CSE 473: Artificial Intelligence

Bayesian Networks - Learning

Dieter Fox

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

1

Space of ML Problems

Type of Supervision
(eg, Experience, Feedback)

What is Being Learned?

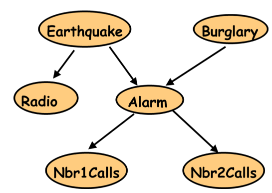
	Labeled Examples	Reward	Nothing
Discrete Function	Classification		Clustering
Continuous Function	Regression		
Policy	Learning from Demonstration	Reinforcement Learning	

2

Learning Topics

- Learning Parameters for a Bayesian Network
  - Fully observable
  - Hidden variables (EM algorithm)
- Learning Structure of Bayesian Networks

Parameter Estimation and Bayesian Networks

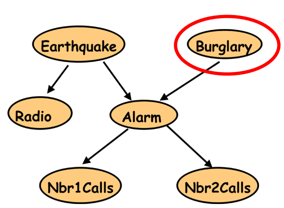


E	B	R	A	J	M
T	F	T	T	F	T
F	F	F	F	F	T
F	T	F	T	T	T
F	F	F	T	T	T
F	T	F	F	F	F
...					

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

4

Parameter Estimation and Bayesian Networks



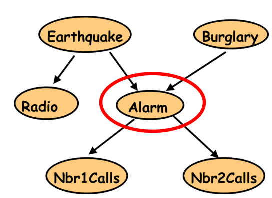
B
F
F
T
F
T

P(B) = ? = 0.4

P(¬B) = 1- P(B) = 0.6

5

Parameter Estimation and Bayesian Networks



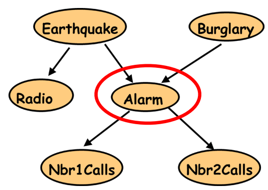
E	B
T	F
F	F
F	T
F	F
F	T
...	

A
T
F
T
T
F

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

6

Parameter Estimation and Bayesian Networks



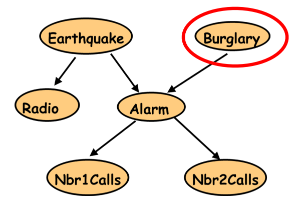
E	B
T	F
F	F
F	T
F	F
F	T
...	

A
T
F
T
T
F

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

7

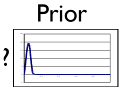
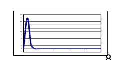
Parameter Estimation and Bayesian Networks



B
F
F
T
F
T

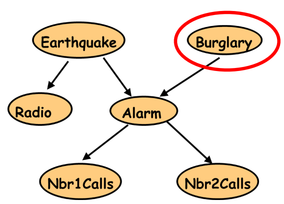
Now compute either MAP or Bayesian estimate

Prior

P(B|data) = ?  + data = 

8

Parameter Estimation and Bayesian Networks



B
F
F
T
F
T

Prior

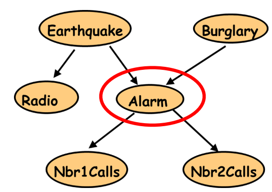
P(B|data) = **Beta(1,4)** "+ data" = (3,7)

B	*¬B*
.3	.7

Prior P(B)= 1/(1+4) = 20% with equivalent sample size 5

9

Parameter Estimation and Bayesian Networks



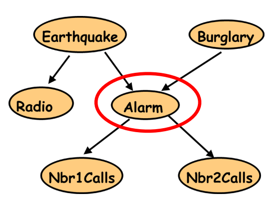
E	B
T	F
F	F
F	T
F	F
F	T
...	

A
T
F
T
T
F

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

10

Parameter Estimation and Bayesian Networks



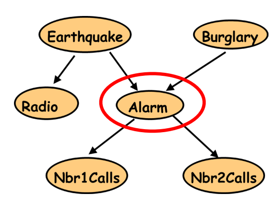
E	B
T	F
F	F
F	T
F	F
F	T
...	

