EM:

Expectation Maximization

Machine Learning – 10701/15781

Carlos Guestrin

Carnegie Mellon University

April 9th, 2007

Gaussian Bayes Classifier Reminder

$$\underbrace{P(y=i\mid \mathbf{x}_j) = \frac{\underline{p(\mathbf{x}_j\mid y=i)}\,P(y=i)}{p(\mathbf{x}_j)}}$$

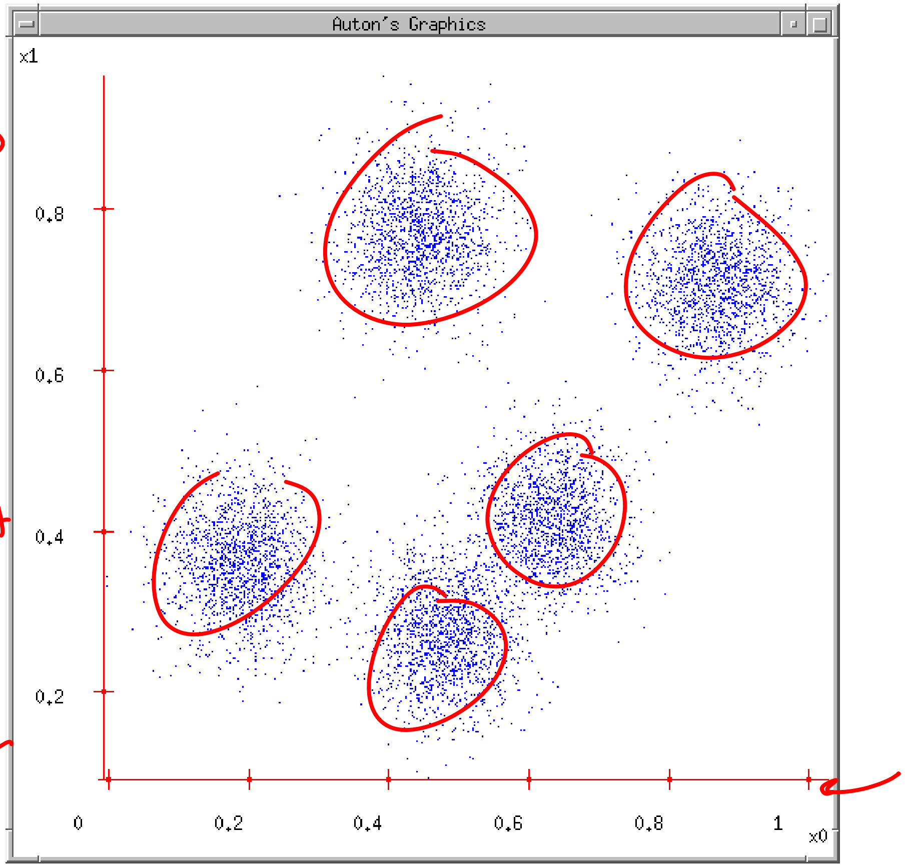
$$P(y=i\mid \mathbf{x}_j) \propto \underbrace{\frac{1}{(2\pi)^{m/2}\,\|\Sigma_i\|^{1/2}}\exp\left[-\frac{1}{2}\left(\mathbf{x}_j-\overset{\text{class mean}}{\mu_i}\right)^T \overset{\text{class covariance}}{\Sigma_i^{-1}}\left(\mathbf{x}_j-\mu_i\right)\right]}_{\text{Gaussian likelihood}}\underbrace{P(y=i)}_{\text{prior}}$$

Next… back to Density Estimation

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

want to represent P(X)

$$P(x) = \sum_i \underbrace{P(X|y=i)}_{\text{Gaussian}} \cdot \underbrace{P(y=i)}_{\text{weight}}$$



$X = \{x_1, x_2\}$

Marginal likelihood for general case

$$\underbrace{P(y=i\mid\mathbf{x}_j)}_{\text{observed data}} \propto \frac{1}{(2\pi)^{m/2}\,\|\Sigma_i\|^{1/2}}\exp\left[-\frac{1}{2}\left(\mathbf{x}_j-\mu_i\right)^T\Sigma_i^{-1}\left(\mathbf{x}_j-\mu_i\right)\right]P(y=i)$$

defn. of GMM

x_j ← observed

assumption : $x_j \sim$ GMM

$$P(x_j) = \sum_i P(y=i)\cdot P(x_j\mid y=i)$$

Gaussian

don't observe y_j

⇒ max $P(x_j)$

- Marginal likelihood:

average over class labels

$$\log\prod_{j=1}^{m}P(\mathbf{x}_j) = \log\prod_{j=1}^{m}\sum_{i=1}^{k}P(\mathbf{x}_j, y=i)$$

$$= \log\prod_{j=1}^{m}\sum_{i=1}^{k}\frac{1}{(2\pi)^{m/2}\,\|\Sigma_i\|^{1/2}}\exp\left[-\frac{1}{2}\left(\mathbf{x}_j-\mu_i\right)^T\Sigma_i^{-1}\left(\mathbf{x}_j-\mu_i\right)\right]P(y=i)$$

$$= \sum_{j=1}^{m}\log\sum_{i=1}^{k}\frac{1}{(2\pi)^{m/2}\,\|\Sigma_i\|^{1/2}}\,e^{-\frac{1}{2}(x_j-\mu_i)^T\Sigma_i^{-1}(x_j-\mu_i)}\,P(y=i)$$

sum over data points

©2005-2007 Carlos Guestrin

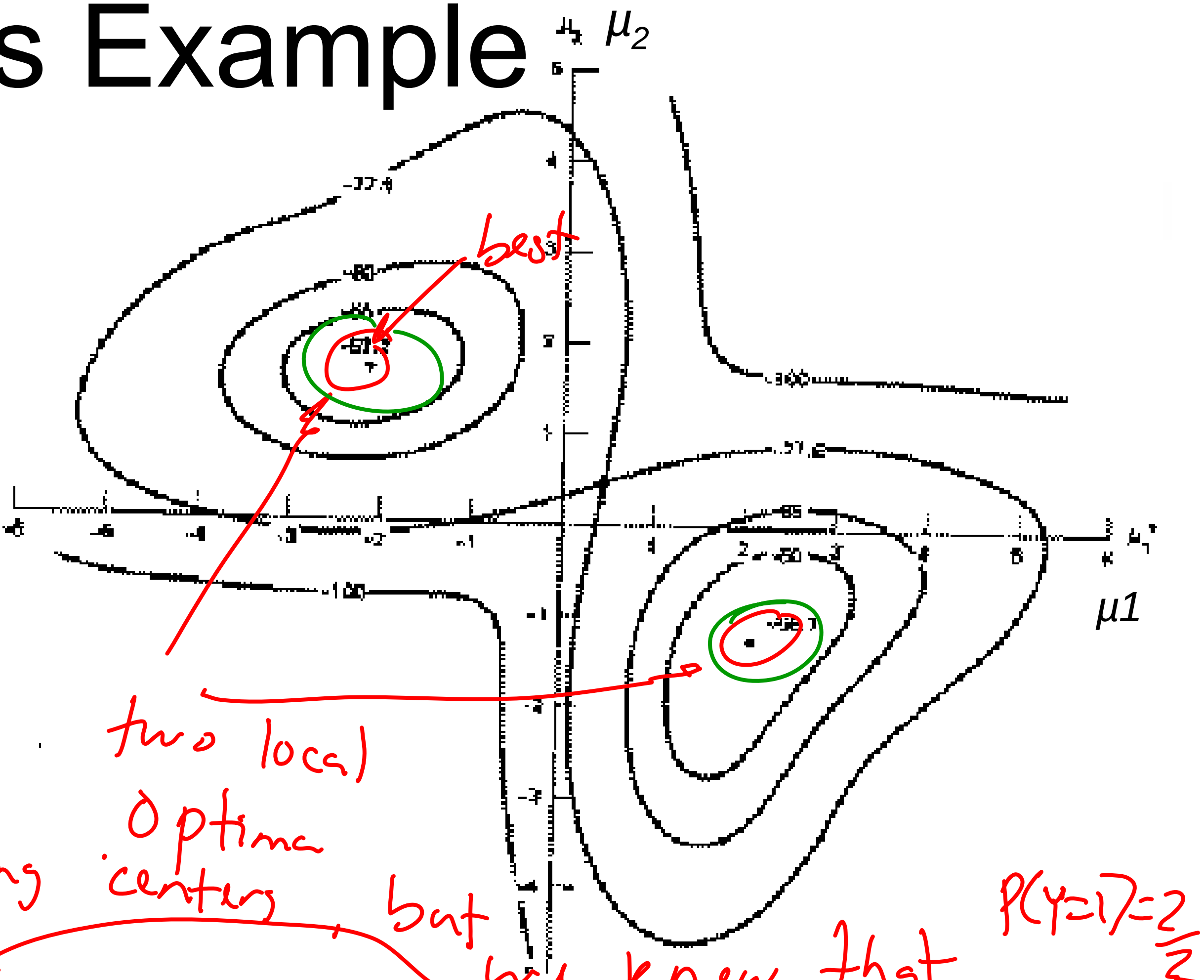
Duda & Hart's Example

$\max_{\mu} P(X|\mu)$

Graph of

log P(x_1, x_2 .. x_{25} | μ_1, μ_2)

against μ_1 (→) and μ_2 (↑)



best

two local optima

flipping centers, but we knew that

P(Y=1)=2/3

P(Y=2)=1/3

Max likelihood = (μ_1 =-2.13, μ_2 =1.668)

Local minimum, but very close to global at (μ_1 =2.085, μ_2 =-1.257)*

* corresponds to switching y_1 with y_2.

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

$\prod_j P(x_j | \mu_1 \ldots \mu_k)$

We can compute P(data | $\boldsymbol{\mu}_1,\boldsymbol{\mu}_2..\boldsymbol{\mu}_k$)

How do we find the $\boldsymbol{\mu}_i$‘s which give max. likelihood?

not a convex problem

(some versions are NP-hard)

- The normal max likelihood trick:

  Set $\frac{\partial}{\partial \mu_i}$ log Prob (….) = 0

  and solve for μ_i‘s.

