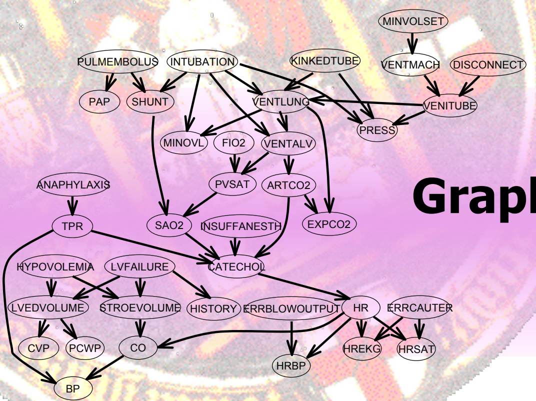
Based on J. A. Bilmes,"A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models", TR-97-021, U.C. Berkeley, April 1998; G. J. McLachlan, T. Krishnan, „The EM Algorithm and Extensions", John Wiley & Sons, Inc., 1997; D. Koller, course CS-228 handouts, Stanford University, 2001., N. Friedman & D. Koller's NIPS'99.

Advanced I WS 06/07



Graphical Models - Learning -

Parameter Estimation

Wolfram Burgard, Luc De Raedt, Kristian Kersting, Bernhard Nebel

Albert-Ludwigs University Freiburg, Germany

Advanced I WS 06/07

Outline

- Introduction
- Reminder: Probability theory
- Basics of Bayesian Networks
- Modeling Bayesian networks
- Inference (VE, Junction tree)
- [Excourse: Markov Networks]
- Learning Bayesian networks
- Relational Models

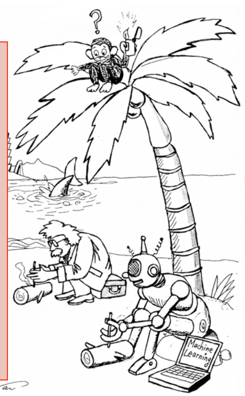
Advanced I WS 06/07

What is Learning?

- Agents are said to learn if they improve their performance over time based on experience.

The problem of understanding intelligence is said to be the greatest problem in science today and "the" problem for this century – as deciphering the genetic code was for the second half of the last one…the problem of learning represents a gateway to understanding intelligence in man and machines.

-- Tomasso Poggio and Steven Smale, 2003.



Advanced I WS 06/07

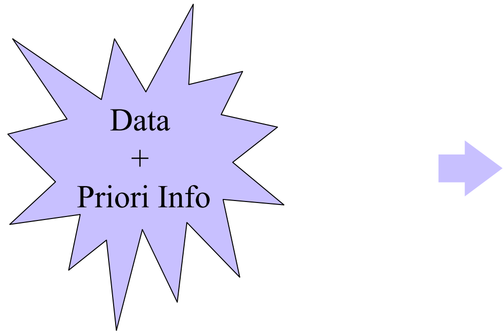
Why bothering with learning?

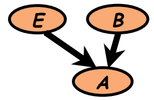
- **Bottleneck of knowledge aquisition**
  - Expensive, difficult
  - Normally, no expert is around
- **Data is cheap !**
  - Huge amount of data avaible, e.g.
    - Clinical tests
    - Web mining, e.g. log files
    - ….

Why Learning Bayesian Networks?

- Conditional independencies & graphical language capture structure of many real-world distributions
- Graph structure provides much insight into domain
  - Allows "knowledge discovery"
- Learned model can be used for many tasks
- Supports all the features of probabilistic learning
  - Model selection criteria
  - Dealing with missing data & hidden variables

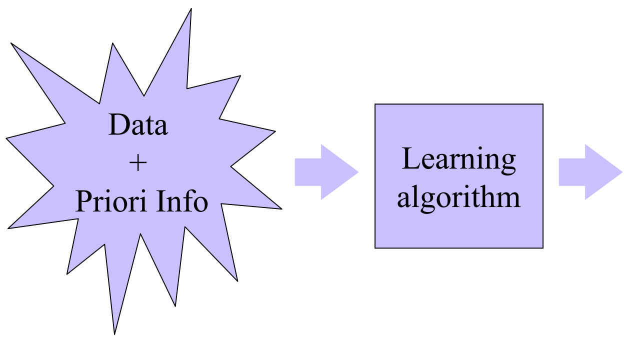
Learning With Bayesian Networks

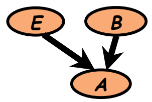




E	B	P(A | E,B)	
e	b	.9	.1
e	$\bar{b}$	.7	.3
$\bar{e}$	b	.8	.2
$\bar{e}$	$\bar{b}$	.99	.01

Learning With Bayesian Networks





E	B	P(A | E,B)	
e	b	.9	.1
e	$\bar{b}$	.7	.3
$\bar{e}$	b	.8	.2
$\bar{e}$	$\bar{b}$	.99	.01

What does the data look like?

attributes/variables

complete data set

A1	A2	A3	A4	A5	A6	
true	true	false	true	false	false	X1
false	true	true	true	false	false	X2
...	...	...	...	...	...	⋮
true	false	false	false	true	true	XM

data cases

What does the data look like?

incomplete data set

A1	A2	A3	A4	A5	A6
true	true	?	true	false	false
?	true	?	?	false	false
...	...	...	...	...	...
true	false	?	false	true	?

Real-world data: states of some random variables are missing
- E.g. medical diagnose: not all patient are subjects to all test
- Parameter reduction, e.g. clustering, ...

What does the data look like?

incomplete data set

A1	A2	A3	A4	A5	A6
true	true	?	true	false	false
?	true	?	?	false	false
...	...	...	...	...	...
true	false	?	false	true	?

missing value

Real-world data: states of some random variables are missing
- E.g. medical diagnose: not all patient are subjects to all test
- Parameter reduction, e.g. clustering, ...

What does the data look like?

hidden/ latent

incomplete data set

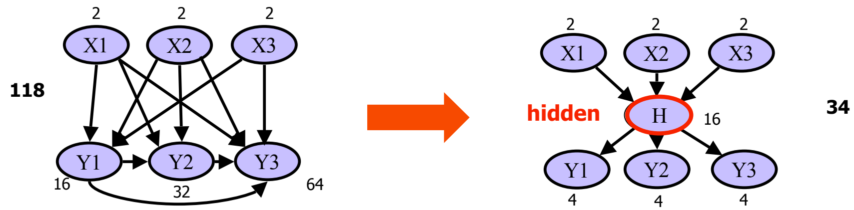
A1	A2	A3	A4	A5	A6
true	true	?	true	false	false
?	true	?	?	false	false
...	...	...	...	...	...
true	false	?	false	true	?

missing value

Real-world data: states of some random variables are missing
- E.g. medical diagnose: not all patient are subjects to all test
- Parameter reduction, e.g. clustering, ...

