CS440/ECE448 Lecture 14: Bayesian Networks

Mark Hasegawa-Johnson

Outline

- Review: Bayesian classifier
- The Los Angeles burglar alarm example
- Bayesian network: A better way to represent knowledge
- Inference using a Bayesian network
- Independence and Conditional independence

Review: Bayesian Classifier

- Class label $Y = y$, drawn from some set of labels
- Observation $X = x$, drawn from some set of features
- Bayesian classifier: choose the class label, y, that minimizes your probability of making a mistake:

$$f(x) = \underset{y}{\operatorname{argmax}}\, P(Y = y|X = x)$$

Today: What if P(X,Y) is complicated, and the naïve Bayes assumption is unreasonable?

- Example: Y is a scalar, but $X = [X_1, \ldots, X_{100}]^T$ is a vector
- Then, even if every variable is binary, $P(Y = y|X = \boldsymbol{x})$ is a table with 2^{101} numbers. Hard to learn from data; hard to use.
- The naïve Bayes assumption simplified the problem as
$$P(X_1, \ldots, X_{100}|Y) \approx \prod_{i=1}^{100} P(X_i|Y)$$
- … but what if that assumption is unreasonable? Do we then have no alternative besides learning all 2^{101} probabilities?
- Today: an alternative called a Bayesian network

Outline

- Review: Bayesian classifier
- The Los Angeles burglar alarm example
- Bayesian network: A better way to represent knowledge
- Inference using a Bayesian network
- Independence and Conditional independence

The Los Angeles burglar alarm example

- Suppose I have a house in LA. I'm in Champaign.
- My phone beeps in class: I have messages from both of my LA neighbors, John and Mary.
- Does getting messages from both John and Mary mean that my burglar alarm is going off?
- If my burglar alarm is going off, does that mean my house is being robbed, or is it just an earthquake?

Variables

- $B = \top$ if my house is being burglarized, else $B = \bot$
- $E = \top$ if there's an earthquake in LA right now, else $E = \bot$
- $A = \top$ if my alarm is going off right now, else $A = \bot$
- $J = \top$ if John is texting me, else $J = \bot$
- $M = \top$ if Mary is texting me, else $M = \bot$

Inference Problem

- Given that $J = \top$ and $M = \top$, I want to know what is the probability that I'm being burglarized
- In other words, what is $P(B = \top | M = \top, J = \top)$
- How on Earth would I estimate that probability? I don't know how to estimate that.

Available Knowledge

- LA has 1 million houses & 41 burglaries/day: $\Pr(B = \top) = \frac{41}{1000000}$
- There are ~20 earthquakes/year: $P(E = \top) = \frac{20}{365}$
- My burglar alarm is pretty good:

	$B = \bot, E = \bot$	$B = \bot, E = \top$	$B = \top, E = \bot$	$B = \top, E = \top$
$P(A = \top \mid B, E)$	$\frac{1}{100}$	$\frac{3}{5}$	$\frac{99}{100}$	$\frac{99}{100}$

- John would text if there was an alarm: $P(J = \top \mid A = \top) = \frac{9}{10}$
- On days with no alarm, he often sends cat videos: $P(J = \top \mid A = \bot) = \frac{1}{2}$

Combining the Available Knowledge

Putting it all together, we have … well, we have a big mess. And that's not including the variable M:

	$B = \bot$	$B = \top$
$P(B, E = \bot, A = \bot, J = \bot)$	$\left(\frac{999959}{1000000}\right)\left(\frac{345}{365}\right)\left(\frac{99}{100}\right)\left(\frac{1}{2}\right)$	$\left(\frac{41}{1000000}\right)\left(\frac{345}{365}\right)\left(\frac{99}{100}\right)\left(\frac{1}{2}\right)$
$P(B, E = \bot, A = \bot, J = \top)$	$\left(\frac{999959}{1000000}\right)\left(\frac{345}{365}\right)\left(\frac{99}{100}\right)\left(\frac{1}{2}\right)$	$\left(\frac{41}{1000000}\right)\left(\frac{345}{365}\right)\left(\frac{99}{100}\right)\left(\frac{1}{2}\right)$
$P(B, E = \bot, A = \top, J = \bot)$	$\left(\frac{999959}{1000000}\right)\left(\frac{345}{365}\right)\left(\frac{1}{100}\right)\left(\frac{1}{10}\right)$	$\left(\frac{41}{1000000}\right)\left(\frac{345}{365}\right)\left(\frac{1}{100}\right)\left(\frac{1}{10}\right)$
$P(B, E = \bot, A = \top, J = T)$	$\left(\frac{999959}{1000000}\right)\left(\frac{345}{365}\right)\left(\frac{1}{100}\right)\left(\frac{9}{10}\right)$	$\left(\frac{41}{1000000}\right)\left(\frac{345}{365}\right)\left(\frac{99}{100}\right)\left(\frac{9}{10}\right)$
⋮	⋮	⋮