  # Here you get non-linear non-analytically-solvable equations

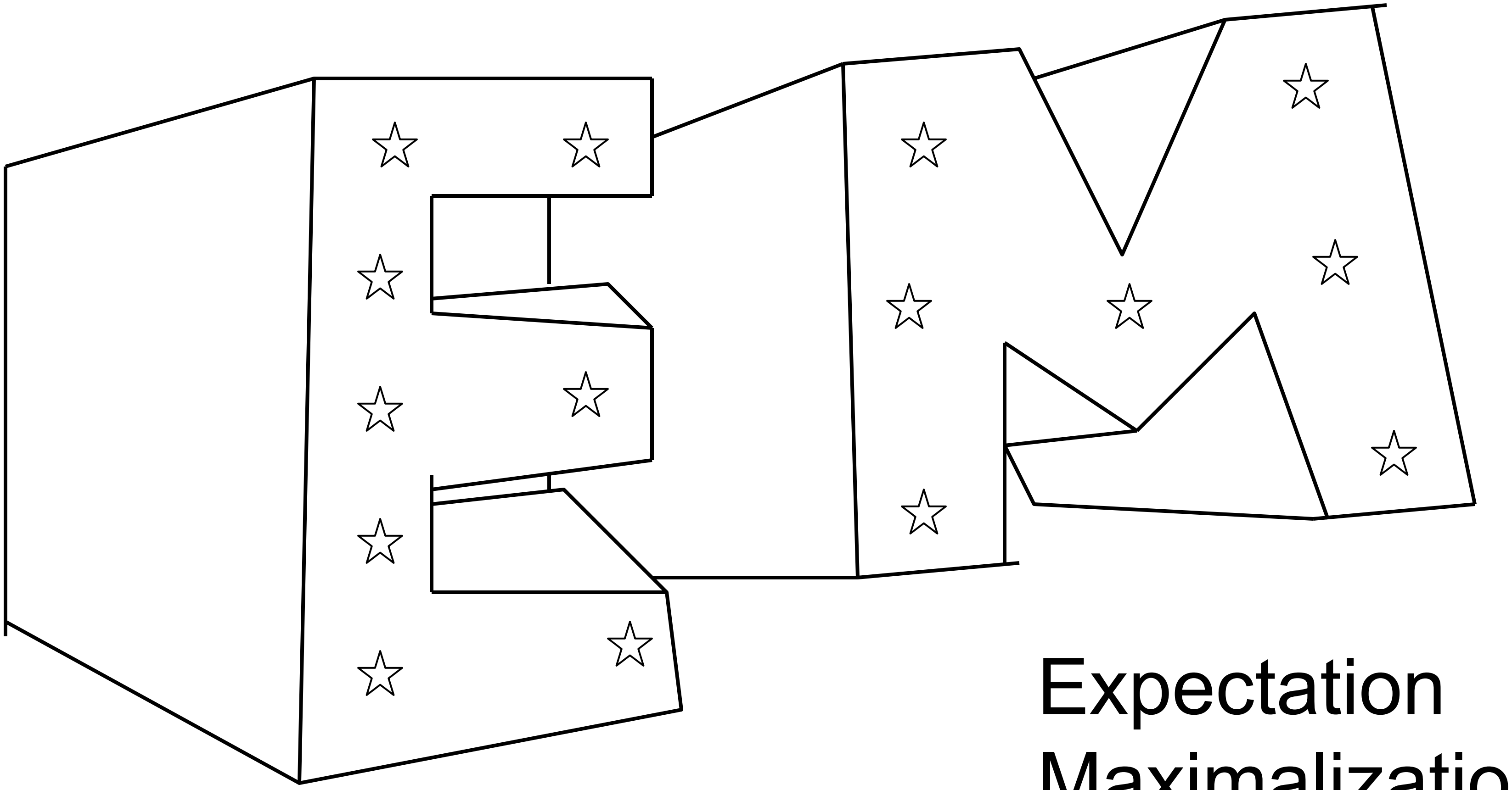
- Use gradient descent

  Slow but doable

  there is some debate as to what is better

- Use a much faster, cuter, and recently very popular method…

  EM alg.



Expectation
Maximalization

The E.M. Algorithm

DETOUR

- 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 learning BNs with hidden data

Silly Example

Let events be “grades in a class”

w_1 = Gets an A   P(A) = ½

w_2 = Gets a B   P(B) = μ

w_3 = Gets a C   P(C) = 2μ

w_4 = Gets a D   P(D) = ½-3μ

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

in terms of μ

what is μ?

MLE sense

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

Counts
a → A
b → B
c → C
d → D

P(A) = ½ P(B) = μ P(C) = 2μ P(D) = ½-3μ

P(*a*,*b*,*c*,*d* | μ) = K(½)^a(μ)^b(2μ)^c(½-3μ)^d ← log likelihood

log P(*a*,*b*,*c*,*d* | μ) = log K + *a*log ½ + *b*log μ + *c*log 2μ + *d*log (½-3μ) ← take log

FOR MAX LIKE μ, SET $\frac{\partial \mathrm{LogP}}{\partial \mu} = 0$ ← set to zero

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

move things around

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

So if class got

A	B	C	D
14	6	9	10

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

Boring, but true!

Same Problem with Hidden Information

h people got A's or B's

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) = ½

P(B) = μ

P(C) = 2μ

P(D) = ½-3μ

Same Problem with Hidden Information

REMEMBER

$P(A) = ½$

$P(B) = \mu$

$P(C) = 2\mu$

$P(D) = ½-3\mu$

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*

Bayes rule

Since the ratio a:b should be the same as the ratio ½ : μ

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

expected # of A's & B's

MAXIMIZATION

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

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

if we had exact data

E.M. for our Trivial Problem

REMEMBER

P(A) = ½

P(B) = μ

P(C) = 2μ

P(D) = ½-3μ

We begin with a guess for μ

We iterate between EXPECTATION and MAXIM~~AL~~IZATION 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

$$\mu^{(0)} = \text{initial guess} \quad \longleftarrow \text{e.g., } \mu = 1/8$$

$$b^{(t)} = \frac{\mu^{(t)} h}{{}^{1}\!/_{2} + \mu^{(t)}} = \underbrace{\mathrm{E}\left[b \mid \mu^{(t)}\right]}$$

E-step

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

$$= \text{max like est. of } \mu \text{ given } b^{(t)}$$

M-step

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

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

$h = 20$
$c = 10$
$d = 10$
$\mu^{(0)} = 0$

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 Unsupervised Learning of GMMs – a simple case

A simple case:

We have unlabeled data $\mathbf{x}_1\ \mathbf{x}_2 \ldots \mathbf{x}_m$

We know there are k classes

3

We know $P(y_1)\ P(y_2)\ P(y_3) \ldots P(y_k)$

proportions of cluster sizes

We don't know $\boldsymbol{\mu}_1\ \boldsymbol{\mu}_2 \ldots \boldsymbol{\mu}_k$

centers

We can write P(data | $\boldsymbol{\mu}_1 \ldots \boldsymbol{\mu}_k$)

$$= \mathrm{p}(x_1 \ldots x_m | \mu_1 \ldots \mu_k)$$

iid

$$= \prod_{j=1}^{m} \mathrm{p}(x_j | \mu_1 \ldots \mu_k)$$

MM

$$= \prod_{j=1}^{m} \sum_{i=1}^{k} \mathrm{p}(x_j | \mu_i) \mathrm{P}(y = i)$$

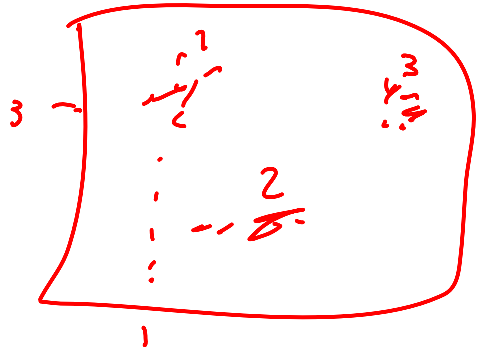
GMM

$$\propto \prod_{j=1}^{m} \sum_{i=1}^{k} \exp\left(-\frac{1}{2\sigma^2} \left\| x_j - \mu_i \right\|^2 \right) \mathrm{P}(y = i)$$

EM for simple case of GMMs: The E-step

- If we know $\mu_1,\ldots,\mu_k$ → easily compute prob. point x_j belongs to class y=i

$$p\left(y=i \middle| x_j,\mu_1\ldots\mu_k\right) \propto \underbrace{\exp\left(-\frac{1}{2\sigma^2}\left\|x_j-\mu_i\right\|^2\right)P(y=i)}$$



for $x_j=(1,3)$

compute $P(Y=1|X_j=(1,3),\mu) \propto 0.3$

$P(Y=2|X_j=(1,3),\mu) \propto 0.1$

$P(Y=3|X_j=(1,3),\mu) \propto 0.2$

renormalize

$\frac{0.3}{0.6}=0.5$

$=\frac{1}{6}$

$=\frac{1}{3}$

EM for simple case of GMMs: The M-step

- If we know prob. point x_j belongs to class y=i

  → MLE for μ_i is weighted average

  - □ imagine k copies of each x_j, each with weight $P(y=i|x_j)$:

$$\mu_i = \frac{\sum_{j=1}^{m} P(y=i|x_j)\, x_j}{\sum_{j=1}^{m} P(y=i|x_j)}$$

in K means, $P(Y=i|X_j)$ is zero or one

otherwise, equation is the same.

E.M. for GMMs (simple case)

E-step

Compute “expected” classes of all datapoints for each class

$$p\left(y=i \middle| x_j,\mu_1...\mu_k\right) \propto \exp\left(-\frac{1}{2\sigma^2}\left\|x_j-\mu_i\right\|^2\right)P(y=i)$$