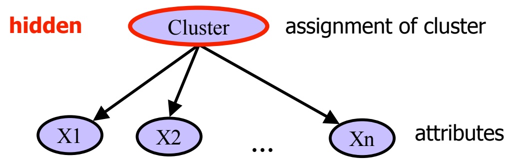
Hidden variable – Examples

1. Parameter reduction:

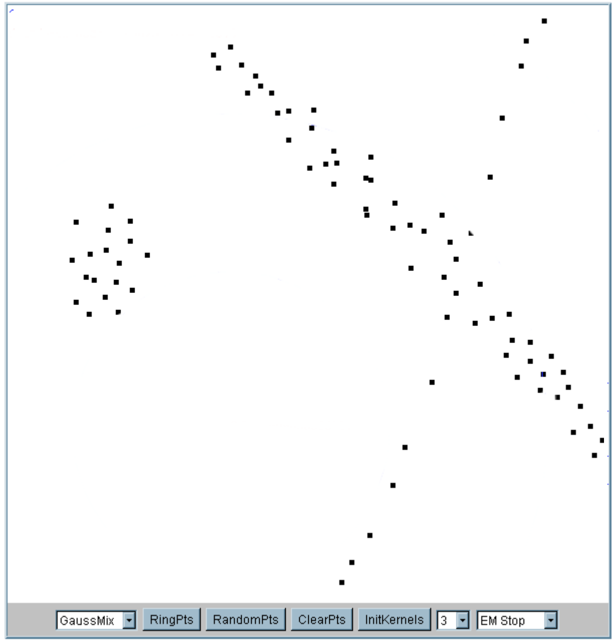


Hidden variable – Examples

2. Clustering:

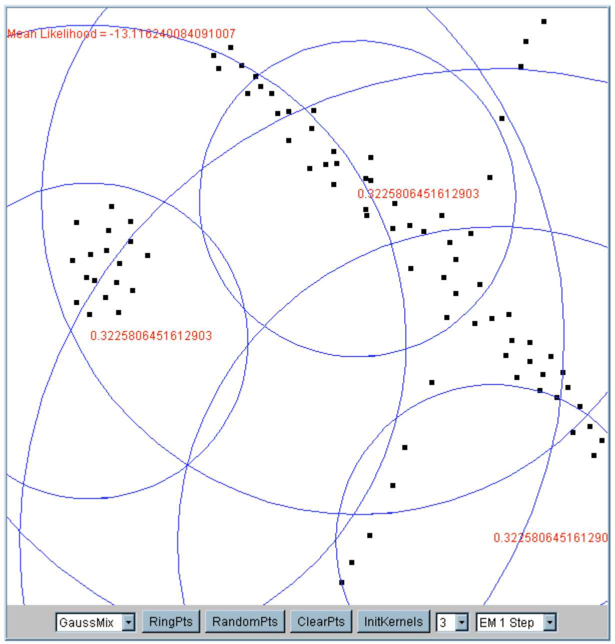


- Hidden variables also appear in **clustering**
- **Autoclass** model:
  - Hidden variables assigns class labels
  - Observed attributes are independent given the class