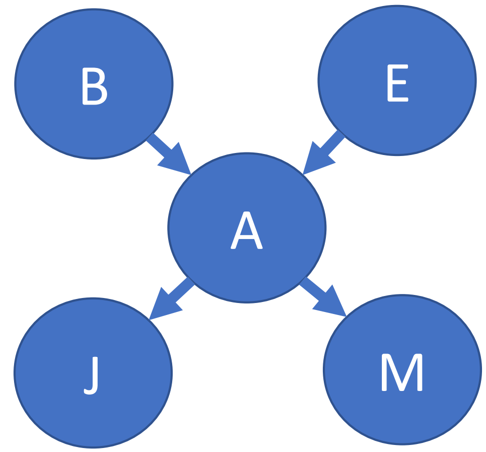
Outline

- Review: Bayesian classifier
- The Los Angeles burglar alarm example
- Bayesian network: A better way to represent knowledge
- Inference using a Bayesian network
- Independence and Conditional independence

Bayesian network: A better way to represent knowledge

A Bayesian network is a graph in which:

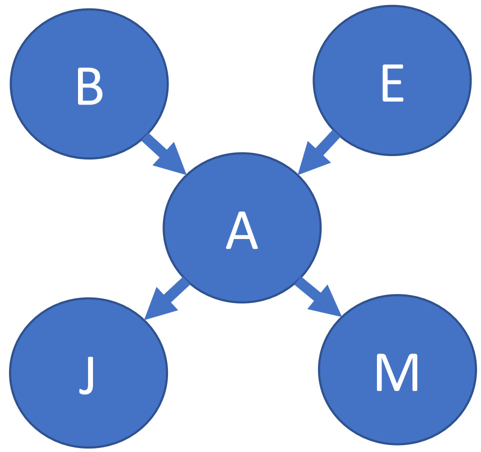
- Each variable is a node.
- An arrow between two nodes means that the child depends on the parent.
- If the child has no direct dependence on the parent, then there is no arrow.



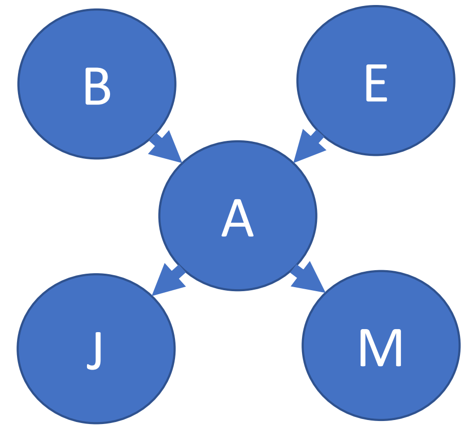
Bayesian network: A better way to represent knowledge

For example, this graph shows my knowledge that:

- My alarm rings if there is a burglary or an earthquake.
- John is more likely to call if my alarm is going off.
- Mary is more likely to call if my alarm is going off.



Complete description of my knowledge about the burglar alarm



$P(B = \top)$	$\frac{41}{1000000}$

$P(E = \top)$	$\frac{20}{365}$

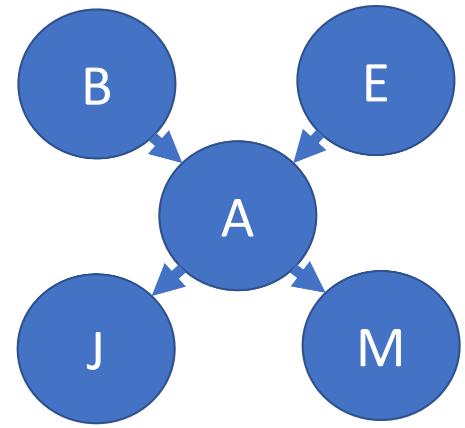
	$B = \bot, E = \bot$	$B = \bot, E = \top$	$B = \top, E = \bot$	$B = \top, E = \top$
$P(A = \top \mid B, E)$	$\frac{1}{100}$	$\frac{3}{5}$	$\frac{99}{100}$	$\frac{99}{100}$

