CS440/ECE448 Lecture 16: Parameter and Structure Learning for Bayesian Networks

By Mark Hasegawa-Johnson, 2/2020

With some slides by Svetlana Lazebnik, 9/2017

Parameter and Structure Learning for Bayesian Networks

- Parameter Learning
  - from Fully Observed data: Maximum Likelihood
  - from Partially Observed data: Expectation Maximization
- Structure Learning
  - The usual method: knowledge engineering
  - An interesting recent method: causal analysis

Outline

- Parameter Learning
  - from Fully Observed data: Maximum Likelihood
  - from Partially Observed data: Expectation Maximization
- Structure Learning
  - The usual method: knowledge engineering
  - An interesting recent method: causal analysis

Flying cows

The scenario:

Central Illinois has recently had a problem with flying cows.

Farmers have called the university to complain that their cows flew away.

Flying cows

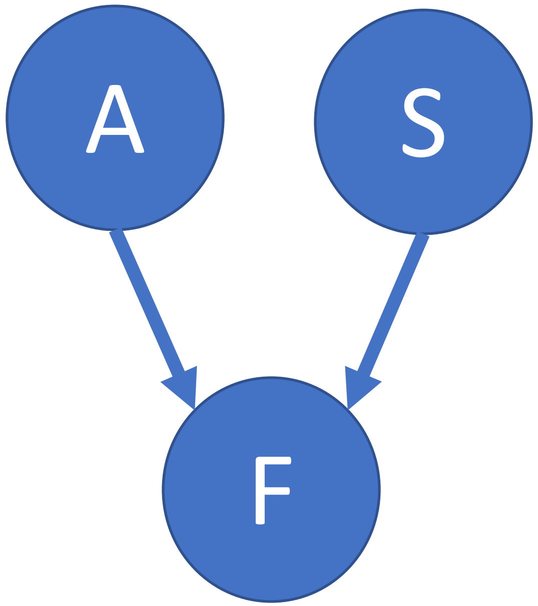
The university dispatched a team of expert vaccavolatologists. They determined that almost all flying cows were explained by one or both of the following causes:

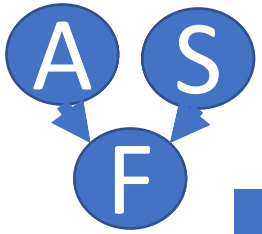
- **Smart cows**. The cows learned how to fly, on their own, without help.
- **Alien intervention**. UFOs taught the cows how to fly.

Flying cows

The vaccavolatologists created a Bayes net, to help them predict any future instances of cow flying:

- P(A) = Probability that aliens teach the cow.
- P(S) = Probability that a cow is smart enough to figure out how to fly on its own.
- P(F|S,A) = Probability that a cow learns to fly.





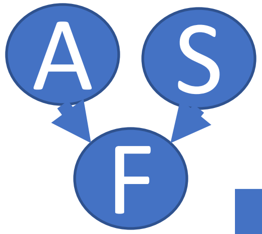
Flying cows

They went out to watch a nearby pasture for ten days.

- They reported the number of days on which A, S, and/or F occurred.
- Their results are shown in the table at left (True is marked as "T"; False is shown with a blank).

Day	A	S	F
1			
2		T	T
3			
4	T	T	T
5	T		
6			
7	T		T
8			
9			T
10			

Flying cows



The vaccavolatologists now wish to estimate the parameters of their Bayes net

- P(A)
- P(S)
- P(F|S,A)

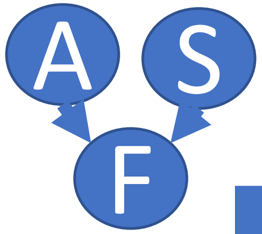
…so that they will be better able to testify before Congress about the relative dangers of aliens versus smart cows.

Day	A	S	F
1			
2		T	T
3			
4	T	T	T
5	T		
6			
7	T		T
8			
9			T
10			

Outline

- Parameter Learning
  - from Fully Observed data: Maximum Likelihood
  - from Partially Observed data: Expectation Maximization
- Structure Learning
  - The usual method: knowledge engineering
  - An interesting recent method: causal analysis

Maximum Likelihood Estimation

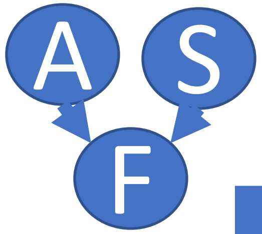


Suppose we have n training examples, $1 \le i \le n$, with known values for each of the random variables:

