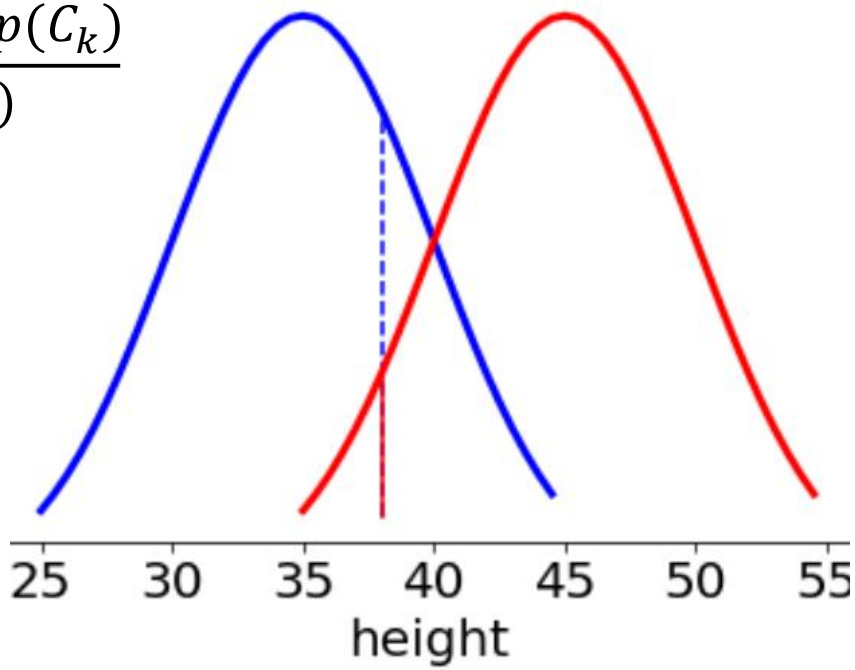
Guess the dog breed: Calculations

$$p(C_k|38) = \frac{p(38|C_k)\, p(C_k)}{p(38)}$$

$$p(GS|38) = \frac{p(38|GS)\, p(GS)}{p(38)}$$

$$\propto \frac{1}{\sqrt{50\pi}} \exp\left(-\frac{(38-35)^2}{50}\right) * 0.5$$

$$= 0.067 * 0.5 = \mathbf{0.034}$$



$$p(D|38) = \frac{p(38|D)\, p(D)}{p(38)}$$

$$\propto \frac{1}{\sqrt{50\pi}} \exp\left(-\frac{(38-45)^2}{50}\right) * 0.5$$

$$= 0.03 * 0.5 = 0.015$$

$p(D|38) < p(GS|38)$, therefore, dog $\tilde{d}$ is more likely to be a German Shepherd.

Announcements

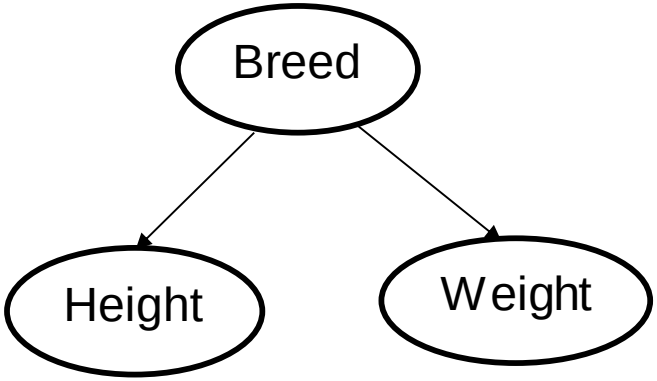
- Checkpoint 2 due on Sunday, Feb 10
- Discussion section on Friday, Feb 8
  - Concept and example related to Naïve Bayes and Bayesian Networks will be covered
- In class activity 2 on Monday, Feb 11
- We encourage to NOT post your solutions on Piazza
- HW0 grades have been released
  - Regrade requests will be taken only till Monday, Feb 11
  - Please email TAs or contact them in office hours/discussion section

- Today's Topic: Bayesian Networks

Naïve Bayes example revisited

- 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 example revisited

- Using NB graph structure and CPTs, we answered the following question – what is the most probably breed of a dog given the height and weight?