A
T
F
T
T
F

P(A|E,B) = ? Prior
P(A|E,¬B) = ?
P(A|¬E,B) = ? **Beta(2,3)**
P(A|¬E,¬B) = ?

11

Parameter Estimation and Bayesian Networks



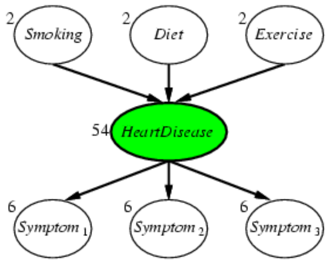
E	B
T	F
F	F
F	T
F	F
F	T
...	

A
T
F
T
T
F

P(A|E,B) = ? Prior
P(A|E,¬B) = ?
P(A|¬E,B) = ? **Beta(2,3)** + data= **(3,4)**
P(A|¬E,¬B) = ?

12

Hidden Variables

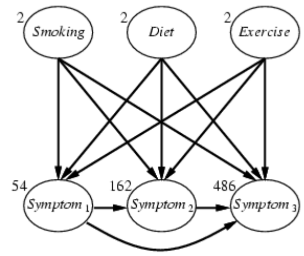


- But we can't observe the disease variable
- Can't we learn without it?

13

We *–could-*

- But we'd get a fully-connected network

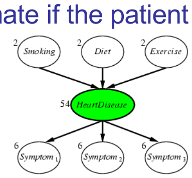


With 708 parameters (vs. 78)
Much harder to learn!

14

Chicken & Egg Problem

- If we knew that a training instance (patient) had the disease, then it'd be easy to learn P(symptom | disease)
- If we knew params, e.g. P(symptom | disease) then it'd be easy to estimate if the patient had the disease



1977: The EM Algorithm

- Dempster, Laird, and Rubin
  - General framework for likelihood-based parameter estimation with missing data
    - start with initial guesses of parameters
    - E-step: estimate memberships given params
    - M-step: estimate params given memberships
    - Repeat until convergence
  - Converges to a **local** maximum of likelihood
  - E-step and M-step are often computationally simple
  - Can incorporate priors over parameters

16

Expectation Maximization (EM)

(high-level version)

- 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
- Iterate until convergence!

17

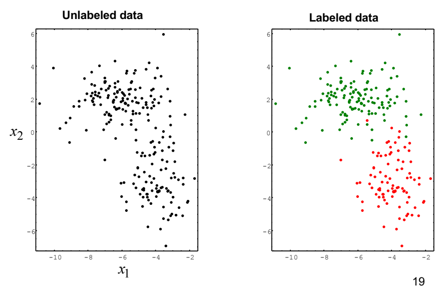
Expectation Maximization and Gaussian Mixtures

CSE 473

18

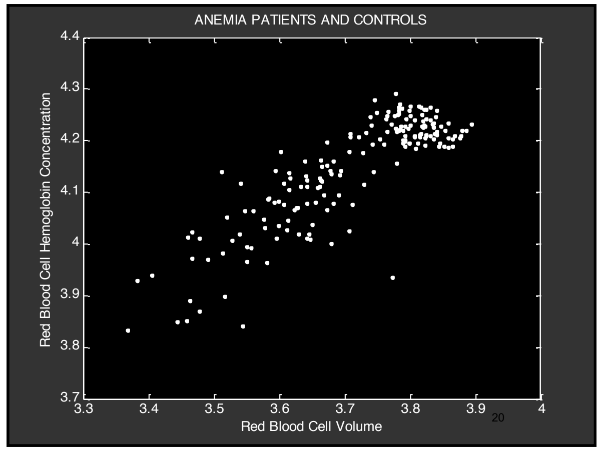
The problem of finding labels for unlabeled data

In nature, items often do not come with labels. How can we learn labels without a teacher?

