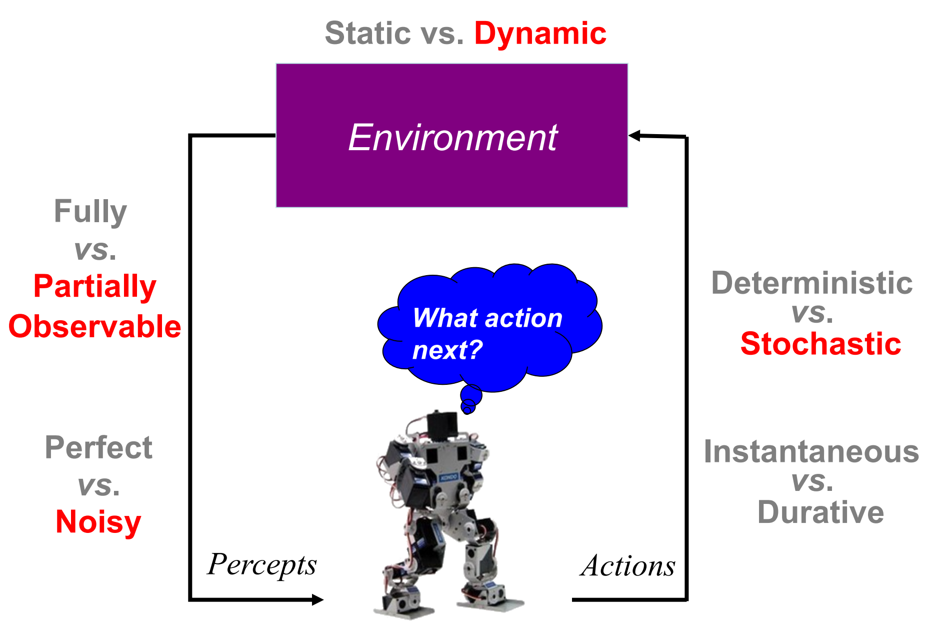
CSE 473: Artificial Intelligence

Bayesian Networks - Learning

Dan Weld

Slides adapted from Jack Breese, Dan Klein, Daphne Koller, Stuart Russell, Andrew Moore & Luke Zettlemoyer

1



2

AI Topics

- Search
  - Problem Spaces
  - BFS, DFS, UCS, A* (tree and graph)
  - Completeness and Optimality
  - Heuristics: admissibility, consistency & creation
  - Pattern databases
- Games
  - Minimax, Alpha-beta pruning, Expectimax, Evaluation Functions
- MDPs
  - Bellman equations
  - Value iteration & policy iteration
  - RTDP,
  - POMDPs
- Reinforcement Learning
  - Exploration *vs.* Exploitation
  - Model-based *vs.* model-free
  - Q-learning
  - Linear value function approx.
- Hidden Markov Models
  - Markov chains
  - Forward algorithm
  - Particle Filter
- Bayesian Networks
  - Basic definition, independence (d-sep)
  - Variable elimination
  - Gibbs sampling
- Learning
  - BN parameters with data complete & incomplete (Expectation Maximization)
  - Structure learning as search

3

Search thru a
Problem Space / State Space

Ex. Proving a trig identity, e.g. $\sin^2(x) = ½ - ½ \cos(2x)$

- Input:
  - Set of states
  - Operators [and costs]
  - Start state
  - Goal state [test]
- Output:
  - Path: start $\Rightarrow$ a state satisfying goal test
  - [May require shortest path]
  - [Sometimes just need state passing test]

4

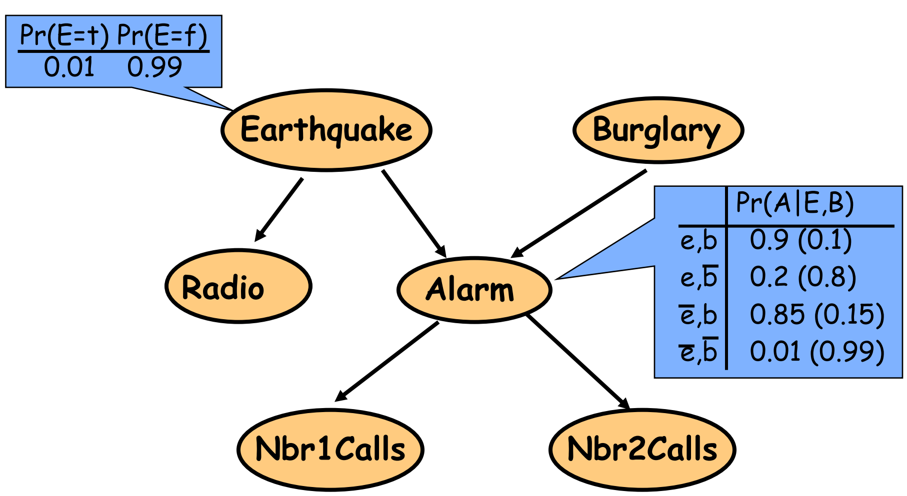
Today

- Bonus Topic – Hybrid Bayes Nets
- Learning
  - Parameter Learning & Priors
  - Expectation Maximization
  - Structure Learning

5

5

Bayes Nets



 6

6

Continuous Variables

Pr(E=t)	Pr(E=f)
0.01	0.99

Earthquake

So far: assuming variables have discrete values, e.g. True / False
Could also allow continuous values, E $\in \mathbb{R}$, eg 7.9 (*on the Richter scale*)
How specify probabilities? (explicit CPT would be infinitely large)

