CS440/ECE448 Lecture 5: 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?

Binary Event Notation

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

Inference Problem

- Given J and M, I want to know what is the probability that I'm being burglarized
- In other words, what is $P(B|M, J)$
- 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: $\mathrm{P}(B) = \frac{41}{1000000}$
- There are ~20 earthquakes/year: $P(E) = \frac{20}{365}$
- My burglar alarm is pretty good:

	$\neg B, \neg E$	$\neg B, E$	$B, \neg E$	B, E
$P(A \mid B?, E?)$	$\frac{1}{100}$	$\frac{3}{5}$	$\frac{99}{100}$	$\frac{99}{100}$

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

Combining the Available Knowledge

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

	$\neg B$	B
$P(B?, \neg E, \neg A, \neg J)$	$\left(\frac{999959}{1000000}\right)\left(\frac{345}{365}\right)\left(\frac{99}{100}\right)\left(\frac{1}{2}\right)$	$\left(\frac{41}{1000000}\right)\left(\frac{345}{365}\right)\left(\frac{99}{100}\right)\left(\frac{1}{2}\right)$
$P(B?, \neg E, \neg A, J)$	$\left(\frac{999959}{1000000}\right)\left(\frac{345}{365}\right)\left(\frac{99}{100}\right)\left(\frac{1}{2}\right)$	$\left(\frac{41}{1000000}\right)\left(\frac{345}{365}\right)\left(\frac{99}{100}\right)\left(\frac{1}{2}\right)$
$P(B?, \neg E, A, \neg J)$	$\left(\frac{999959}{1000000}\right)\left(\frac{345}{365}\right)\left(\frac{1}{100}\right)\left(\frac{1}{10}\right)$	$\left(\frac{41}{1000000}\right)\left(\frac{345}{365}\right)\left(\frac{1}{100}\right)\left(\frac{1}{10}\right)$
$P(B?, \neg E, A, J)$	$\left(\frac{999959}{1000000}\right)\left(\frac{345}{365}\right)\left(\frac{1}{100}\right)\left(\frac{9}{10}\right)$	$\left(\frac{41}{1000000}\right)\left(\frac{345}{365}\right)\left(\frac{99}{100}\right)\left(\frac{9}{10}\right)$
$\vdots$	$\vdots$	$\vdots$

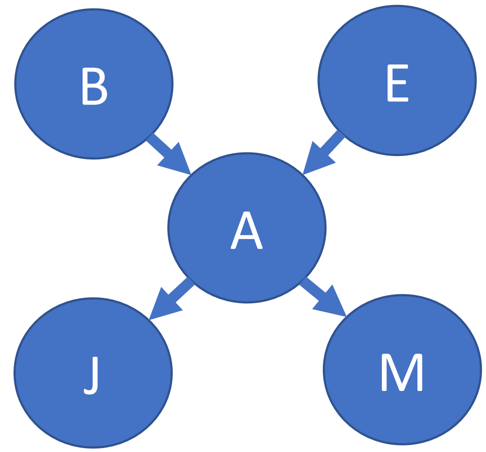
Outline

- 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:

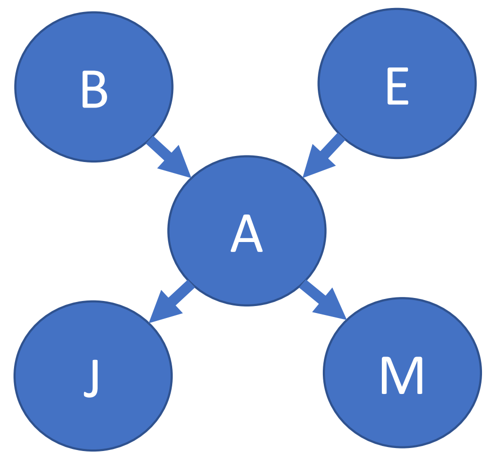
- Each variable is a node.
- An arrow between two nodes means that the child depends on the parent.
- If the child has no direct dependence on the parent, then there is no arrow.