- A_i or $\neg A_i$
- S_i or $\neg S_i$
- F_i or $\neg F_i$

Day	A	S	F
1	$\neg A_1$	$\neg S_1$	$\neg F_1$
2	$\neg A_2$	S_2	F_2
3	$\neg A_3$	$\neg S_3$	$\neg F_3$
4	A_4	S_4	F_4
5	A_5	$\neg S_5$	$\neg F_5$
6	$\neg A_6$	$\neg S_6$	$\neg F_6$
7	A_7	$\neg S_7$	F_7
8	$\neg A_8$	$\neg S_8$	$\neg F_8$
9	$\neg A_9$	$\neg S_9$	F_9
10	$\neg A_{10}$	$\neg S_{10}$	$\neg F_{10}$

Maximum Likelihood Estimation

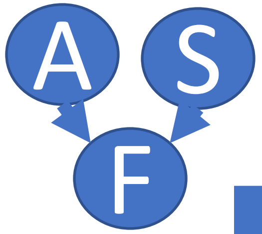


We can estimate model parameters to be the values that maximize the likelihood of the observations, subject to the constraints that

$$P(A) + P(\neg A) = 1$$
$$P(S) + P(\neg S) = 1$$
$$P(F|S,A) + P(\neg F|S,A) = 1$$

Day	A	S	F
1	$\neg A_1$	$\neg S_1$	$\neg F_1$
2	$\neg A_2$	S_2	F_2
3	$\neg A_3$	$\neg S_3$	$\neg F_3$
4	A_4	S_4	F_4
5	A_5	$\neg S_5$	$\neg F_5$
6	$\neg A_6$	$\neg S_6$	$\neg F_6$
7	A_7	$\neg S_7$	F_7
8	$\neg A_8$	$\neg S_8$	$\neg F_8$
9	$\neg A_9$	$\neg S_9$	F_9
10	$\neg A_{10}$	$\neg S_{10}$	$\neg F_{10}$

Maximum Likelihood Estimation



The maximum likelihood parameters are

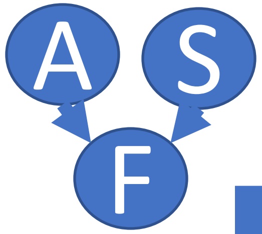
$$P(A) = \frac{\text{\# days on which } A_i}{\text{\# days total}}$$

$$P(S) = \frac{\text{\# days on which } S_i}{\text{\# days total}}$$

$$P(F|s,a) = \frac{\text{\# days (A=}a\text{,S=}s\text{, F)}}{\text{\# days (A=}a\text{,S=}s\text{)}}$$

Day	A	S	F
1	$\neg A_1$	$\neg S_1$	$\neg F_1$
2	$\neg A_2$	S_2	F_2
3	$\neg A_3$	$\neg S_3$	$\neg F_3$
4	A_4	S_4	F_4
5	A_5	$\neg S_5$	$\neg F_5$
6	$\neg A_6$	$\neg S_6$	$\neg F_6$
7	A_7	$\neg S_7$	F_7
8	$\neg A_8$	$\neg S_8$	$\neg F_8$
9	$\neg A_9$	$\neg S_9$	F_9
10	$\neg A_{10}$	$\neg S_{10}$	$\neg F_{10}$

Maximum Likelihood Estimation



The maximum likelihood parameters are

$$P(A) = \frac{3}{10}, \qquad P(S) = \frac{2}{10}$$

a	s	$P(F\|s,a)$
F	F	1/6
F	T	1
T	F	1/2
T	T	1

Day	A	S	F
1	$\neg A_1$	$\neg S_1$	$\neg F_1$
2	$\neg A_2$	S_2	F_2
3	$\neg A_3$	$\neg S_3$	$\neg F_3$
4	A_4	S_4	F_4
5	A_5	$\neg S_5$	$\neg F_5$
6	$\neg A_6$	$\neg S_6$	$\neg F_6$
7	A_7	$\neg S_7$	F_7
8	$\neg A_8$	$\neg S_8$	$\neg F_8$
9	$\neg A_9$	$\neg S_9$	F_9
10	$\neg A_{10}$	$\neg S_{10}$	$\neg F_{10}$

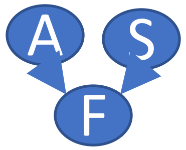
Conclusions: maximum likelihood estimation

- Smart cows are far more dangerous than aliens.
- Maximum likelihood estimation is very easy to use, IF you have training data in which the values of ALL variables are observed.

- …but what if some of the variables can't be observed?
- For example: the cows decide to stop responding to written surveys. Therefore, it's impossible to **observe**, on any given day, how smart the cows are. We don't know if $s_i = T$ or $s_i = F$…

Outline

- Parameter Learning
  - from Fully Observed data: Maximum Likelihood
  - from Partially Observed data: Expectation Maximization
- Structure Learning
  - The usual method: knowledge engineering
  - An interesting recent method: causal analysis



Partially observed data

Suppose that we have the following observations:

- We know whether A=True or False.
- We know whether F=True or False.
- We don't know whether S is True or False (shown as ”?”).