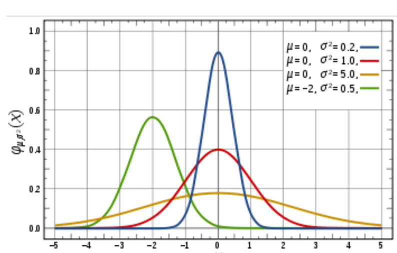
7

Continuous Variables

Pr(E=t)	Pr(E=f)
0.01	0.99

Earthquake

So far: assuming variables have discrete values
Could also allow continuous values, E $\in \mathbb{R}$,
Specify probabilities with a pre-defined continuous distribution, eg Gaussian



$$P(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

8

Continuous Variables

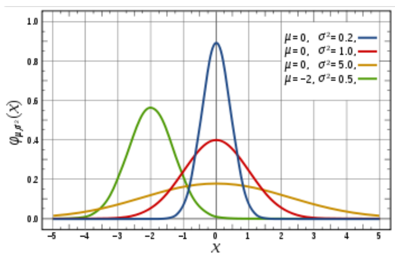
Earthquake

Pr(E=x)

mean: μ = 6

variance: σ = 2

So far: assuming variables have discrete values
Could also allow continuous values, E $\in \mathbb{R}$,
And specify probabilities using a continuous distribution, such as a Gaussian



$$P(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

9

Continuous Variables with Discrete Parents

Pr(A=t)	Pr(A=f)
0.01	0.99

Aliens

Earthquake

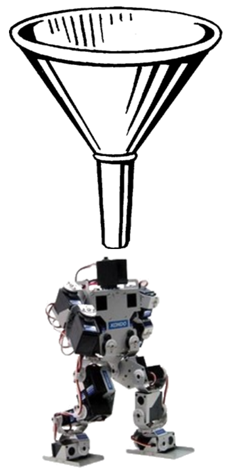
	Pr(E|A)
a	μ = 6
	σ = 2
$\bar{a}$	μ = 1
	σ = 3

10

End Bonus Topic…

Back to:

Learning



11

What is Machine Learning?

Study of algorithms that

- improve their ***performance***
- at some ***task***
- with ***experience***

12

12

Space of ML Problems

Type of Supervision
(eg, Experience, Feedback)

What is Being Learned?

	Labeled Examples	Reward	Nothing
Discrete Function	Classification		Clustering
Continuous Function	Regression		
Policy	Apprenticeship Learning	Reinforcement Learning	

14

14

Supremacy of Machine Learning

- Machine learning is preferred approach to
  - Speech recognition, Natural language processing
  - Web search – result ranking
  - Computer vision
  - Medical outcomes analysis
  - Robot control
  - Computational biology
  - Sensor networks
  - …
- This trend is accelerating
  - Improved machine learning algorithms
  - Improved data capture, networking, faster computers
  - Software too complex to write by hand
  - New sensors / IO devices
  - Demand for self-customization to user, environment

15

15

Reinforcement Learning

Study of algorithms that
- improve their performance
- at some task
- with experience

Ability to accumulate reward

Executing actions

Executing actions

17

17

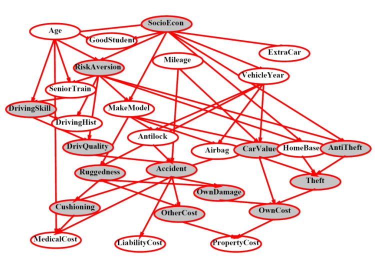
Bayes Net Learning

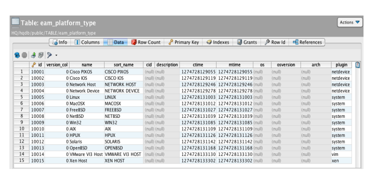
Study of algorithms that
- improve their performance
- at some task
- with experience

Prediction accuracy

Answering probabilistic queries

Seeing labeled data





18

18

Learning Bayes Networks

- Learning Parameters for a Bayesian Network
  - Fully observable variables
    - Maximum Likelihood (ML), MAP & Bayesian estimation
    - Example: Naïve Bayes for text classification
  - Hidden variables
    - Expectation Maximization (EM)
- Learning the Structure of Bayesian Networks

19

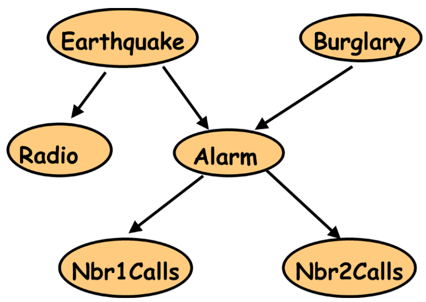
Learning Bayes Nets

Suppose …

1. Know structure & get complete observations of every var
2. Know structure & get observations only of ***some*** vars
   Others are hidden (learn with EM)
3. Don't even know structure!

21

Parameter Estimation and Bayesian Networks

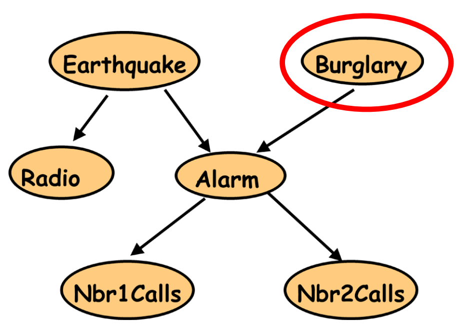


E	B	R	A	J	M
T	F	T	T	F	T
F	F	F	F	F	T
F	T	F	T	T	T
F	F	F	T	T	T
F	T	F	F	F	F
...					