Just evaluate a Gaussian at x_j

iterate

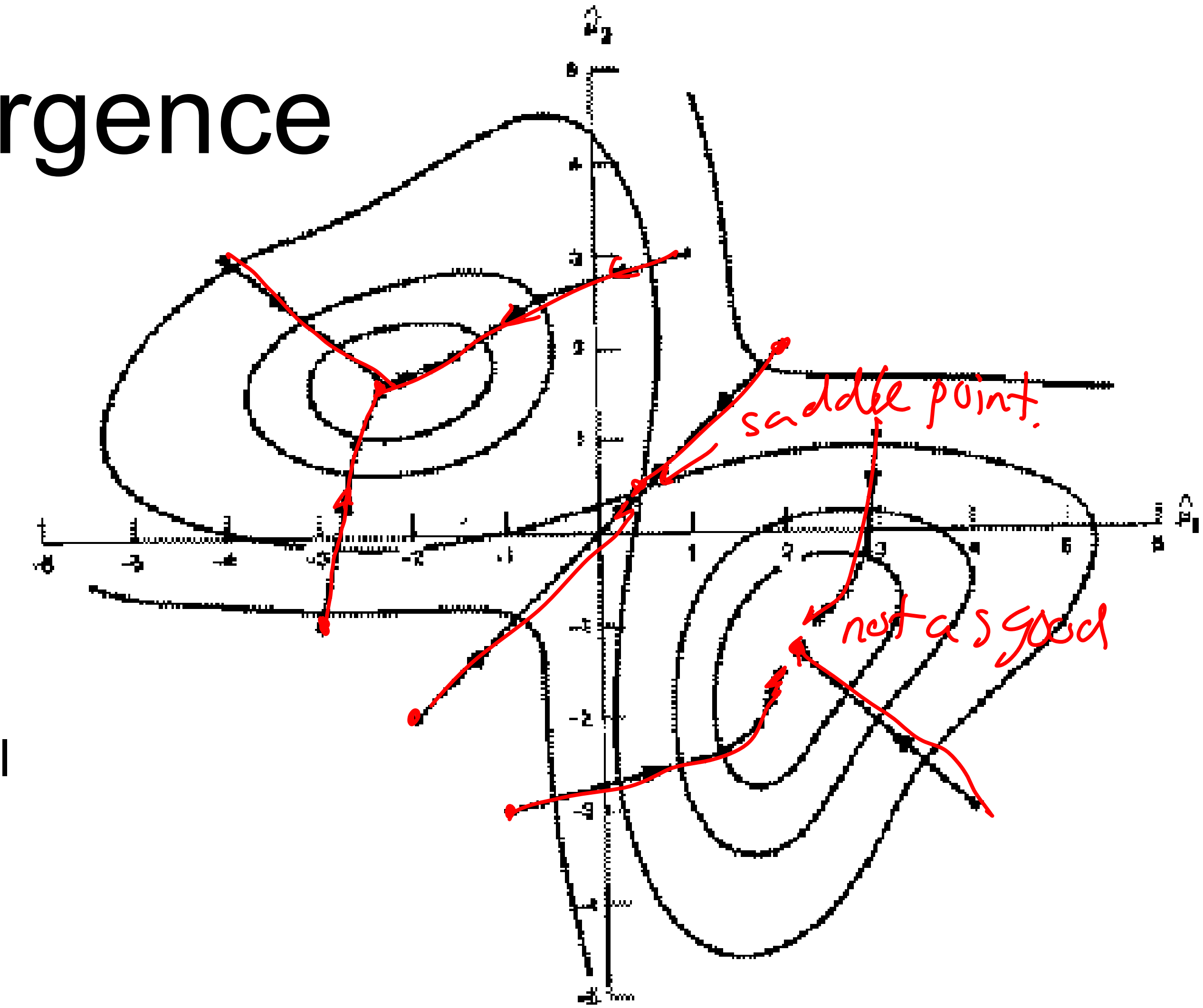
M-step

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

$$\mu_i = \frac{\sum_{j=1}^{m} P\left(y=i \middle| x_j\right)x_j}{\sum_{j=1}^{m} P\left(y=i \middle| x_j\right)}$$

E.M. Convergence

- EM is coordinate ascent on an interesting potential function
- Coord. ascent for bounded pot. func. → convergence to a local optimum guaranteed
- See Neal & Hinton reading on class webpage



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

E.M. for General GMMs

$p_i^{(t)}$ is shorthand for estimate of $P(y=i)$ on t'th iteration

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

$$\underbrace{\lambda_t}_{\text{set of parameters}} = \{ \underbrace{\mu_1^{(t)}, \mu_2^{(t)} \ldots \mu_k^{(t)}}_{\text{means}}, \underbrace{\Sigma_1^{(t)}, \Sigma_2^{(t)} \ldots \Sigma_k^{(t)}}_{\text{covariances}}, \underbrace{p_1^{(t)}, p_2^{(t)} \ldots p_k^{(t)}}_{\text{class probs. for prior}} \}$$

E-step

Compute "expected" classes of all datapoints for each class

parameters

$$\underline{P(y=i \mid x_j, \lambda_t)} \propto p_i^{(t)} p(x_j \mid \mu_i^{(t)}, \Sigma_i^{(t)})$$

Just evaluate a Gaussian at x_j

& renormalize

M-step

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

$$\mu_i^{(t+1)} = \frac{\sum_j P(y=i \mid x_j, \lambda_t) x_j}{\sum_j P(y=i \mid x_j, \lambda_t)}$$

weighted means

$$\Sigma_i^{(t+1)} = \frac{\sum_j P(y=i \mid x_j, \lambda_t) \left[x_j - \mu_i^{(t+1)}\right]\left[x_j - \mu_i^{(t+1)}\right]^T}{\sum_j P(y=i \mid x_j, \lambda_t)}$$

Same as normal covariances, but w. weighted data

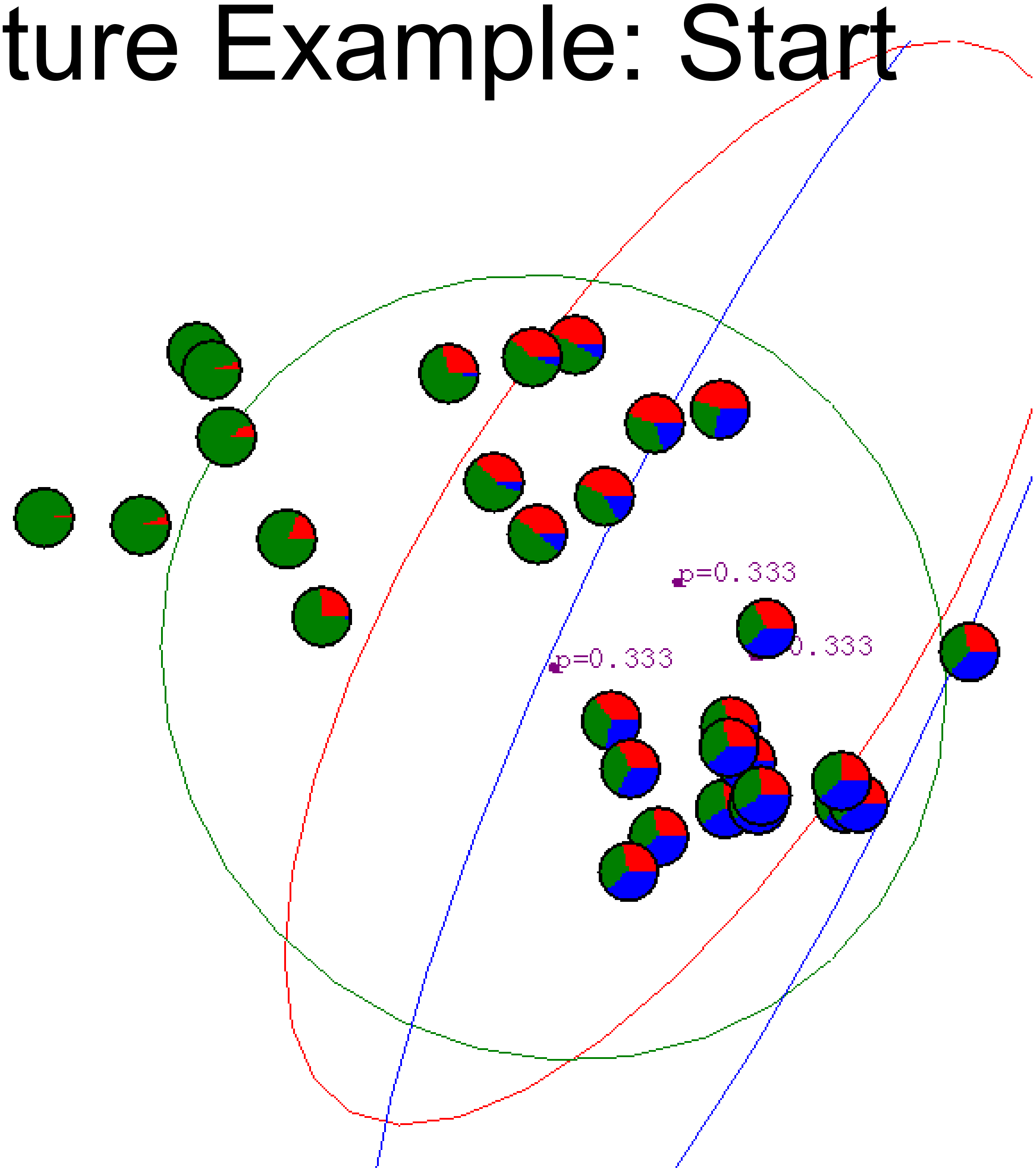
$$p_i^{(t+1)} = \frac{\sum_j P(y=i \mid x_j, \lambda_t)}{m}$$

class probs.