Day	A	S	F
1		?	
2		?	T
3		?	
4	T	?	T
5	T	?	
6		?	
7	T	?	T
8		?	
9		?	T
10		?	

Expectation Maximization (EM): Main idea

Remember that maximum likelihood estimation counts examples:

$$P(F|S=s, A=a) = \frac{\#\text{ days } \; S=s, A=a, \; F}{\#\text{ days } \; S=s, A=a}$$

Expectation maximization is similar, but using “expected counts” instead of actual counts:

$$P(F|S=s, A=a) = \frac{E[\#\text{ days } S=s, \; A=a, \; F]}{E[\#\text{ days } S=s, \; A=a]}$$

Where E[X] means “expected value of X”.

Expectation Maximization (EM): overview

INITIALIZE: Make some **initial guess** what might be the values of P(A), P(S), and P(F|A,S).

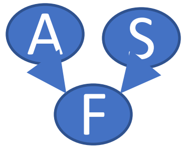
ITERATE until convergence:

1. **Partial days**: compute $P(S_i|A_i, F_i)$ for each of the days in your training corpus ($1 \leq i \leq n$).
2. **Expected counts**:

$$E[\#\text{ days } S, A_i = a, F_i = f] = \sum_{i:a_i=a, f_i=f} P(S|a, f)$$

1. **Re-estimate** the probabilities P(A), P(S), and P(F|A,S):

$$P(F|S = s, A = a) = \frac{E[\#\text{ days } S = s,\ \ A = a,\ \ \ F]}{E[\#\text{ days } S = s,\ \ A = a]}$$

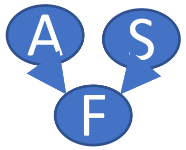


Example: Initialize

Marilyn Modigliani is a professional vaccavolatologist. She gives us these initial guesses about the possible model parameters (her guesses are probably not quite right, but they are as good a guess as anybody else's):

$$P(A) = \frac{1}{4}, \qquad P(S) = \frac{1}{4}$$

a	s	$P(F\|s,a)$
F	F	0
F	T	1/2
T	F	1/2
T	T	1

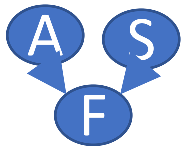


Partial days

Based on Marilyn’s model, we calculate $P(S = s|a_i, f_i)$ for each day, as shown in the table at right.

Day	A	$P(S\|a_i, f_i)$	$P(\neg S\|a_i, f_i)$	F
1		1/7	6/7	
2		1	0	T
3		1/7	6/7	
4	T	2/5	3/5	T
5	T	0	1	
6		1/7	6/7	
7	T	2/5	3/5	T
8		1/7	6/7	
9		1	0	T
10		1/7	6/7	

Expected counts

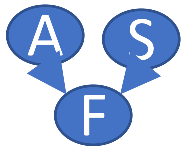


The expected counts are

$$E[\#\text{ days } S = s, A = a, F = f] = \sum_{i:a_i=a, f_i=f} P(S = s|a, f)$$

a	f	$E[\# \; days \; S\vert a, f]$	$E[\# \; days \; \neg S\vert a, f]$
F	F	$\frac{1}{7}+\frac{1}{7}+\frac{1}{7}+\frac{1}{7}+\frac{1}{7}=\frac{5}{7}$	$\frac{6}{7}+\frac{6}{7}+\frac{6}{7}+\frac{6}{7}+\frac{6}{7}=\frac{30}{7}$
F	T	$1+1=2$	0+0=0
T	F	0	1
T	T	$\frac{2}{5}+\frac{2}{5}=\frac{4}{5}$	$\frac{3}{5}+\frac{3}{5}=\frac{6}{5}$

Re-estimated probabilities



The re-estimated probabilities are

$$P(A) = \frac{\#\text{ days } A}{\#\text{ days total}} = \frac{3}{10}$$

$$P(S) = \frac{E[\#\text{ days } S]}{\#\text{ days total}} = \frac{\frac{5}{7} + 2 + 0 + \frac{4}{5}}{10} = \frac{123}{350}$$

$$P(F|S = s, A = a) = \frac{E[\#\text{ days } S = s,\ \ A = a,\ \ F]}{E[\#\text{ days } S = s,\ \ A = a]}$$

a	s	$P(F\|S = s, A = a)$
F	F	$\frac{0}{\frac{30}{7} + 0} = 0$
F	T	$\frac{2}{\frac{5}{7} + 2} = \frac{14}{19}$
T	F	$\frac{6/5}{1 + \frac{6}{5}} = \frac{6}{11}$
T	T	$\frac{4/5}{0 + 4/5} = 1$

Expectation Maximization (EM): review

INITIALIZE: Make some **initial guess** what might be the values of P(A)=0.25, P(S)=0.25, and $P(F|A,S) = \left(0, \frac{1}{2}, \frac{1}{2}, 1\right)$.