We have:
- Bayes Net structure and observations
- We need: Bayes Net parameters

22

Parameter Estimation and Bayesian Networks



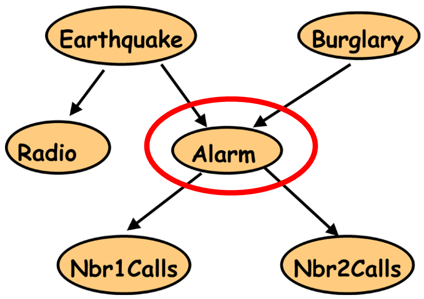
B
F
F
T
F
T

P(B) = ? = 0.4

P(¬B) = 1- P(B) = 0.6

23

Parameter Estimation and Bayesian Networks



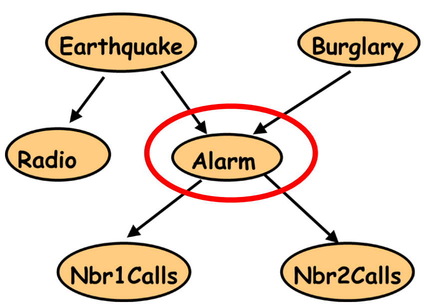
E	B
T	F
F	F
F	T
F	F
F	T
…	

A
T
F
T
T
F

P(A|E,B) = ?
P(A|E,¬B) = ?
P(A|¬E,B) = ?
P(A|¬E,¬B) = 0.5

24

Parameter Estimation and Bayesian Networks



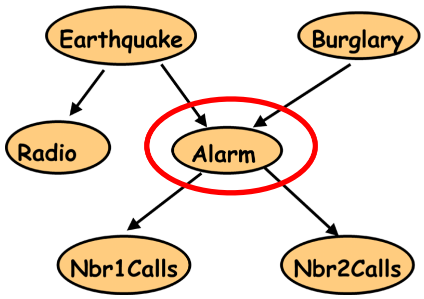
E	B
T	F
F	F
F	T
F	F
F	T
…	

A
T
F
T
T
F

P(A|E,B) = ?
P(A|E,¬B) = 1.0 ?
P(A|¬E,B) = ?
P(A|¬E,¬B) = ?

25

Parameter Estimation and Bayesian Networks



E	B	A
T	F	T
F	F	F
F	T	T
F	F	T
F	T	F
…		

P(A|E,B) = ?
P(A|E,¬B) = ?
P(A|¬E,B) = ?
P(A|¬E,¬B) = ?

26

Estimation: Laplace Smoothing

- Laplace's estimate:
  pretend you saw every outcome once more than you actually did

H H T

$$P_{LAP}(x) = \frac{c(x)+1}{\sum_x [c(x)+1]}$$

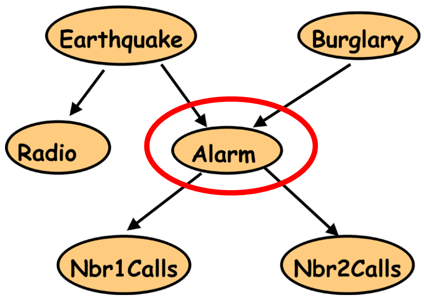
$$= \frac{c(x)+1}{N+|X|}$$

$P_{LAP}(H)$ = (2+1) / (3+2)
= 3/5

Another name for computing the MAP estimate with Dirichlet priors
(Bayesian justification)

41

Parameter Estimation and Bayesian Networks



E	B	A
T	F	T
F	F	F
F	T	T
F	F	T
F	T	F
…		

P(A|E,B) = ?

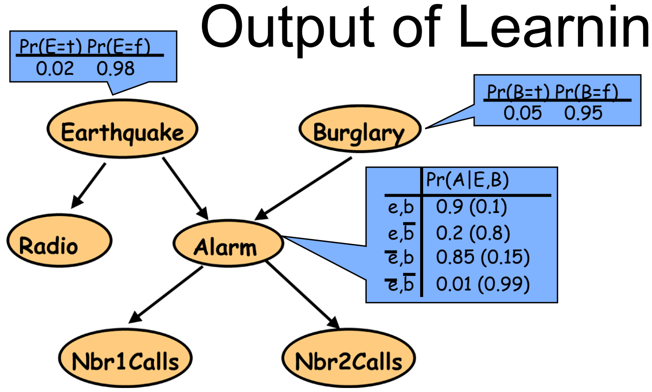
P(A|E,¬B) = 2 / 3. Laplacian smoothing: imaginary T, F

P(A|¬E,B) = ?

P(A|¬E,¬B) = ?

42

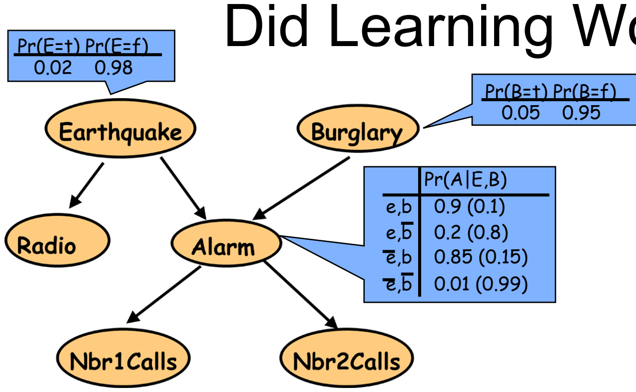
Output of Learning



E	B	R	A	J	M
T	F	T	T	F	T
F	F	F	F	F	T
F	T	F	T	T	T
F	F	F	T	T	T
F	T	F	F	F	F
…					

52

Did Learning Work Well?