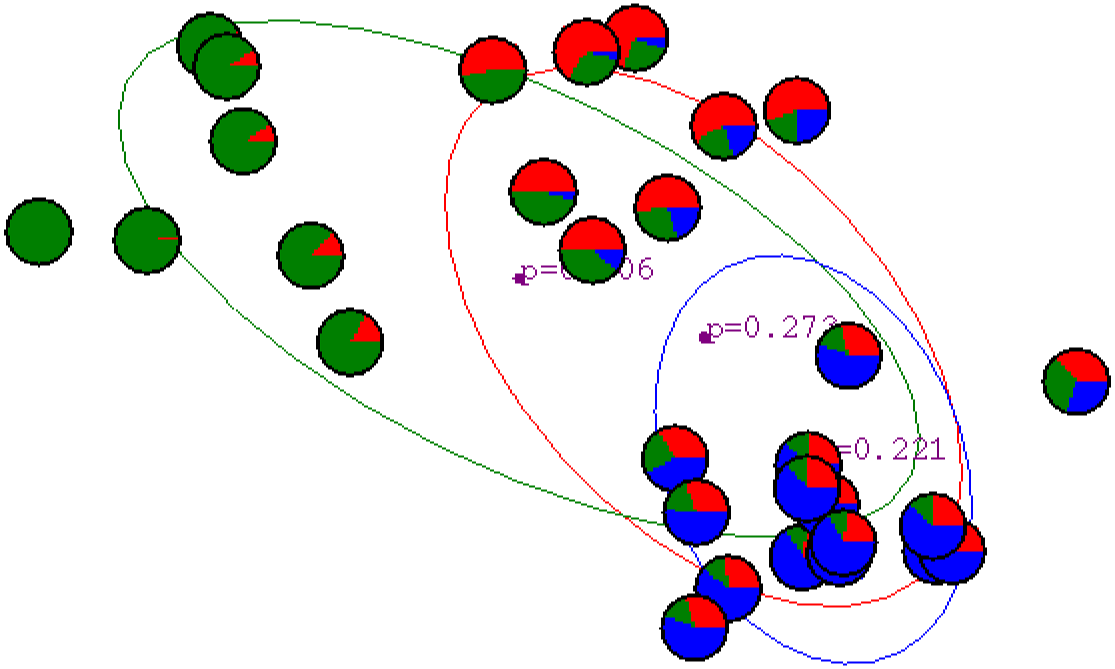
Prior

m = #records

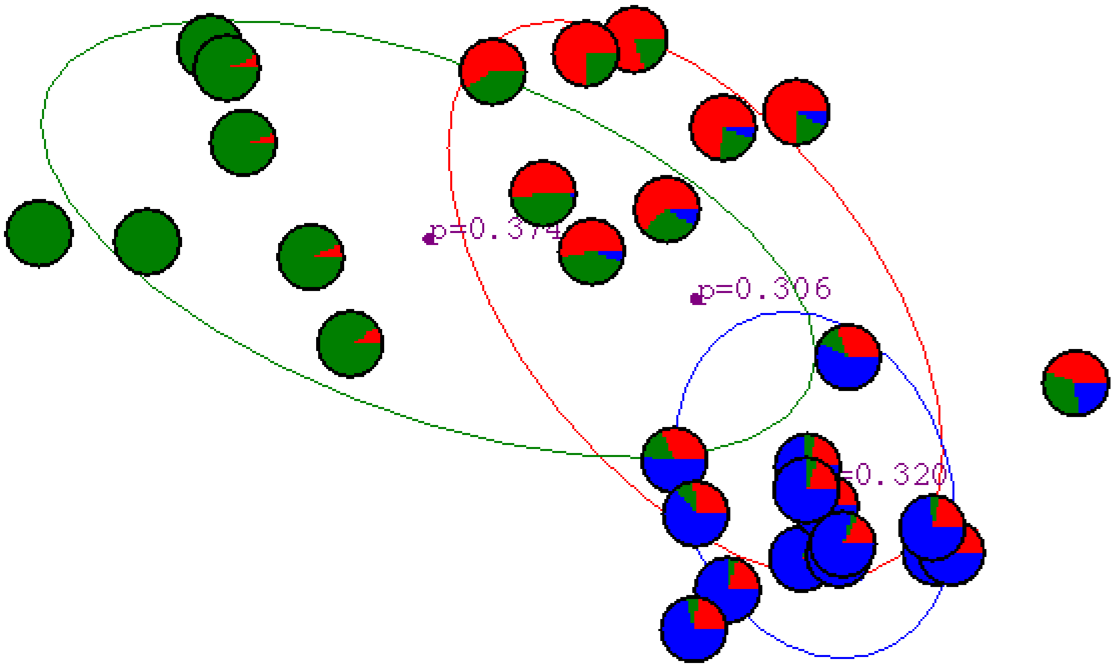
Gaussian Mixture Example: Start



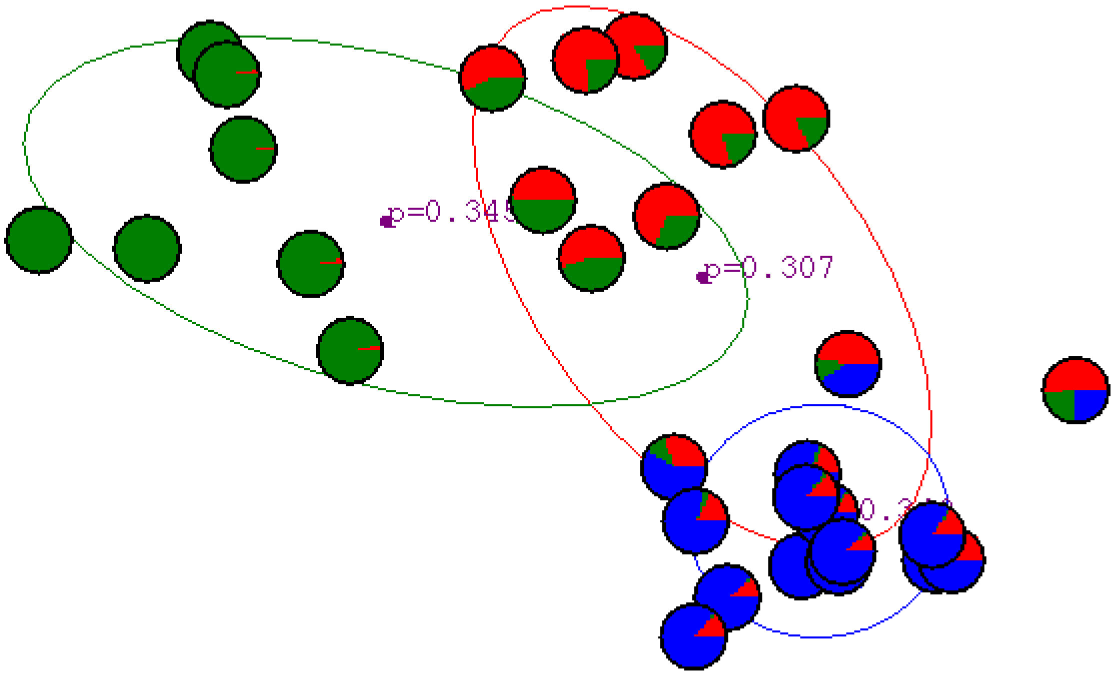
After first iteration



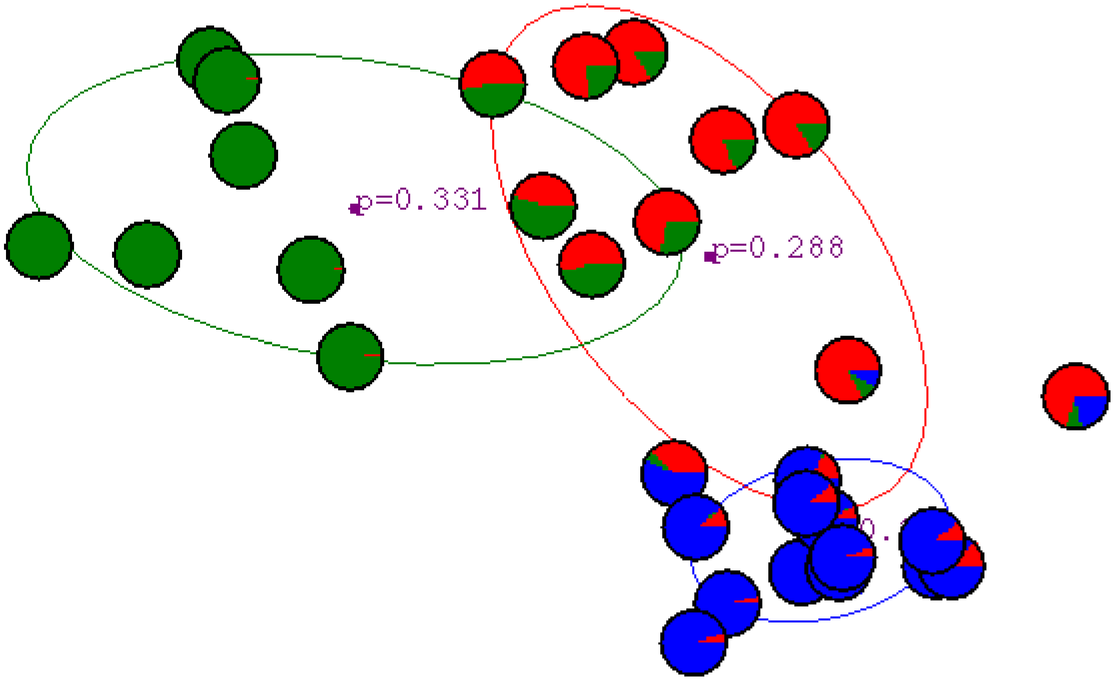
After 2nd iteration



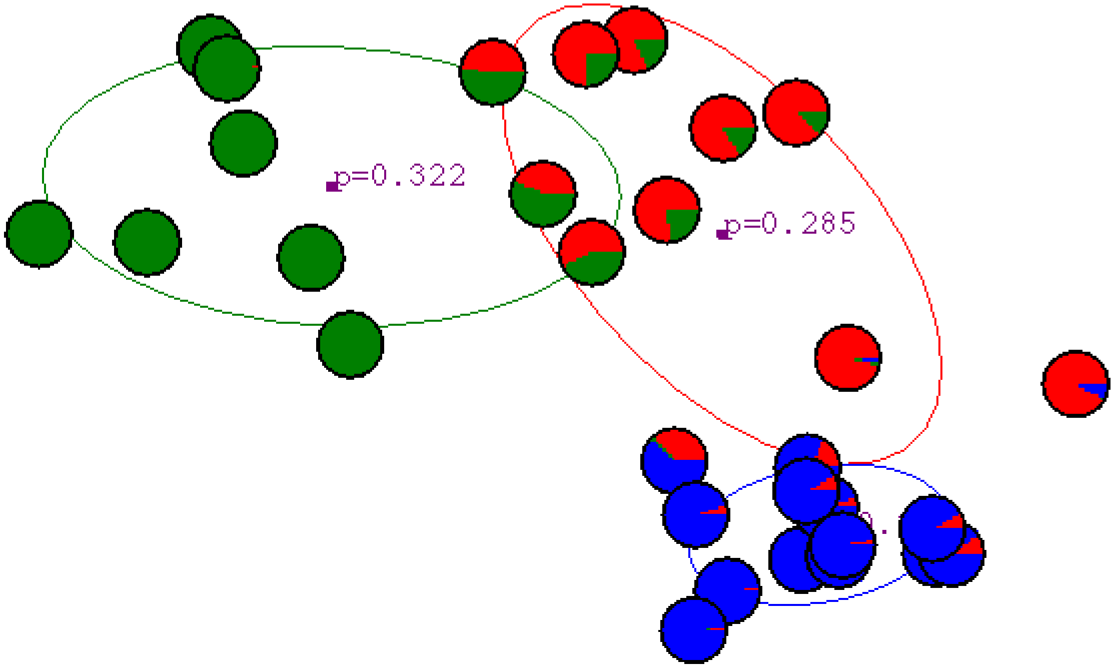
After 3rd iteration



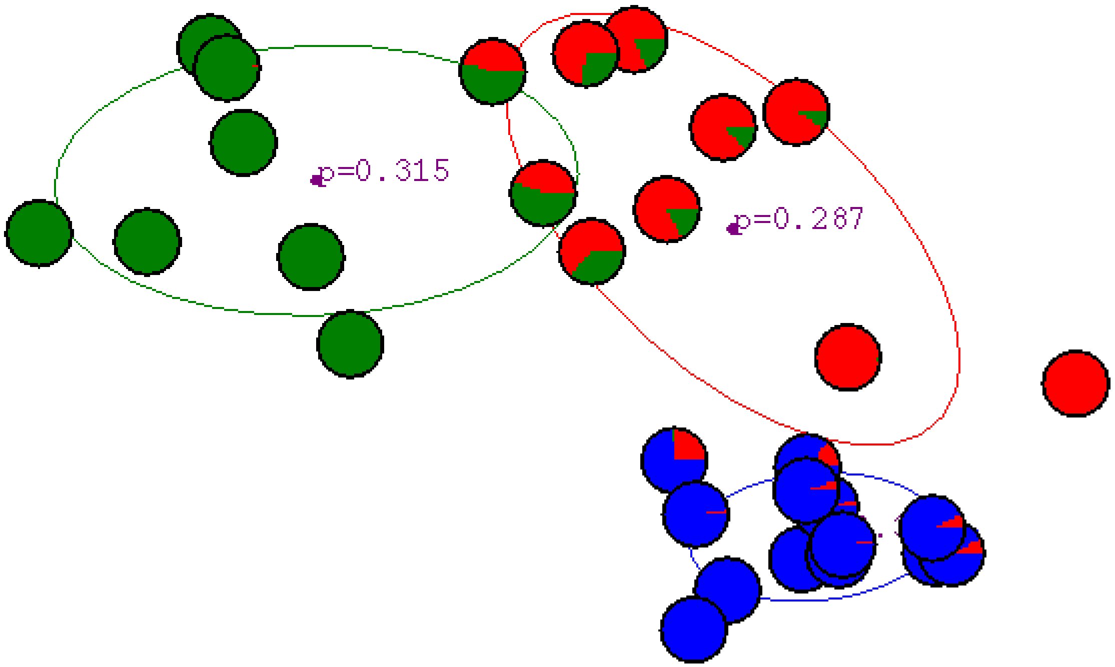
After 4th iteration



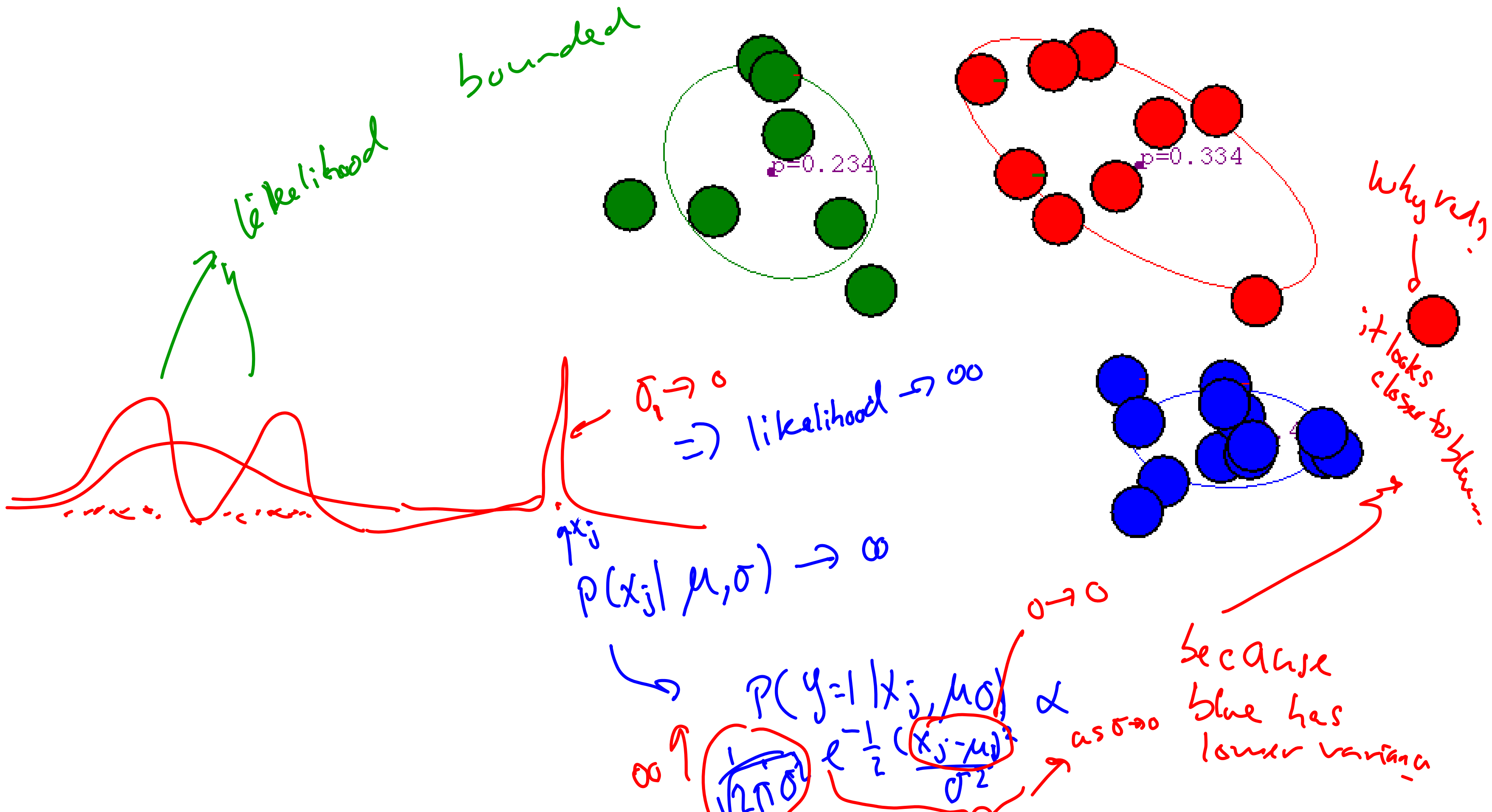