	$A = \bot$	$\boldsymbol{A} = \top$
$P(J = \top \mid A)$	$\frac{1}{2}$	$\frac{9}{10}$

	$A = \bot$	$\boldsymbol{A} = \top$
$P(M = \top \mid A)$	$\frac{1}{8}$	$\frac{7}{8}$

Space complexity

- Without the Bayes network, space complexity is $\mathcal{O}\{v^n\}$
  - v = max cardinality of each variable
  - n = total # of variables
- With the Bayes network, space complexity is $\mathcal{O}\{nv^p\}$
  - p = max # parents any variable is allowed to have



Space complexity

- This is a Bayes network to help diagnose problems with your car's audio system.
- Naïve method: 41 binary variables, so the distribution is a table with $2^{41} \approx 2\times10^{12}$ entries.
- Bayes network: each variable has at most four parents, so the whole distribution can be described by less than $41\times2^4 = 656$ numbers.

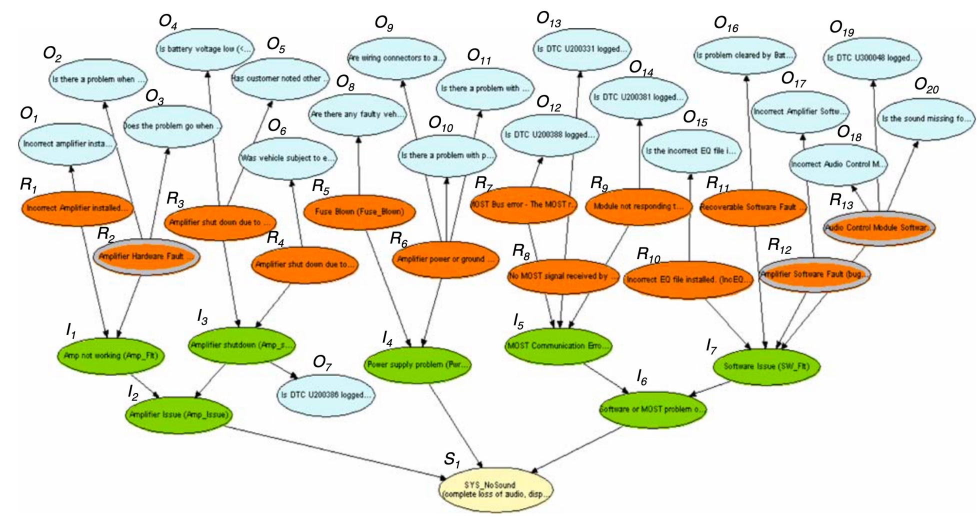


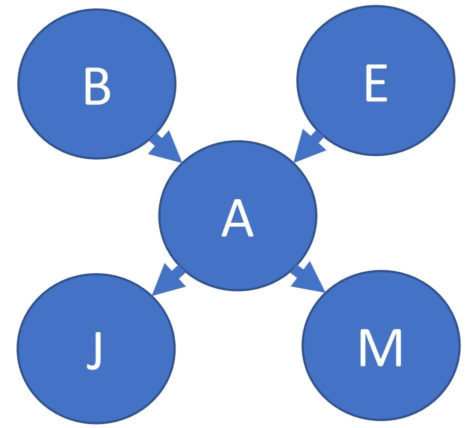
Fig. 6 Bayesian diagnostic model for the symptom "no sound"

Huang, McMurran, Dhadyalla & Jones, "Probability-based vehicle fault diagnosis: Bayesian network method," 2008

Outline

- Review: Bayesian classifier
- The Los Angeles burglar alarm example
- Bayesian network: A better way to represent knowledge
- Inference using a Bayesian network
- Independence and Conditional independence

Inference



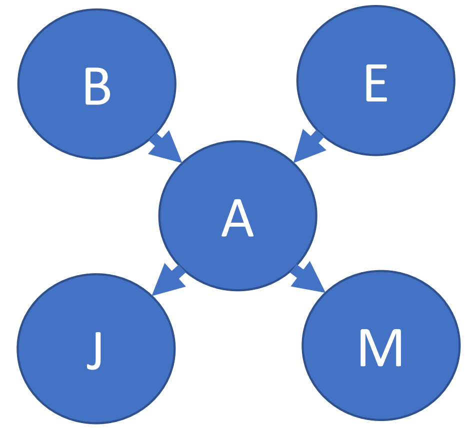
Both John and Mary texted me. Am I being burglarized?