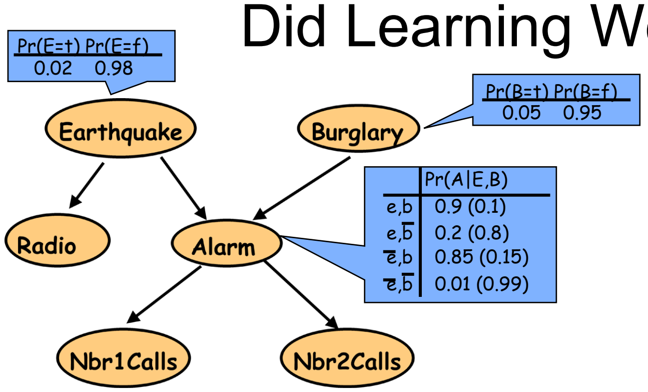


E	B	R	A	J	M
T	F	T	T	F	T
F	F	F	F	F	T
F	T	F	T	T	T
F	F	F	T	T	T
F	T	F	F	F	F
…					

Calculate P(data)
Assuming learned parameters

53

Did Learning Work Well?



E	B	R	A	J	M
T	F	T	T	F	T

Calculate P(data)
Assuming learned parameters

0.02 * 0.95 * 0.2 *….

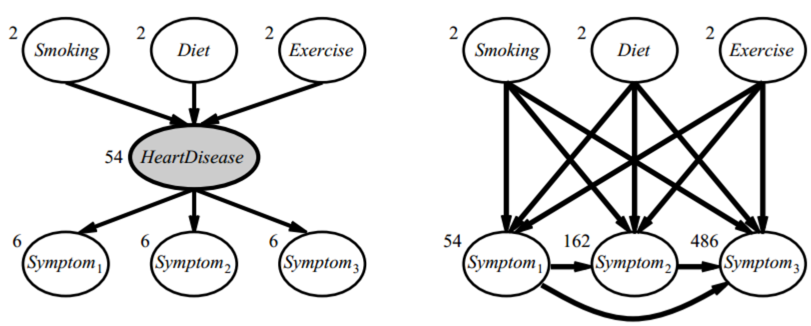
54

Topics

- Another Useful Bayes Net
  - Hybrid Discrete / Continuous
- Learning Parameters for a Bayesian Network
  - Fully observable
  - Hidden variables (EM algorithm)
- Learning Structure of Bayesian Networks

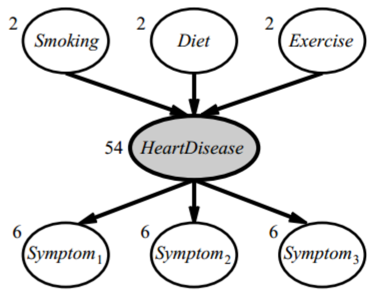
55

Why Learn Hidden Variables?



56

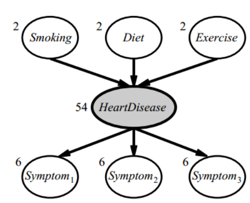
How Learn Hidden Variables?



57

Chicken & Egg Problem

- If we knew whether patient had disease
  - It would be easy to learn CPTs
    Fully observable!
  - But we can't observe states, so we don't!



- If we knew CPTs
  - It would be easy to predict if patient had disease
  - But we don't, so we can't!