After 5th iteration



After 6th iteration



After 20th iteration



$\sigma_1 \to 0$

$\Rightarrow$ likelihood $\to \infty$

$p(x_j \mid \mu, \sigma) \to \infty$

$P(y=1 \mid x_j, \mu\sigma) \propto \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}}}_{\infty} \, e^{-\frac{1}{2}\frac{(x_j-\mu_i)^2}{\sigma^2}}$

$0 \to 0$

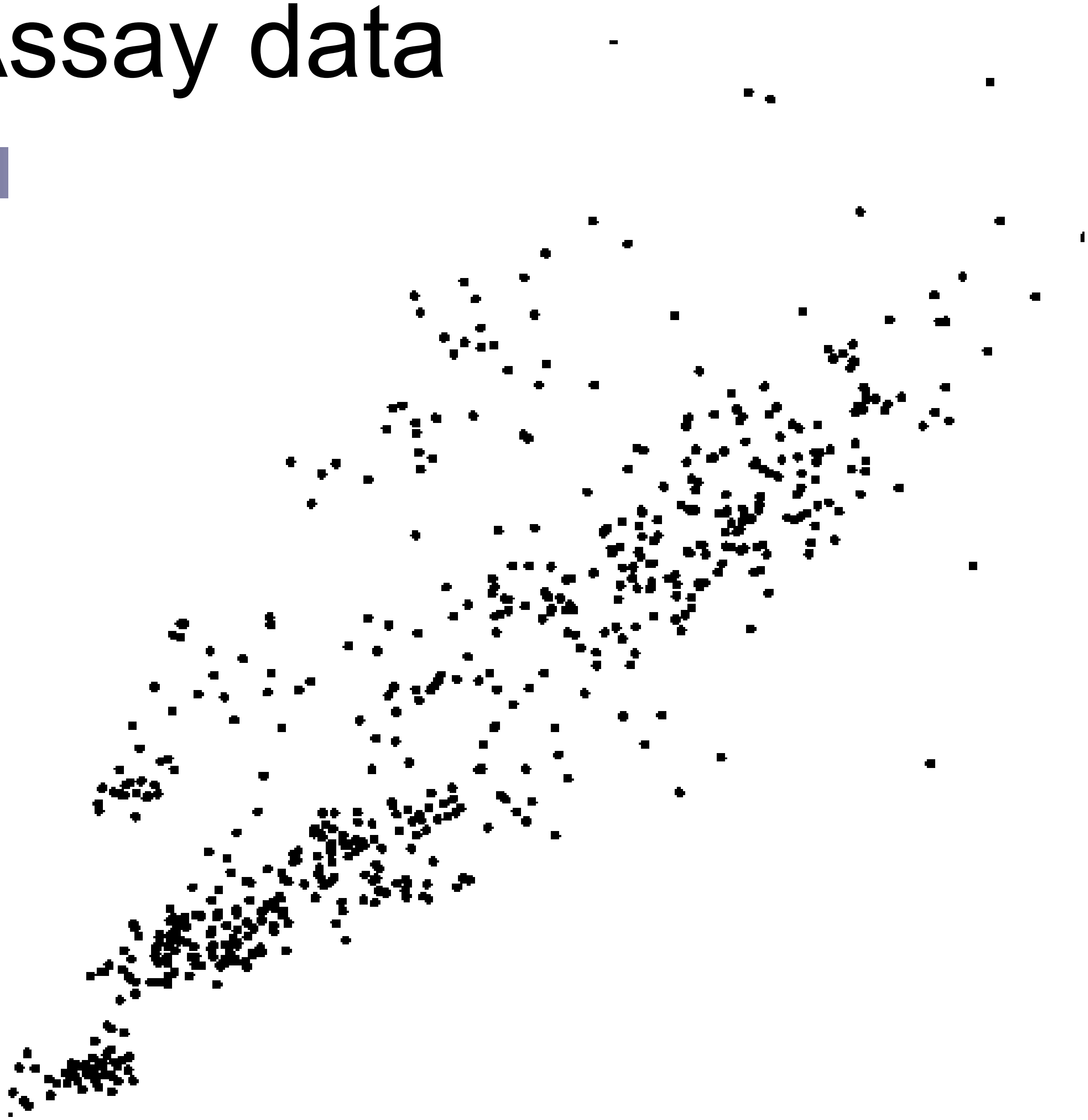
as $\sigma \to 0$

why red?