$$P(breed|height, weight) = \frac{P(height, weight|breed)P(breed)}{P(height, weight)}$$

$$\propto P(height|breed)P(weight|breed)P(breed)$$

- Can ask other inference questions?
- Example 1: How likely is it that the dog is tall given than it is GS?
  - $P(tall|GS)$: We can get the probability from the CPT
- Example 2: How likely is it that the dog is heavy given that it is tall regardless of the class? Note that height and weight are independent given the class (breed), but not in general
  - $P(heavy|tall) = \frac{P(heavy,tall|GS)P(GS)+P(heavy,tall|D)P(D)}{P(tall)}$

$$= \frac{P(heavy|GS)P(tall|GS)P(GS) + P(heavy|D)P(tall|D)P(D)}{P(tall)}$$

We can get each of the probabilities from the CPT

Joint distribution

- The questions we answered in the previous examples (and any other inference questions) can be computed from the **joint probability distribution** over the **variables**

$$P(weight, height, breed)$$

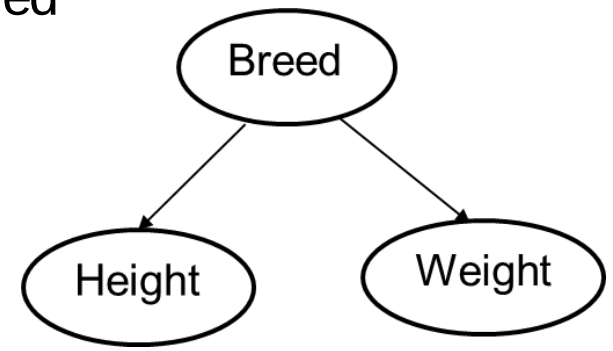
- In a general case, the number of parameters to specify the above joint distribution is 2 x 2 x 2 - 1 = 7
- The value of Naïve Bayes is in that, if the assumptions are true, then the joint distribution can be specified with fewer parameters; in this case 5*

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

*Following parameters are required: $P(tall|GS)$, $P(tall|D)$, $P(heavy|GS)$, $P(heavy|D)$, $P(GS)$

Joint distribution

- The structure of the Naïve Bayes **graph** helped in determining a factorization of the joint distribution
- In general, structure of the graph represents the set of **conditional independence** assumptions about a distribution



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

- By assuming class conditional independence in Naïve Bayes, you have a much simpler graph
- If we relax the class conditional independence assumption we can handle more realistic situations and yet maintain a tractable joint distribution (i.e., fewer parameters) e.g., Bayesian Network

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 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:

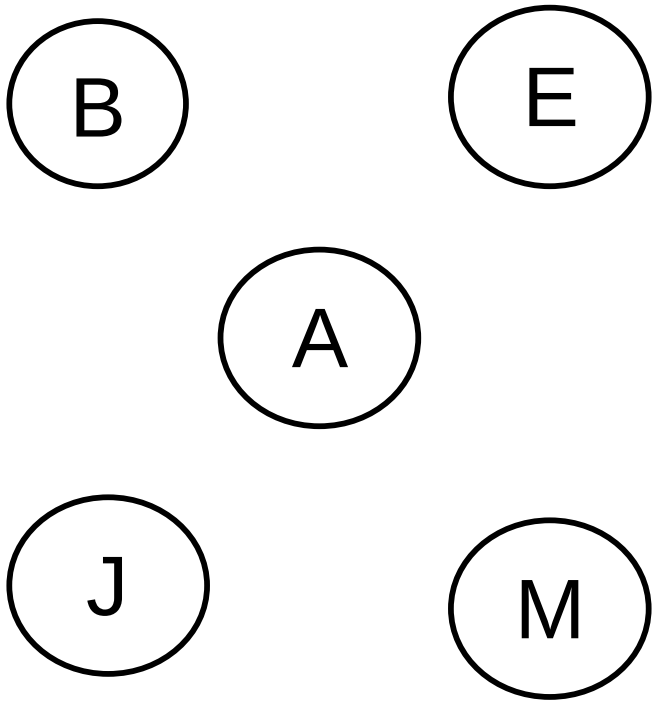
1. Graph representation: This gave as how to factorize the joint distribution
   - Variables as nodes
   - Direct influence of one variable on another represented by edge
2. CPTs: These were extracted from the training data
   - Or provided by an oracle

