

Note to other teachers and users of these slides. Andrew would be delighted if you found this source material useful in giving your own lectures. Feel free to use these slides verbatim, or to modify them to fit your own needs. PowerPoint originals are available. If you make use of a significant portion of these slides in your own lecture, please include this message, or the following link to the source repository of Andrew's tutorials: <http://www.cs.cmu.edu/~awm/tutorials> . Comments and corrections gratefully received.

Clustering with Gaussian Mixtures

Andrew W. Moore
Professor
School of Computer Science
Carnegie Mellon University

www.cs.cmu.edu/~awm

awm@cs.cmu.edu

412-268-7599

Copyright © 2001, 2004, Andrew W. Moore

Unsupervised Learning

- You walk into a bar.
A stranger approaches and tells you:
"I've got data from k classes. Each class produces observations with a normal distribution and variance $\sigma^2 I$. Standard simple multivariate gaussian assumptions. I can tell you all the $P(w_i)$'s."
- So far, looks straightforward.
"I need a maximum likelihood estimate of the μ 's."
- No problem:
"There's just one thing. None of the data are labeled. I have datapoints, but I don't know what class they're from (any of them!)"
- Uh oh!!

Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 2

Gaussian Bayes Classifier Reminder

$$P(y = i | \mathbf{x}) = \frac{p(\mathbf{x} | y = i)P(y = i)}{p(\mathbf{x})}$$

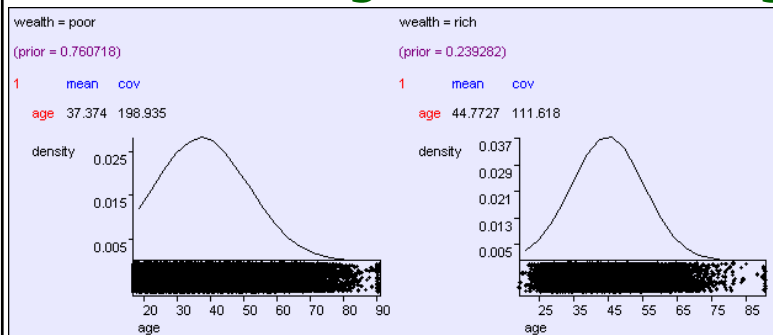
$$P(y = i | \mathbf{x}) = \frac{\frac{1}{(2\pi)^{m/2} \|\Sigma_i\|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}_k - \boldsymbol{\mu}_i)^T \Sigma_i (\mathbf{x}_k - \boldsymbol{\mu}_i)\right] p_i}{p(\mathbf{x})}$$

How do we deal with that?

Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 3

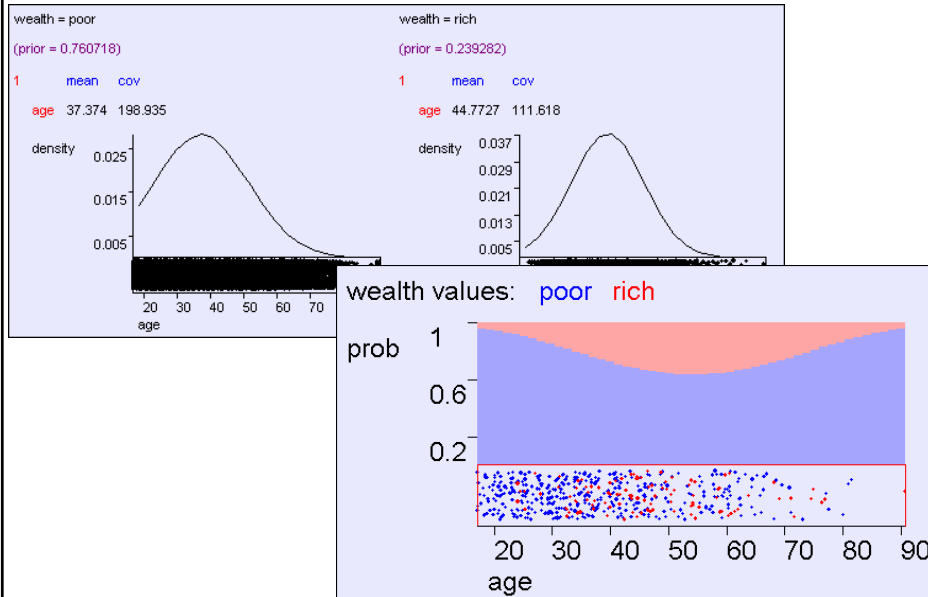
Predicting wealth from age



Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 4

Predicting wealth from age

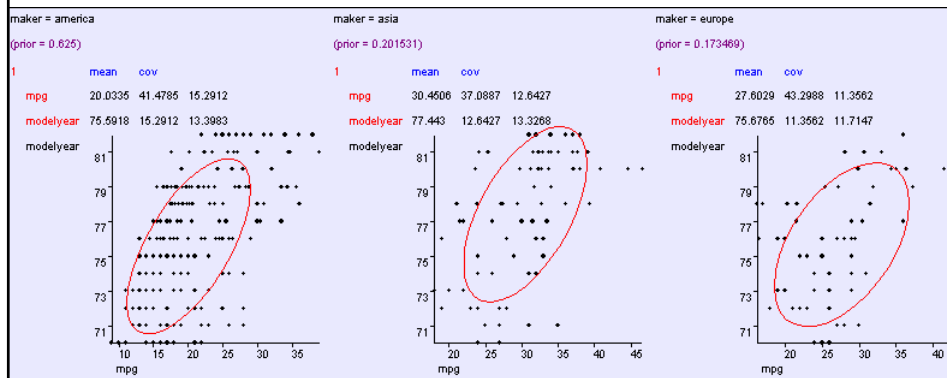


Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 5

Learning modelyear , mpg ---> maker

$$\Sigma = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{12} & \sigma_{22}^2 & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1m} & \sigma_{2m} & \cdots & \sigma_{mm}^2 \end{pmatrix}$$

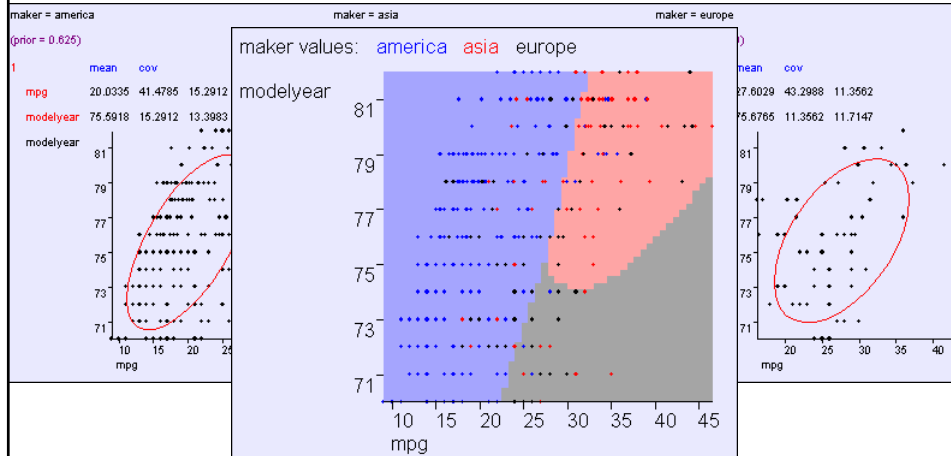


Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 6

General: $O(m^2)$ parameters

$$\Sigma = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{12} & \sigma_{22}^2 & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1m} & \sigma_{2m} & \cdots & \sigma_{mm}^2 \end{pmatrix}$$