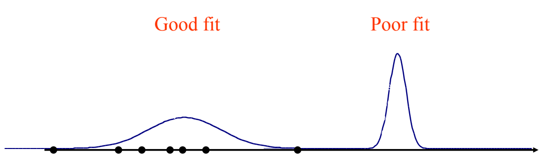


19

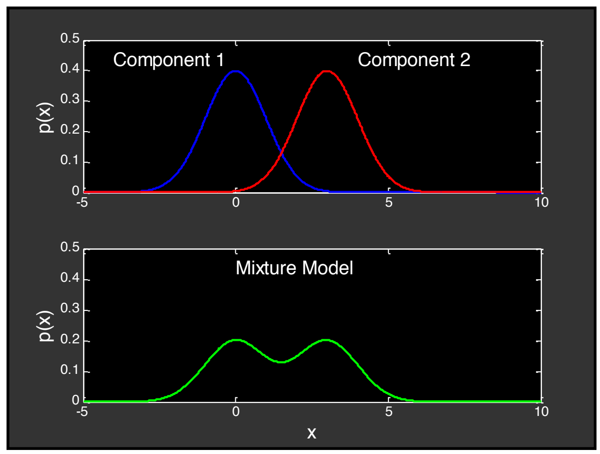
From Shadmehr & Diedrichsen

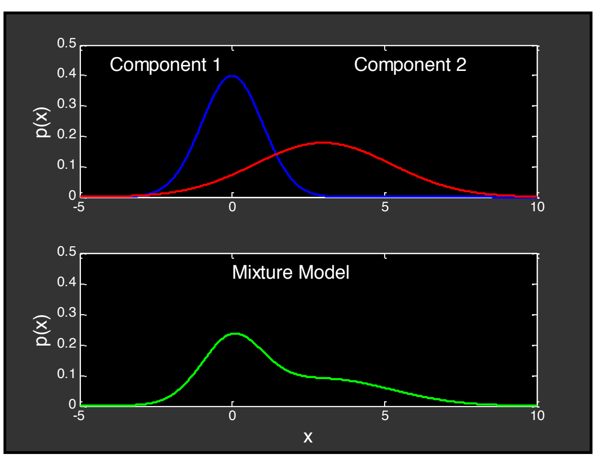


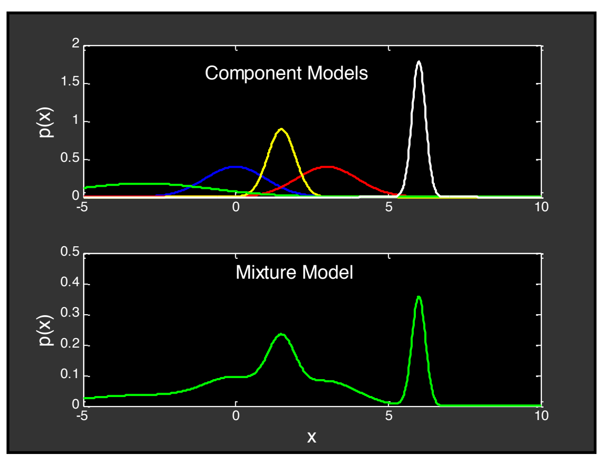
Fitting a Gaussian PDF to Data



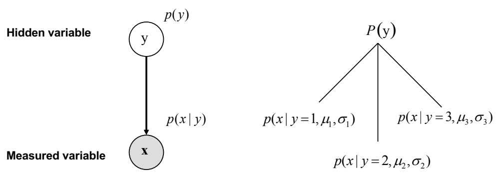
21







Bayes Net for Gaussian Mixtures



$$p(x) = \sum_{i=1}^{3} p(y=i)\, p(x \mid y=i, \mu_i, \sigma_i)$$

25

Learning of mixture models

26

Learning Mixtures from Data

Consider fixed K = 2

e.g., unknown parameters $\Theta = \{\mu_1, \sigma_1, \mu_2, \sigma_2, \alpha_1\}$

Given data D = $\{x_1, \ldots\ldots, x_N\}$, we want to find the parameters Θ that "best fit" the data