ITERATE until convergence:

1. **Partial days**: compute $P(S|a_i, f_i)$
2. **Expected counts**: $E[\#\text{ days } S, A = a, F = f]$.
3. **Re-estimate.** After the first iteration, we have

$$\mathrm{P(A)} = \frac{3}{10}, \mathrm{P(S)} = \frac{123}{350}, \text{ and } P(F|A,S) = \left(0, \frac{14}{19}, \frac{6}{11}, 1\right).$$

Continue the iteration, shown above, until P(A), P(S), and P(F|A,S) stop changing.

Properties of the EM algorithm

- It always converges.
- The parameters it converges to (P(A), P(S), and P(F|A,S)):
  - are guaranteed to be **at least as good as** your initial guess, but
  - They depend on your initial guess. Different initial guesses may result in different results, after the algorithm converges.
  - For example, Marilyn's initial guess was $P(F|\neg S, \neg A) = \mathbf{0}$. Notice that we ended up with the same value! According to the fully observed data we saw earlier, that might not be the best possible parameter for these data.

Outline

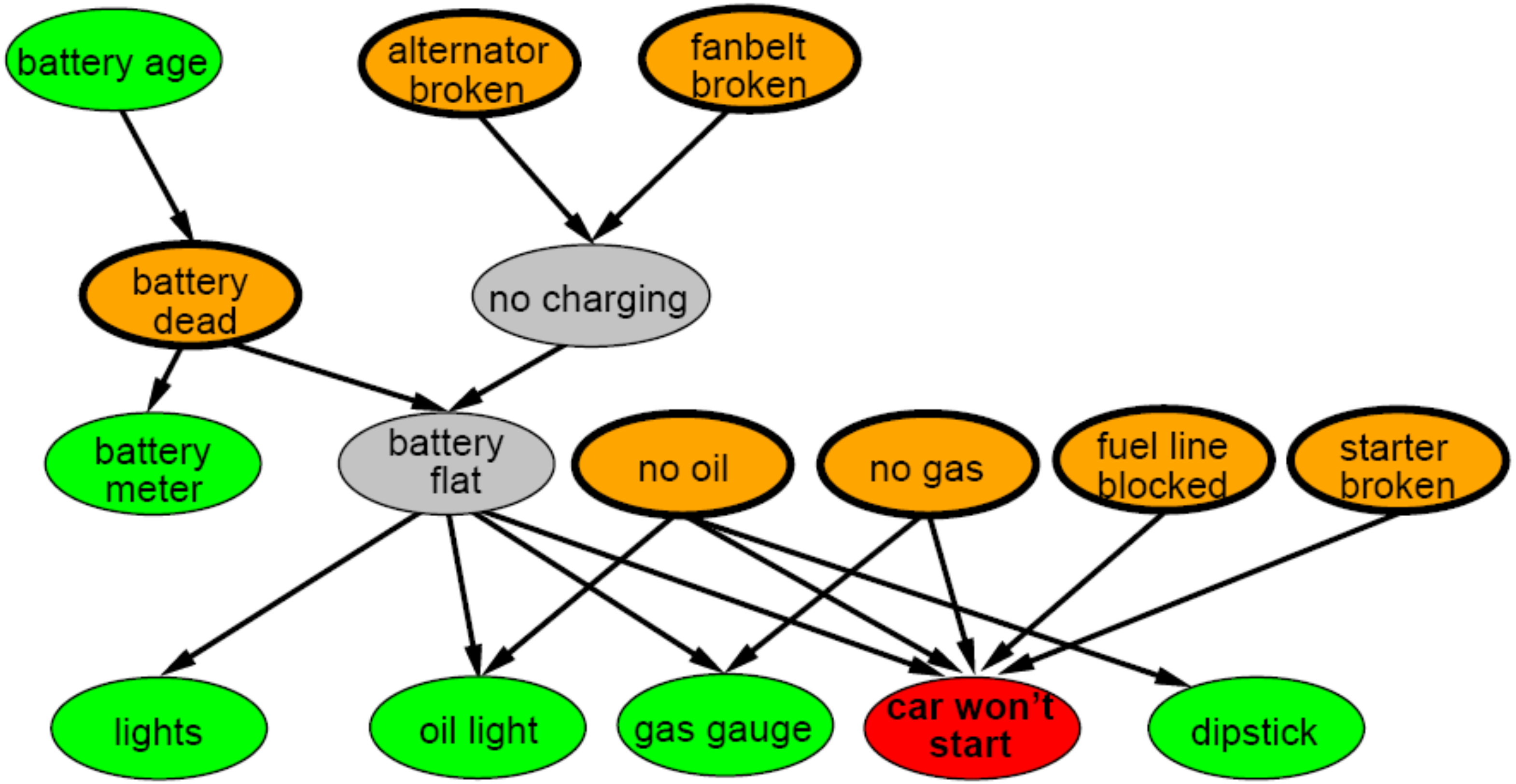
- Parameter Learning
  - from Fully Observed data: Maximum Likelihood
  - from Partially Observed data: Expectation Maximization
- Structure Learning
  - The usual method: knowledge engineering
  - An interesting recent method: causal analysis