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



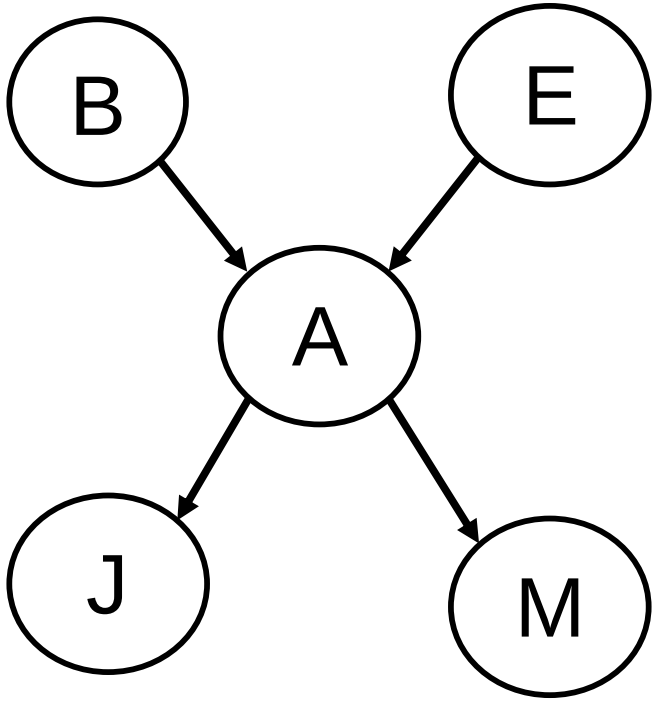
Burglary Network Example

How are **edges** drawn?

- Edges represent a direct influence in the direction of the edge
- The resulting directed graph must be acyclic
- The **topology** of the graph reflects "causal" knowledge

For the burglary example:

- A burglar can set the alarm off
- An earthquake can set the alarm off
- The alarm can cause Mary to call
- The alarm can cause John to call