27

EM for Mixture of Gaussians

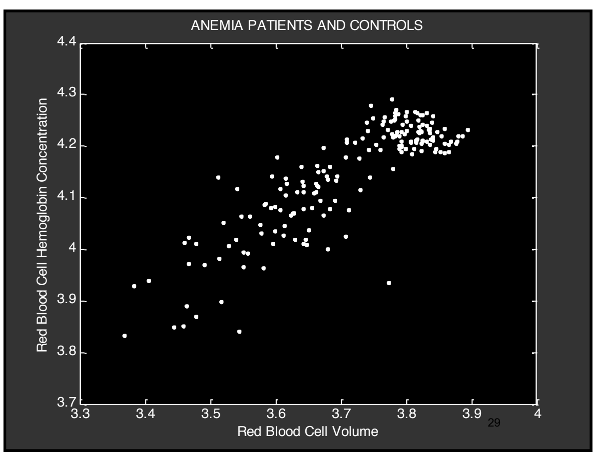
- E-step: Compute probability that point x_j was generated by component i:

$$p_{ij} = \alpha\, P(x_j \mid C=i)\, P(C=i)$$
$$p_i = \sum_j p_{ij}$$

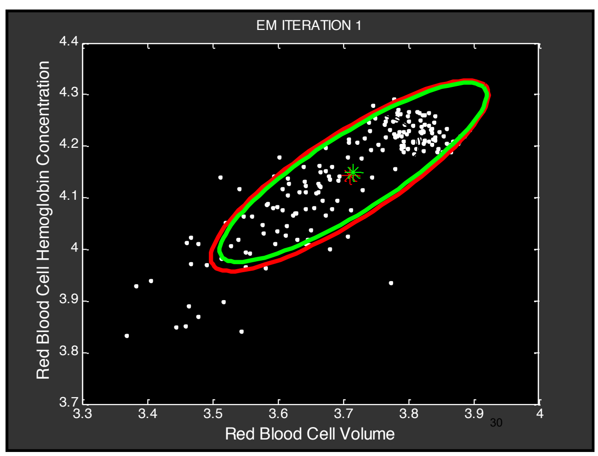
- M-step: Compute new mean, covariance, and component weights:

$$\mu_i \leftarrow \sum_j p_{ij} x_j / p_i$$
$$\sigma^2 \leftarrow \sum_j p_{ij} (x_j - \mu_i)^2 / p_i$$
$$w_i \leftarrow p_i$$

