CS440/ECE448 Lecture 21: Parameter Learning for Bayesian Networks

By Mark Hasegawa-Johnson, 3/2022

Parameter Learning for Bayesian Networks

- From observed data: Maximum likelihood
- From observed data: Laplace smoothing
- From partially observed data: Expectation maximization

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

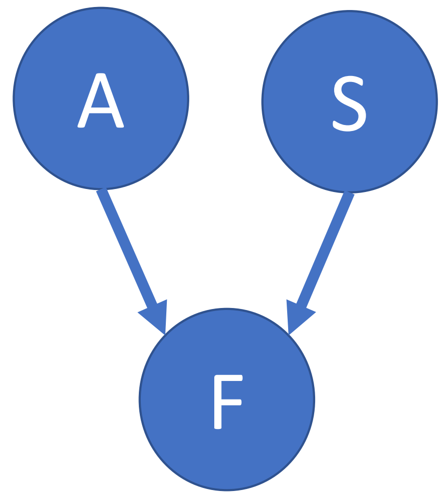
The university dispatched a team of expert vaccavolatologists. They determined that almost all flying cows were explained by one or both of the following causes:

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

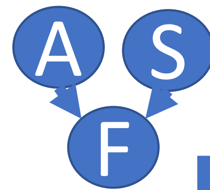
Flying cows

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

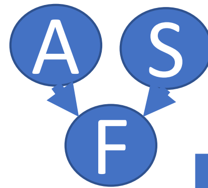


They went out to watch a nearby pasture for ten days.

- 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



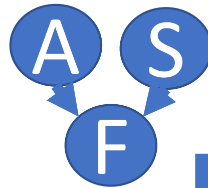
The vaccavolatologists now wish to estimate the parameters of their Bayes net

- 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			

Maximum Likelihood Estimation

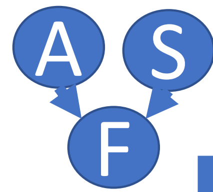


Suppose we have n training examples, $1 \leq i \leq n$, with known values for each of the random variables:

- $a_i = T$ or $a_i = F$
- $s_i = T$ or $s_i = F$
- $f_i = T$ or $f_i = F$

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

Maximum Likelihood Estimation

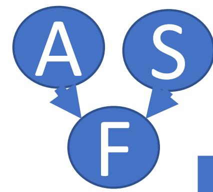


We can estimate model parameters to be the values that maximize the likelihood of the observations, subject to the constraints that

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

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

Maximum Likelihood Estimation



The maximum likelihood parameters are

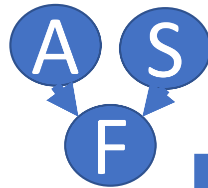
$$P(A = T) = \frac{\text{\# days on which } a_i = T}{\text{\# days total}}$$

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

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

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

Maximum Likelihood Estimation



The maximum likelihood parameters are

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

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

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

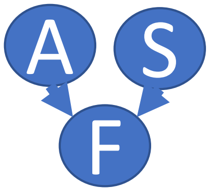
Conclusions: maximum likelihood estimation

- 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.

Parameter Learning for Bayesian Networks

- From observed data: Maximum likelihood
- From observed data: Laplace smoothing
- From partially observed data: Expectation maximization

Laplace smoothing



Laplace smoothing adds an extra count of k to both categories.

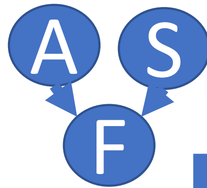
Unlike naïve Bayes, we assume that we know the cardinality of each RV in advance, so the denominator uses $k\times$ the known cardinality (no OOVs).

$$P(A = T) = \frac{(\#\text{ days on which } a_i = T) + k}{(\#\text{ days total})+2\mathrm{k}}$$

$$P(S = T) = \frac{(\#\text{ days on which } s_i = T) + k}{(\#\text{ days total}) + 2k}$$

$$P(F = F|s,a) = \frac{(\#\text{ days (A=}a\text{,S=}s\text{, F=T))}+k}{(\#\text{ days (A=}a\text{,S=}s\text{))} + 2k}$$

Laplace smoothing



Laplace-smoothed parameters:

$$P(A = T) = \frac{3 + k}{10 + 2k},$$

$$P(S = T) = \frac{2 + k}{10 + 2k}$$

a	s	$P(F = T \mid s, a)$
F	F	(1+k)/(6+2k)
F	T	(1+k)/(1+2k)
T	F	(1+k)/(2+2k)
T	T	(1+k)/(1+2k)

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

Conclusions: Laplace smoothing

Just like in naïve Bayes:

- Laplace smoothing makes it possible for things to happen in the test data that never happened in the training data. For example, maximum likelihood resulted in $P(F = F|S = T, A = T) = 0$, but with Laplace smoothing, we smooth that parameter to $P(F = F|S = T, A = T) = \frac{k}{1+2k}$
- This smoothing improves generalization from training data to test data.