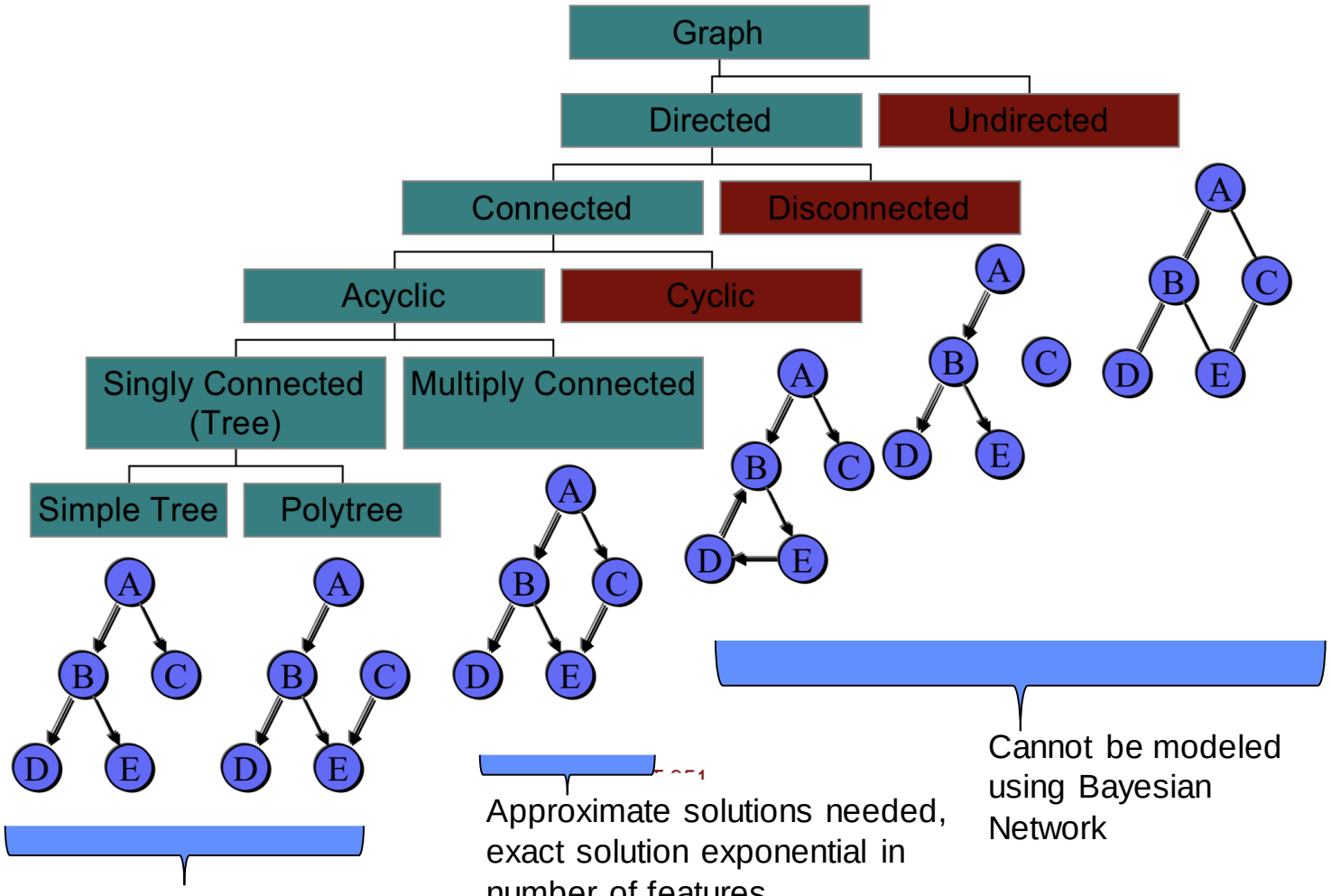


- **Parent**:
  - Burglary is a parent of Alarm
- **Descendant**:
  - Mary is a descendant of Alarm
  - John is descendant of Earthquake
- **Non-descendent:**
  - Burglary is a non-descendant of Mary

Bayesian Network

- The burglary graph is a Bayesian Network
- Definition: A Bayesian Network 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

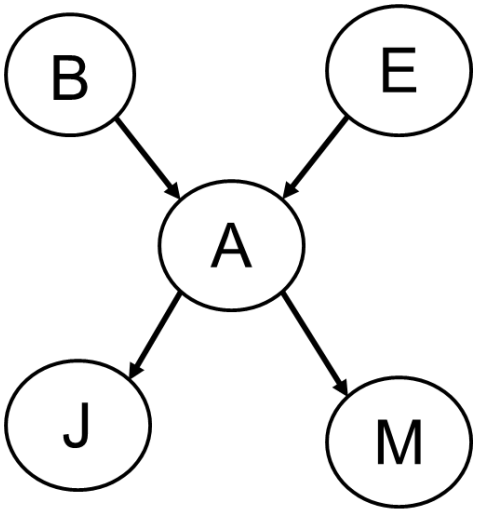


Factorizing the Joint distribution

- How did we go from a joint to its factorized distribution in Naïve Bayes?
  - Applied chain rule
  - Applied conditional independence using the graph structure
- The above steps are followed for a BN

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

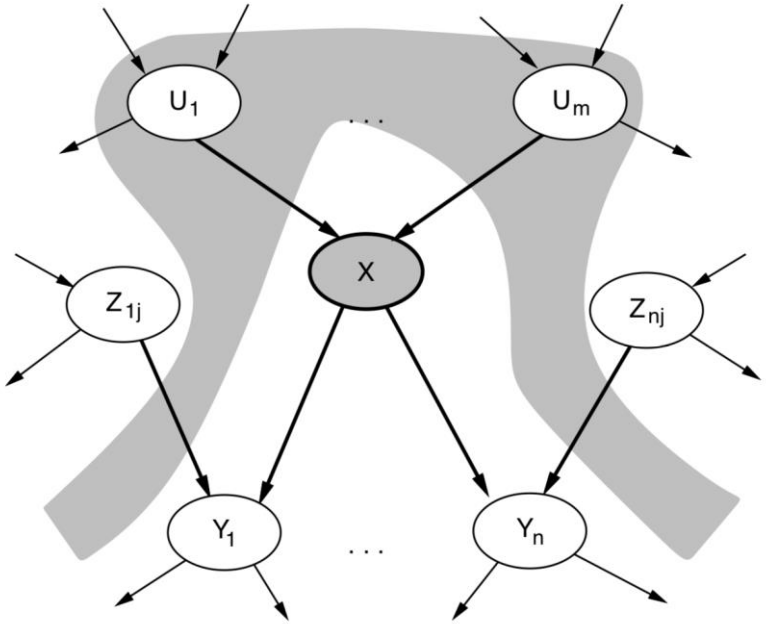
- **Note:** The conditional probabilities are written in such a way that the parents of a node are the ones conditioned on.
- Example: $P(A|E, B)$ is preferred over $P(B|E, A)$



Factorizing the Joint distribution

- We have applied chain rule. Now use graph structure to apply **conditional independence** to factorize the joint distribution
- Conditional independence was easy in NB but for a general BN it is more involved. Use local semantics.