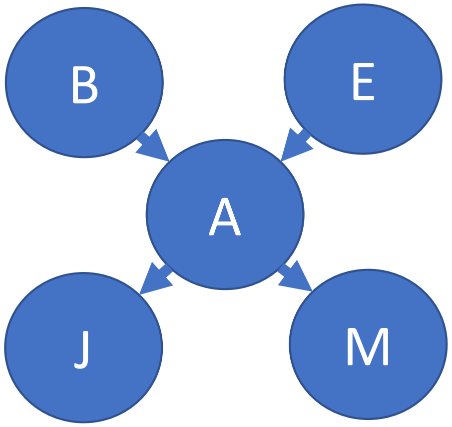
Bayesian network: A better way to represent knowledge

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



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

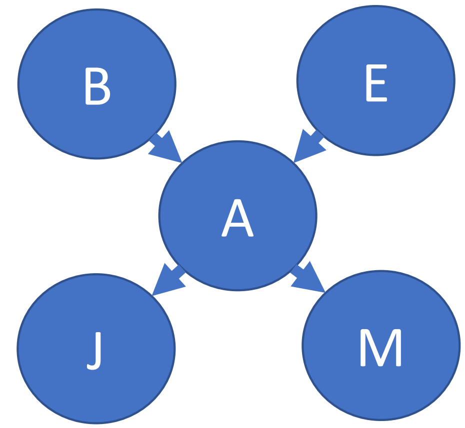
$P(E)$	$\frac{20}{365}$

	$\neg B, \neg E$	$\neg B, E$	$B, \neg E$	B, E
$P(A \mid B?, E?)$	$\frac{1}{100}$	$\frac{3}{5}$	$\frac{99}{100}$	$\frac{99}{100}$

	$\neg A$	$\boldsymbol{A}$
$P(J \mid A?)$	$\frac{1}{2}$	$\frac{9}{10}$

	$\neg A$	$\boldsymbol{A}$
$P(M \mid A?)$	$\frac{1}{8}$	$\frac{7}{8}$

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
  - Burglar alarm example: n=5, v=2, p=2, so it's $\mathcal{O}\{5\times 2^2\} = \mathcal{O}\{20\}$ parameters. Actual # parameters (previous slide) was 10.

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 $2^{41} \approx 2{\times}10^{12}$ entries.
- Bayes network: each variable has at most four parents, so the whole distribution can be described by less than $41{\times}2^{4} = 656$ numbers.