Knowledge engineering

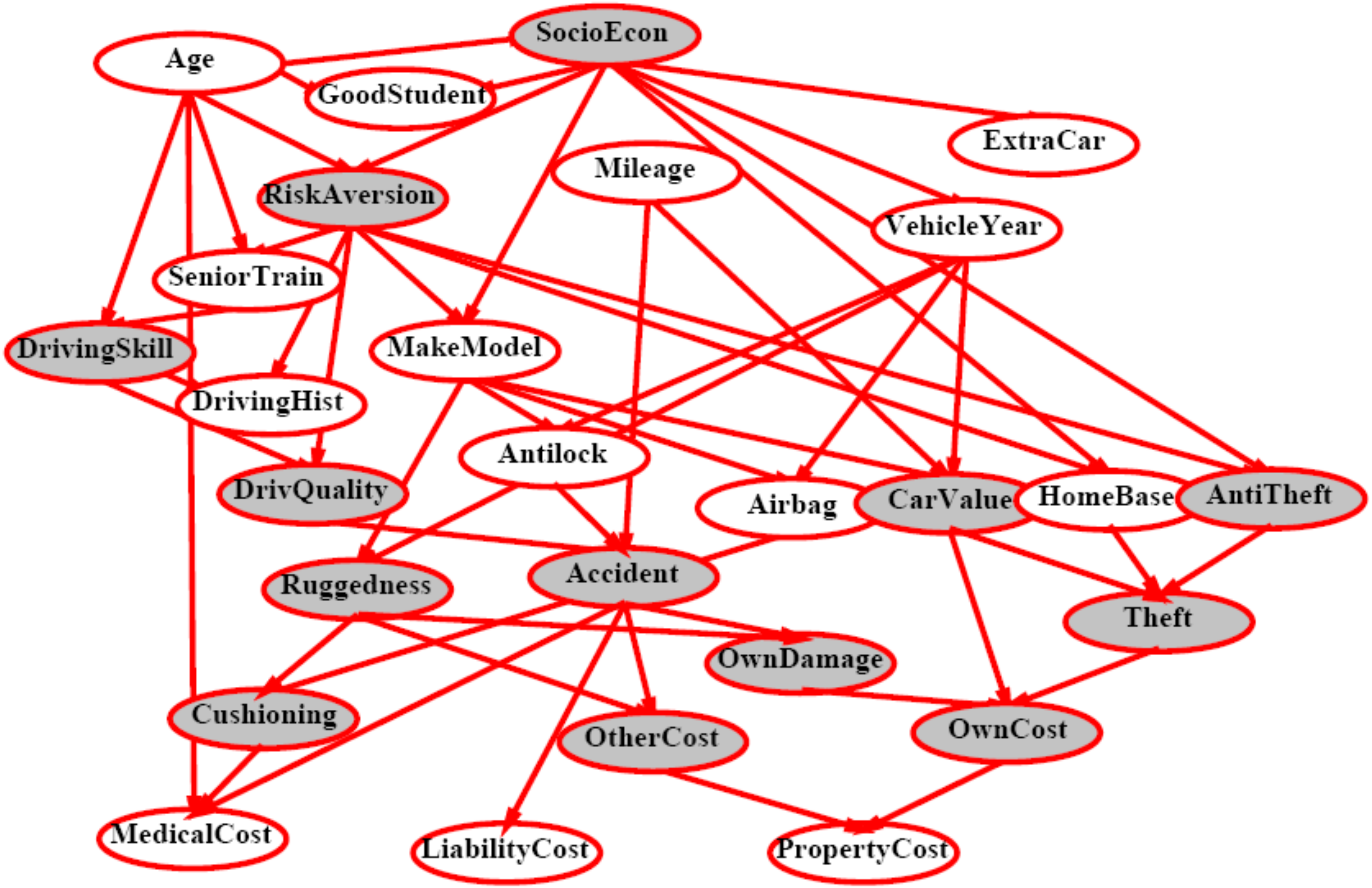
1. Find somebody who knows a lot about the problem you're trying to model (flying cows, or burglars in Los Angeles, or whatever).
2. Get her to tell you which variables depend on which others.
3. Draw corresponding circles and arrows.
4. Done! Proceed to parameter estimation.