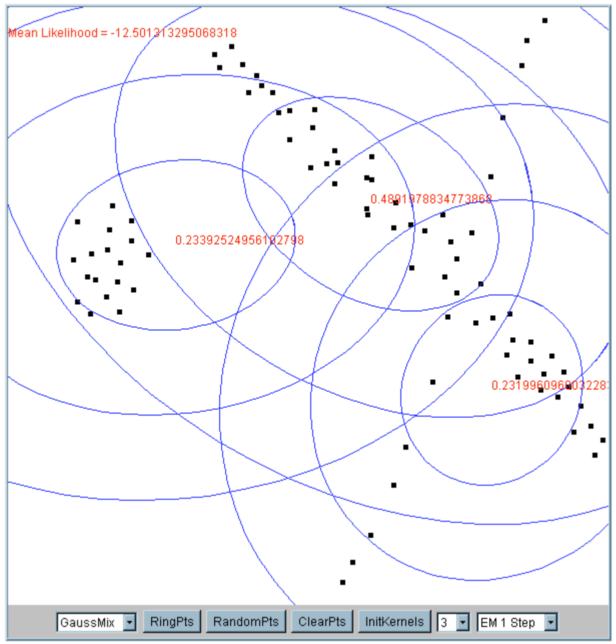


Iteration 1

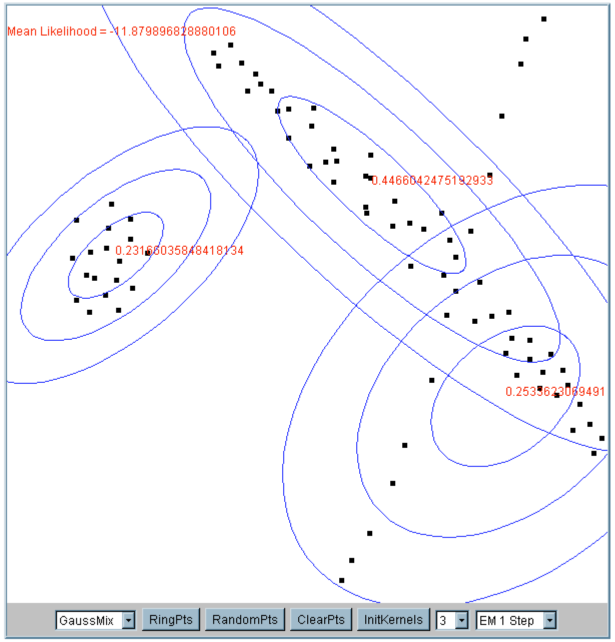
The cluster means are randomly assigned



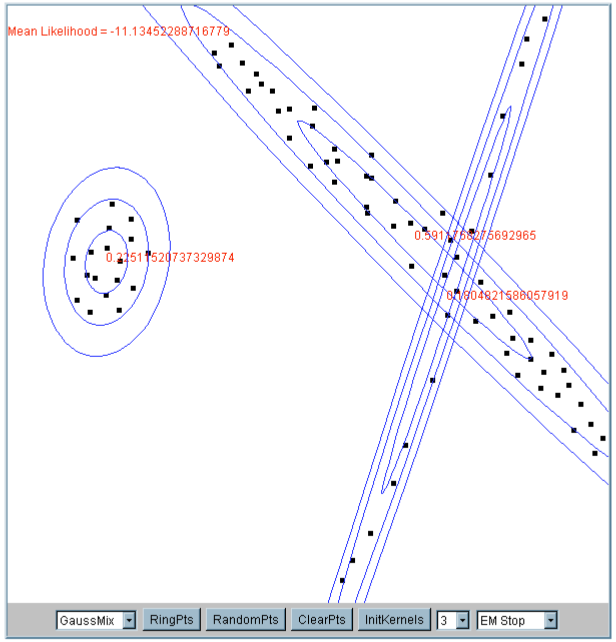
Iteration 2



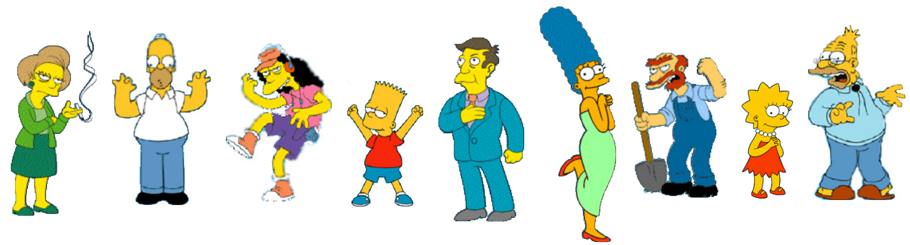
Iteration 5



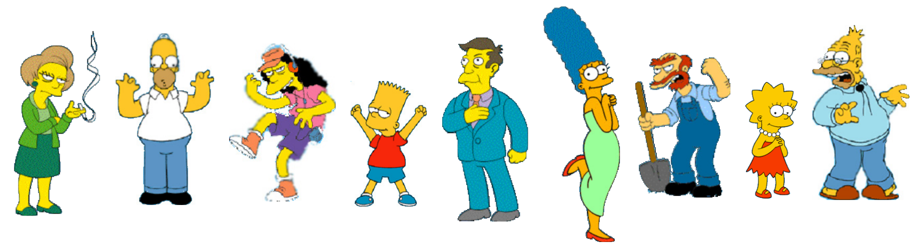
Iteration 25



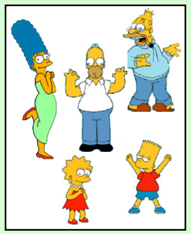
What is a natural grouping among these objects?



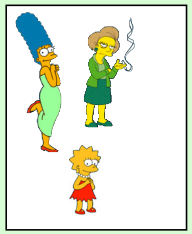
What is a natural grouping among these objects?



Clustering is subjective


Simpson's Family


School Employees


Females


Males

Learning With Bayesian Networks

		Fixed structure $A \rightarrow B$	Fixed variables A ? B	Hidden variables A ? B ? H
observed	fully	Easiest problem counting	Selection of arcs New domain with no domain expert Data mining	
observed	Partially	Numerical, nonlinear optimization, Multiple calls to BNs, Difficult for large networks	Encompasses to difficult subproblem, „Only" Structural EM is known	Scientific discouvery

Parameter Estimation

- Let $\mathcal{X} = \{X_1, X_2, \ldots, X_n\}$ be set of data over m RVs
- $X_i \in \mathcal{X}$ is called a *data case*
- **iid** - assumption:
  - All data cases are **i**ndependently sampled from **i**dentical **d**istributions

Find:

Parameters Θ of CPDs which match the data best

Maximum Likelihood - Parameter Estimation

- What does „best matching" mean ?

Maximum Likelihood - Parameter Estimation

- What does „best matching" mean ?

Find paramteres Θ which have most likely produced the data

Maximum Likelihood - Parameter Estimation

- What does „best matching" mean ?
  1. MAP parameters $\Theta^* = \arg\max_\Theta P(\Theta|\mathcal{X})$

$$= \arg\max_\Theta P(\mathcal{X}|\Theta) \cdot \frac{P(\Theta)}{P(\mathcal{X})}$$

Maximum Likelihood - Parameter Estimation

- What does „best matching" mean ?
  1. MAP parameters $\Theta^* = \arg\max_\Theta P(\Theta|\mathcal{X})$

$$= \arg\max_\Theta P(\mathcal{X}|\Theta) \cdot \frac{P(\Theta)}{\cancel{P(\mathcal{X})}}$$

  2. Data is equally likely for all parameters.

Maximum Likelihood - Parameter Estimation

- What does „best matching" mean ?
  1. MAP parameters $\Theta^* = \arg\max_\Theta P(\Theta|\mathcal{X})$

$$= \arg\max_\Theta P(\mathcal{X}|\Theta) \cdot \frac{\cancel{P(\Theta)}}{\cancel{P(\mathcal{X})}}$$

  2. Data is equally likely for all parameters
  3. All parameters are apriori equally likely

Maximum Likelihood - Parameter Estimation

- What does „best matching" mean ?

Find:

ML parameters

$$\Theta^* = \arg\max_\Theta P(\mathcal{X}|\Theta)$$

Maximum Likelihood - Parameter Estimation

- What does „best matching" mean ?

Find:

ML parameters

$$\Theta^* = \arg\max_\Theta P(\mathcal{X}|\Theta)$$

Likelihood $\mathcal{L}(\Theta|\mathcal{X})$ of the paramteres given the data

Advanced I WS 06/07

Maximum Likelihood - Parameter Estimation

- What does „best matching" mean ?

Find:

ML parameters

$$\Theta^* = \arg\max_\Theta P(\mathcal{X}|\Theta)$$

Likelihood $\mathcal{L}(\Theta|\mathcal{X})$ of the paramteres given the data

$$\Theta^* = \arg\max_\Theta \log P(\mathcal{X}|\Theta)$$

Log-Likelihood $\mathcal{LL}(\Theta|\mathcal{X})$ of the paramteres given the data

Bayesian Networks - Learning

Advanced I WS 06/07

Maximum Likelihood

- This is one of the **most commonly used** estimators in statistics
- **Intuitively appealing**
- **Consistent:** estimate converges to best possible value as the number of examples grow
- **Asymptotic efficiency:** estimate is as close to the true value as possible given a particular training set

Bayesian Networks - Learning

Advanced I WS 06/07

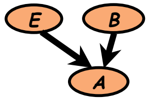
Learning With Bayesian Networks

		Fixed structure $A \rightarrow B$	**Fixed variables** A ? B	**Hidden variables** A ? B ? H
observed	**fully**	Easiest problem counting **?**	Selection of arcs New domain with no domain expert Data mining	
	Partially	Numerical, nonlinear optimization, Multiple calls to BNs, Difficult for large networks	Encompasses to difficult subproblem, „Only" Structural EM is known	Scientific discouvery

Bayesian Networks - Learning

Known Structure, Complete Data

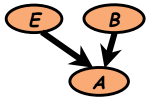
E, B, A
<Y,N,N>
<Y,N,Y>
<N,N,Y>
<N,Y,Y>
.
.
<N,Y,Y>



E	*B*	*P(A | E,B)*	
e	b	?	?
e	$\bar{b}$	?	?
$\bar{e}$	b	?	?
$\bar{e}$	$\bar{b}$	?	?

Learning algorithm

- Network structure is specified
  - Learner needs to estimate parameters
- Data does not contain missing values



E	*B*	*P(A | E,B)*	
e	b	.9	.1
e	$\bar{b}$	.7	.3
$\bar{e}$	b	.8	.2
$\bar{e}$	$\bar{b}$	.99	.01

ML Parameter Estimation

$$\mathcal{LL}(\Theta|\mathcal{X}) = \log P(X_1, X_2, \ldots, X_n|\Theta)$$

A1	A2	A3	A4	A5	A6
true	true	false	true	false	false
false	true	true	true	false	false
...	...	...	...	...	...
true	false	false	false	true	true

ML Parameter Estimation

$$\mathcal{LL}(\Theta|\mathcal{X}) = \log P(X_1, X_2, \ldots, X_n|\Theta)$$

(iid) $$= \log \prod_{i=1}^{n} P(X_i|\Theta)$$

A1	A2	A3	A4	A5	A6
true	true	false	true	false	false
false	true	true	true	false	false
...	...	...	...	...	...
true	false	false	false	true	true