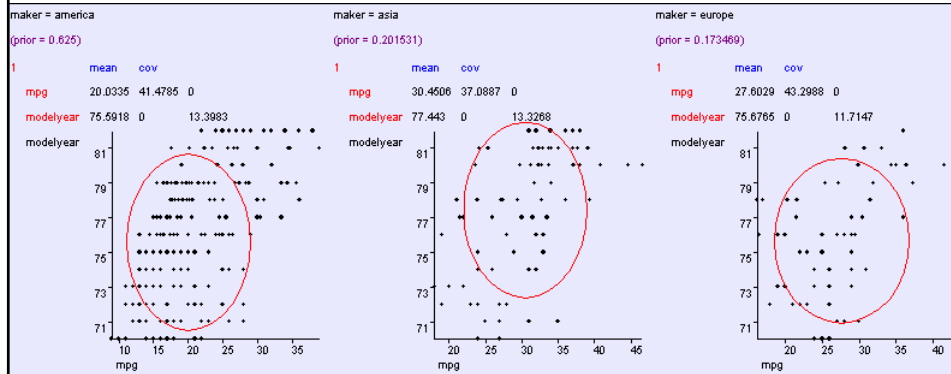


Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 7

Aligned: $O(m)$ parameters

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \sigma_2^2 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \sigma_3^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma_{m-1}^2 & 0 \\ 0 & 0 & 0 & \cdots & 0 & \sigma_m^2 \end{pmatrix}$$

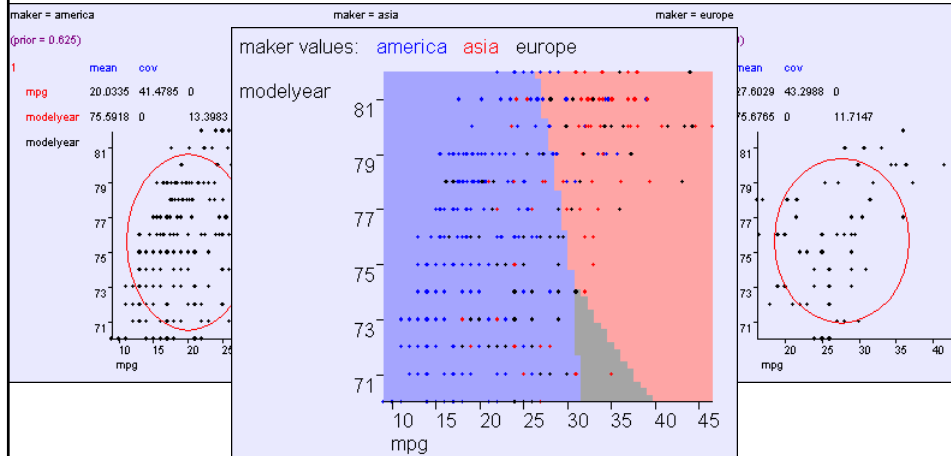


Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 8

Aligned: $O(m)$ parameters

$$\Sigma = \begin{pmatrix} \sigma^2_1 & 0 & 0 & \dots & 0 & 0 \\ 0 & \sigma^2_2 & 0 & \dots & 0 & 0 \\ 0 & 0 & \sigma^2_3 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \sigma^2_{m-1} & 0 \\ 0 & 0 & 0 & \dots & 0 & \sigma^2_m \end{pmatrix}$$

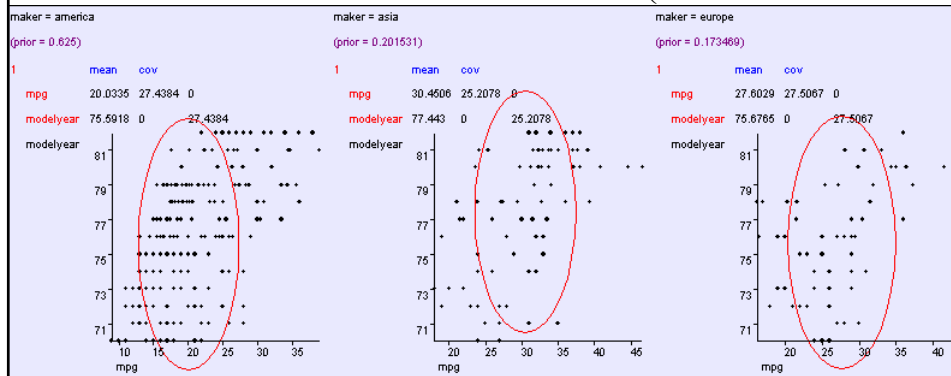


Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 9

Spherical: $O(1)$ cov parameters

$$\Sigma = \begin{pmatrix} \sigma^2 & 0 & 0 & \dots & 0 & 0 \\ 0 & \sigma^2 & 0 & \dots & 0 & 0 \\ 0 & 0 & \sigma^2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \sigma^2 & 0 \\ 0 & 0 & 0 & \dots & 0 & \sigma^2 \end{pmatrix}$$

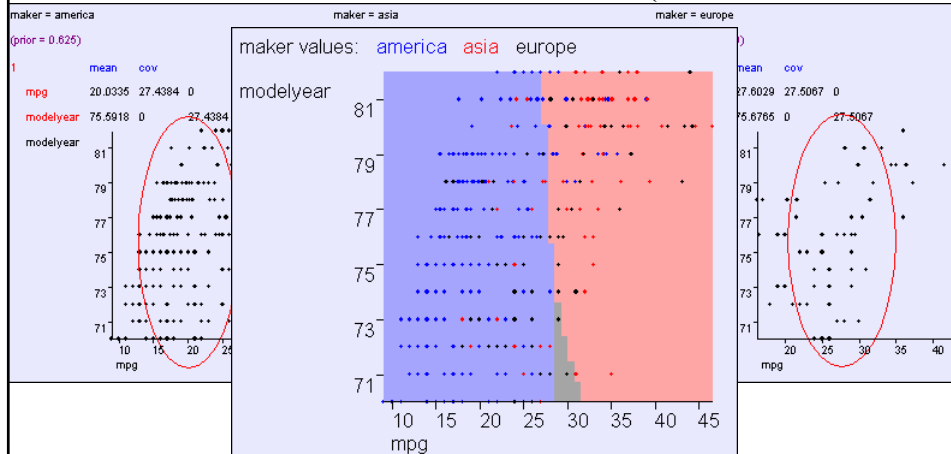


Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 10

Spherical: $O(1)$ cov parameters

$$\Sigma = \begin{pmatrix} \sigma^2 & 0 & 0 & \dots & 0 & 0 \\ 0 & \sigma^2 & 0 & \dots & 0 & 0 \\ 0 & 0 & \sigma^2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \sigma^2 & 0 \\ 0 & 0 & 0 & \dots & 0 & \sigma^2 \end{pmatrix}$$



Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 11

Making a Classifier from a Density Estimator

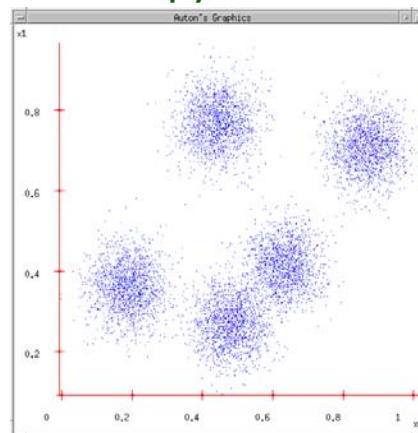
		Categorical inputs only	Real-valued inputs only	Mixed Real / Cat okay
Inputs	Classifier → Predict category	Joint BC Naïve BC	Gauss BC	Dec Tree
Inputs	Density Estimator → Prob- ability	Joint DE Naïve DE	Gauss DE	
Inputs	Regressor → Predict real no.			

Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 12

Next... back to Density Estimation

What if we want to do density estimation with multimodal or clumpy data?

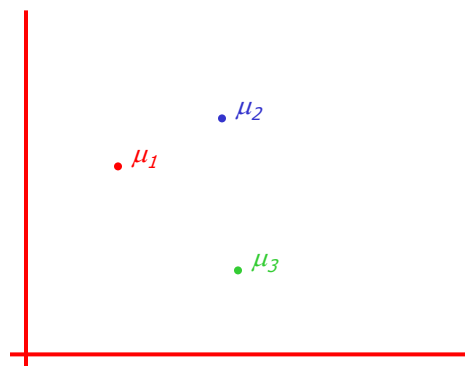


Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 13

The GMM assumption

- There are k components. The i 'th component is called ω_i
- Component ω_i has an associated mean vector μ_i



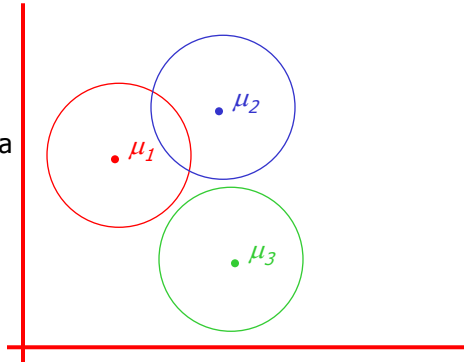
Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 14

The GMM assumption

- There are k components. The i 'th component is called ω_i
- Component ω_i has an associated mean vector μ_i
- Each component generates data from a Gaussian with mean μ_i and covariance matrix $\sigma^2 I$

Assume that each datapoint is generated according to the following recipe:



Copyright © 2001, 2004, Andrew W. Moore

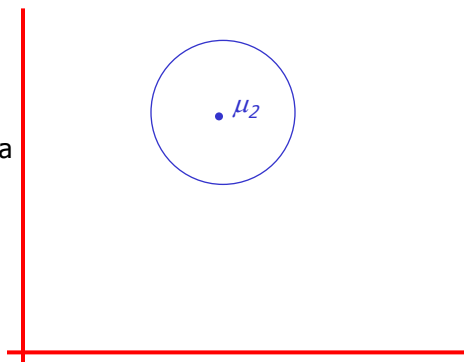
Clustering with Gaussian Mixtures: Slide 15

The GMM assumption

- There are k components. The i 'th component is called ω_i
- Component ω_i has an associated mean vector μ_i
- Each component generates data from a Gaussian with mean μ_i and covariance matrix $\sigma^2 I$

Assume that each datapoint is generated according to the following recipe:

1. Pick a component at random. Choose component i with probability $P(\omega_i)$.



Copyright © 2001, 2004, Andrew W. Moore

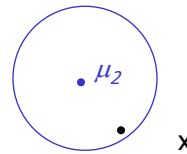
Clustering with Gaussian Mixtures: Slide 16

The GMM assumption

- There are k components. The i 'th component is called ω_i
- Component ω_i has an associated mean vector μ_i
- Each component generates data from a Gaussian with mean μ_i and covariance matrix $\sigma^2 I$

Assume that each datapoint is generated according to the following recipe:

1. Pick a component at random. Choose component i with probability $P(\omega_i)$.
2. Datapoint $\sim N(\mu_i, \sigma^2 I)$



Copyright © 2001, 2004, Andrew W. Moore

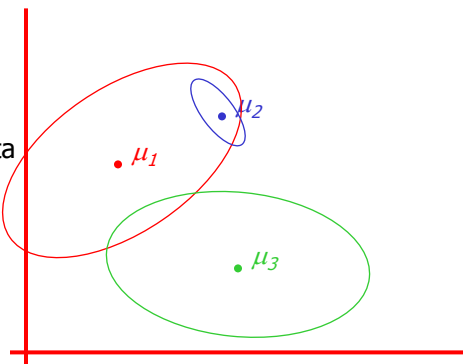
Clustering with Gaussian Mixtures: Slide 17

The General GMM assumption

- There are k components. The i 'th component is called ω_i
- Component ω_i has an associated mean vector μ_i
- Each component generates data from a Gaussian with mean μ_i and covariance matrix Σ_i

Assume that each datapoint is generated according to the following recipe:

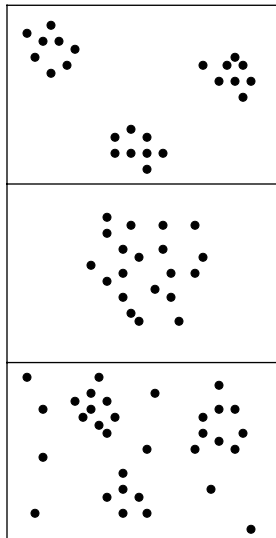
1. Pick a component at random. Choose component i with probability $P(\omega_i)$.
2. Datapoint $\sim N(\mu_i, \Sigma_i)$



Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 18

Unsupervised Learning: not as hard as it looks



Sometimes easy

Sometimes impossible

and sometimes
in between

*IN CASE YOU'RE
WONDERING WHAT
THESE DIAGRAMS ARE,
THEY SHOW 2-d
UNLABELED DATA (X
VECTORS)
DISTRIBUTED IN 2-d
SPACE. THE TOP ONE
HAS THREE VERY
CLEAR GAUSSIAN
CENTERS*

Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 19

Computing likelihoods in unsupervised case

We have $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$

We know $P(w_1) P(w_2) \dots P(w_k)$

We know σ

$P(\mathbf{x} | w_i, \boldsymbol{\mu}_i, \dots, \boldsymbol{\mu}_k) =$ Prob that an observation from class w_i would have value $\mathbf{x}$ given class means $\boldsymbol{\mu}_1 \dots \boldsymbol{\mu}_k$

Can we write an expression for that?

Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 20

likelihoods in unsupervised case

We have $x_1 x_2 \dots x_n$

We have $P(w_1) \dots P(w_k)$. We have σ .

We can define, for any x , $P(x|w_i, \mu_1, \mu_2 \dots \mu_k)$

Can we define $P(x | \mu_1, \mu_2 \dots \mu_k)$?

Can we define $P(x_1, x_2 \dots x_n | \mu_1, \mu_2 \dots \mu_k)$?

[YES, IF WE ASSUME THE x_i 'S WERE DRAWN INDEPENDENTLY]

Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 21

Unsupervised Learning: Mediumly Good News

We now have a procedure s.t. if you give me a guess at $\mu_1, \mu_2 \dots \mu_k$,
I can tell you the prob of the unlabeled data given those μ 's.

Suppose x 's are 1-dimensional.

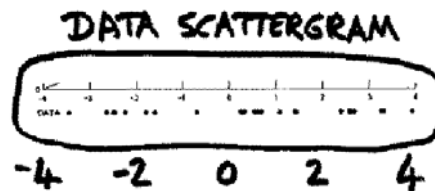
(From Duda and Hart)

There are two classes; w_1 and w_2

$P(w_1) = 1/3$ $P(w_2) = 2/3$ $\sigma = 1$.

There are 25 unlabeled datapoints

$x_1 = 0.608$
 $x_2 = -1.590$
 $x_3 = 0.235$
 $x_4 = 3.949$
 $\vdots$
 $x_{25} = -0.712$

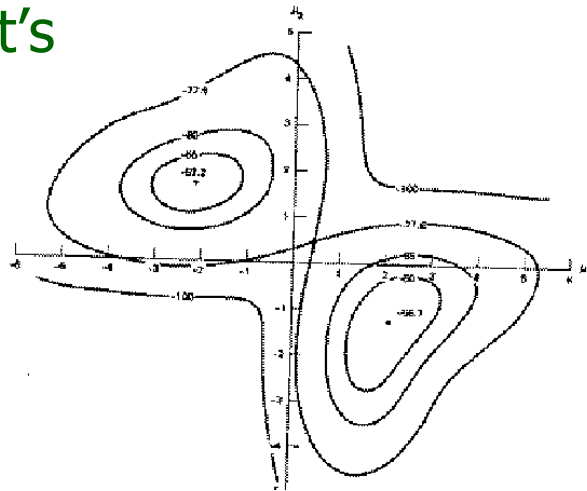


Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 22

Duda & Hart's Example

Graph of
 $\log P(x_1, x_2 \dots x_{25} | \mu_1, \mu_2)$
 against $\mu_1 (\rightarrow)$ and $\mu_2 (\uparrow)$



Max likelihood = $(\mu_1 = -2.13, \mu_2 = 1.668)$

Local minimum, but very close to global at $(\mu_1 = 2.085, \mu_2 = -1.257)^*$

* corresponds to switching $w_1 + w_2$.

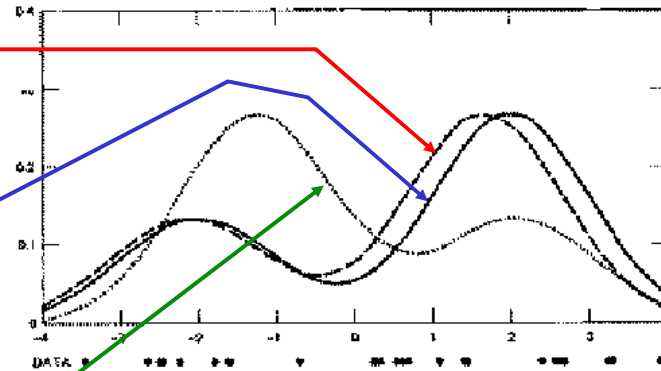
Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 23

Duda & Hart's Example

We can graph the
 prob. dist. function
 of data given our
 μ_1 and μ_2
 estimates.

We can also graph the
 true function from
 which the data was
 randomly generated.



- They are close. Good.
- The 2nd solution tries to put the "2/3" hump where the "1/3" hump should go, and vice versa.
- In this example unsupervised is almost as good as supervised. If the $x_1 \dots x_{25}$ are given the class which was used to learn them, then the results are $(\mu_1 = -2.176, \mu_2 = 1.684)$. Unsupervised got $(\mu_1 = -2.13, \mu_2 = 1.668)$.

Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 24

Finding the max likelihood $\mu_1, \mu_2 \dots \mu_k$

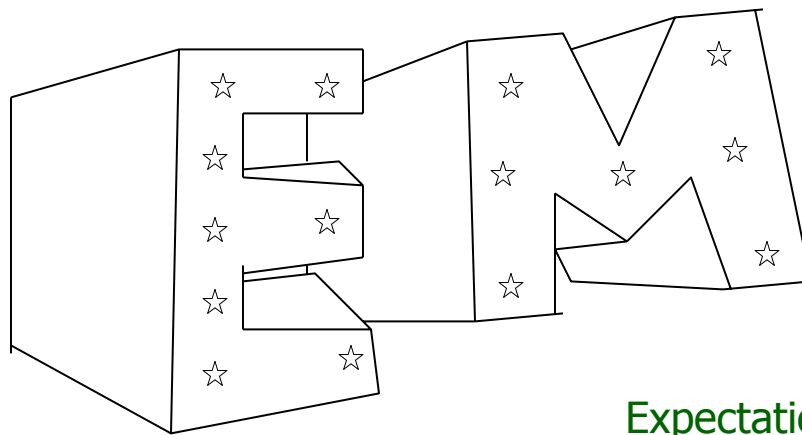
We can compute $P(\text{data} \mid \mu_1, \mu_2 \dots \mu_k)$

How do we find the μ_j 's which give max. likelihood?

- The normal max likelihood trick:
Set $\frac{\partial}{\partial \mu_j} \log \text{Prob}(\dots) = 0$
and solve for μ_j 's.
Here you get non-linear non-analytically-solvable equations
- Use gradient descent
Slow but doable
- Use a much faster, cuter, and recently very popular method...

Copyright © 2001, 2004, Andrew W. Moore

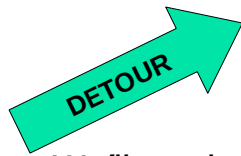
Clustering with Gaussian Mixtures: Slide 25



Expectation
Maximalization

Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 26



The E.M. Algorithm

- We'll get back to unsupervised learning soon.
- But now we'll look at an even simpler case with hidden information.
- The EM algorithm
 - Can do trivial things, such as the contents of the next few slides.
 - An excellent way of doing our unsupervised learning problem, as we'll see.
 - Many, many other uses, including inference of Hidden Markov Models (future lecture).

Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 27

Silly Example

Let events be "grades in a class"

w_1	= Gets an A	$P(A) = 1/2$
w_2	= Gets a B	$P(B) = \mu$
w_3	= Gets a C	$P(C) = 2\mu$
w_4	= Gets a D	$P(D) = 1/2 - 3\mu$

(Note $0 \leq \mu \leq 1/6$)

Assume we want to estimate μ from data. In a given class there were

a	A's
b	B's
c	C's
d	D's

What's the maximum likelihood estimate of μ given a, b, c, d ?

Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 28

Silly Example

Let events be "grades in a class"

w_1 = Gets an A

w_2 = Gets a B

w_3 = Gets a C

w_4 = Gets a D

$P(A) = 1/2$

$P(B) = \mu$

$P(C) = 2\mu$

$P(D) = 1/2 - 3\mu$

(Note $0 \leq \mu \leq 1/6$)

Assume we want to estimate μ from data. In a given class there were

a A's

b B's

c C's

d D's

What's the maximum likelihood estimate of μ given a,b,c,d ?

Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 29

Trivial Statistics

$P(A) = 1/2$ $P(B) = \mu$ $P(C) = 2\mu$ $P(D) = 1/2 - 3\mu$

$P(a,b,c,d | \mu) = K(1/2)^a(\mu)^b(2\mu)^c(1/2 - 3\mu)^d$

$\log P(a,b,c,d | \mu) = \log K + a \log 1/2 + b \log \mu + c \log 2\mu + d \log (1/2 - 3\mu)$

FOR MAX LIKE μ , SET $\frac{\partial \log P}{\partial \mu} = 0$

$$\frac{\partial \log P}{\partial \mu} = \frac{b}{\mu} + \frac{2c}{2\mu} - \frac{3d}{1/2 - 3\mu} = 0$$

$$\text{Gives max like } \mu = \frac{b + c}{6(b + c + d)}$$

So if class got

A	B	C	D
14	6	9	10

$$\text{Max like } \mu = \frac{1}{10}$$

Boring, but true!

Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 30

Same Problem with Hidden Information

Someone tells us that

Number of High grades (A's + B's) = h

Number of C's = c

Number of D's = d

What is the max. like estimate of μ now?

REMEMBER

$$P(A) = 1/2$$

$$P(B) = \mu$$

$$P(C) = 2\mu$$

$$P(D) = 1/2 - 3\mu$$

Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 31

Same Problem with Hidden Information

Someone tells us that

Number of High grades (A's + B's) = h

Number of C's = c

Number of D's = d

What is the max. like estimate of μ now?

We can answer this question circularly:

EXPECTATION

If we know the value of μ we could compute the expected value of a and b

Since the ratio $a:b$ should be the same as the ratio $1/2 : \mu$

$$a = \frac{1/2}{1/2 + \mu} h \quad b = \frac{\mu}{1/2 + \mu} h$$

MAXIMIZATION

If we know the expected values of a and b we could compute the maximum likelihood value of μ

$$\mu = \frac{b + c}{6(b + c + d)}$$

Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 32

E.M. for our Trivial Problem

We begin with a guess for μ

We iterate between EXPECTATION and MAXIMALIZATION to improve our estimates of μ and a and b .

REMEMBER

$$P(A) = 1/2$$

$$P(B) = \mu$$

$$P(C) = 2\mu$$

$$P(D) = 1/2 - 3\mu$$

Define $\mu(t)$ the estimate of μ on the t 'th iteration

$b(t)$ the estimate of b on t 'th iteration

$\mu(0)$ = initial guess

$$b(t) = \frac{\mu(t)h}{1/2 + \mu(t)} = E[b | \mu(t)]$$

E-step

$$\mu(t+1) = \frac{b(t) + c}{6(b(t) + c + d)}$$

M-step

= max like est of μ given $b(t)$

Continue iterating until converged.

Good news: Converging to local optimum is assured.

Bad news: I said "local" optimum.

Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 33

E.M. Convergence

- Convergence proof based on fact that Prob(data | μ) must increase or remain same between each iteration [NOT OBVIOUS]
 - But it can never exceed 1 [OBVIOUS]
- So it must therefore converge [OBVIOUS]

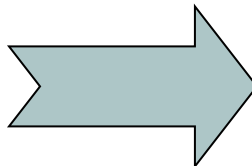
In our example,
suppose we had

$$h = 20$$

$$c = 10$$

$$d = 10$$

$$\mu(0) = 0$$



Convergence is generally linear: error decreases by a constant factor each time step.

t	$\mu(t)$	$b(t)$
0	0	0
1	0.0833	2.857
2	0.0937	3.158
3	0.0947	3.185
4	0.0948	3.187
5	0.0948	3.187
6	0.0948	3.187

Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 34

Back to Unsupervised Learning of GMMs

Remember:

We have unlabeled data $x_1 x_2 \dots x_R$
 We know there are k classes
 We know $P(w_1) P(w_2) P(w_3) \dots P(w_k)$
 We don't know $\mu_1 \mu_2 \dots \mu_k$

We can write $P(\text{data} \mid \mu_1 \dots \mu_k)$

$$\begin{aligned} &= P(x_1 \dots x_R \mid \mu_1 \dots \mu_k) \\ &= \prod_{i=1}^R P(x_i \mid \mu_1 \dots \mu_k) \\ &= \prod_{i=1}^R \sum_{j=1}^k P(x_i \mid w_j, \mu_1 \dots \mu_k) P(w_j) \\ &= \prod_{i=1}^R \sum_{j=1}^k K \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu_j)^2\right) P(w_j) \end{aligned}$$

Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 35

E.M. for GMMs

For Max likelihood we know $\frac{\partial}{\partial \mu_i} \log \text{Pr ob}(\text{data} \mid \mu_1 \dots \mu_k) = 0$

Some wild'n'crazy algebra turns this into : "For Max likelihood, for each j ,

$$\mu_j = \frac{\sum_{i=1}^R P(w_j \mid x_i, \mu_1 \dots \mu_k) x_i}{\sum_{i=1}^R P(w_j \mid x_i, \mu_1 \dots \mu_k)}$$

See

<http://www.cs.cmu.edu/~awm/doc/gmm-algebra.pdf>

This is n nonlinear equations in μ_j 's."

If, for each x_i we knew that for each w_j the prob that μ_j was in class w_j is $P(w_j \mid x_i, \mu_1 \dots \mu_k)$ Then... we would easily compute μ_j .

If we knew each μ_j then we could easily compute $P(w_j \mid x_i, \mu_1 \dots \mu_k)$ for each w_j and x_i .

...I feel an EM experience coming on!!

Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 36

E.M. for GMMs

Iterate. On the t th iteration let our estimates be

$$\lambda_t = \{ \mu_1(t), \mu_2(t) \dots \mu_d(t) \}$$

E-step

Compute "expected" classes of all datapoints for each class

Just evaluate
a Gaussian at
 x_k

$$P(w_i | x_k, \lambda_t) = \frac{p(x_k | w_i, \lambda_t) P(w_i | \lambda_t)}{p(x_k | \lambda_t)} = \frac{p(x_k | w_i, \mu_i(t), \sigma^2 \mathbf{I}) p_i(t)}{\sum_{j=1}^c p(x_k | w_j, \mu_j(t), \sigma^2 \mathbf{I}) p_j(t)}$$

M-step.