Slide by Daniel S. Weld

58

58

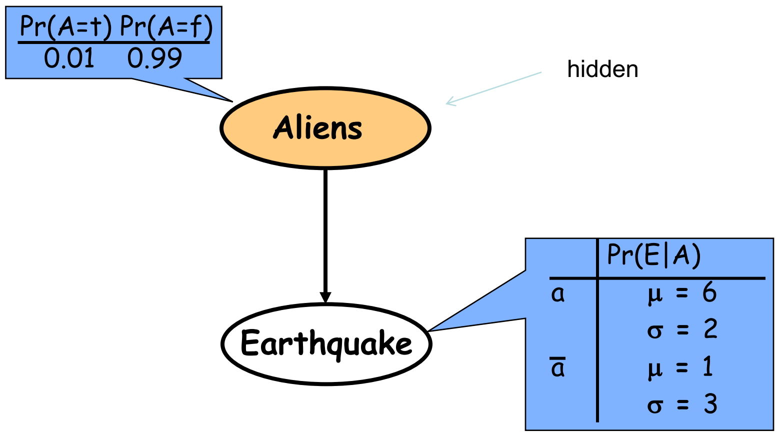
Face It…

59

59

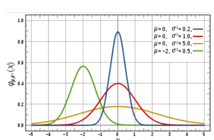
60

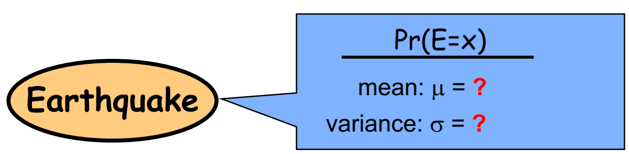
Continuous Variables



61

Learning with Continuous Variables



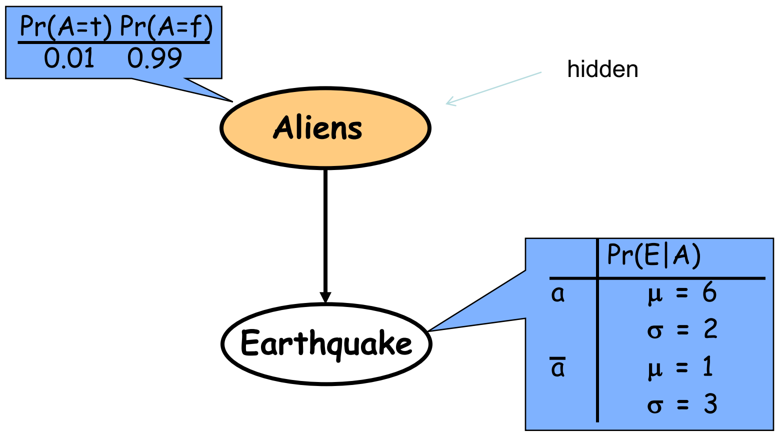


$$\widehat{\mu}_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\widehat{\sigma}^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N} (x_i - \widehat{\mu})^2$$

62

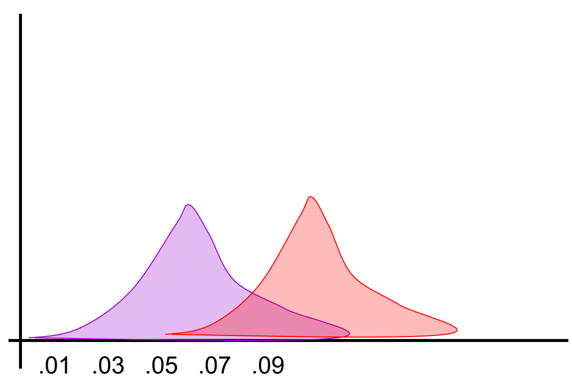
Continuous Variables



63

Simplest Version

- Mixture of two distributions



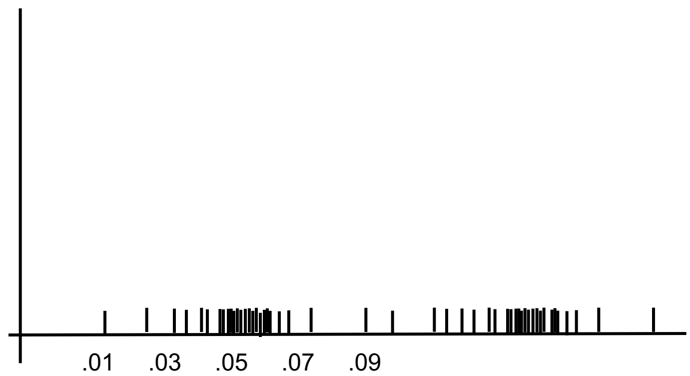
- Know: form of distribution & variance, $\sigma = .5$
- Just need *mean* of each distribution

Slide by Daniel S. Weld

64

64

Input Looks Like

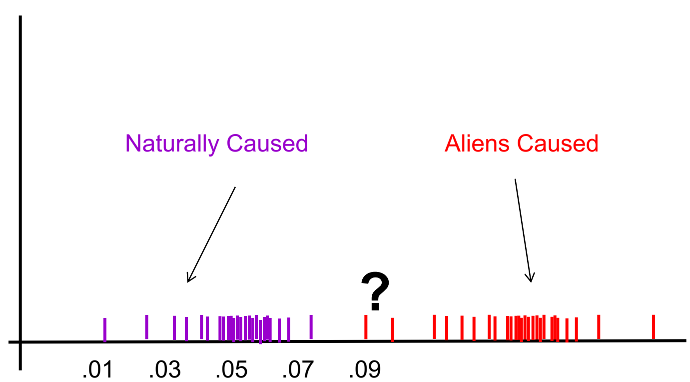


Slide by Daniel S. Weld

65

65

We Want to Predict



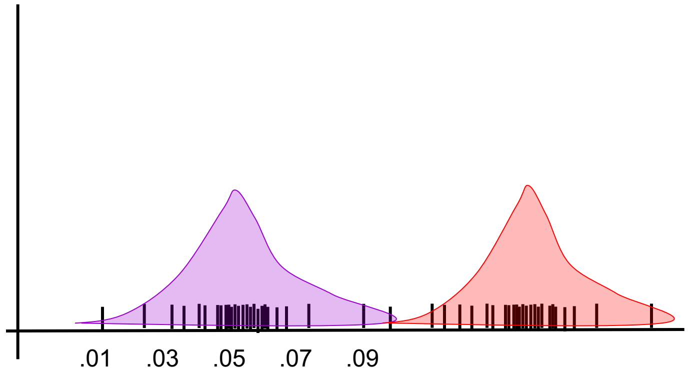
Slide by Daniel S. Weld

66

66

Chicken & Egg

Note that coloring instances would be easy *if* we knew Gausians….