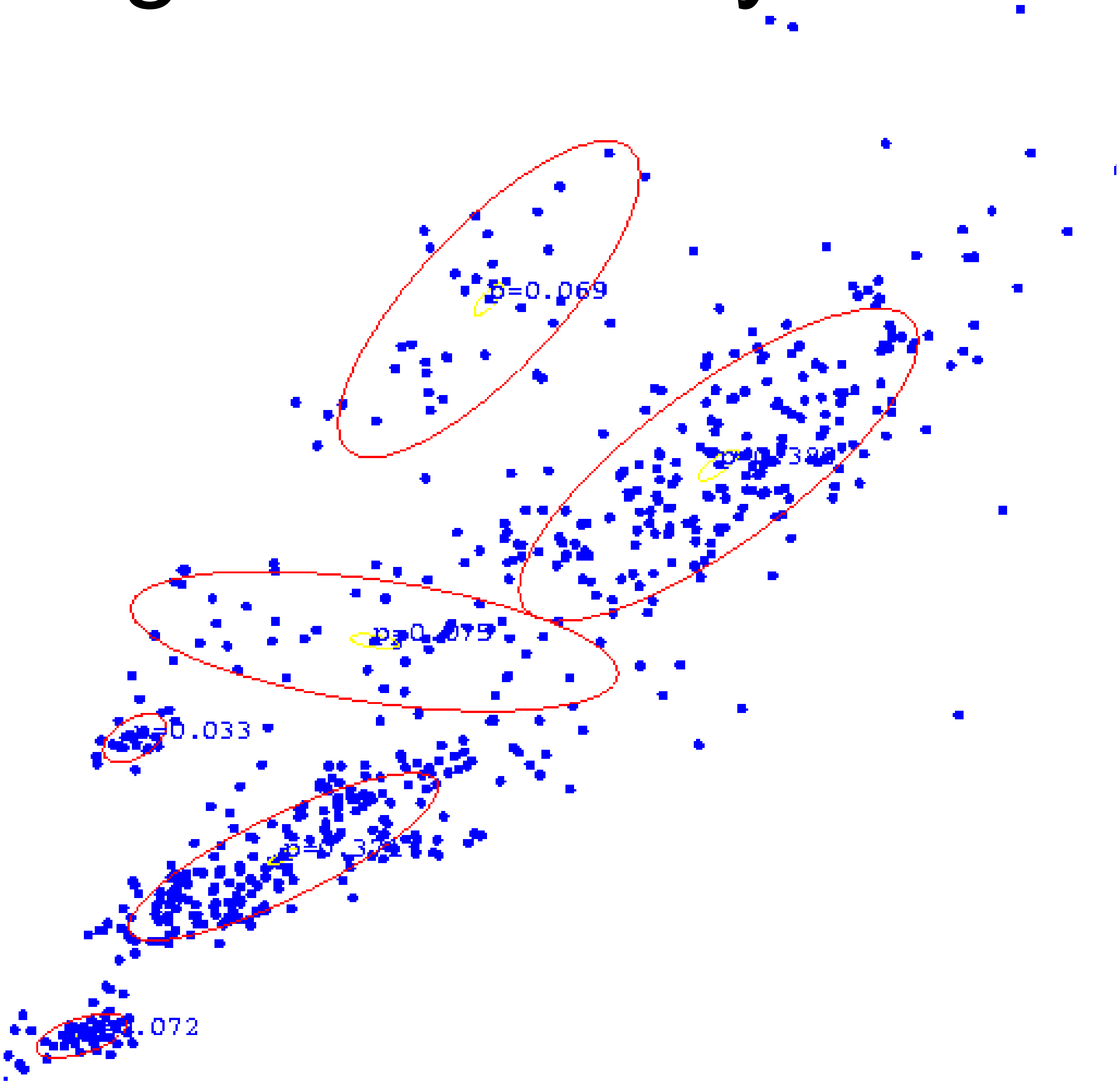
it looks closer to blue...

because blue has lower variance

Some Bio Assay data

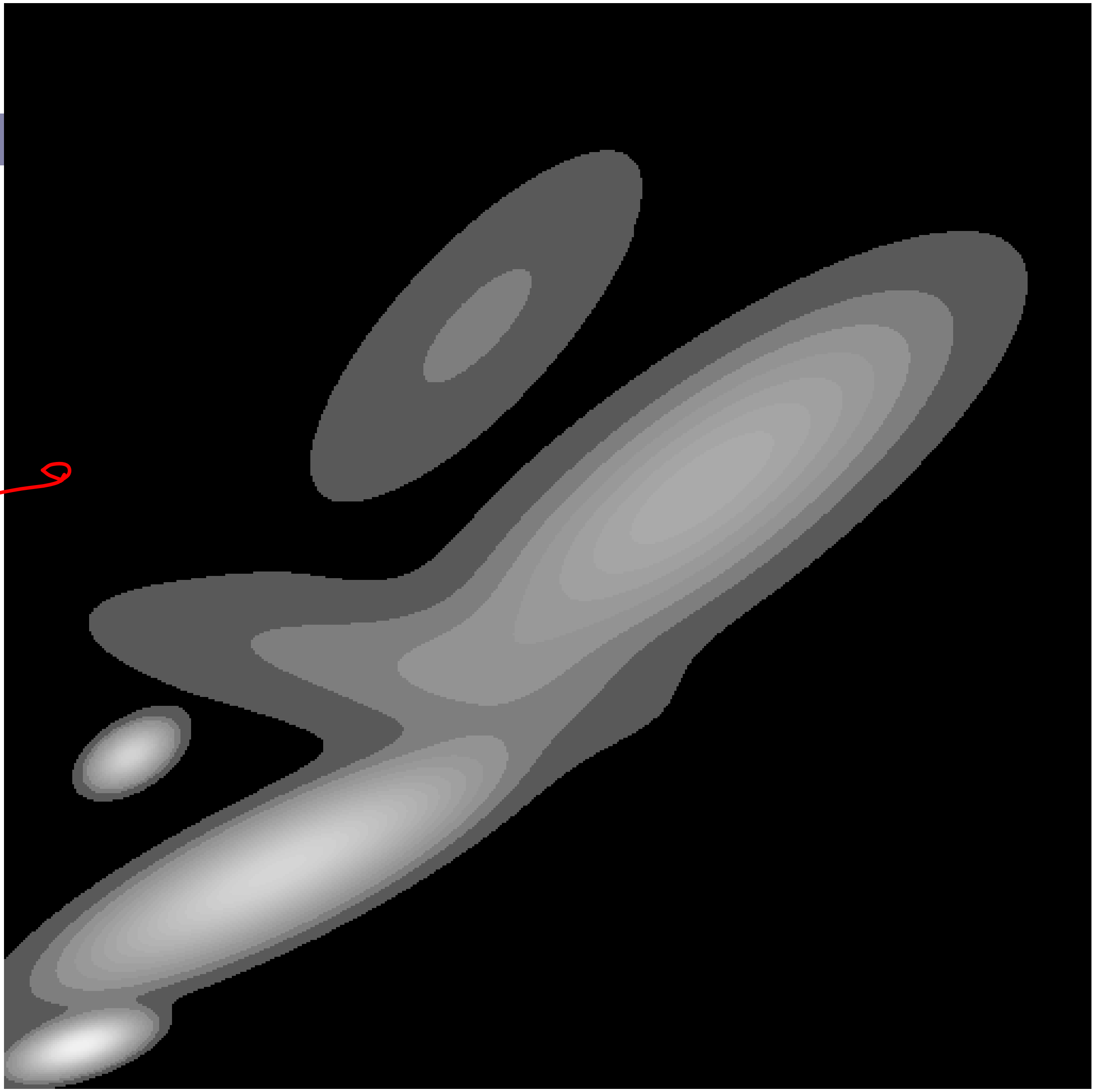


GMM clustering of the assay data



Resulting Density Estimator

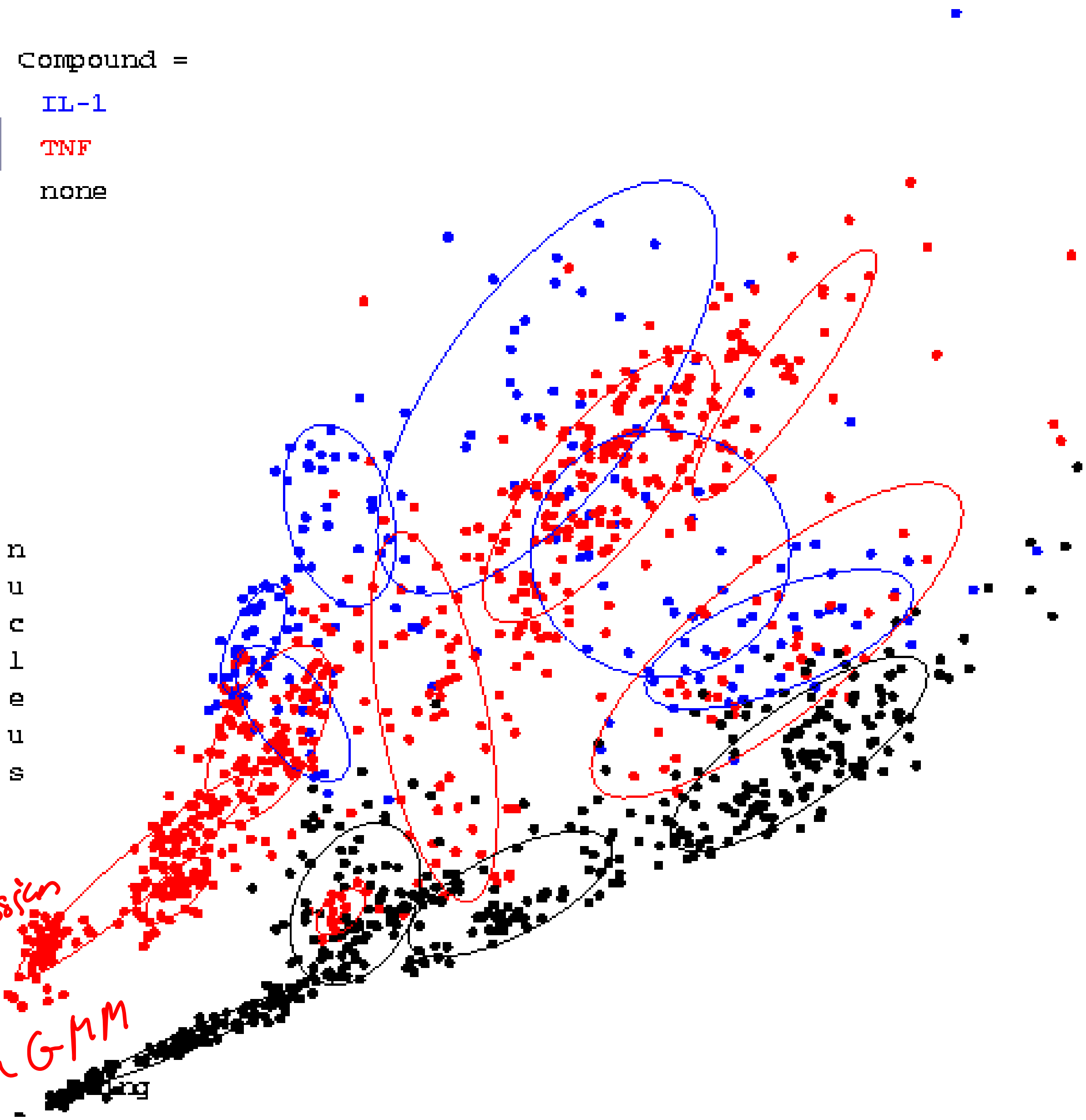
$P(x)$



©2005-2007 Carlos Guestrin

Three classes of assay

(each learned with it's own mixture model)