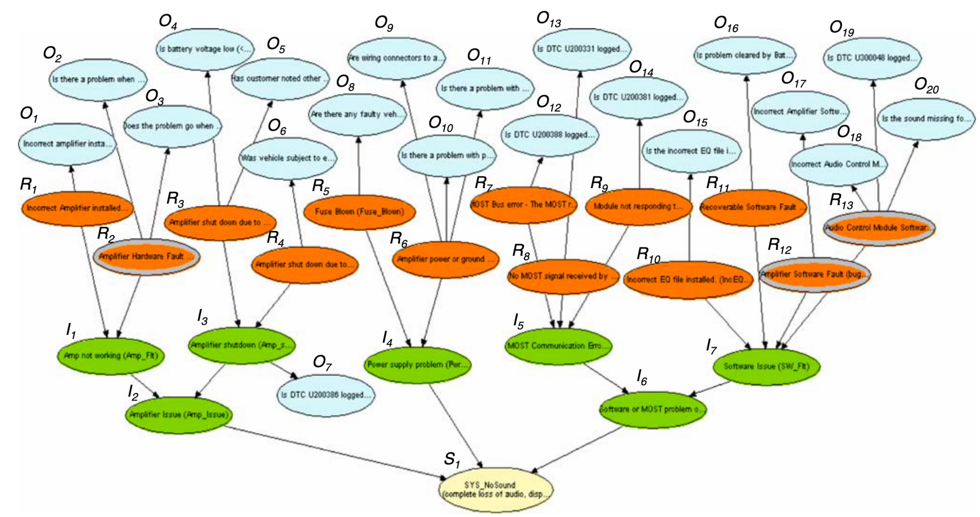


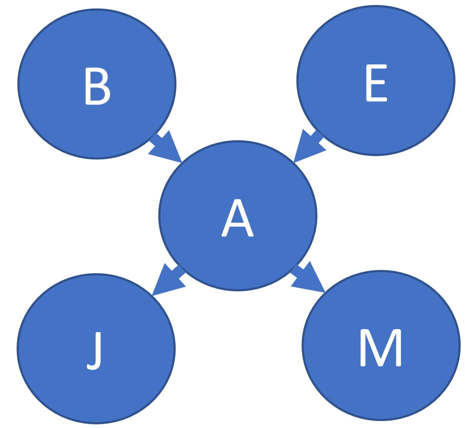
Fig. 6 Bayesian diagnostic model for the symptom "no sound"

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



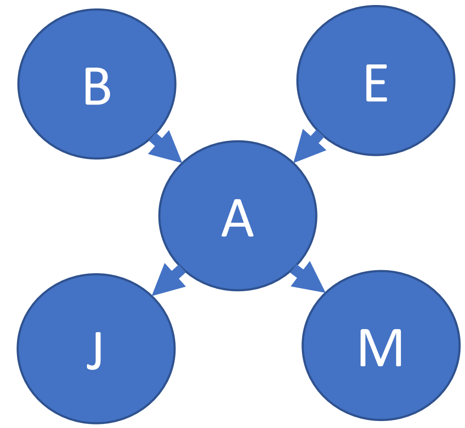
Both John and Mary texted me. Am I being burglarized?

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

$$P(B,J,M) = \sum_{E?} \sum_{A?} P(B, E?, A?, J, M)$$

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

Time Complexity



- Using a Bayes network doesn't usually change the time complexity of a problem.
- If computing $P(B|J, M)$ 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

$$P(B|J, M) < P(\neg B|J, M)$$

- The probability of an earthquake is higher!

$$P(B|J, M) < P(E|J, M)$$

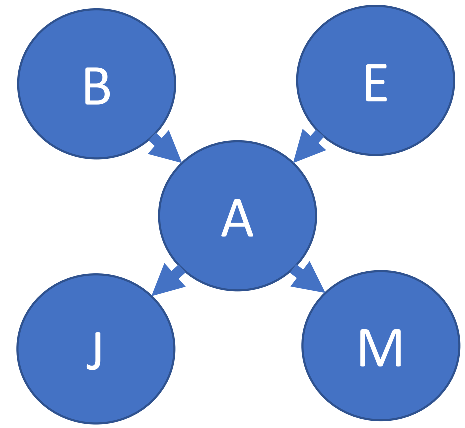
Quiz

Try the quiz!

Outline

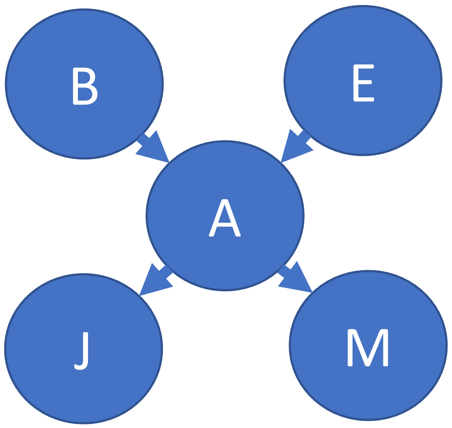
- Review: Bayesian classifier
- The Los Angeles burglar alarm example
- Bayesian network: A better way to represent knowledge
- Inference using a Bayesian network
- Independence and Conditional independence