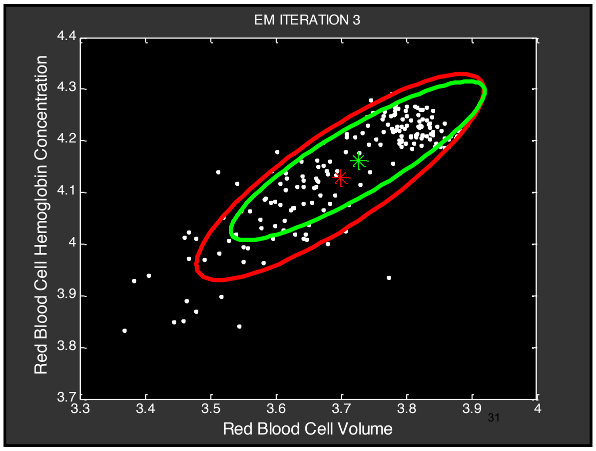
28

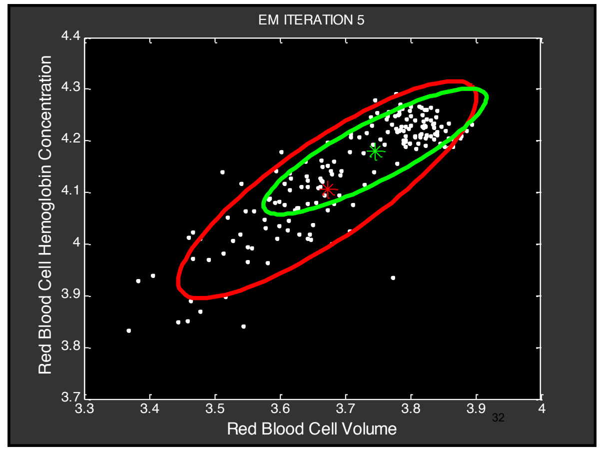


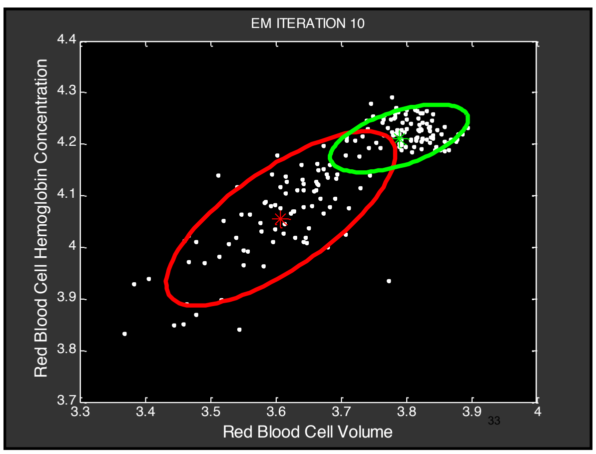
29

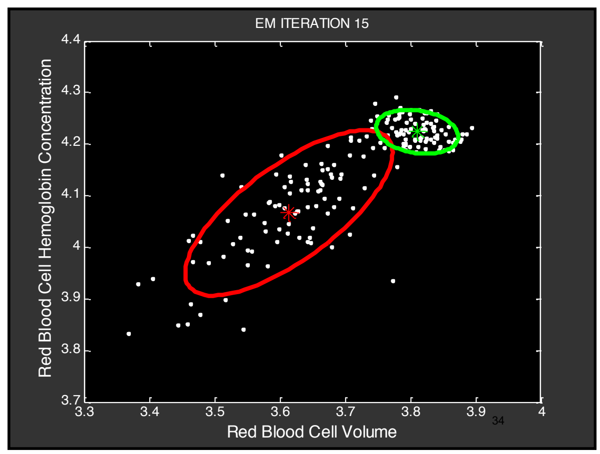


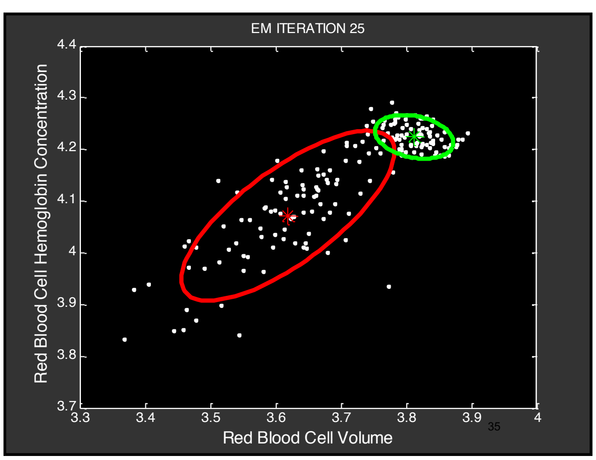
30

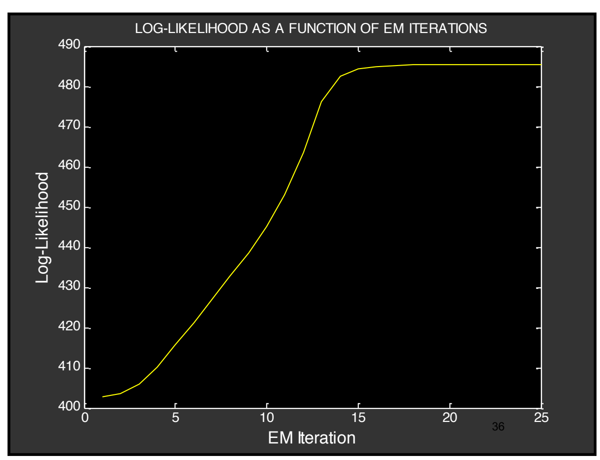

EM ITERATION 3
Red Blood Cell Hemoglobin Concentration
Red Blood Cell Volume
31

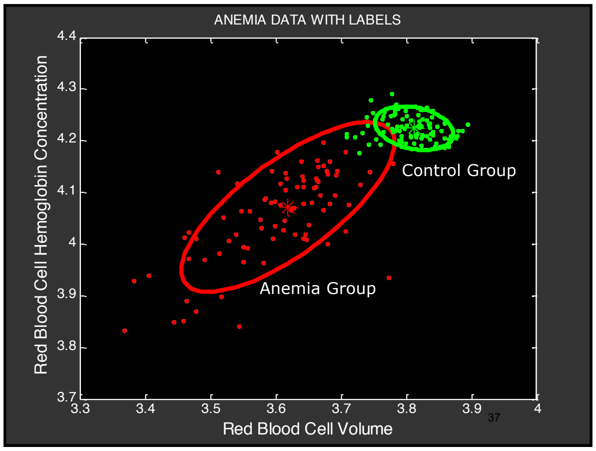

EM ITERATION 5
Red Blood Cell Hemoglobin Concentration
Red Blood Cell Volume
32


EM ITERATION 10
Red Blood Cell Hemoglobin Concentration
Red Blood Cell Volume
33

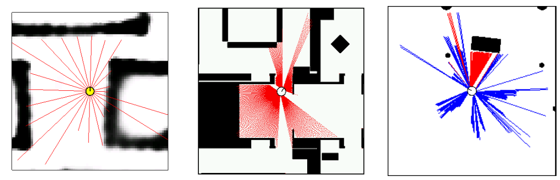

EM ITERATION 15
Red Blood Cell Hemoglobin Concentration
Red Blood Cell Volume
34


EM ITERATION 25
Red Blood Cell Hemoglobin Concentration
Red Blood Cell Volume
35


LOG-LIKELIHOOD AS A FUNCTION OF EM ITERATIONS