Compute Max. like μ given our data's class membership distributions

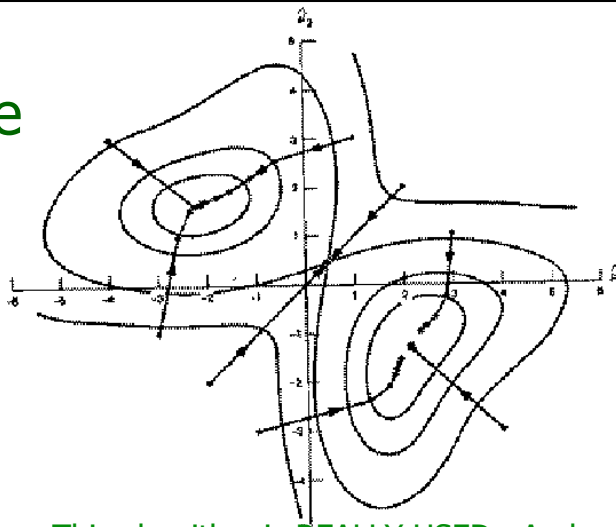
$$\mu_i(t+1) = \frac{\sum_k P(w_i | x_k, \lambda_t) x_k}{\sum_k P(w_i | x_k, \lambda_t)}$$

Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 37

E.M. Convergence

- Your lecturer will (unless out of time) give you a nice intuitive explanation of why this rule works.
- As with all EM procedures, convergence to a local optimum guaranteed.



- This algorithm is REALLY USED. And in high dimensional state spaces, too. E.G. Vector Quantization for Speech Data

Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 38

E.M. for General GMMs

Iterate. On the t 'th iteration let our estimates be

$$\lambda_t = \{ \mu_1(t), \mu_2(t) \dots \mu_d(t), \Sigma_1(t), \Sigma_2(t) \dots \Sigma_d(t), p_1(t), p_2(t) \dots p_d(t) \}$$

$p_i(t)$ is shorthand for estimate of $P(\omega_i)$ on t 'th iteration

E-step

Compute "expected" classes of all datapoints for each class

Just evaluate a Gaussian at x_k

$$P(w_i | x_k, \lambda_t) = \frac{p(x_k | w_i, \lambda_t) P(w_i | \lambda_t)}{p(x_k | \lambda_t)} = \frac{p(x_k | w_i, \mu_i(t), \Sigma_i(t)) p_i(t)}{\sum_{j=1}^c p(x_k | w_j, \mu_j(t), \Sigma_j(t)) p_j(t)}$$

M-step.

Compute Max. like μ given our data's class membership distributions

$$\mu_i(t+1) = \frac{\sum_k P(w_i | x_k, \lambda_t) x_k}{\sum_k P(w_i | x_k, \lambda_t)} \quad \Sigma_i(t+1) = \frac{\sum_k P(w_i | x_k, \lambda_t) [x_k - \mu_i(t+1)][x_k - \mu_i(t+1)]^T}{\sum_k P(w_i | x_k, \lambda_t)}$$

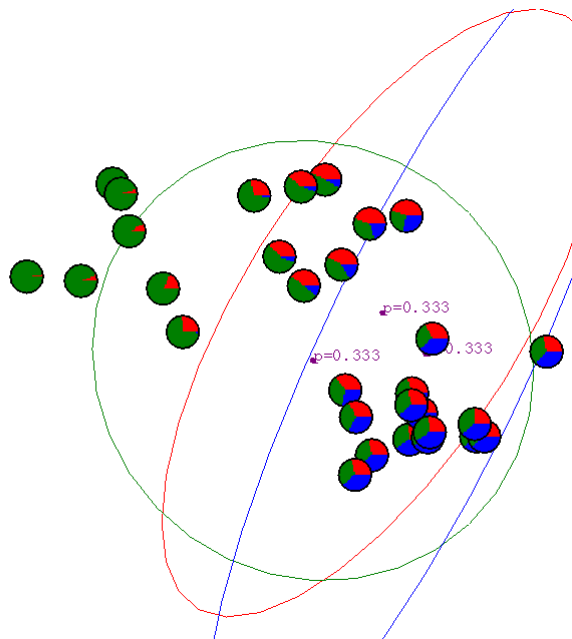
$$p_i(t+1) = \frac{\sum_k P(w_i | x_k, \lambda_t)}{R}$$