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

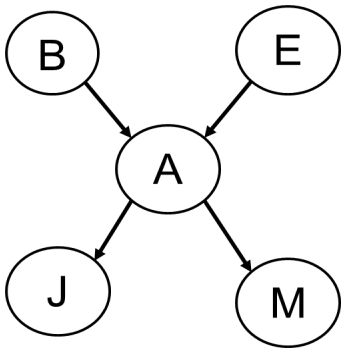


Application of Local Semantics

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

- 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

…



Similarly, applying local semantics simplifies the joint distribution.

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

Solution formulation

- Using a **graph** specific to this problem, we reduced the number of parameters needed

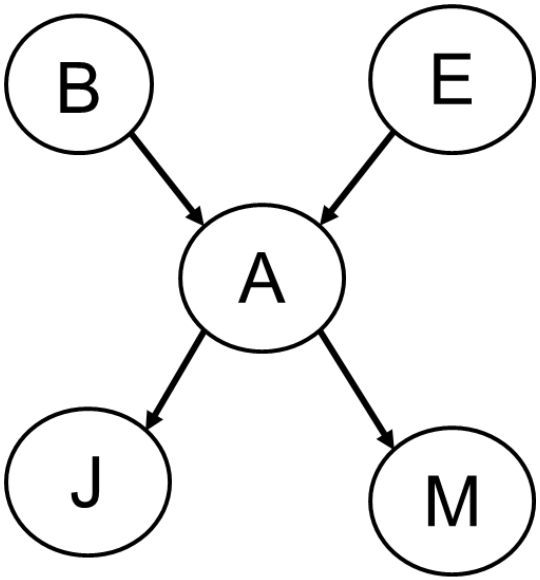
- As in the NB case:

1. Graph representation: This gave as how to factorize the joint distribution
   - Variables as nodes
   - Direct influence of one variable on another represented by edge
2. CPTs: These were extracted from the training data
   - Or provided by an oracle

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

- However, is that the only inference we can do?
  - No! Now that we have the joint and the CPTs, we can answer other inference questions too
- 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 = False).

From BN, we get:

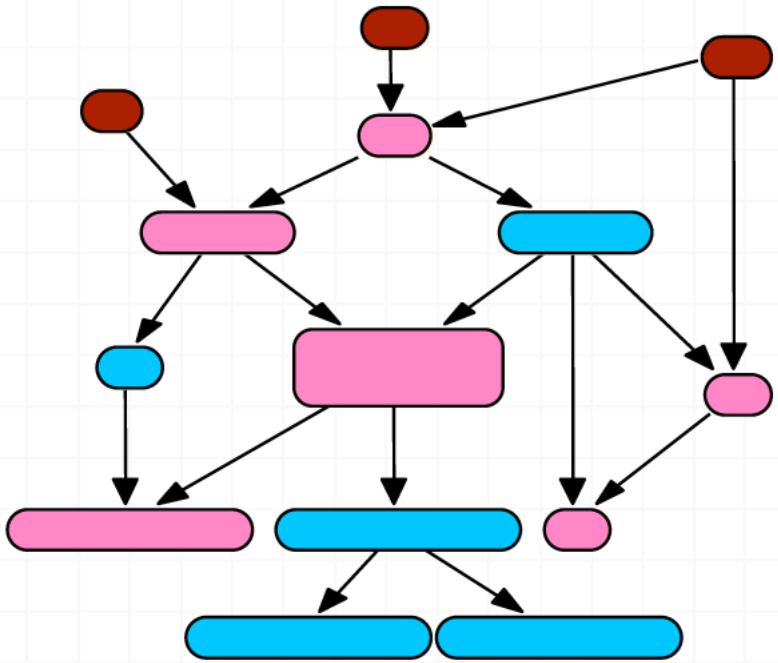
$$P(J, M, A, E, B) = P(J|A)P(M|A)P(A|E, B)P(E)P(B)$$

Substituting values from the CPT, we get:

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

Curse of Dimensionality

- Network Size = number of paramete
- Network grows exponentially with number of nodes ~ 2^N
- Each additional node doubles the siz of the network!
- A network with **100 nodes => 2^{100}-1 parameters! =>** Impractical!
- BN – reduces this complexity



Joint size = 2^{14} = 16K

BBN size = 3*2 + 5*4 + 6*8 = 74

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:*

$$\boldsymbol{P(H_i|X = k)}$$

- 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

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