Log-Likelihood
EM Iteration
36



Beam-based Sensor Model



$$P(z \mid x, m) = \prod_{k=1}^{K} P(z_k \mid x, m)$$

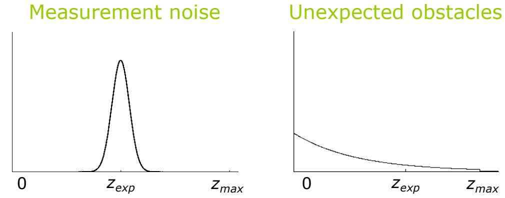
Proximity Measurement

- Measurement can be caused by …
  - a known obstacle.
  - cross-talk.
  - an unexpected obstacle (people, furniture, …).
  - missing all obstacles (total reflection, glass, …).
- Noise is due to uncertainty …
  - in measuring distance to known obstacle.
  - in position of known obstacles.
  - in position of additional obstacles.
  - whether obstacle is missed.

Beam-based Proximity Model

Measurement noise

Unexpected obstacles



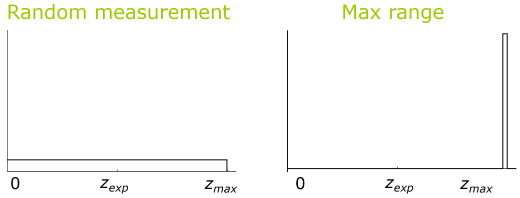
$$P_{hit}(z \mid x, m) = \eta \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(z-z_{exp})^2}{\sigma^2}}$$