$R = \text{\#records}$

Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 39

Gaussian Mixture Example: Start

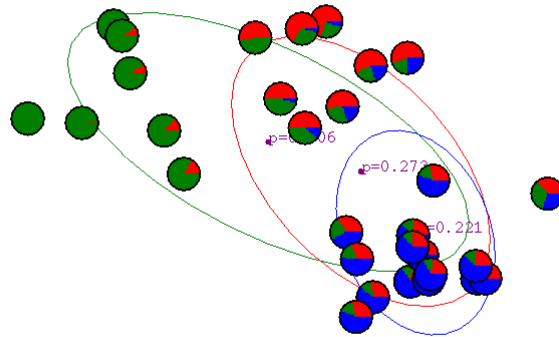


Advance apologies: in Black and White this example will be incomprehensible

Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 40

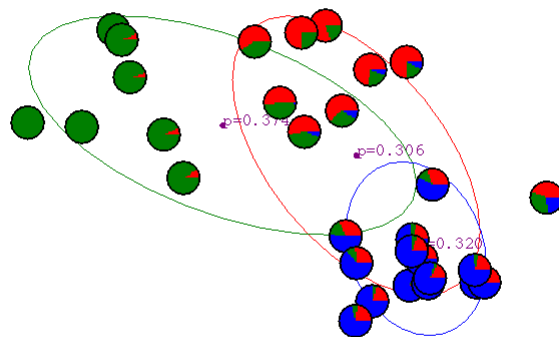
After first
iteration



Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 41

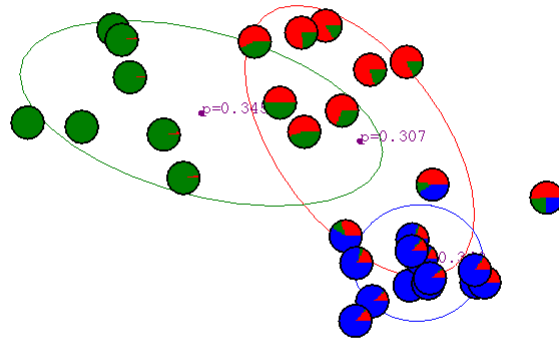
After 2nd
iteration



Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 42

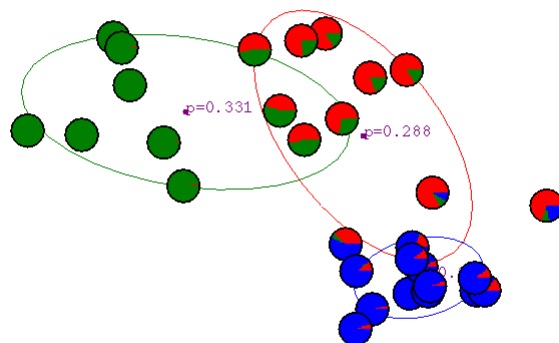
After 3rd
iteration



Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 43

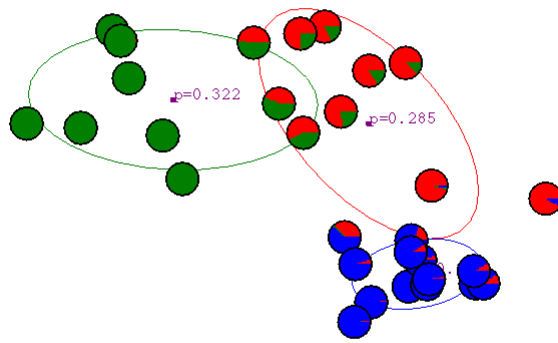
After 4th
iteration



Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 44

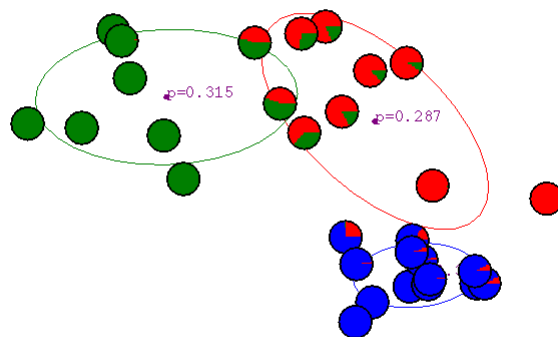
After 5th
iteration



Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 45

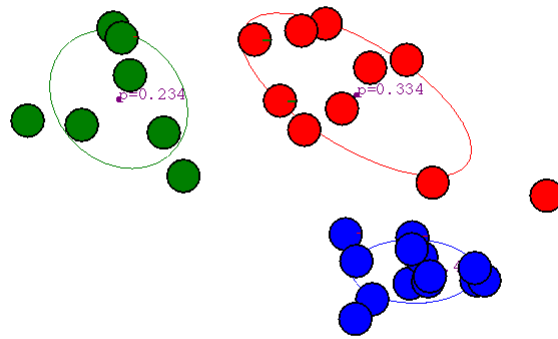
After 6th
iteration



Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 46

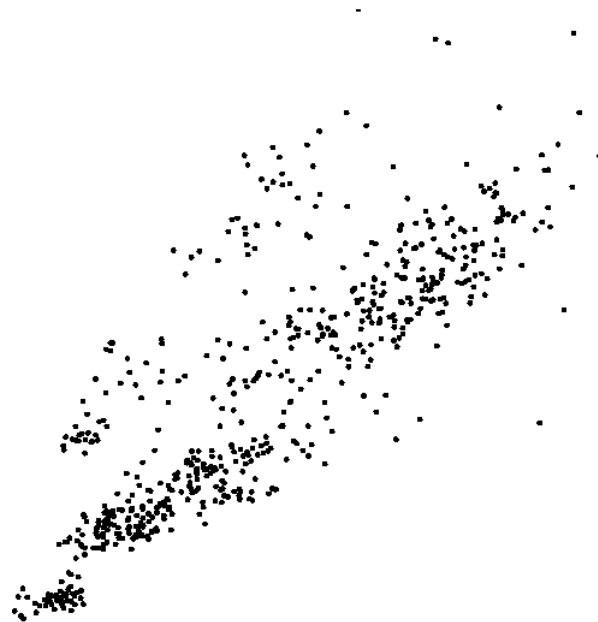
After 20th
iteration



Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 47

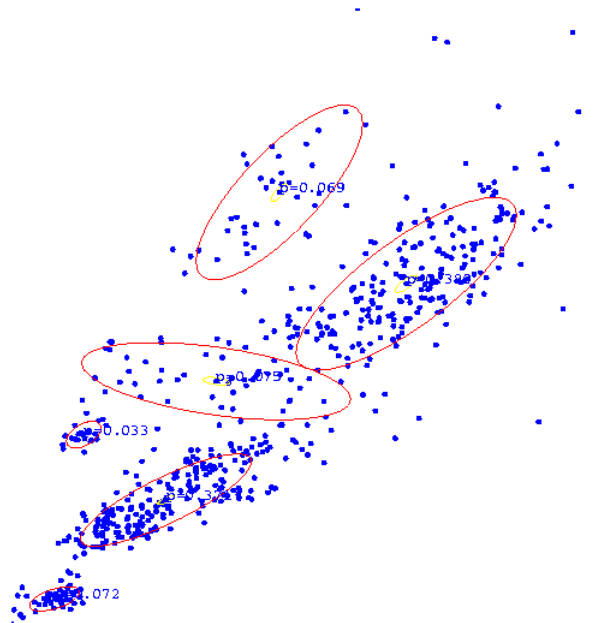
Some Bio
Assay
data



Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 48

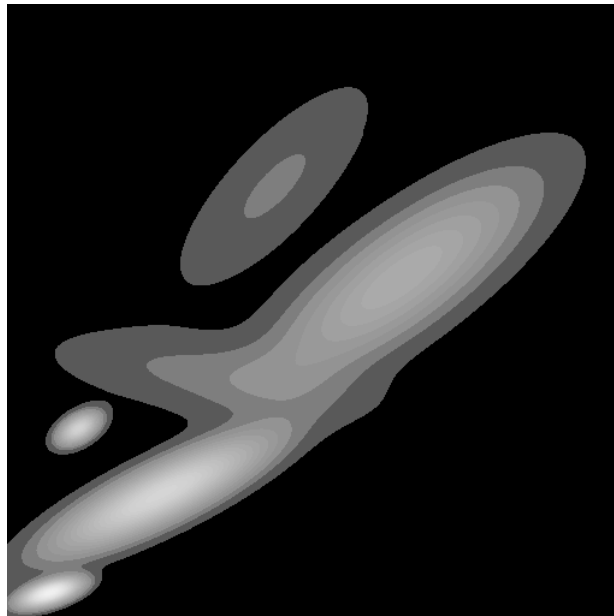
GMM clustering of the assay data



Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 49

Resulting Density Estimator



Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 50

Where are we now?

Inputs	Inference Engine Learn	$P(E_1 E_2)$	Joint DE, Bayes Net Structure Learning
Inputs	Classifier	Predict category	Dec Tree, Sigmoid Perceptron, Sigmoid N.Net, Gauss/Joint BC, Gauss Naïve BC, N.Neigh, Bayes Net Based BC, Cascade Correlation
Inputs	Density Estimator	Probability	Joint DE, Naïve DE, Gauss/Joint DE, Gauss Naïve DE, Bayes Net Structure Learning, GMMs
Inputs	Regressor	Predict real no.	Linear Regression, Polynomial Regression, Perceptron, Neural Net, N.Neigh, Kernel, LWR, RBFs, Robust Regression, Cascade Correlation, Regression Trees, GMDH, Multilinear Interp, MARS

Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 51

The old trick...