Y ← 3 class

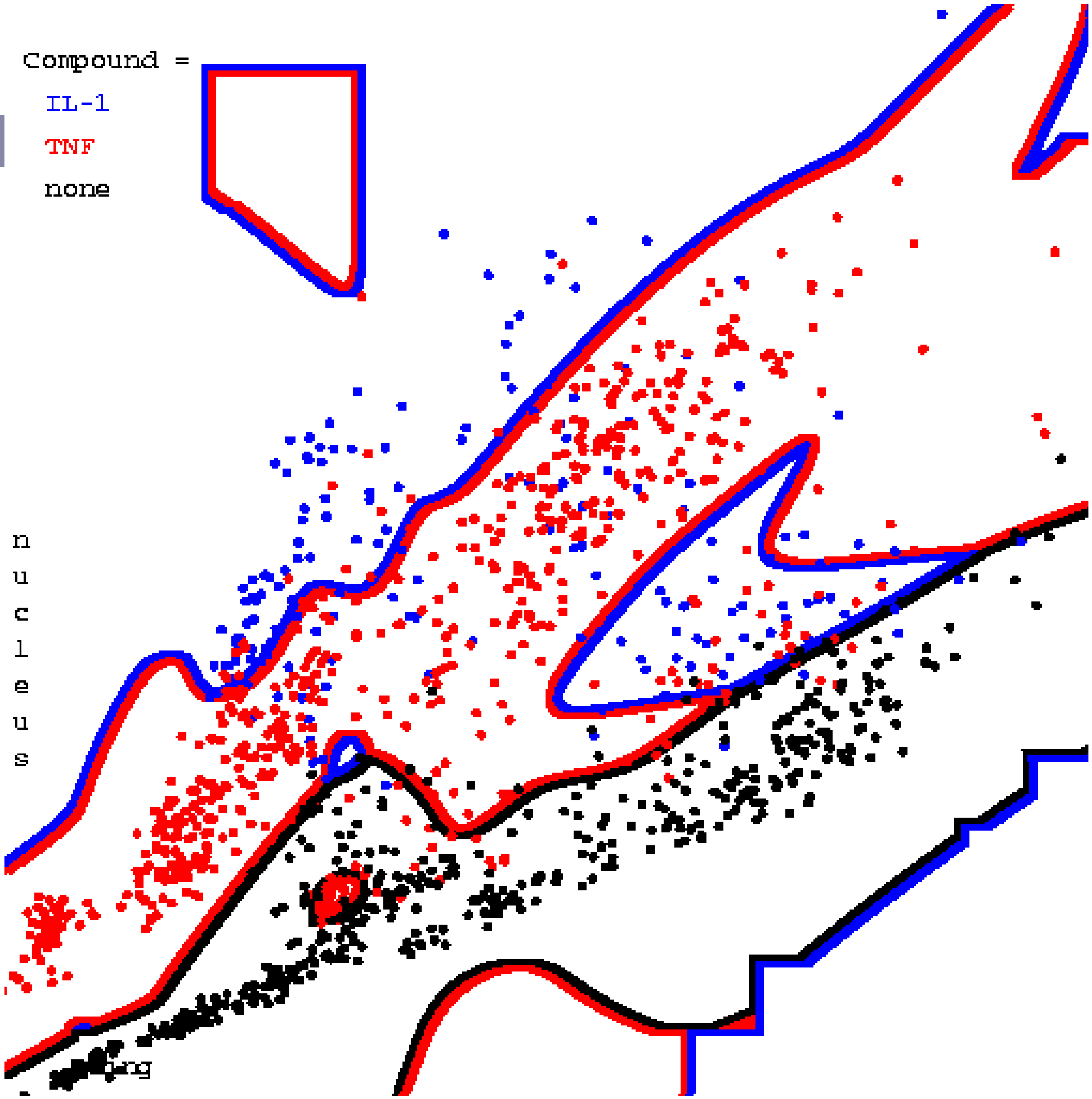
↓

X

instead of Cond. Gaussian use GMM

P(X|Y) is a GMM

Resulting Bayes Classifier

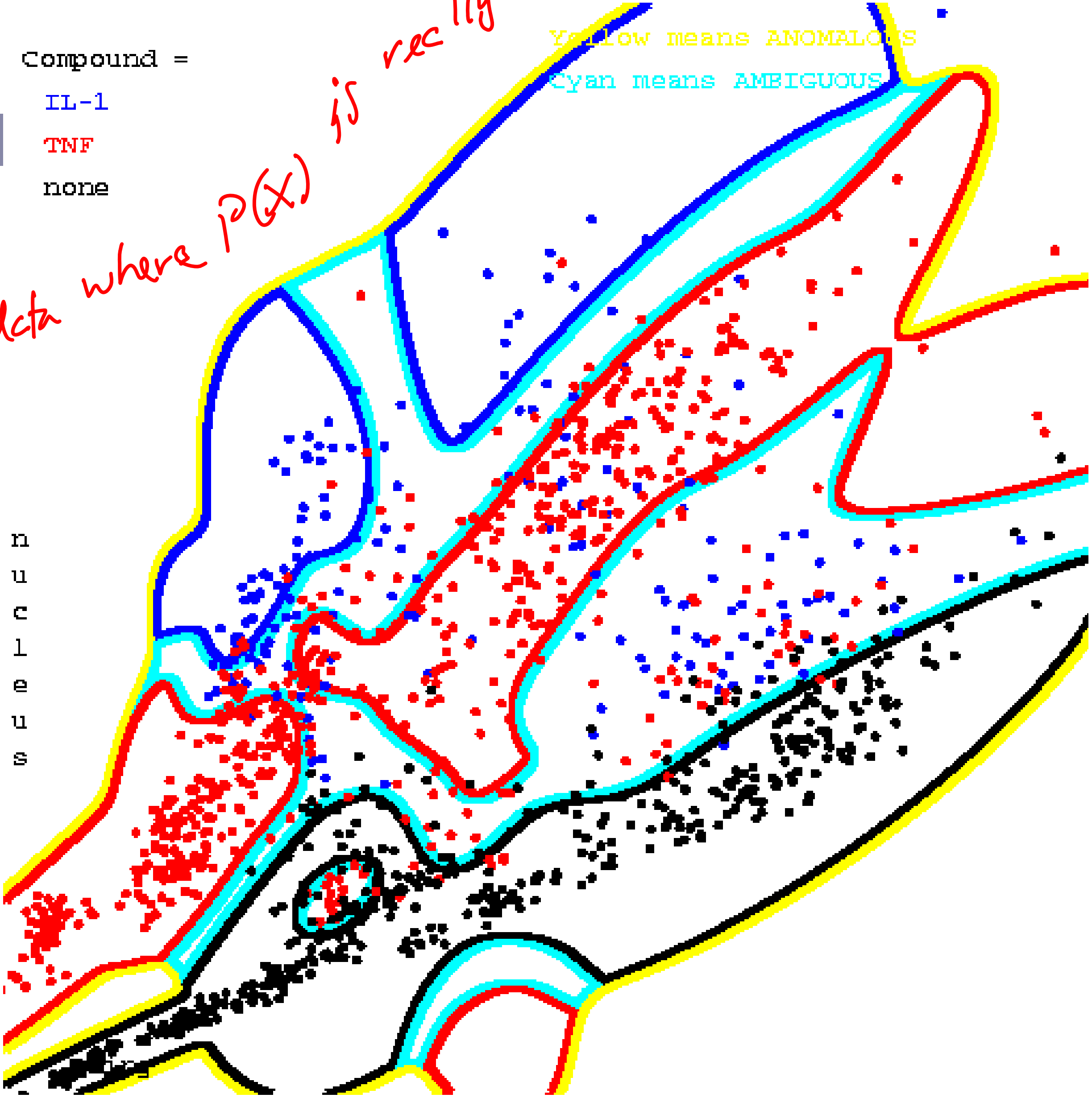


Resulting Bayes Classifier, using posterior probabilities to alert about ambiguity and anomalousness

a data where p(x) is really small

Yellow means anomalous

Cyan means ambiguous



The general learning problem with missing data

- Marginal likelihood – **x** is observed, **z** is missing:

$$
\begin{aligned}
\underbrace{\ell(\theta:\mathcal{D})} &= \log \prod_{j=1}^{m} \underline{P(\mathbf{x}_j \mid \theta)}\ \text{observed parts} \\
&= \sum_{j=1}^{m} \log P(\mathbf{x}_j \mid \theta) \\
&= \underbrace{\sum_{j=1}^{m} \log \sum_{\mathbf{z}} P(\mathbf{x}_j, \mathbf{z} \mid \theta)}_{\uparrow\ \text{sum over hidden vars (marginalize out)}}
\end{aligned}
$$

E-step

- **x** is observed, **z** is missing

at t's iteration

- Compute probability of missing data given current choice of θ
  - □ Q(**z**|**x**$_j$) for each **x**$_j$
    - e.g., probability computed during classification step
    - corresponds to "classification step" in K-means

iteration

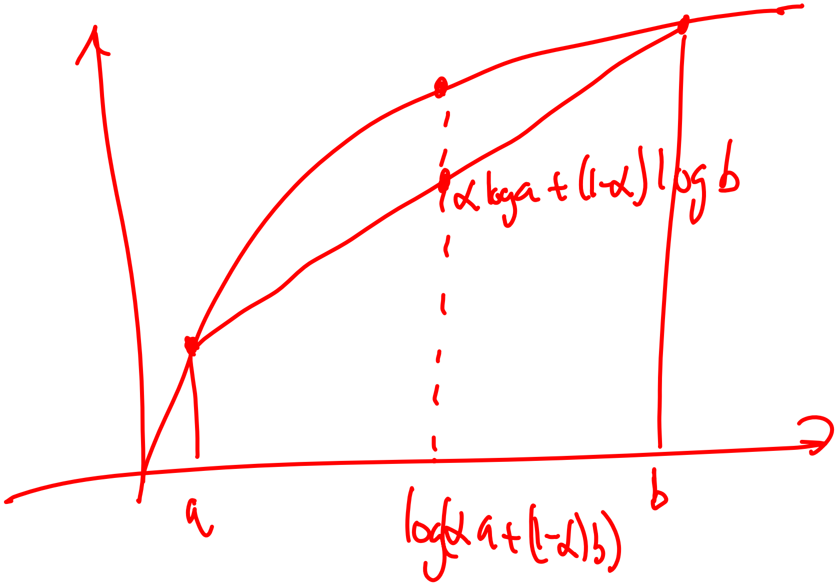
$$\underbrace{Q^{(t+1)}(\mathbf{z} \mid \mathbf{x}_j) \;=\; P(\mathbf{z} \mid \mathbf{x}_j, \theta^{(t)})}$$

Jensen's inequality

$\log 2 = \log(1+1) \geq \underset{0}{\log 1} + \underset{0}{\log 1}$

$$\ell(\theta : \mathcal{D}) = \sum_{j=1}^{m} \log \sum_{\mathbf{z}} P(\mathbf{z} \mid \mathbf{x}_j) P(\mathbf{x}_j \mid \theta)$$

- **Theorem**: log $\sum_{\mathbf{z}}$ P(**z**) f(**z**) ≥ $\sum_{\mathbf{z}}$ P(**z**) log f(**z**)



