

cmssc726 Spring 2006: Probabilistic Clustering and EM

material from: Andrew Moore, Sebastian Thrun

Families of Clustering Algorithms

- Partition-based methods
 - e.g., K-means
- Hierarchical clustering
 - e.g., hierarchical agglomerative clustering
- Probabilistic model-based clustering
 - e.g., mixture models, Gaussian Mixture Models
 - expectation maximization
- Spectral Clustering

AWM's cartoon of GMM

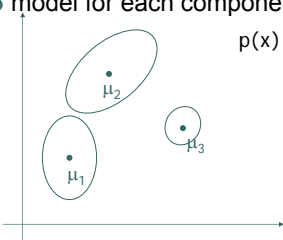
- You walk into a bar.
A stranger approaches and tells you:
"I've got data from k classes. Each class produces observations with normal distribution and variance σ^2 . Standard simple multivariate gaussian assumptions. I can tell you all the $P(w_j)$'s."
- So far, looks straightforward.
"I need a maximum 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!!

Probabilistic Model-based Clustering

- Assume a probability model for each component cluster
- Mixture Model: $p(x) = \sum_{k=1}^K w_k f_k(x; \theta_k)$
 - where f_k are component distributions
 - components: Gaussian, Poisson, exponential
 - Most common: Gaussian mixture model (GMM)

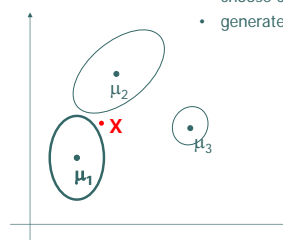
Gaussian Mixture Models (GMM)

- K components
 - model for each component cluster $N(\mu_k, \sigma_k)$
- $$p(x) = \sum_{k=1}^K w_k f(x; \mu_k, \sigma_k)$$



GMM cont.

- Generative Model
 - choose component with probability w_k
 - generate $X \sim N(\mu_k, \sigma_k)$



● ● ● | GMM

- Model
- Likelihood

- Problem: we have a bunch on non-linear non-analytically-solvable equations
- One solution: gradient descent.... slow
- instead....

● ● ● | Expectation Maximization (EM)

- Dempster, Laird, and Rubin, 1977
- **extremely** popular recently
- applicable in a wide range of problems
- many uses besides clustering: hidden markov models, Bayesian networks
- basic idea is quite simple...

● ● ● | Silly Example

Let events be "grades in a class"

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

(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 ?

● ● ● | Silly Example

Let events be "grades in a class"

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

(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 ?

● ● ● | Trivial Statistics

$$P(A) = \frac{1}{2} \quad P(B) = \mu \quad P(C) = 2\mu \quad P(D) = \frac{1}{2} - 3\mu$$

$$P(a,b,c,d) = (\frac{1}{2})^a (\mu)^b (2\mu)^c (\frac{1}{2} - 3\mu)^d$$

$$\log P(a,b,c,d) = a \log \frac{1}{2} + b \log \mu + c \log 2\mu + d \log (\frac{1}{2} - 3\mu)$$

$$\text{FOR MLE } \mu, \text{ SET } \frac{\partial \log P}{\partial \mu} = 0$$

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

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

So if class got

A	B	C	D
14	6	9	10

$$\text{MLE } \mu = \frac{1}{10}$$

● ● ● | 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

$P(A) = \frac{1}{2}$

$P(B) = \mu$

$P(C) = 2\mu$

$P(D) = \frac{1}{2} - 3\mu$

● ● ● | 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:

REMEMBER

$P(A) = \frac{1}{2}$
 $P(B) = \mu$
 $P(C) = 2\mu$
 $P(D) = \frac{1}{2} - 3\mu$

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 $\frac{1}{2} : \mu$

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

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

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

● ● ● | E.M. for our Trivial Problem

REMEMBER

$P(A) = \frac{1}{2}$
 $P(B) = \mu$
 $P(C) = 2\mu$
 $P(D) = \frac{1}{2} - 3\mu$

We begin with a guess for μ

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

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

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

M-step
 $\mu(t+1) = \frac{b(t) + c}{6(b(t) + c + d)}$
 = max like est of μ given $b(t)$

Continue iterating until converged.
 Good news: Converging to local optimum is assured.
 Bad news: "local" optimum.


● ● ● | E.M. Convergence

- Convergence proof based on fact that $\text{Prob}(\text{data} | \mu)$ 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