Inputs	Inference Engine Learn	$P(E_1 E_2)$	Joint DE, Bayes Net Structure Learning
Inputs	Classifier	Predict category	Dec Tree, Sigmoid Perceptron, Sigmoid N.Net, Gauss/Joint BC, Gauss Naïve BC, N.Neigh, Bayes Net Based BC, Cascade Correlation, GMM-BC
Inputs	Density Estimator	Probability	Joint DE, Naïve DE, Gauss/Joint DE, Gauss Naïve DE, Bayes Net Structure Learning, GMMs
Inputs	Regressor	Predict real no.	Linear Regression, Polynomial Regression, Perceptron, Neural Net, N.Neigh, Kernel, LWR, RBFs, Robust Regression, Cascade Correlation, Regression Trees, GMDH, Multilinear Interp, MARS

Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 52

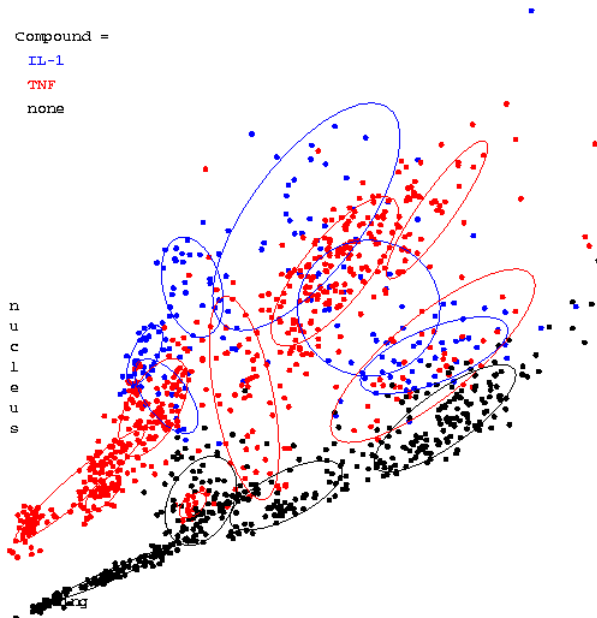
Three classes of assay

(each learned with its own mixture model)

(Sorry, this will again be semi-useless in black and white)

Compound =
IL-1
TNF
none

n
u
c
l
e
u
s



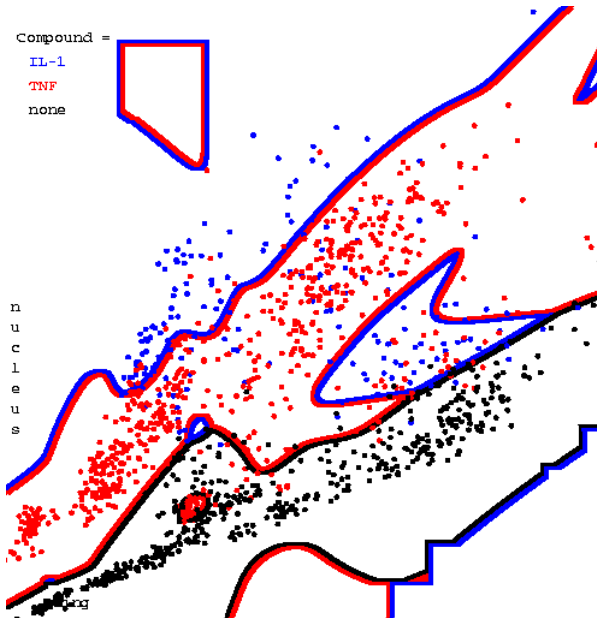
Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 53

Resulting Bayes Classifier

Compound =
IL-1
TNF
none

n
u
c
l
e
u
s



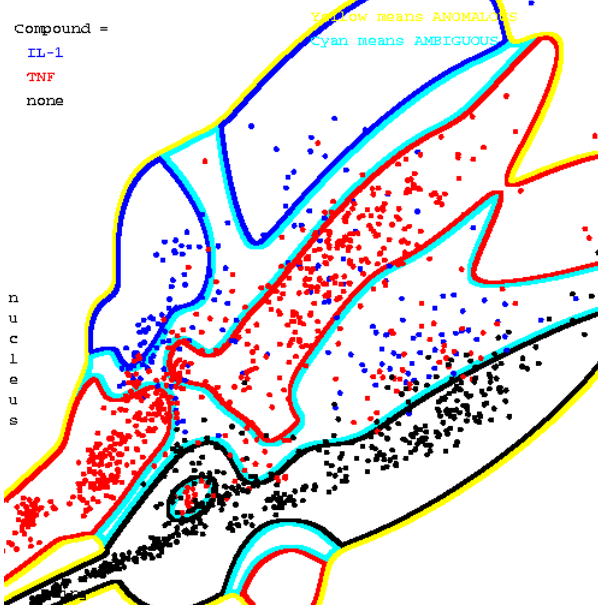
Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 54

Resulting Bayes Classifier, using posterior probabilities to alert about ambiguity and anomalousness

Yellow means anomalous

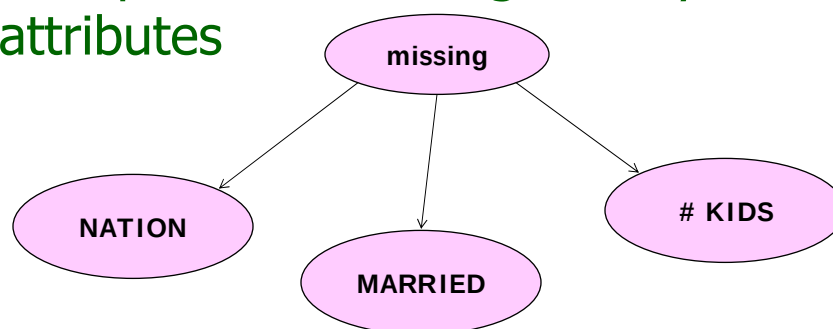
Cyan means ambiguous



Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 55

Unsupervised learning with symbolic attributes



It's just a "learning Bayes net with known structure but hidden values" problem.

Can use Gradient Descent.

EASY, fun exercise to do an EM formulation for this case too.

Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 56

Final Comments

- Remember, E.M. can get stuck in local minima, and empirically it DOES.
- Our unsupervised learning example assumed $P(w_i)$'s known, and variances fixed and known. Easy to relax this.
- It's possible to do Bayesian unsupervised learning instead of max. likelihood.
- There are other algorithms for unsupervised learning. We'll visit K-means soon. Hierarchical clustering is also interesting.
- Neural-net algorithms called "competitive learning" turn out to have interesting parallels with the EM method we saw.

Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 57

What you should know

- How to "learn" maximum likelihood parameters (locally max. like.) in the case of unlabeled data.
- Be happy with this kind of probabilistic analysis.
- Understand the two examples of E.M. given in these notes.

For more info, see Duda + Hart. It's a great book. There's much more in the book than in your handout.

Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 58

Other unsupervised learning methods

- K-means (see next lecture)
- Hierarchical clustering (e.g. Minimum spanning trees) (see next lecture)
- Principal Component Analysis
simple, useful tool
- Non-linear PCA
Neural Auto-Associators
Locally weighted PCA
Others...