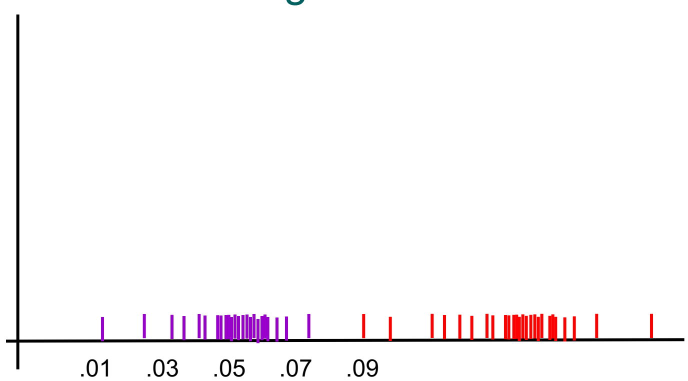
Slide by Daniel S. Weld

67

67

Chicken & Egg

And finding Gausian parameters would be easy *If* we knew the coloring



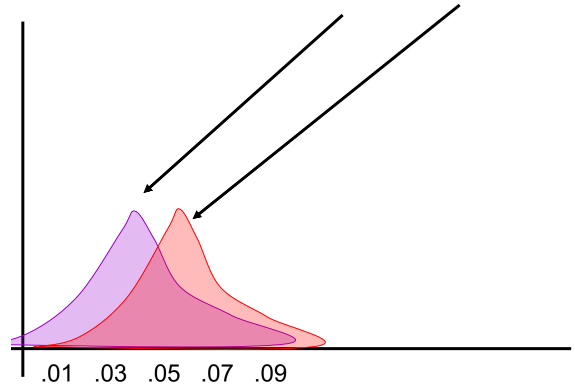
Slide by Daniel S. Weld

68

68

Expectation Maximization (EM)

- Pretend we *do* know the parameters
  - Initialize randomly: set θ_1=?; θ_2=?



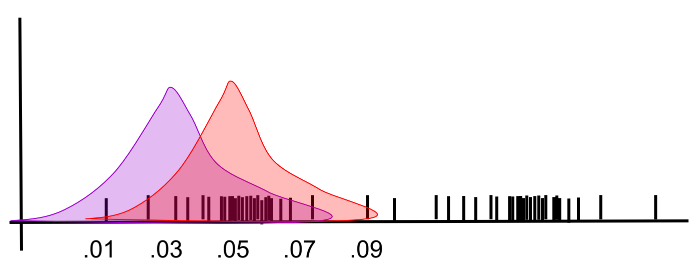
Slide by Daniel S. Weld

69

69

Expectation Maximization (EM)

- Pretend we *do* know the parameters
  - Initialize randomly
- [E step] Compute probability of instance having each possible value of the hidden variable



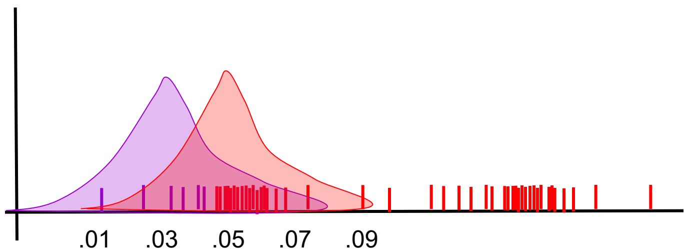
Slide by Daniel S. Weld

70

70

Expectation Maximization (EM)

- Pretend we *do* know the parameters
  - Initialize randomly
- [E step] Compute probability of instance having each possible value of the hidden variable



Slide by Daniel S. Weld

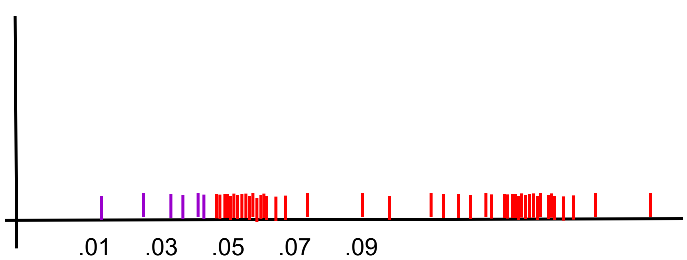
71

71

Expectation Maximization (EM)

- Pretend we *do* know the parameters
  - Initialize randomly
- [E step] Compute probability of instance having each possible value of the hidden variable

[M step] Treating each instance as ***fractionally*** having **both** values compute the new parameter values



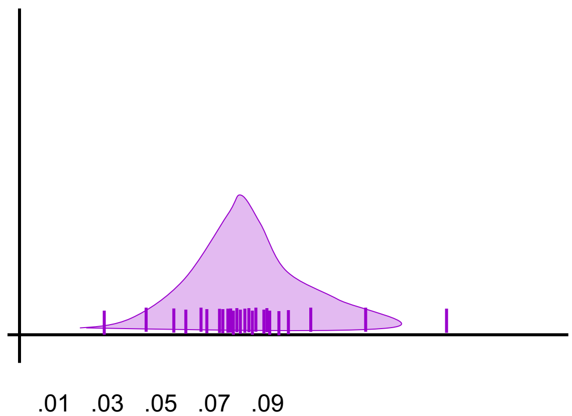
Slide by Daniel S. Weld

72

72

ML Mean of Single Gaussian

$U_{ml} = \text{argmin}_u \sum_i (x_i - u)^2$



Slide by Daniel S. Weld

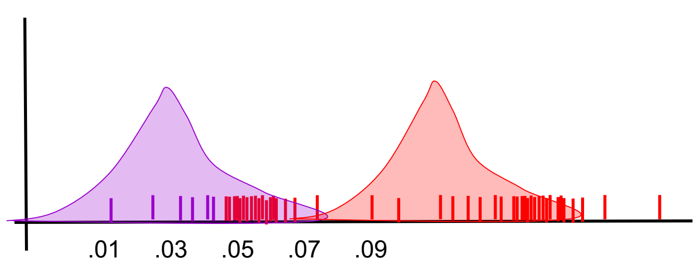
73

73

Expectation Maximization (EM)

-

[M step] Treating each instance as fractionally having **both** values compute the new parameter values