Conclusions: Laplace smoothing

Unlike naïve Bayes:

- In Bayesian networks, we assume that we know the cardinality of each random variable in advance, so no extra probability mass is kept aside for OOV events.

$$P(X = x|H = h) = \frac{(\#\ \text{observations of}\ (\text{H}{=}h, \text{X}{=}x)){+}k}{(\#\ \text{observations of}\ (\text{H}{=}h)) + k \cdot (\#\ \text{distinct values of}\ X)}$$

Parameter Learning for Bayesian Networks

- From observed data: Maximum likelihood
- From observed data: Laplace smoothing
- From partially observed data: Expectation maximization

Partially observed data

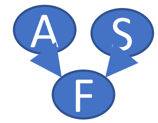
- Maximum likelihood estimation is very easy to use, IF you have training data in which the values of ALL variables are observed.

- …but what if some of the variables can't be observed?
- For example: after the 6th day, 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 $s_i = F$…

Partially observed data

Suppose that we have the following observations:

- We know whether A=True or False.
- We know whether F=True or False.
- After the 6th day, we don't know whether S is True or False (shown as "?").



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

Expectation Maximization (EM): Main idea

Remember that maximum likelihood estimation counts examples:

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

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

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

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

Definition of Expectation

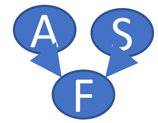
The expected value of a random variable is its weighted average value, with weights equal to the probabilities.

$$E[\#\text{days } A = a, S = s, F = T] = \sum_{i \in \text{Days}} P(A_i = a, S_i = s, F_i = T)$$

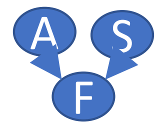
Expectation

$$E[\#\text{days } A = F, S = T, F = T]$$

$$= \sum_{i=1}^{10} P(A_i = F, S_i = T, F_i = T)$$



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



Expectation

$$E[\#\text{days } A = F, S = T, F = T]$$

$$= \sum_{i=1}^{10} \begin{matrix} P(A_i = F, F_i = T)\times \\ P(S_i = T | A_i = F, F_i = T) \end{matrix}$$

$P(A_i = F, F_i = T)$ is either 0 or 1, depending on whether the event certainly occurred (days 2 and 9) or certainly did not occur (every other day).

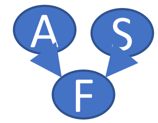
Day	A	S	F
1			
2		T	T
3			
4	T	T	T
5	T		
6			
7	T	?	T
8		?	
9		?	T
10		?	

Expectation

$$E[\#\text{days } A = F, S = T, F = T]$$

$$= \sum_{i=1}^{10} \begin{matrix} P(A_i = F, F_i = T) \times \\ P(S_i = T | A_i = F, F_i = T) \end{matrix}$$

- $P(S_i = T | A_i = F, F_i = T) = 1$ on day 2, because the event certainly occurred.
- $P(S_i = T | A_i = F, F_i = T)$ is unknown on day 9



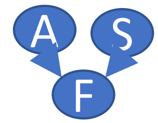
Day	A	S	F
1			
2		T	T
3			
4	T	T	T
5	T		
6			
7	T	?	T
8		?	
9		?	T
10		?	

Expectation

$$E[\#\text{days } A = F, S = T, F = T]$$

$$= 1 + P(S_8 = T | A_8 = F, F_8 = T)$$

- How can we compute $P(S_9 = T | A_9 = F, F_9 = T)$?
- In order to compute it, we need the model parameters
- The model parameters are the thing we're trying to estimate!



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

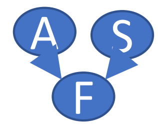
Expectation Maximization (EM) is iterative

INITIALIZE: **guess** the model parameters.

ITERATE until convergence:

1. **E-Step**: $E[\#\ \text{days}\ S = s, A = a, F = f] = \sum_{i:a_i=a, f_i=f} P(S = s|a, f)$
2. **M-Step:** $P(F = f|S = s, A = a) = \frac{E[\#\ \text{days}\ S=s, A=a, F=f]}{E[\#\ \text{days}\ S=s,\ A=a]}$

Continue the iteration, shown above, until the model parameters stop changing.