Example Bayes Network: Car diagnosis

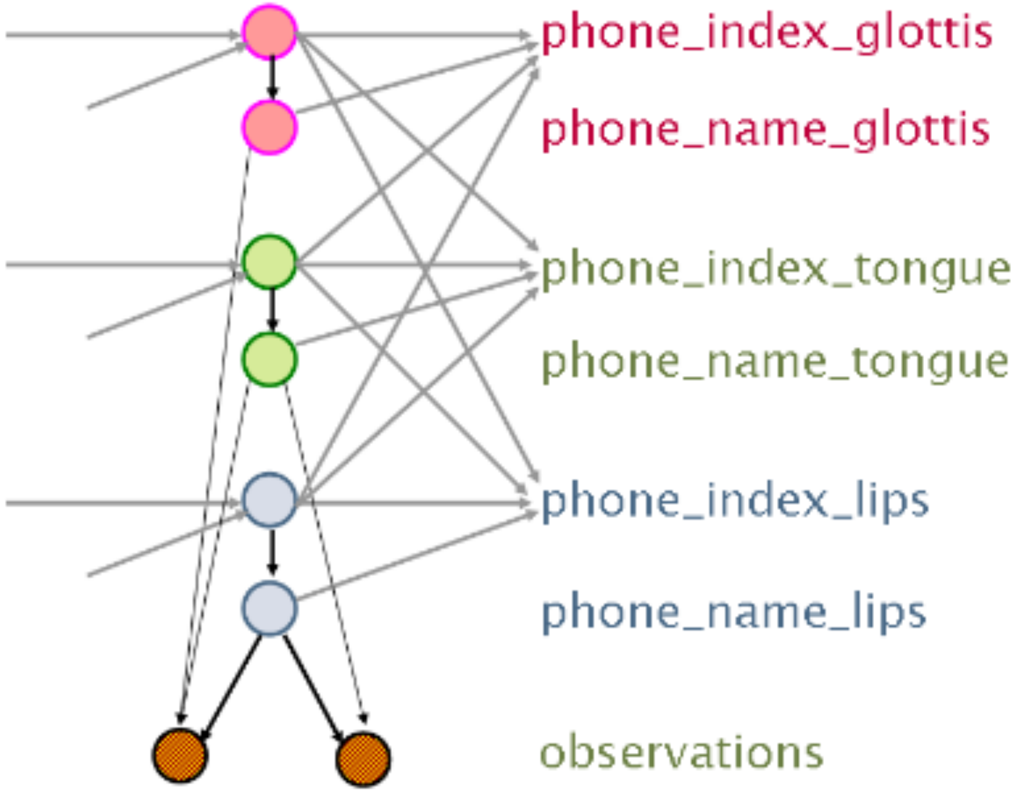
- Initial observation: car won't start
- Orange: "broken, so fix it" nodes
- Green: testable evidence
- Gray: "hidden variables" to ensure sparse structure, reduce parameters



Example Bayes Network: Cost of Car insurance



Example Bayes Network: speech acoustics and speech appearance depend on glottis, tongue, and lip positions



[Audiovisual Speech Recognition with Articulator Positions as Hidden Variables]

Mark Hasegawa-Johnson, Karen Livescu, Partha Lal and Kate Saenko

International Congress on Phonetic Sciences 1719:299-302, 2007

Outline

- Parameter Learning
  - from Fully Observed data: Maximum Likelihood
  - from Partially Observed data: Expectation Maximization
- Structure Learning
  - The usual method: knowledge engineering
  - An interesting recent method: causal analysis

Causal analysis

Suppose you know that you have V variables $X_1, \ldots, X_V$, but you don't know which variables depend on which others. You can learn this from the data:

For every possible ordering of the variables (there are V! possible orderings):

1. Create a blank initial network
2. For each variable in this ordering, i = 1 to V:
   a. add variable X_i to the network
   b. Check your training data. If there is any variable $X_1, \ldots, X_{i-1}$ that CHANGES the probability of X_i=1, then add that variable to the set Parents(X_i) such that $P(X_i \mid \text{Parents}(X_i)) = P(X_i \mid X_1, \ldots X_{i-1})$
3. Count the number of edges in the graph with this ordering.

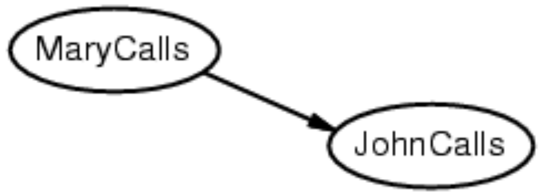
Choose the graph with the smallest number of edges.