ML Parameter Estimation

$$\mathcal{LL}(\Theta|\mathcal{X}) = \log P(X_1, X_2, \ldots, X_n|\Theta)$$

(iid) $$= \log \prod_{i=1}^{n} P(X_i|\Theta)$$

$$= \sum_{i=1}^{n} \log P(X_i|\Theta) = \sum_{i=1}^{n} \log P(x_i^1, x_i^2, \ldots, x_i^m|\Theta)$$

$$\log \prod = \sum \log$$

A1	A2	A3	A4	A5	A6
true	true	false	true	false	false
false	true	true	true	false	false
...	...	...	...	...	...
true	false	false	false	true	true

ML Parameter Estimation

A1	A2	A3	A4	A5	A6
true	true	false	true	false	false
false	true	true	true	false	false
...	...	...	...	...	...
true	false	false	false	true	true

$$\mathcal{LL}(\Theta|\mathcal{X}) = \log P(X_1, X_2, \ldots, X_n|\Theta)$$

(iid) $= \log \prod_{i=1}^{n} P(X_i|\Theta)$

$\log \prod = \sum \log$

$$= \sum_{i=1}^{n} \log P(X_i|\Theta) = \sum_{i=1}^{n} \log P(x_i^1, x_i^2, \ldots, x_i^m|\Theta)$$

$$= \sum_{i=1}^{n} \log \left(\prod_{j=1}^{m} P(x_i^j \,|\, \mathrm{pa}(x_i^j), \Theta)\right)$$ **(BN semantics)**

Advanced I WS 06/07

ML Parameter Estimation

A1	A2	A3	A4	A5	A6
true	true	false	true	false	false
false	true	true	true	false	false
...	...	...	...	...	...
true	false	false	false	true	true

$$\mathcal{LL}(\Theta|\mathcal{X}) = \log P(X_1, X_2, \ldots, X_n|\Theta)$$

(iid) $= \log \prod_{i=1}^{n} P(X_i|\Theta)$

$\log \prod = \sum \log$

$$= \sum_{i=1}^{n} \log P(X_i|\Theta) = \sum_{i=1}^{n} \log P(x_i^1, x_i^2, \ldots, x_i^m|\Theta)$$

$$= \sum_{i=1}^{n} \log \left(\prod_{j=1}^{m} P(x_i^j \,|\, \mathrm{pa}(x_i^j), \Theta)\right)$$ **(BN semantics)**

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} \log P(x_i^j \,|\, \mathrm{pa}(x_i^j), \Theta)$$

Bayesian Networks - Learning

Advanced I WS 06/07

ML Parameter Estimation

A1	A2	A3	A4	A5	A6
true	true	false	true	false	false
false	true	true	true	false	false
...	...	...	...	...	...
true	false	false	false	true	true

$$\mathcal{LL}(\Theta|\mathcal{X}) = \log P(X_1, X_2, \ldots, X_n|\Theta)$$

(iid) $= \log \prod_{i=1}^{n} P(X_i|\Theta)$

$\log \prod = \sum \log$

$$= \sum_{i=1}^{n} \log P(X_i|\Theta) = \sum_{i=1}^{n} \log P(x_i^1, x_i^2, \ldots, x_i^m|\Theta)$$

$$= \sum_{i=1}^{n} \log \left(\prod_{j=1}^{m} P(x_i^j \,|\, \mathrm{pa}(x_i^j), \Theta)\right)$$ **(BN semantics)**

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} \log P(x_i^j \,|\, \mathrm{pa}(x_i^j), \Theta)$$

$$= \sum_{j=1}^{m} \sum_{i=1}^{n} \log P(x_i^j \,|\, \mathrm{pa}(x_i^j), \Theta_j)$$

Bayesian Networks - Learning

Advanced I WS 06/07

ML Parameter Estimation

A1	A2	A3	A4	A5	A6
true	true	false	true	false	false
false	true	true	true	false	false
...	...	...	...	...	...
true	false	false	false	true	true

$$\mathcal{LL}(\Theta|\mathcal{X}) = \log P(X_1, X_2, \ldots, X_n|\Theta)$$

(iid) $= \log \prod_{i=1}^{n} P(X_i|\Theta)$

$\log \prod = \sum \log$

$$= \sum_{i=1}^{n} \log P(X_i|\Theta) = \sum_{i=1}^{n} \log P(x_i^1, x_i^2, \ldots, x_i^m|\Theta)$$

$$= \sum_{i=1}^{n} \log \left(\prod_{j=1}^{m} P(x_i^j \,|\, \mathrm{pa}(x_i^j), \Theta)\right)$$ **(BN semantics)**

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} \log P(x_i^j \,|\, \mathrm{pa}(x_i^j), \Theta)$$

$$= \sum_{j=1}^{m} \sum_{i=1}^{n} \log P(x_i^j \,|\, \mathrm{pa}(x_i^j), \boxed{\Theta_j})$$

Only local parameters of family of Aj involved

Bayesian Networks - Learning

ML Parameter Estimation

A1	A2	A3	A4	A5	A6
true	true	false	true	false	false
false	true	true	true	false	false
...	...	...	...	...	...
true	false	false	false	true	true

$$\mathcal{LL}(\Theta|\mathcal{X}) = \log P(X_1, X_2, \ldots, X_n|\Theta)$$

(iid) $$= \log \prod_{i=1}^{n} P(X_i|\Theta)$$

$\log \prod$ $= \sum \log$ $$= \sum_{i=1}^{n} \log P(X_i|\Theta) = \sum_{i=1}^{n} \log P(x_i^1, x_i^2, \ldots, x_i^m|\Theta)$$

$$= \sum_{i=1}^{n} \log \left(\prod_{j=1}^{m} P(x_i^j \,|\, \mathrm{pa}(x_i^j), \Theta) \right)$$ **(BN semantics)**

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} \log P(x_i^j \,|\, \mathrm{pa}(x_i^j), \Theta)$$

$$= \sum_{j=1}^{m} \sum_{i=1}^{n} \log P(x_i^j \,|\, \mathrm{pa}(x_i^j), \boxed{\Theta_j})$$

Only local parameters of family of Aj involved

$$= \sum_{j=1}^{m} \mathcal{LL}(\Theta_j|\mathcal{X})$$

Each factor individually !!

Advanced I WS 06/07

ML Parameter Estimation

A1	A2	A3	A4	A5	A6
true	true	false	true	false	false
false	true	true	true	false	false
...	...	...	...	...	...
true	false	false	false	true	true

$$\mathcal{LL}(\Theta|\mathcal{X}) = \log P(X_1, X_2, \ldots, X_n|\Theta)$$

(iid) $$= \log \prod_{i=1}^{n} P(X_i|\Theta)$$

$\log \prod$ $= \sum \log$ $$= \sum_{i=1}^{n} \log P(X_i|\Theta) = \sum_{i=1}^{n} \log P(x_i^1, x_i^2, \ldots, x_i^m|\Theta)$$

Decomposability of the likelihood

$$= \sum_{j=1}^{m} \sum_{i=1}^{n} \log P(x_i^j \,|\, \mathrm{pa}(x_i^j), \boxed{\Theta_j})$$

parameters of family of Aj involved

$$= \sum_{j=1}^{m} \mathcal{LL}(\Theta_j|\mathcal{X})$$

Each factor individually !!