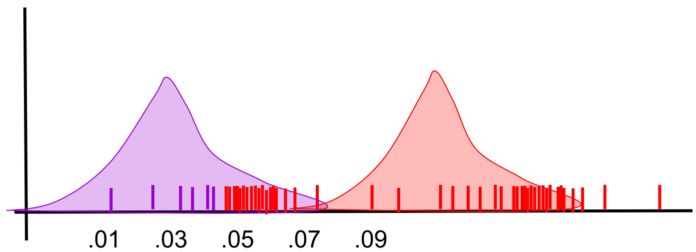
Slide by Daniel S. Weld

74

74

Expectation Maximization (EM)

- [E step] Compute probability of instance having each possible value of the hidden variable



Slide by Daniel S. Weld

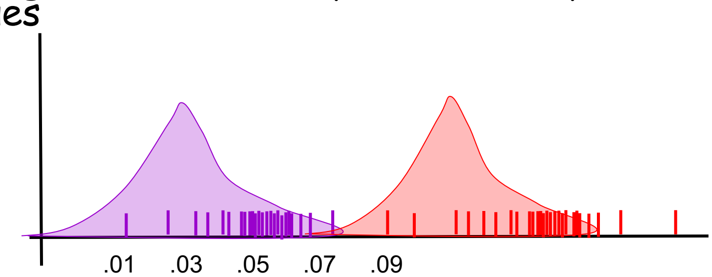
75

75

Expectation Maximization (EM)

- [E step] Compute probability of instance having each possible value of the hidden variable

[M step] Treating each instance as fractionally having both values compute the new parameter values



Slide by Daniel S. Weld

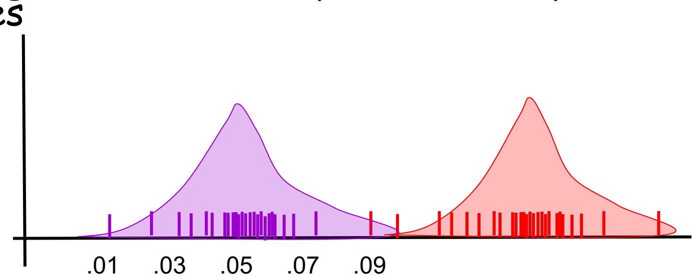
76

76

Expectation Maximization (EM)

- [E step] Compute probability of instance having each possible value of the hidden variable

[M step] Treating each instance as fractionally having both values compute the new parameter values



Slide by Daniel S. Weld

77

77

Expectation Maximization

- Guaranteed to converge to fixed point solution
- NOT guaranteed to find optimal solution (one with highest likelihood given data)
- Used everywhere!

78

78

Topics

- Another Useful Bayes Net
  - Hybrid Discrete / Continuous
- Learning Parameters for a Bayesian Network
  - Fully observable
    - Maximum Likelihood (ML),
    - Maximum A Posteriori (MAP)
  - Hidden variables (EM algorithm)
- Learning Structure of Bayesian Networks

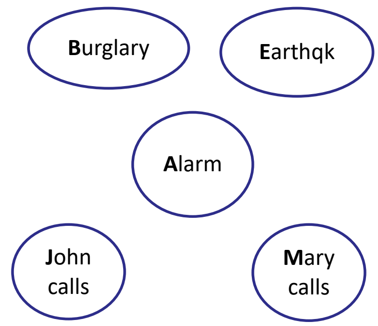
79

What if we ***don't*** know structure?

80

Learning The Structure of Bayesian Networks

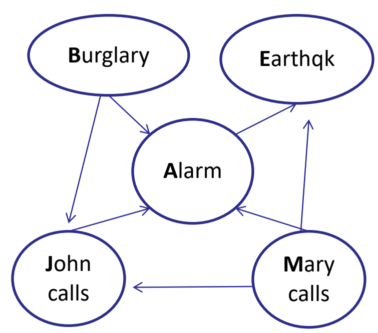
E	B	R	A	J	M
T	F	T	T	F	T
F	F	F	F	F	T
F	T	F	T	T	T
F	F	F	T	T	T
F	T	F	F	F	F
...					



81

Learning The Structure of Bayesian Networks

E	B	R	A	J	M
T	F	T	T	F	T
F	F	F	F	F	T
F	T	F	T	T	T
F	F	F	T	T	T
F	T	F	F	F	F
...					

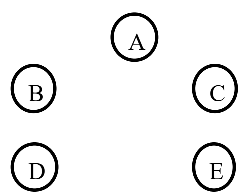


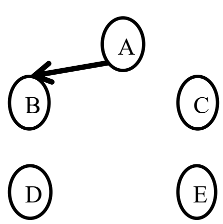
82

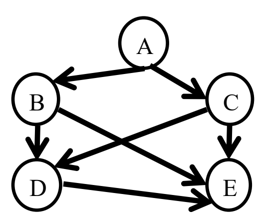
Learning The Structure of Bayesian Networks

- Search thru the space…
  - of possible network structures!
- For each structure, learn parameters
  - As just shown…
- Pick the one that fits observed data best
  - Learn best parameter values for that structure
  - Calculate P(data)

83







Two problems:
- Fully connected will be most probable
- Exponential number of structures