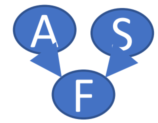


Example: Initialize

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):

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

a	s	$P(F=T\|s,\boldsymbol{a})$
F	F	0
F	T	1/2
T	F	1/2
T	T	1

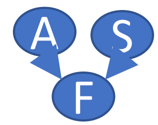


E-Step

Based on Marilyn's model, we calculate $P(S = T|a_i, f_i)$ for each of the missing days, as shown in the table at right.

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

E-Step



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	$\boldsymbol{E[\#\ days\ S = T \| a, f]}$	$\boldsymbol{E[\#\ days\ S = F \| a, f]}$
F	F	$0 + 0 + 0 + \frac{1}{7} + \frac{1}{7} = \frac{2}{7}$	$1 + 1 + 1 + \frac{6}{7} + \frac{6}{7} = \frac{33}{7}$
F	T	$1 + 1 = 2$	0+0=0
T	F	0	1
T	T	$1 + \frac{2}{5} = \frac{7}{5}$	$0 + \frac{3}{5} = \frac{3}{5}$

a	f	$E[\# \; days \; S = T \mid a, f]$	$E[\# \; days \; S = F \mid a, f]$
F	F	$0+0+0+\frac{1}{7}+\frac{1}{7}=\frac{2}{7}$	$1+1+1+\frac{6}{7}+\frac{6}{7}=\frac{33}{7}$
F	T	$1+1=2$	0+0=0
T	F	0	1
T	T	$1+\frac{2}{5}=\frac{7}{5}$	$0+\frac{3}{5}=\frac{3}{5}$

M-Step

Now let's re-estimate the model parameters. For example,

$$P(F=T \mid S=F, A=F) = \frac{E[\# \text{ days } S=F, A=F, F=T]}{E[\# \text{ days } S=F, \quad A=F]}$$

$$= \frac{0}{\frac{33}{7}+0} = 0$$

a	f	$E[\# \; days \; S \mid a, f]$	$E[\# \; days \; \neg S \mid a, f]$
F	F	$0+0+0+\frac{1}{7}+\frac{1}{7}=\frac{2}{7}$	$1+1+1+\frac{6}{7}+\frac{6}{7}=\frac{33}{7}$
F	T	$1+1=2$	0+0=0
T	F	0	1
T	T	$1+\frac{2}{5}=\frac{7}{5}$	$0+\frac{3}{5}=\frac{3}{5}$

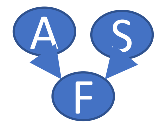
M-Step

Now let's re-estimate the model parameters. For example,

$$P(F=T \mid S=T, A=F) = \frac{E[\# \text{ days } S=T, A=F, F=T]}{E[\# \text{ days } S=T, \quad A=F]}$$

$$= \frac{2}{\frac{2}{7}+2} = \frac{7}{8}$$

M-Step



The re-estimated probabilities are

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

$$P(S=T)=\frac{E[\#\text{days } S=T]}{\#\text{ days total}}=\frac{\frac{2}{7}+2+0+\frac{7}{5}}{10}$$
$$=\frac{94}{350}$$

a	s	$P(F \mid S=s, A=a)$
F	F	$\frac{0}{\frac{33}{7}+0}=0$
F	T	$\frac{2}{\frac{2}{7}+2}=\frac{7}{8}$
T	F	$\frac{3/5}{1+\frac{3}{5}}=\frac{3}{8}$
T	T	$\frac{7/5}{0+7/5}=1$

Expectation Maximization (EM): review

INITIALIZE: **guess** the model parameters.

ITERATE until convergence:

1. **E-Step**: $E[\#\ \text{days}\ S = s, A = a, F = f] = \sum_{i:a_i=a, f_i=f} P(S = s|a, f)$
2. **M-Step:** $P(F = f|S = s, A = a) = \frac{E[\#\ \text{days}\ S=s, A=a, F=f]}{E[\#\ \text{days}\ S=s,\ A=a]}$

Continue the iteration, shown above, until the model parameters 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 $P(F = T|S = F, A = F) = \mathbf{0}$. 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.

Parameter Learning for Bayesian Networks

- Maximum Likelihood (ML):

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

- Laplace Smoothing:

$$P(F = T|S = s, A = a) = \frac{\text{\# days (A=}a\text{, S=}s\text{, F=T)} + k}{\text{\# days (A=}a\text{, S=}s\text{)} + 2k}$$

- Expectation Maximization (EM):

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