Bayesian Networks - Learning

Decomposability of Likelihood

If the data set if **complete** (no question marks)

- we can maximize each local likelihood function **independently**, and
- then **combine** the solutions to get an MLE solution.

decomposition of the **global problem** to **independent, local sub-problems.** This allows efficient solutions to the MLE problem.

Likelihood for Multinominals

- Random variable V with 1,...,K values

$$P(V=k) = \theta_k \qquad \sum_{k=1}^{K} \theta_k = 1$$

This constraint implies that the choice on θ_I influences the choice on θ_j (i<>j)

- $$\mathcal{LL}(\Theta_v|\mathcal{X}) = \sum_{k=1}^{K} \log \theta_k^{N_k} = \sum_{k=1}^{K} N_k \cdot \log \theta_k$$

where Nk is the counts of state k in data

Likelihood for Binominals (2 states only)

- **Compute partial derivative**

$$\frac{\partial}{\partial \theta_i} \mathcal{LL}(\Theta_v|\mathcal{X}) = \frac{\partial}{\partial \theta_i}(N_1 \log \theta_1 + N_2 \log(1-\theta_1))$$
$$= \frac{N_1}{\theta_1} + \frac{N_2}{1-\theta_1}$$

$\theta_1 + \theta_2 = 1$

- **Set partial derivative zero**

$$\frac{\partial}{\partial \theta_i} \mathcal{LL}(\Theta_v|\mathcal{X}) = 0 \Leftrightarrow \frac{N_1}{\theta_1} + \frac{N_2}{1-\theta_1} = 0$$

=> MLE is $\theta_1^* = \frac{N_1}{N_1 + N_2}$

Likelihood for Binominals (2 states only)

- **Compute partial derivative**

$$\frac{\partial}{\partial \theta_i} \mathcal{LL}(\Theta_v|\mathcal{X}) = \frac{\partial}{\partial \theta_i}(N_1 \log \theta_1 + N_2 \log(1-\theta_1))$$
$$= \frac{N_1}{\theta_1} + \frac{N_2}{1-\theta_1}$$

$\theta_1 + \theta_2 = 1$

- **Set partial derivative zero**

In general, for multinomials (>2 states), the MLE is

$$\theta_i^* = \frac{N_i}{\sum_j N_j}$$

Likelihood for Conditional Multinominals

- $P(V = k|\,\mathrm{pa}(V) = \mathrm{pa})$ multinomial for each joint state pa of the parents of V:

$$P(k|1,1), P(k|1,2), P(k|2,1), P(k|2,2)$$

- $\mathcal{LL}(\Theta_v|\mathcal{X})$

$$= \sum_{\mathrm{pa}} \sum_{k=1}^{K} \log \theta_{k|\mathrm{pa}}^{N_{k,\mathrm{pa}}} = \sum_{\mathrm{pa}} \sum_{k=1}^{K} N_{k,\mathrm{pa}} \cdot \theta_{k|\mathrm{pa}}$$

- MLE

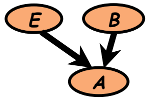
$$\theta_{k|\mathrm{pa}}^* = \frac{N_{k,\mathrm{pa}}}{N_{\mathrm{pa}}}$$

Learning With Bayesian Networks

		Fixed structure A → B	**Fixed variables** A ? B	**Hidden variables** A ? B ? H
observed	fully	Easiest problem counting	Selection of arcs New domain with no domain expert Data mining	
	Partially	Numerical, nonlinear optimization, Multiple calls to BNs, Difficult for large networks ?	Encompasses to difficult subproblem, „Only" Structural EM is known	Scientific discouvery

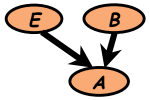
Known Structure, Incomplete Data

E, B, A
<Y,?,N>
<Y,N,?>
<N,N,Y>
<N,Y,Y>
.
.
<?,Y,Y>



E	B	P(A | E,B)	
e	b	?	?
e	$\bar{b}$	?	?
$\bar{e}$	b	?	?
$\bar{e}$	$\bar{b}$	?	?

→ Learning algorithm →



E	B	P(A | E,B)	
e	b	.9	.1
e	$\bar{b}$	.7	.3
$\bar{e}$	b	.8	.2
$\bar{e}$	$\bar{b}$	.99	.01

- Network structure is specified
- Data contains missing values
  - Need to consider assignments to missing values

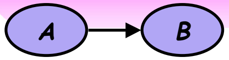
EM Idea

- In the case of complete data, ML parameter estimation is easy:
  - **simply counting** (1 iteration)

Incomplete data ?

1. **Complete data** (Imputation)
   - most probable?, average?, ... value
2. **Count**
3. **Iterate**

EM Idea: complete the data

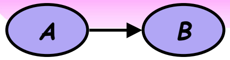


$\theta_{A=\texttt{true}} = \frac{1}{2}$
$\theta_{B=\texttt{true}|A=\texttt{true}} = \frac{1}{2}$
$\theta_{B=\texttt{true}|A=\texttt{false}} = \frac{1}{2}$

incomplete data

A	B
true	true
true	?
false	true
true	false
false	?

EM Idea: complete the data



$\theta_{A=\texttt{true}} = \frac{1}{2}$
$\theta_{B=\texttt{true}|A=\texttt{true}} = \frac{1}{2}$
$\theta_{B=\texttt{true}|A=\texttt{false}} = \frac{1}{2}$

incomplete data

A	B
true	true
true	?
false	true
true	false
false	?

complete

$P(B=\texttt{true}|A=\texttt{true}) = 0.5$
$P(B=\texttt{true}|A=\texttt{false}) = 0.5$

complete data

expected counts

A	B	N
true	true	1.5
true	false	1.5
false	true	1.5
false	false	0.5

EM Idea: complete the data

$A \rightarrow B$

$\theta_{A=\texttt{true}} = \frac{1}{2}$
$\theta_{B=\texttt{true}|A=\texttt{true}} = \frac{1}{2}$
$\theta_{B=\texttt{true}|A=\texttt{false}} = \frac{1}{2}$

incomplete data

A	B
true	true
true	?
false	true
true	false
false	?

complete

$$P(B=\texttt{true}|A=\texttt{true}) = 0.5$$
$$P(B=\texttt{true}|A=\texttt{false}) = 0.5$$

complete data

expected counts

A	B	N
true	true	1.5
true	false	1.5
false	true	1.5
false	false	0.5

A	B	N
true	true	1.0
true	?=true	0.5
true	?=false	0.5
false	true	1.0
true	false	1.0
false	?=true	0.5
false	?=false	0.5

EM Idea: complete the data

$A \rightarrow B$

$\theta_{A=\texttt{true}} = \frac{1}{2}$
$\theta_{B=\texttt{true}|A=\texttt{true}} = \frac{1}{2}$
$\theta_{B=\texttt{true}|A=\texttt{false}} = \frac{1}{2}$

incomplete data

A	B
true	true
true	?
false	true
true	false
false	?