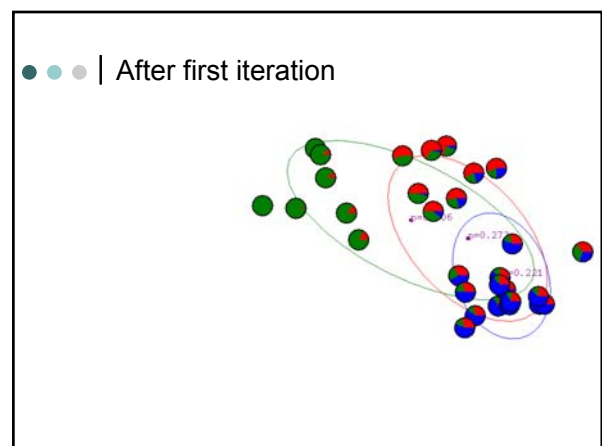
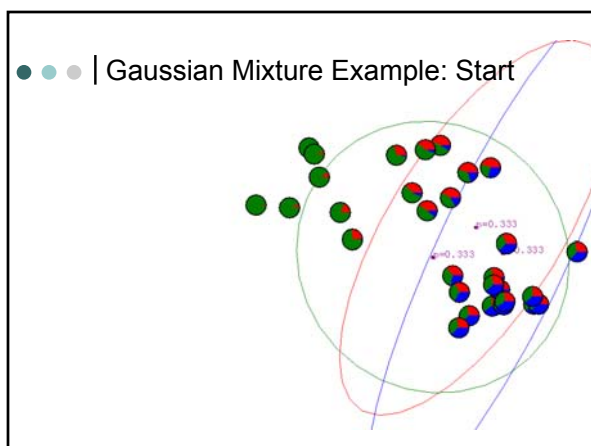
- $h = 20$
- $c = 10$
- $d = 10$
- $\mu(0) = 0$

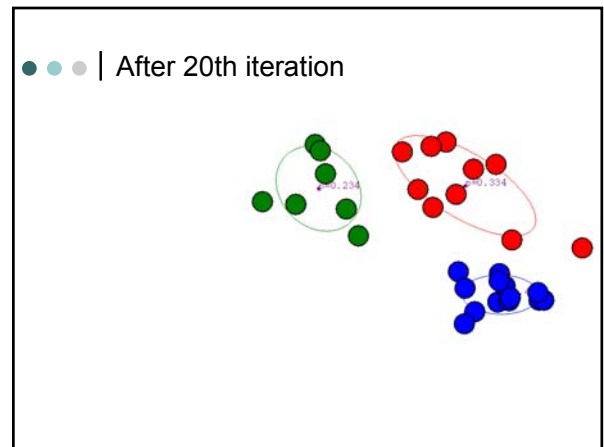
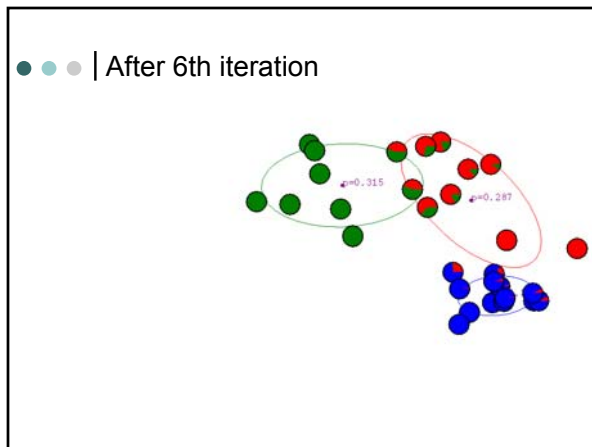
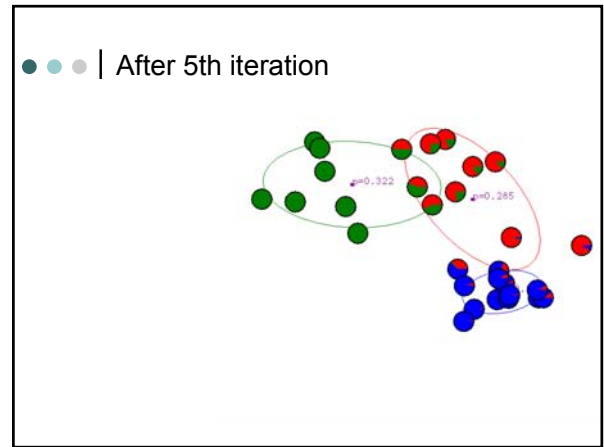
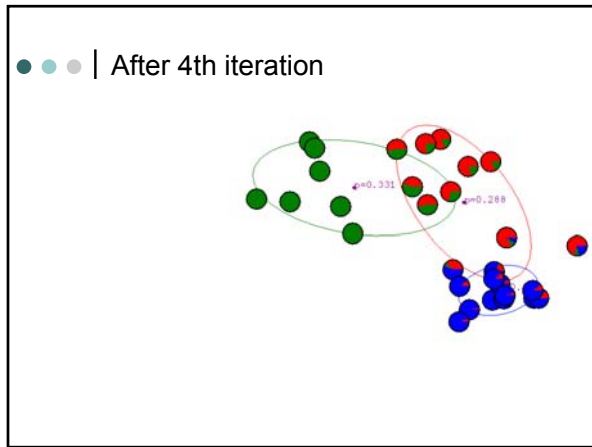
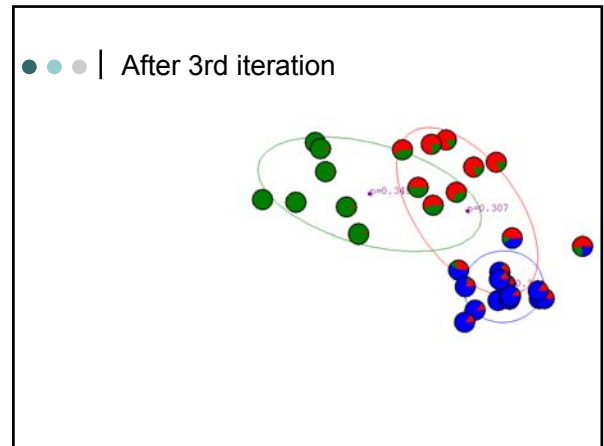
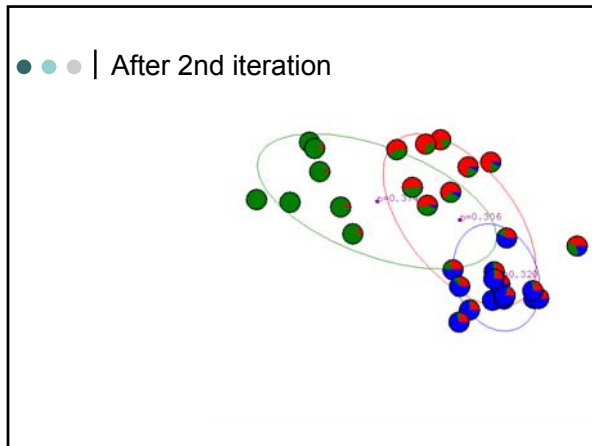