Example: The Los Angeles burglar alarm

- Suppose we choose the ordering M, J, A, B, E

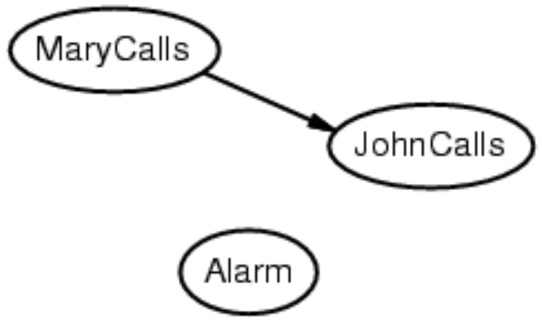
Example: The Los Angeles burglar alarm

- Suppose we choose the ordering M, J, A, B, E



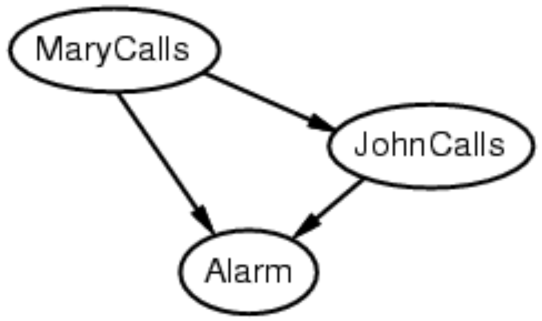
Example: The Los Angeles burglar alarm

- Suppose we choose the ordering M, J, A, B, E



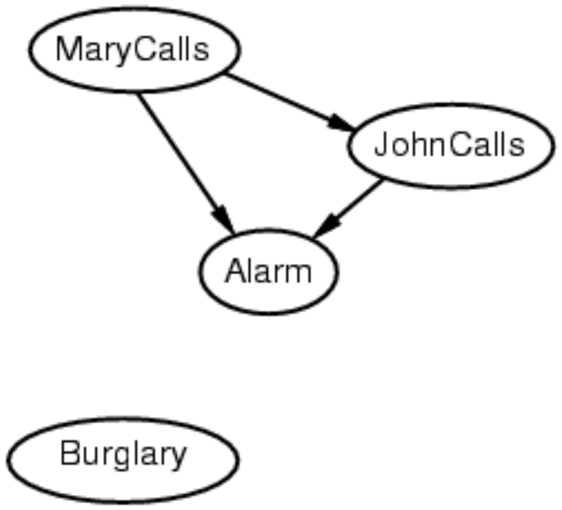
Example: The Los Angeles burglar alarm

- Suppose we choose the ordering M, J, A, B, E



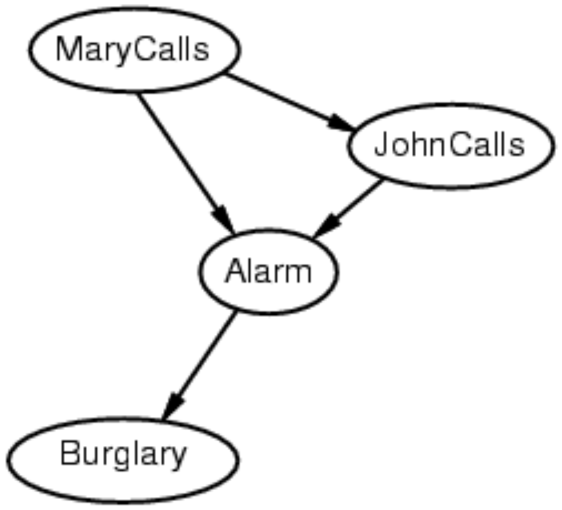
Example: The Los Angeles burglar alarm

- Suppose we choose the ordering M, J, A, B, E



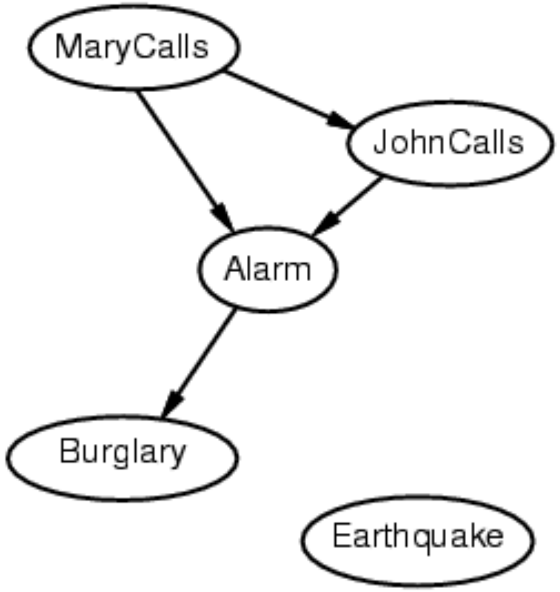
Example: The Los Angeles burglar alarm

- Suppose we choose the ordering M, J, A, B, E



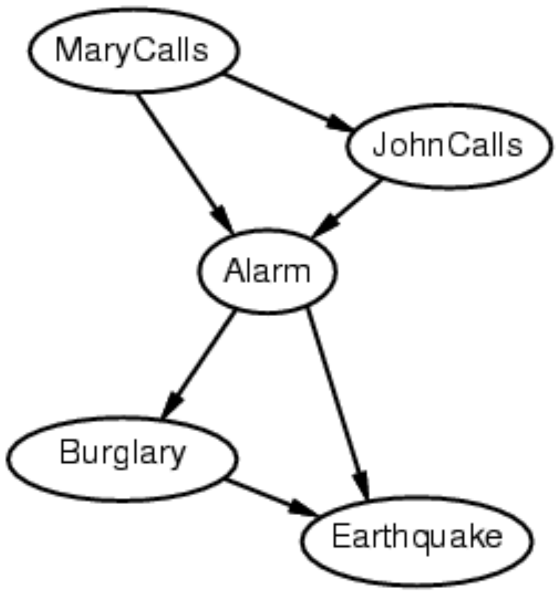
Example: The Los Angeles burglar alarm

- Suppose we choose the ordering M, J, A, B, E

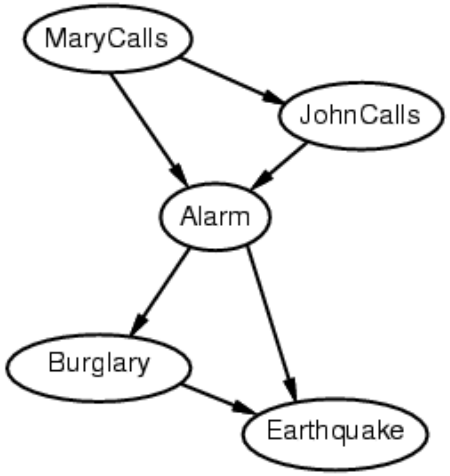


Example: The Los Angeles burglar alarm

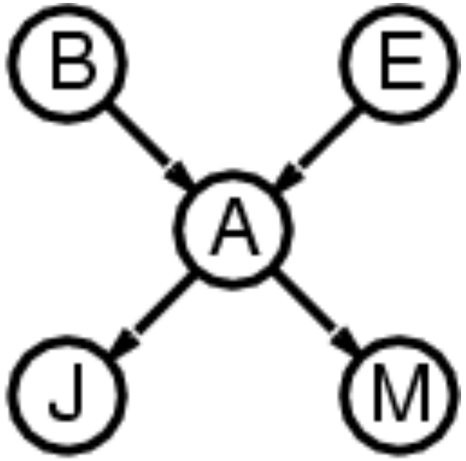
- Suppose we choose the ordering M, J, A, B, E



Example: The Los Angeles burglar alarm



versus

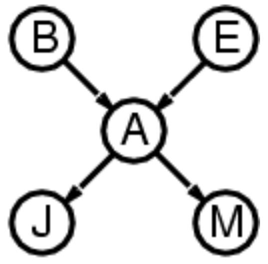


- Deciding conditional independence is hard in noncausal directions