complete

$$P(B=\texttt{true}|A=\texttt{true}) = 0.5$$
$$P(B=\texttt{true}|A=\texttt{false}) = 0.5$$

complete data

expected counts

A	B	N
true	true	1.5
true	false	1.5
false	true	1.5
false	false	0.5

maximize

$$\theta_{A=\texttt{true}} = \frac{1.5+1.5}{1.5+1.5+1.5+0.5} = 0.6$$
$$\theta_{B=\texttt{true}|A=\texttt{true}} = \frac{1.5}{1.5+1.5} = 0.5$$
$$\theta_{B=\texttt{true}|A=\texttt{false}} = \frac{1.5}{1.5+0.5} = 0.75$$

EM Idea: complete the data

$A \rightarrow B$

$\theta_{A=\texttt{true}} = \frac{1}{2}$
$\theta_{B=\texttt{true}|A=\texttt{true}} = \frac{1}{2}$
$\theta_{B=\texttt{true}|A=\texttt{false}} = \frac{1}{2}$

incomplete data

A	B
true	true
true	?
false	true
true	false
false	?

complete

$$P(B=\texttt{true}|A=\texttt{true}) = 0.5$$
$$P(B=\texttt{true}|A=\texttt{false}) = 0.5$$

iterate

complete data

expected counts

A	B	N
true	true	1.5
true	false	1.5
false	true	1.5
false	false	0.5

maximize

$$\theta_{A=\texttt{true}} = \frac{1.5+1.5}{1.5+1.5+1.5+0.5} = 0.6$$
$$\theta_{B=\texttt{true}|A=\texttt{true}} = \frac{1.5}{1.5+1.5} = 0.5$$
$$\theta_{B=\texttt{true}|A=\texttt{false}} = \frac{1.5}{1.5+0.5} = 0.75$$

EM Idea: complete the data

$A \rightarrow B$

$\theta_{A=\texttt{true}} = 0.6$
$\theta_{B=\texttt{true}|A=\texttt{true}} = 0.5$
$\theta_{B=\texttt{true}|A=\texttt{false}} = 0.75$

incomplete data

A	B
true	true
true	?
false	true
true	false
false	?

EM Idea: complete the data

A → B

$\theta_{A=\mathtt{true}} = 0.6$
$\theta_{B=\mathtt{true}|A=\mathtt{true}} = 0.5$
$\theta_{B=\mathtt{true}|A=\mathtt{false}} = 0.75$

incomplete data

A	B
true	true
true	?
false	true
true	false
false	?

complete

$P(B=\mathtt{true}|A=\mathtt{true}) = 0.5$
$P(B=\mathtt{true}|A=\mathtt{false}) = 0.75$

complete data

expected counts

A	B	N
true	true	1.5
true	false	1.5
false	true	1.75
false	false	0.25

maximize

$$\theta_{A=\mathtt{true}} = \frac{1.5+1.5}{1.5+1.5+1.75+0.25} = 0.6$$
$$\theta_{B=\mathtt{true}|A=\mathtt{true}} = \frac{1.5}{1.5+1.5} = 0.5$$
$$\theta_{B=\mathtt{true}|A=\mathtt{false}} = \frac{1.75}{1.75+0.25} = 0.875$$

iterate

EM Idea: complete the data

A → B

$\theta_{A=\mathtt{true}} = 0.6$
$\theta_{B=\mathtt{true}|A=\mathtt{true}} = 0.5$
$\theta_{B=\mathtt{true}|A=\mathtt{false}} = 0.875$

incomplete data

A	B
true	true
true	?
false	true
true	false
false	?

EM Idea: complete the data

A → B

$\theta_{A=\mathtt{true}} = 0.6$
$\theta_{B=\mathtt{true}|A=\mathtt{true}} = 0.5$
$\theta_{B=\mathtt{true}|A=\mathtt{false}} = 0.875$

incomplete data

A	B
true	true
true	?
false	true
true	false
false	?

complete

$P(B=\mathtt{true}|A=\mathtt{true}) = 0.5$
$P(B=\mathtt{true}|A=\mathtt{false}) = 0.875$

complete data

expected counts

A	B	N
true	true	1.5
true	false	1.5
false	true	1.875
false	false	0.125

maximize

$$\theta_{A=\mathtt{true}} = \frac{1.5+1.5}{1.5+1.5+1.875+0.125} = 0.6$$
$$\theta_{B=\mathtt{true}|A=\mathtt{true}} = \frac{1.5}{1.5+1.5} = 0.5$$
$$\theta_{B=\mathtt{true}|A=\mathtt{false}} = \frac{1.875}{1.875+0.125} = 0.9375$$

iterate

Complete-data likelihood

incomplete-data likelihood

$$\Theta^* = \arg\max_\Theta \mathcal{L}(\Theta|\mathcal{X})$$

A1	A2	A3	A4	A5	A6
true	true	?	true	false	false
?	true	?	?	false	false
...	...	...	...	...	...
true	false	?	false	true	?

Assume complete data $\mathcal{Z} = (\mathcal{X}, \mathcal{Y})$ exists with

$$P(\mathcal{Z}|\Theta) = P(\mathcal{X},\mathcal{Y}|\Theta) = P(\mathcal{Y}|\mathcal{X},\Theta) \cdot P(\mathcal{X}|\Theta)$$

complete-data likelihood

$$\mathcal{L}(\Theta|\mathcal{Z}) = \mathcal{L}(\Theta|\mathcal{X},\mathcal{Y}) = P(\mathcal{X},\mathcal{Y}|\Theta)$$
$$\mathcal{LL}(\Theta|\mathcal{Z}) = \mathcal{LL}(\Theta|\mathcal{X},\mathcal{Y}) = \log P(\mathcal{X},\mathcal{Y}|\Theta)$$

EM Algorithm - Abstract

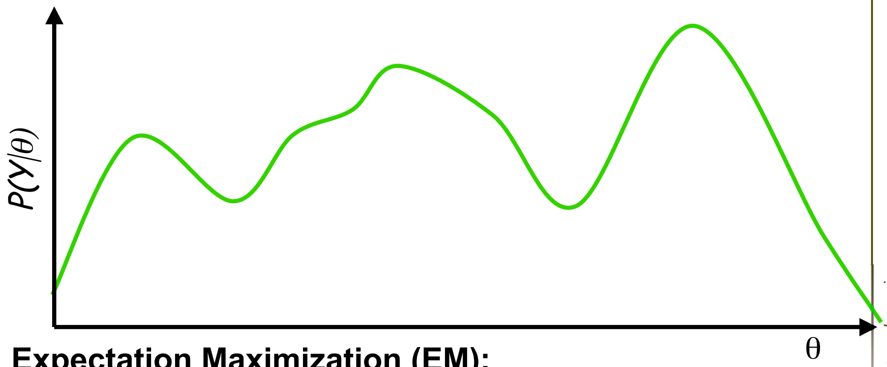
Expectation Step

$$\mathcal{Q}(\Theta, \Theta^{i-1}) = E\left[\mathcal{L}(\mathcal{Z}|\Theta)|\mathcal{X}, \Theta^{i-1}\right]$$

Maximization Step

$$\Theta^{i} = \arg\max_{\Theta} \mathcal{Q}(\Theta, \Theta^{i-1})$$

EM Algorithm - Principle

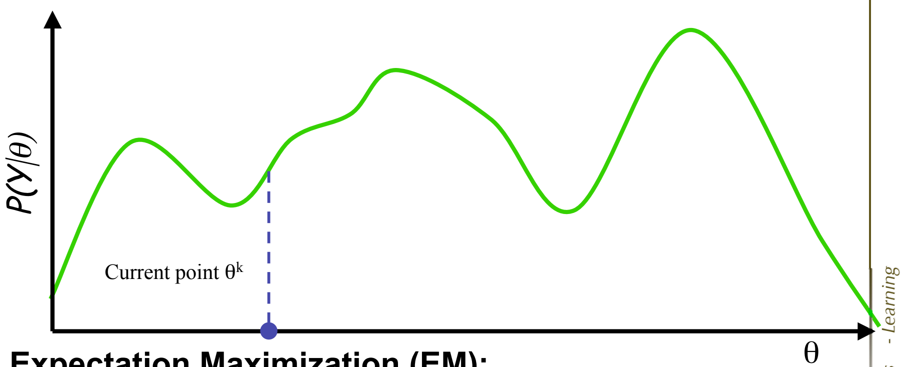


Expectation Maximization (EM):

Construct an new function based on the “current point” (which “behaves well”)

Property: The maximum of the new function has a better scoring then the current point.