$$P(B = \top | J = \top, M = \top) = \frac{P(B = \top, J = \top, M = \top)}{P(B = \top, J = \top, M = \top) + P(B = \bot, J = \top, M = \top)}$$

$$P(B = \top, J = \top, M = \top) = \sum_{e=\top}^{\bot} \sum_{a=\top}^{\bot} P(B = \top, E = e, A = a, J = \top, M = \top)$$

$$= \sum_{e=\top}^{\bot} \sum_{a=\top}^{\bot} P(B = \top) P(E = e) P(A = a | B = \top, E = e) P(J = \top | A = a) P(M = \top | A = a)$$

Time Complexity



- Using a Bayes network doesn't usually change the time complexity of a problem.
- If computing $P(B = \top | J = \top, M = \top)$ required considering $\mathcal{O}\{v^n\}$ possibilities without a Bayes network, it still requires considering $\mathcal{O}\{v^n\}$ possibilities

Some unexpected conclusions

- Burglary is so unlikely that, even if both Mary and John call, it is still more probable that a burglary didn't happen

$$P(B = \top | J = \top, M = \top) < P(B = \bot | J = \top, M = \top)$$

- The probability of an earthquake is higher!

$$P(B = \top | J = \top, M = \top) < P(E = \top | J = \top, M = \top)$$

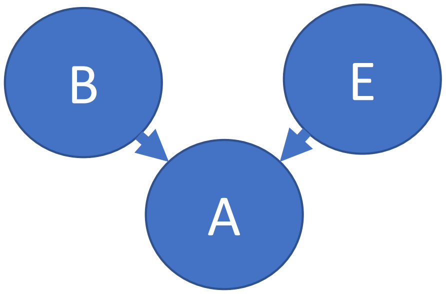
Quiz

Try the quiz!

Outline

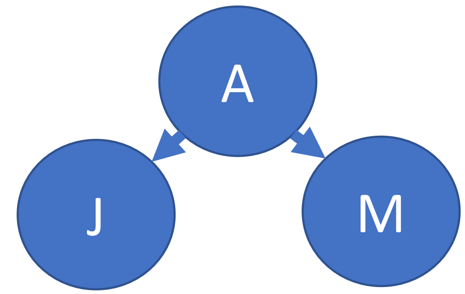
- Review: Bayesian classifier
- The Los Angeles burglar alarm example
- Bayesian network: A better way to represent knowledge
- Inference using a Bayesian network
- Independence and Conditional independence

Independence: No shared ancestors



- The variables B and E are independent
- Days with earthquakes and days w/o earthquakes have the same number of burglaries: $P(B = \top | E = \top) = P(B = \top | E = \bot) = P(B = \top)$.

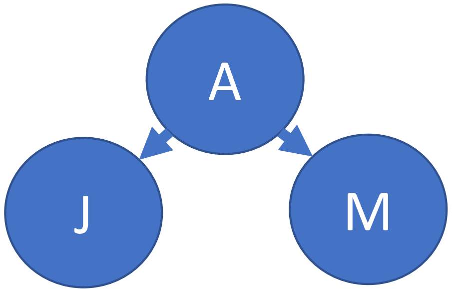
Shared ancestor = Not independent



- The variables J and M are not independent!
- If you know that John texted, that tells you that there was probably an alarm. Knowing that there was an alarm tells you that Mary will probably text you too:

$$P(M = \top | J = \top) \neq P(M = \top | J = \bot)$$

Conditional Independence if the Connection is Cut

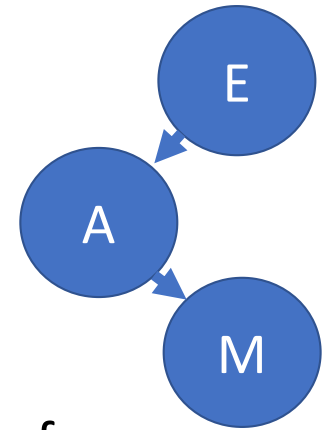


- The variables J and M are conditionally independent of one another given knowledge of A
- If you know that there was an alarm, then knowing that John texted gives no extra knowledge about whether Mary will text:

$$P(M=\top|J=\top,A=\top)=P(M=\top|J=\bot,A=\top)=P(M=\top|A=\top)$$

- Our knowledge of A "cuts the connection" between J and M

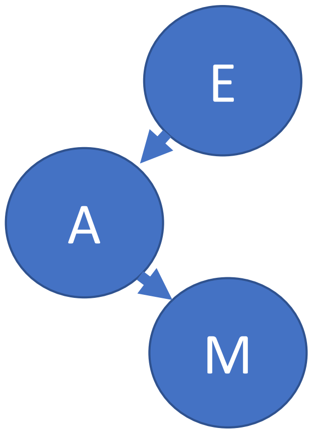
Shared ancestor = Not independent



- The ”shared ancestor” rule also applies when the shared ancestor of one variable is the descendant of the other
- For example, the variables E and M are not independent! M’s ancestor, A, is the descendant of E.
- If you know that Mary texted, that tells you that there was probably an alarm. Knowing that there was an alarm tells you that there is a >50% probability that there was an earthquake:

$$P(E = \top | M = \top) \neq P(E = \top | M = \bot)$$

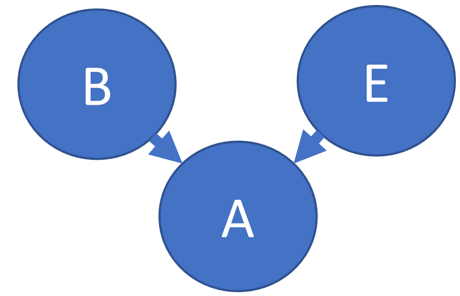
Conditional Independence if the Connection is Cut



- The variables E and M are conditionally independent of one another given knowledge of A
- If you know that there was an alarm, then knowing that Mary texted gives no extra knowledge about the existence of an earthquake:

$$P(E = \top | M = \top, A = \top) = P(E = \top | M = \bot, A = \top) = P(E = \top | A = \top)$$

Independent variables may not be conditionally independent!



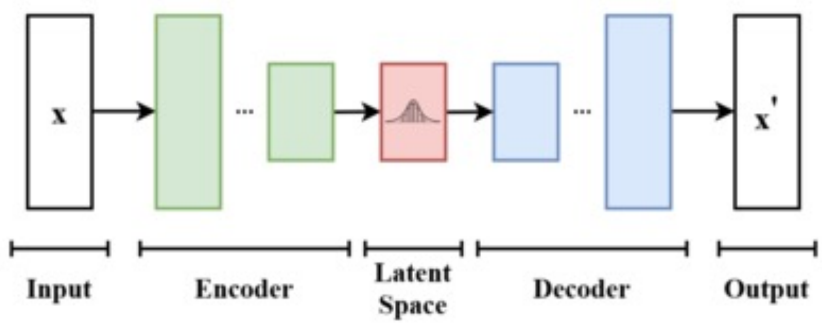
- The variables B and E are not conditionally independent of one another given knowledge of A
- If your alarm is ringing, then you probably have an earthquake ***OR*** a burglary. If there is an earthquake, then the conditional probability of a burglary goes down:

$$P(B = \top | E = \top, A = \top) \neq P(B = \top | E = \bot, A = \top)$$

- This is called the "explaining away" effect. The earthquake "explains away" the alarm, so you become less worried about a burglary.

Knowing about Independence and Conditional Independence can improve time complexity

- Improve time complexity by specifying the value of a shared ancestor: cuts the network into conditionally independent halves
- Example: Variational Autoencoder. Given the latent variable, the encoder and decoder are conditionally independent, can be solved with less time complexity



https://commons.wikimedia.org/wiki/File:VAE_Basic.png

Summary

- Bayesian network: A better way to represent knowledge
  - Reduces space complexity from $\mathcal{O}\{v^n\}$ to $\mathcal{O}\{nv^p\}$ -- huge if $n \gg p$
  - Does not automatically reduce time complexity.
- Key ideas: Independence and Conditional independence

		Shared Ancestor (of at least one)?	
		No	Yes
Shared Descendant (of both)?	No	Independent	Dependent unless shared ancestor value is known
	Yes	Independent unless shared descendant value is known	Dependent unless shared ancestor value known AND shared descendant value unknown