Time Complexity Again…



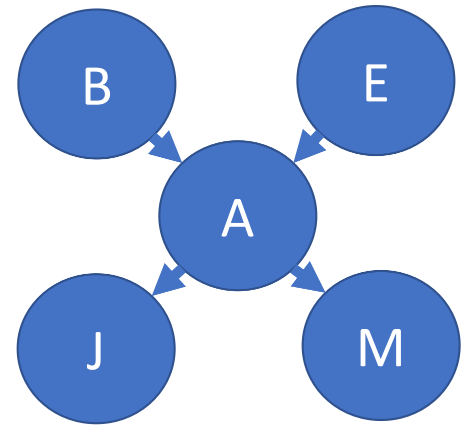
- Bayes networks make it possible to reduce time complexity, but only with additional assumptions.
- Additional assumptions are always about independence and/or conditional independence of some of the variables.

Independence & Conditional Independence



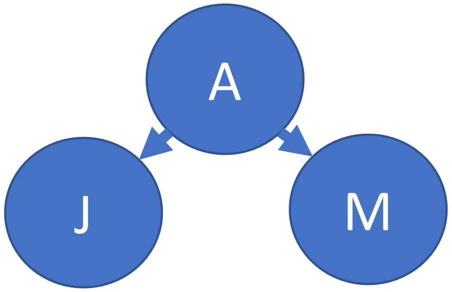
- Remember the basics: if an arrow connects two variables, then they are dependent.
- What about the other variables?

Three basic rules



1. Variables with a shared ancestor are dependent on one another unless the value of the ancestor is known
2. Variables with a shared descendant are dependent on one another unless the value of the descendant is UNKNOWN
3. Variables with no shared ancestor or descendent are independent

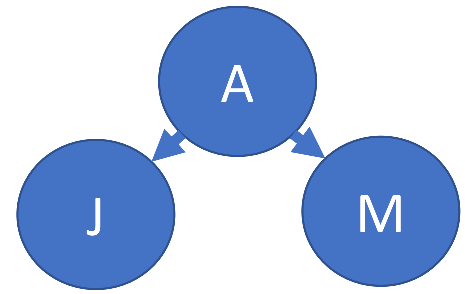
1. Shared ancestor = Not independent



- 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:

$$P(M|J) \neq P(M|\neg J)$$

1. Conditionally Independent given knowledge of the shared ancestor

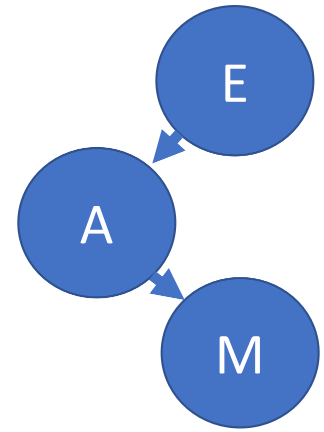


- The variables J and M are conditionally independent of one another given knowledge of A
- If you know that there was an alarm, then knowing that John texted gives no extra knowledge about whether Mary will text:

$$P(M|J, A) = P(M|\neg J, A) = P(M|A)$$

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

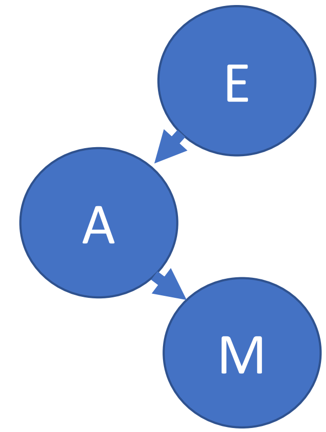
1. I am my own ancestor



- These rules only work if "I am my own ancestor"
- For example, the variables E and M are not independent because E is M's ancestor, and E is also its own ancestor
- 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:

$$P(E|M) \neq P(E|\neg M)$$

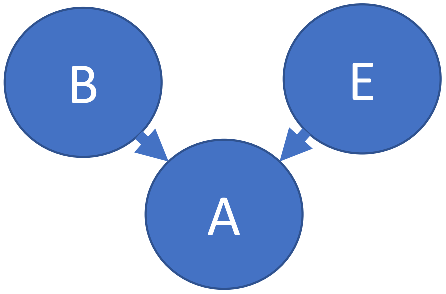
1. Ties of ancestry are broken if an intermediate variable is known