EM Algorithm - Principle

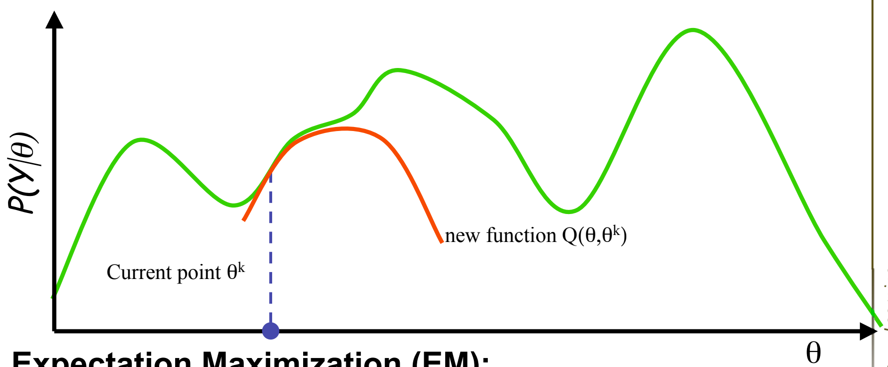


Expectation Maximization (EM):

Construct an new function based on the “current point” (which “behaves well”)

Property: The maximum of the new function has a better scoring then the current point.

EM Algorithm - Principle

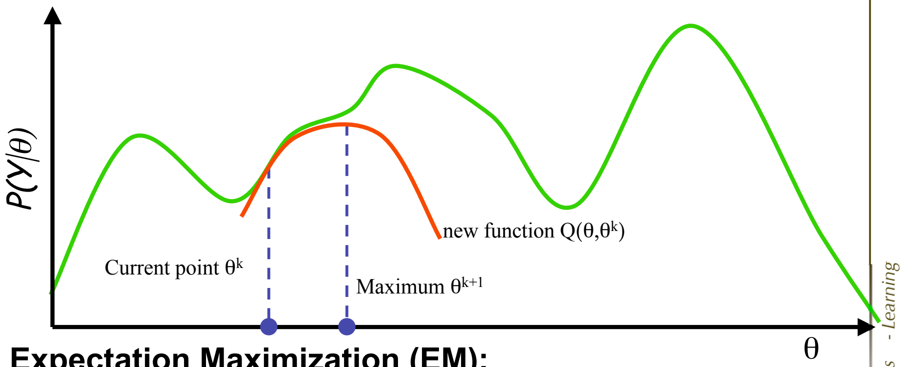


Expectation Maximization (EM):

Construct an new function based on the “current point” (which “behaves well”)

Property: The maximum of the new function has a better scoring then the current point.

EM Algorithm - Principle



Expectation Maximization (EM):
Construct an new function based on the “current point” (which “behaves well”)
Property: The maximum of the new function has a better scoring then the current point.

EM for Multi-Nominals

- Random variable V with 1,...,K values

$$P(V=k)=\theta_k \qquad \sum_{k=1}^{K}\theta_k = 1$$

- $Q(\Theta_v,\Theta') = \sum_{k=1}^{K} \log \theta_k^{EN_k} = \sum_{k=1}^{K} \log EN_k \cdot \theta_k$
  where EN_k are the **expected counts** of state k in the data, i.e.

$$EN_k = \sum_{i=1}^{m} P(k|X_i)$$

- **„MLE“:** $\dfrac{EN_i}{\sum_k EN_k}$

EM for Conditional Multinominals

- $P(V=k|\,\mathrm{pa}(V)=\mathbf{pa})$ multinomial for each joint state pa of the parents of V:

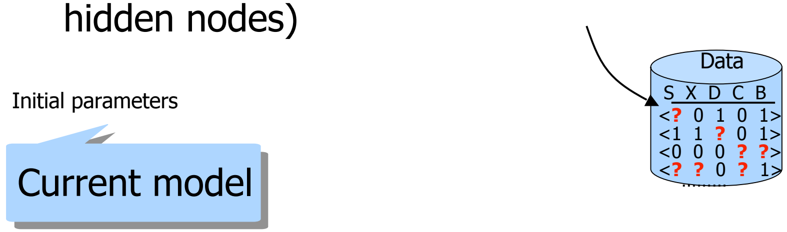
$$P(k|1,1), P(k|1,2), P(k|2,1), P(k|2,2)$$

- $Q(\Theta_v,\Theta')$
  $= \sum_{\mathbf{pa}} \sum_{k=1}^{K} \log \theta_{k|\mathbf{pa}}^{EN_{k,\mathbf{pa}}} = \sum_{\mathbf{pa}} \sum_{k=1}^{K} EN_{k,\mathbf{pa}} \cdot \theta_{k|\mathbf{pa}}$

- **„MLE“** $\theta^*_{k|\mathbf{pa}} = \dfrac{EN_{k,\mathbf{pa}}}{EN_{\mathbf{pa}}}$

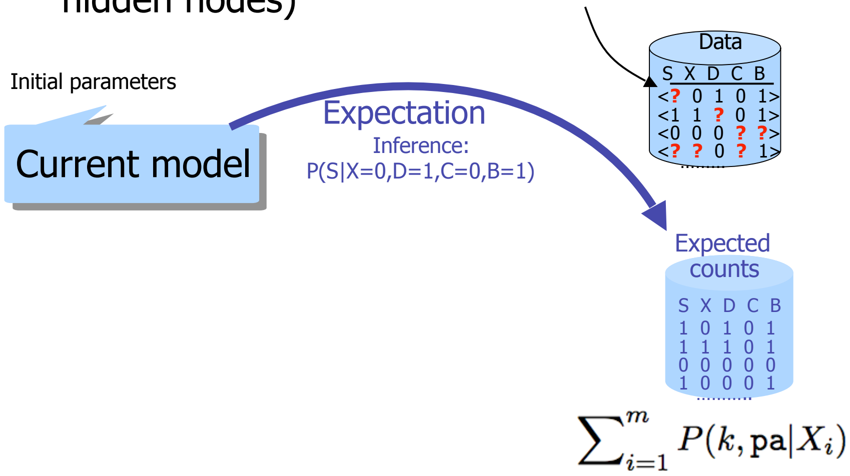
Learning Parameters: incomplete data

Non-decomposable likelihood (missing value, hidden nodes)