t	$\mu(t)$	$b(t)$
0	0	0
1	0.0833	2.857
2	0.0937	3.158
3	0.0947	3.185
4	0.0948	3.187
5	0.0948	3.187
6	0.0948	3.187

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

● ● ● | Back to EM for GMM





● ● ● | Formal EM setup

- Let $D = \{x(1), \dots, x(n)\}$ be n observed data vectors
- Let $Z = \{z(1), \dots, z(n)\}$ be n values of hidden variable (these might be the cluster labels)
- Then the log-likelihood of the observed data is

$$l(\theta) = \log p(D | \theta) = \log \sum_Z p(D, Z | \theta)$$

- both θ and Z are unknown
- Let $Q(Z)$ be any probability distribution for Z .

$$\begin{aligned} l(\theta) &= \log \sum_Z p(D, Z | \theta) \\ &= \log \sum_Z Q(Z) \frac{p(D, Z | \theta)}{Q(Z)} \\ &\geq \sum_Z Q(Z) \log \frac{p(D, Z | \theta)}{Q(Z)} \\ &= \sum_Z Q(Z) \log p(D, Z | \theta) + \sum_Z Q(Z) \log \frac{1}{Q(Z)} \\ &= F(Q, \theta) \end{aligned}$$

lower bound on $l(\theta)$

● ● ● | Jensen's Inequality

- <http://www.engineering.usu.edu/classes/ece/7680/lecture2/node5.html>

● ● ● | EM Algorithm

- EM algorithm alternates between
 - maximize F with respect to dist. Q with θ fixed
E-step: $Q^{k+1} = \arg \max_Q F(Q^k, \theta^k)$
 - maximize F with respect to θ with $Q = p(Z)$ fixed
M-step: $\theta^{k+1} = \arg \max_{\theta} F(Q^{k+1}, \theta^k)$
- Maximum for E step:
 $Q^{k+1} = p(Z | D, \theta^k)$

Intuition:
 • In the E-step, we estimate the distribution on the hidden variables, conditioned on a particular setting of the parameter vector θ^k
 • In the M-step, we choose new set of parameters θ^{k+1} to maximize the expected log-likelihood of observed data

● ● ● | Notes

- Often both the E and M step can be solved in closed form
- Neither the E step nor the M step can decrease the log-likelihood
- Under relatively general conditions the algorithm is guaranteed to converge to a local maximum of log-likelihood
- We must specify a starting point for the algorithm, for example a random choice of θ or Q
- We must specify stopping criteria, or convergence detection
- Computational complexity: number of iterations, time to compute E and M steps

● ● ● | EM Demo

[EM demo](#)

<http://diwww.epfl.ch/mantra/tutorial/english/gaussian/html/>

● ● ● | EM Comments

- complexity of EM for multivariate gaussian mixtures with K components: dominated by calculation of K covariance matrices.
 - With p dimensions, $O(Kp^2)$ covariance parameters to be estimated
 - Each requires summing over n data points and cluster weights, leading to $O(Kp^2n)$ per step
- Often times there are large increases in likelihood over first few iteration and then can slowly converge; likelihood as function of iterations not necessarily concave

● ● ● | and finally...

how do we choose K?

● ● ● | How to choose K

- Choose K that maximizes likelihood?
- NOT.
- As K is increased, the value of the likelihood at maximum cannot decrease
- Problem of scoring models with different complexities
 - Model too flexible \Rightarrow overfit the data \Rightarrow high variance
 - Model too restrictive \Rightarrow can't fit the data \Rightarrow high bias
 - Bias-variance tradeoff: compromise
- Solutions:
 - external validation (use k-fold cross validation, LOOCV)
 - scoring function – MDL, BIC, AIC
 - Bayesian model selection