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

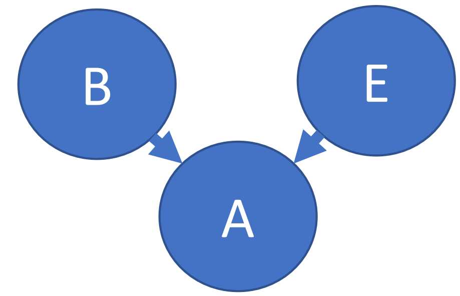
$$P(E|M,A) = P(E|\neg M,A) = P(E|A)$$

2. Shared descendant: can be independent



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

2. Conditionally DEPENDENT given knowledge of a shared DESCENDANT

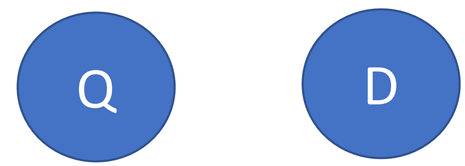


- The variables B and E are conditionally DEPENDENT given knowledge of A
- If your alarm is ringing, then you probably have an earthquake *OR* a burglary. If there is an earthquake, then the conditional probability of a burglary goes down:

$$P(B|E,A) \neq P(B|\neg E,A)$$

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

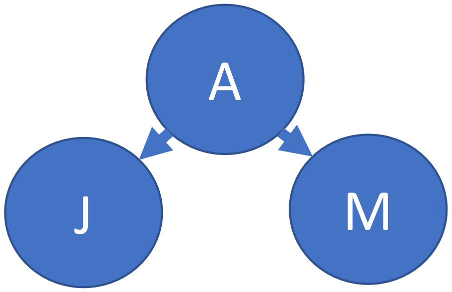
3. No shared ancestor or shared descendant = independent



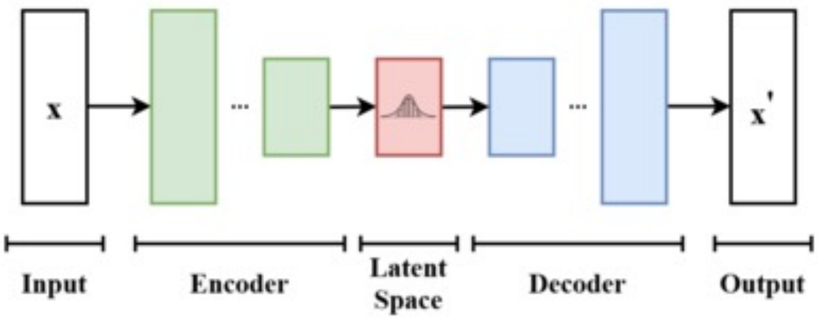
- Suppose I flip a quarter and a dime. Each can either be heads or tails.
- They have no shared ancestor (there is no other variable that affects both coin flips).
- They have no shared descendant (there is no other variable whose distribution is affected by both the quarter and the dime).
- They are totally independent

$$P(Q|D) = P(Q|\neg D)$$

Time Complexity Again…



- In useful situations, you can improve time complexity by specifying the value of a shared ancestor. Example next lecture: HMM.
- Another 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 reduce time complexity without extra assumptions (next lecture).
- Key ideas: Independence and Conditional independence
  1. Shared ancestor $\Longrightarrow$ dependent unless ancestor known
  2. Shared descendant $\Longrightarrow$ dependent unless descendant unknown
  3. No shared ancestor or descendent $\Longrightarrow$ independent