  - The causal direction seems much more natural
- Network is less compact: 1 + 2 + 4 + 2 + 4 = 13 numbers needed (vs. 1+1+4+2+2=10 for the causal ordering)

Why store it in causal order? A: Saves memory

- Suppose we have a Boolean variable X_i with k Boolean parents. How many rows does its conditional probability table have?
  - 2^k rows for all the combinations of parent values
  - Each row requires one number for $P(X_i = \text{true} \mid \text{parent values})$
- If each variable has no more than k parents, how many numbers does the complete network require?
  - $O(n \cdot 2^k)$ numbers – vs. $O(2^n)$ for the full joint distribution
- How many nodes for the burglary network?

  1 + 1 + 4 + 2 + 2 = 10 numbers (vs. 2^5-1 = 31)



Parameter and Structure Learning for Bayesian Networks

- Maximum Likelihood (ML):

$$P(F|S=s, A=a) = \frac{\#\text{ days } (\text{A}=a, \text{S}=s, \text{F})}{\#\text{ days } (\text{A}=a, \text{S}=s)}$$

- Expectation Maximization (EM):

$$P(F|S=s, A=a) = \frac{E[\#\text{ days } A=a, S=s, \ \ F]}{E[\#\text{ days } A=a,\ S=s]}$$

- Knowledge Engineering: ask an expert.
- Causal Analysis: construct all possible graphs, keep the one with the fewest edges.