$$P_{unexp}(z \mid x, m) = \eta \, \lambda \, e^{-\lambda z}$$

Beam-based Proximity Model

Random measurement

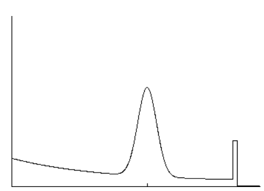
Max range



$$P_{rand}(z \mid x, m) = \eta \frac{1}{z_{max}}$$

$$P_{max}(z \mid x, m) = \eta \frac{1}{z_{small}}$$

Mixture Density

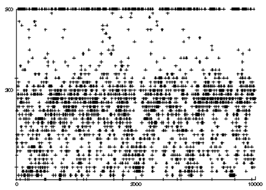


$$P(z \mid x, m) = \begin{pmatrix} \alpha_{hit} \\ \alpha_{unexp} \\ \alpha_{max} \\ \alpha_{rand} \end{pmatrix}^T \cdot \begin{pmatrix} P_{hit}(z \mid x, m) \\ P_{unexp}(z \mid x, m) \\ P_{max}(z \mid x, m) \\ P_{rand}(z \mid x, m) \end{pmatrix}$$

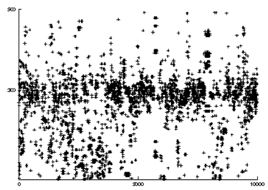
How can we determine the model parameters?

Raw Sensor Data

Measured distances for expected distance of 300 cm.



Sonar



Laser

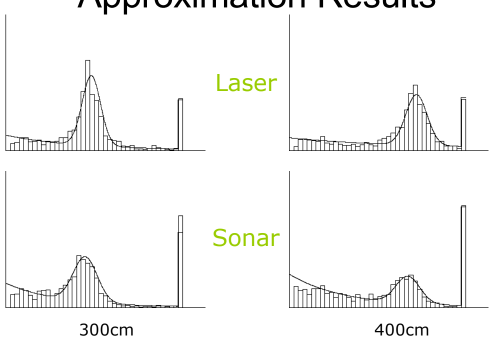
43

Approximation

- Maximize log likelihood of the data z:
  $P(z \mid z_{\exp})$
- Search parameter space.
- EM to find mixture parameters
  - Assign measurements to densities.
  - Estimate densities using assignments.
  - Reassign measurements.

44

Approximation Results



45

What if we *don't* know structure?

46

Learning The Structure of Bayesian Networks

- Search through the space…
  - of possible network structures!
  - (for now, assume we observe all variables)
- For each structure, learn parameters
- Pick the one that fits observed data best
  - Caveat – won't we end up fully connected????

When scoring, add a penalty for model complexity:
Bayesian Information Criterion (BIC)
BIC = -log(P(D | BN)) + penalty
Penalty = ½ (# parameters) log (# data points)

Learning The Structure of Bayesian Networks

- Search through the space
- For each structure, learn parameters
- Pick the one that fits observed data best
  - Penalize complex models
- Problem?

Exponential number of networks!
And we need to learn parameters for each!
Exhaustive search out of the question!

48

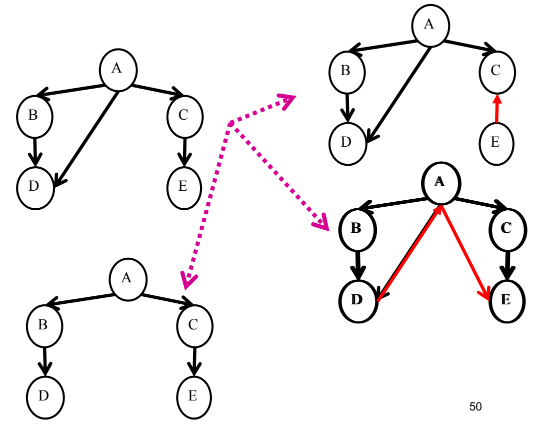
Structure Learning as Search

- Local Search

1. Start with some network structure
2. Try to make a change
   (add or delete or reverse edge)
3. See if the new network is any better

- What should the initial state be?
  - Uniform prior over random networks?
  - Based on prior knowledge?
  - Empty network?
- How do we evaluate networks?

49



50