84

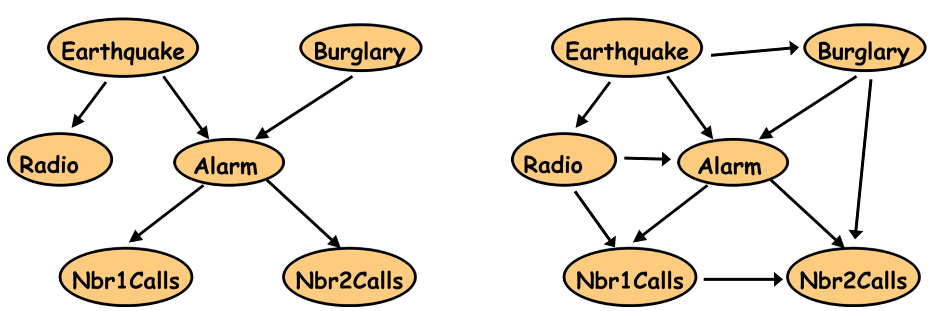
Learning The Structure of Bayesian Networks

- Search thru the space…
  - of possible network structures!
- For each structure, learn parameters
  - As just shown…
- Pick the one that fits observed data best
  - Calculate P(data)

Two problems:
- Fully connected will be most probable
  - Add penalty term (regularization) $\propto$ model complexity
- Exponential number of structures
  - Local search

85

Overfitting



Can represent strictly more P distributions

Can represent NOISE in training data

Often preforms WORSE on test data

86

Augment Score Function

- Bayesian Information Criterion (BIC)
  - P(D | BN) – penalty
  - Penalty = α complexity
  - = α [½ (# parameters) Log (# data points)]

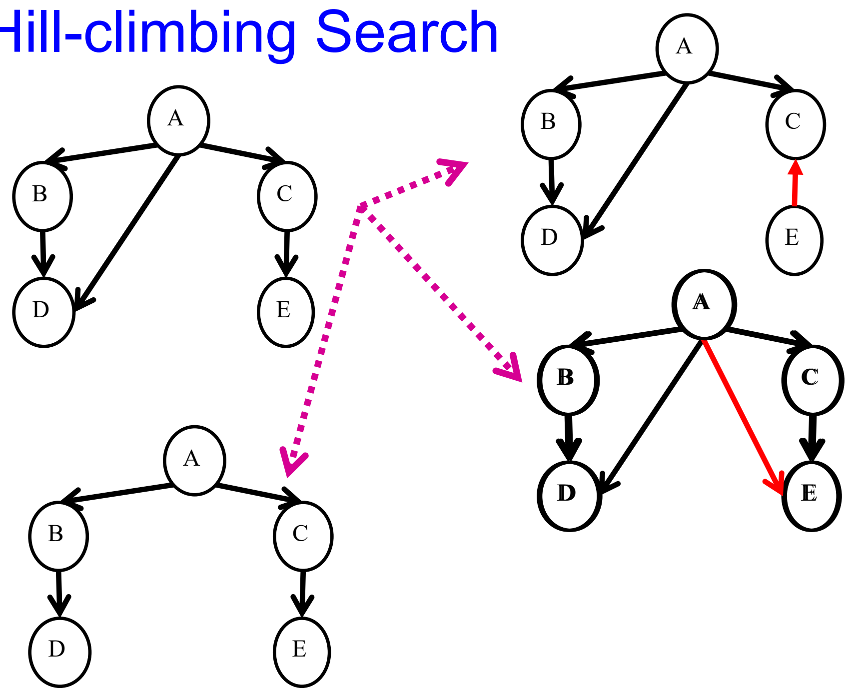
Instance of "***regularization***"

Solves problem of "***overfitting***"

© Daniel S. Weld 87

87

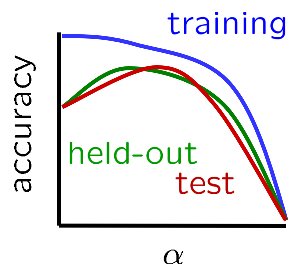
Hill-climbing Search



89

Tuning on Held-Out Data

- Now we've got two kinds of unknowns
  - Parameters: the probabilities P(Y|X), P(Y)
  - Hyperparameters, like
    - the amount of smoothing to do: k, or
    - regularization penalty, α
- Where to learn?
  - Learn parameters from training data
  - Must tune hyperparameters on different data
    - Why?
  - For each value of the hyperparameters, train and test on the held-out data
  - Choose the best value and do a final test on the test data



90

Topics

- Another Useful Bayes Net
  - Hybrid Discrete / Continuous
- Learning Parameters for a Bayesian Network
  - Fully observable
    - Maximum Likelihood (ML),
    - Maximum A Posteriori (MAP)
  - Hidden variables (EM algorithm)
- Learning Structure of Bayesian Networks

92

AI Topics

- Search
  - Problem Spaces
  - BFS, DFS, UCS, A* (tree and graph)
  - Completeness and Optimality
  - Heuristics: admissibility, consistency & creation
  - Pattern databases
- Games
  - Minimax, Alpha-beta pruning, Expectimax, Evaluation Functions
- MDPs
  - Bellman equations
  - Value iteration & policy iteration
  - RTDP,
  - POMDPs
- Reinforcement Learning
  - Exploration *vs.* Exploitation
  - Model-based *vs.* model-free
  - Q-learning
  - Linear value function approx.
- Hidden Markov Models
  - Markov chains
  - Forward algorithm
  - Particle Filter
- Bayesian Networks
  - Basic definition, independence (d-sep)
  - Variable elimination
  - Gibbs sampling
- Learning
  - BN parameters with data complete & incomplete (Expectation Maximization)
  - Structure learning as search

93

Search thru a Problem Space / State Space

- Input:
  - Set of states
  - Operators [and costs]
  - Start state
  - Goal state [test]
- Output:
  - Path: start $\Rightarrow$ a state satisfying goal test
  - [May require shortest path]
  - [Sometimes just need state passing test]

94