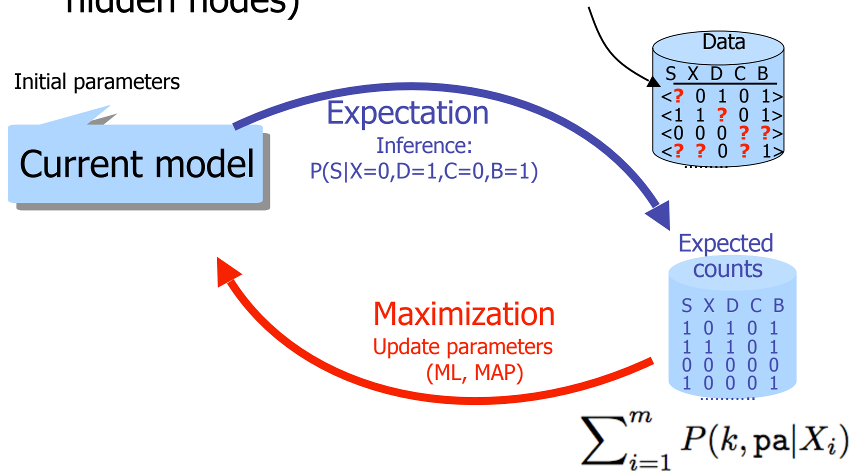
Learning Parameters: incomplete data

Non-decomposable likelihood (missing value, hidden nodes)



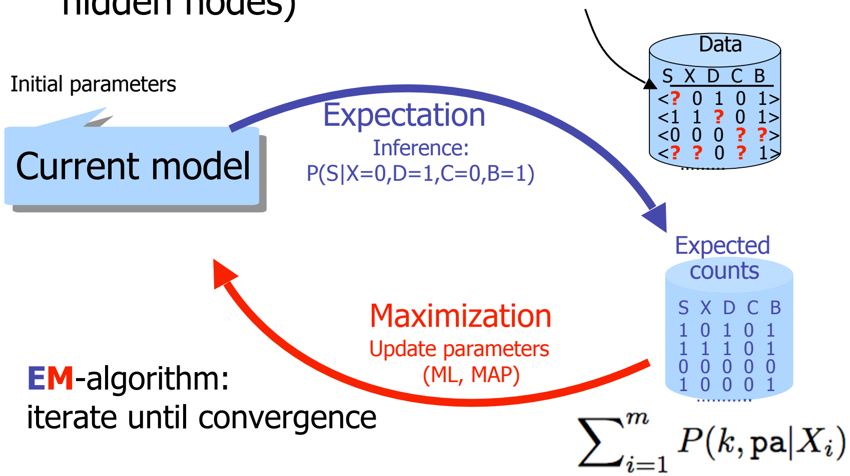
Learning Parameters: incomplete data

Non-decomposable likelihood (missing value, hidden nodes)



Learning Parameters: incomplete data

Non-decomposable likelihood (missing value, hidden nodes)



EM-algorithm:
iterate until convergence

Learning Parameters: incomplete data

1. Initialize parameters
2. Compute pseudo counts for each variable

$$\theta^*_{k|\mathrm{pa}} = \frac{\sum_{i=1}^{m} P(k,\mathrm{pa}|X_i)}{\sum_{i=1}^{m} P(\mathrm{pa}|X_i)}$$

junction tree algorithm

3. Set parameters to the (completed) ML estimates
4. If not converged, iterate to 2

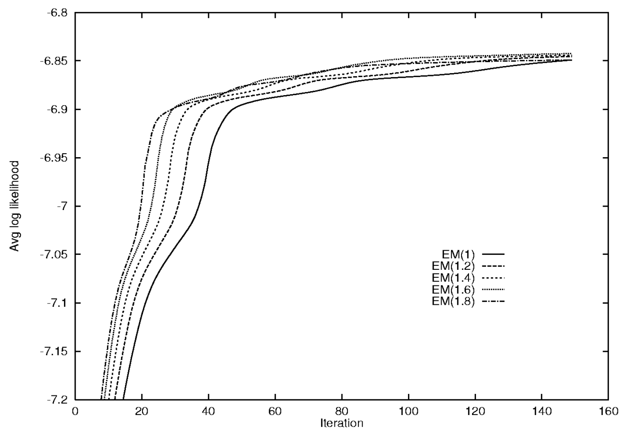
Monotonicity

- (Dempster, Laird, Rubin ´77): the incomplete-data likelihood fuction is not decreased after an EM iteration

$$\mathcal{L}(\Theta^{i}|\mathcal{X}) \geq \mathcal{L}(\Theta^{i-1}|\mathcal{X})$$

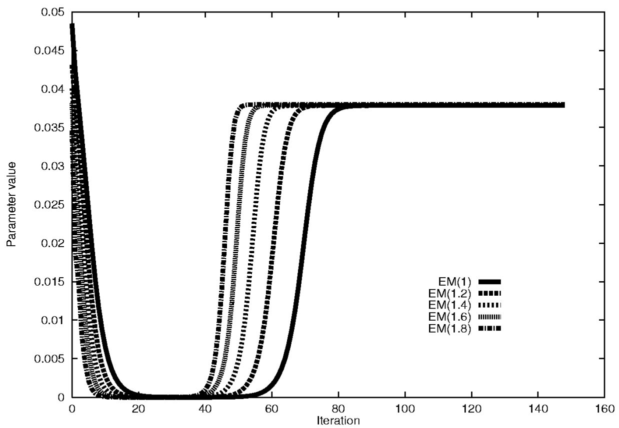
- (discrete) Bayesian networks: for any initial, non-uniform value the EM algorithm converges to a (local or global) maximum.

LL on training set (Alarm)



Experiment by Bauer, Koller and Singer [UAI97]

Parameter value (Alarm)



Experiment by Baur, Koller and Singer [UAI97]

EM in Practice

Initial parameters:
- Random parameters setting
- "Best" guess from other source

Stopping criteria:
- Small change in likelihood of data
- Small change in parameter values

Avoiding bad local maxima:
- Multiple restarts
- Early "pruning" of unpromising ones

Speed up:
- **various methods to speed convergence**

Gradient Ascent

- Main result

$$\frac{\partial \mathcal{LL}(\Theta|\mathcal{X})}{\partial \theta_{k|\mathbf{pa}}} = \frac{1}{\theta_{k|\mathbf{pa}}} \sum_{j=1}^{m} \log P(k, \mathbf{pa}|X_j, \Theta)$$

- Requires same BN inference computations as EM

- **Pros:**
  - Flexible
  - Closely related to methods in neural network training
- **Cons:**
  - Need to project gradient onto space of legal parameters
  - To get reasonable convergence we need to combine with "smart" optimization techniques

Parameter Estimation: Summary

- Parameter estimation is a basic task for learning with Bayesian networks
- Due to missing values non-linear optimization
  - EM, Gradient, ...
- EM for multi-nominal random variables
  - Fully observed data: counting
  - Partially observed data: pseudo counts
- Junction tree to do multiple inference