$\log \frac{a}{b} = \log a - \log b$

Applying Jensen's inequality

$\sum_z Q^{(t+1)}(z|x_j)=1$

$\ell(\theta:D) = \sum_j \log \sum_z P(x,z|\theta)$

- Use: log $\sum_z$ P(**z**) f(**z**) ≥ $\sum_z$ P(**z**) log f(**z**)

$$\ell(\theta^{(t)} : \mathcal{D}) \; = \; \sum_{j=1}^{m} \log \sum_{\mathbf{z}} Q^{(t+1)}(\mathbf{z} \mid \mathbf{x}_j) \frac{P(\mathbf{z}, \mathbf{x}_j \mid \theta^{(t)})}{Q^{(t+1)}(\mathbf{z} \mid \mathbf{x}_j)}$$

$$\overset{\text{Jensens}}{\geq} \sum_{j=1}^{m} \sum_{z} Q^{(t+1)}(z|x_j) \log \frac{P(z,x_j|\theta^{(t)})}{Q^{(t+1)}(z|x_j)}$$

$$= \underbrace{\sum_{j=1}^{m} \sum_{z} Q^{(t+1)}(z|x_j) \log P(z,x_j|\theta^{t})}_{\text{weighted log-likelihood of fully observable data}} - \underbrace{\sum_{j=1}^{m} \sum_{z} Q^{(t+1)}(z|x_j) \log Q^{(t+1)}(z|x_j)}_{m \cdot \hat{H}(Q^{(t+1)})}$$

The M-step maximizes lower bound on weighted data

fix Q in the M step

maximize over θ

fix Q doesn't depend on θ

- Lower bound from Jensen's:

$$\ell(\theta^{\cancel{(t)}} : \mathcal{D}) \geq \max_{\theta} \sum_{j=1}^{m} \sum_{\mathbf{z}} Q^{(t+1)}(\mathbf{z} \mid \mathbf{x}_j) \log P(\mathbf{z}, \mathbf{x}_j \mid \theta^{\cancel{(t)}}) + m.H(Q^{(t+1)})$$

$$\equiv \max_{\theta} \sum_{j=1}^{m} \sum_{z} Q^{(t+1)}(z \mid x_j) \log P(z, x_j \mid \theta)$$

a data point x_j with z hidden

~~repeat~~

- Corresponds to weighted dataset:
  - □ <$\mathbf{x}_1$,**z**=1> with weight $Q^{(t+1)}$(**z**=1|$\mathbf{x}_1$)
  - □ <$\mathbf{x}_1$,**z**=2> with weight $Q^{(t+1)}$(**z**=2|$\mathbf{x}_1$)
  - □ <$\mathbf{x}_1$,**z**=3> with weight $Q^{(t+1)}$(**z**=3|$\mathbf{x}_1$)
  - □ <$\mathbf{x}_2$,**z**=1> with weight $Q^{(t+1)}$(**z**=1|$\mathbf{x}_2$)
  - □ <$\mathbf{x}_2$,**z**=2> with weight $Q^{(t+1)}$(**z**=2|$\mathbf{x}_2$)
  - □ <$\mathbf{x}_2$,**z**=3> with weight $Q^{(t+1)}$(**z**=3|$\mathbf{x}_2$)
  - □ …

a repeat $|Z|$ times

weigh by $Q^{(t+1)}$

The M-step

$$\ell(\theta^{(t)} : \mathcal{D}) \ \geq \ \sum_{j=1}^{m} \sum_{\mathbf{z}} Q^{(t+1)}(\mathbf{z} \mid \mathbf{x}_j) \log P(\mathbf{z}, \mathbf{x}_j \mid \theta^{(t)}) + m.H(Q^{(t+1)})$$

- Maximization step:

$$\underbrace{\theta^{(t+1)} \leftarrow \arg\max_{\theta} \sum_{j=1}^{m} \sum_{\mathbf{z}} Q^{(t+1)}(\mathbf{z} \mid \mathbf{x}_j) \log P(\mathbf{z}, \mathbf{x}_j \mid \theta)}_{}$$

- Use expected counts instead of counts:
  - □ If learning requires Count(**x**,**z**)
  - □ Use $E_{Q(t+1)}$[Count(**x**,**z**)]

like in Boosting

Convergence of EM

- Define potential function F(θ,Q):

$$\ell(\theta : \mathcal{D}) \underset{\overset{\uparrow}{\text{Jensen's}}}{\geq} \underline{\underline{F(\theta, Q)}} = \underbrace{\sum_{j=1}^{m} \sum_{\mathbf{z}} Q(\mathbf{z} \mid \mathbf{x}_j) \log \frac{P(\mathbf{z}, \mathbf{x}_j \mid \theta)}{Q(\mathbf{z} \mid \mathbf{x}_j)}}$$

- EM corresponds to coordinate ascent on F
  - □ Thus, maximizes lower bound on marginal log likelihood

M-step is easy

$$\underline{\theta^{(t+1)}} \leftarrow \arg\max_{\theta} \sum_{j=1}^{m} \sum_{\mathbf{z}} Q^{(t+1)}(\mathbf{z} \mid \mathbf{x}_j) \log P(\mathbf{z}, \mathbf{x}_j \mid \theta)$$

- Using potential function

max this part $\equiv \max_{\theta} F(\theta, Q^{(t+1)})$

$$\underbrace{F(\theta, Q^{(t+1)})}_{\text{fixed Q to } Q^{(t+1)}} = \overbrace{\sum_{j=1}^{m} \sum_{\mathbf{z}} Q^{(t+1)}(\mathbf{z} \mid \mathbf{x}_j) \log P(\mathbf{z}, \mathbf{x}_j \mid \theta)} + \overbrace{m.H(Q^{(t+1)})}^{\text{constant}}$$

E-step also doesn't decrease potential function 1

$$\log \frac{a \cdot b}{c} = \log \frac{a}{c} + \log b$$

- Fixing θ to $\theta^{(t)}$:

fix

maximize over Q

$$\ell(\theta^{(t)}:\mathcal{D}) \;\geq\; F(\theta^{(t)},Q) = \sum_{j=1}^{m}\sum_{\mathbf{z}} Q(\mathbf{z}\mid\mathbf{x}_j)\log\frac{P(\mathbf{z},\mathbf{x}_j\mid\theta^{(t)})}{Q(\mathbf{z}\mid\mathbf{x}_j)}$$

$$= \sum_{j=1}^{m}\sum_{z} Q(z|x_j)\log\frac{P(z|x_j,\theta^{(t)})\cdot P(x_j|\theta^{(t)})}{Q(z|x_j)}$$

chain rule

$$= \underbrace{\sum_{j=1}^{m}\sum_{z} Q(z|x_j)\log\frac{P(z|x_j,\theta^{(t)})}{Q(z|x_j)}}_{-\text{KL-divergence}} + \underbrace{\sum_{j=1}^{m}\sum_{z} Q(z|x_j)\log P(x_j|\theta^{(t)})}_{\underbrace{\sum_{j=1}^{m}\log P(x_j|\theta^{(t)})}_{\ell(\theta^{(t)}:D)}\cdot\left[\cancelto{1}{\sum_{z} Q(z|x_j)}\right]}$$

KL-divergence

- Measures distance between distributions

$$KL(Q||P) = \sum_z Q(z) \log \frac{Q(z)}{P(z)}$$

- KL=zero if and only if Q=P

E-step also doesn't decrease potential function 2

$KL(Q||P)=0 \iff Q=P$

- Fixing θ to θ^(t):

$$\underbrace{\ell(\theta^{(t)}:\mathcal{D})} \geq F(\theta^{(t)},Q) = \ell(\theta^{(t)}:\mathcal{D}) + \sum_{j=1}^{m}\sum_{\mathbf{z}} Q(\mathbf{z}\mid\mathbf{x}_j)\log\frac{P(\mathbf{z}\mid\mathbf{x}_j,\theta^{(t)})}{Q(\mathbf{z}\mid\mathbf{x}_j)}$$

$$= \underbrace{\ell(\theta^{(t)}:\mathcal{D})} - \cancel{m}\sum_{j=1}^{m} \underbrace{KL\left(Q(\mathbf{z}\mid\mathbf{x}_j)||P(\mathbf{z}\mid\mathbf{x}_j,\theta^{(t)})\right)}_{\geq 0 \iff Q=P}$$

$$\Rightarrow Q^{(t+1)}(z|x_j) = P(z|x_j,\theta^{(t)})$$

$$\Rightarrow \ell(\theta^{(t)}:\mathcal{D}) = F(\theta^{(t)}, Q^{(t+1)})$$

E-step also doesn't decrease potential function 3

$$\ell(\theta^{(t)} : \mathcal{D}) \geq F(\theta^{(t)}, Q) \quad = \quad \ell(\theta^{(t)} : \mathcal{D}) - \cancel{m} \sum_{j=1}^{m} KL\left(Q(\mathbf{z} \mid \mathbf{x}_j) || P(\mathbf{z} \mid \mathbf{x}_j, \theta^{(t)})\right)$$

- Fixing θ to $\theta^{(t)}$
- Maximizing $F(\theta^{(t)},Q)$ over Q → set Q to posterior probability:

$$\boxed{Q^{(t+1)}(\mathbf{z} \mid \mathbf{x}_j) \quad \leftarrow \quad P(\mathbf{z} \mid \mathbf{x}_j, \theta^{(t)})}$$

- Note that

$$\boxed{F(\theta^{(t)}, Q^{(t+1)}) \quad = \quad \ell(\theta^{(t)} : \mathcal{D})}$$

EM is coordinate ascent

$$\ell(\theta : \mathcal{D}) \;\geq\; F(\theta, Q) = \sum_{j=1}^{m} \sum_{\mathbf{z}} Q(\mathbf{z} \mid \mathbf{x}_j) \log \frac{P(\mathbf{z}, \mathbf{x}_j \mid \theta)}{Q(\mathbf{z} \mid \mathbf{x}_j)}$$

- **M-step**: Fix Q, maximize F over θ (a lower bound on $\ell(\theta : \mathcal{D})$):

$$\ell(\theta : \mathcal{D}) \;\geq\; F(\theta, Q^{(t)}) = \sum_{j=1}^{m} \sum_{\mathbf{z}} Q^{(t)}(\mathbf{z} \mid \mathbf{x}_j) \log P(\mathbf{z}, \mathbf{x}_j \mid \theta) + m.H(Q^{(t)})$$

- **E-step**: Fix θ, maximize F over Q:

$$\ell(\theta^{(t)} : \mathcal{D}) \geq F(\theta^{(t)}, Q) \;=\; \ell(\theta^{(t)} : \mathcal{D}) - m \sum_{j=1}^{m} KL\left(Q(\mathbf{z} \mid \mathbf{x}_j) || P(\mathbf{z} \mid \mathbf{x}_j, \theta^{(t)})\right)$$

  - "Realigns" F with likelihood:

$$F(\theta^{(t)}, Q^{(t+1)}) \;=\; \ell(\theta^{(t)} : \mathcal{D})$$

What you should know

- K-means for clustering:
  - algorithm
  - converges because it’s coordinate ascent
- EM for mixture of Gaussians:
  - How to “learn” maximum likelihood parameters (locally max. like.) in the case of unlabeled data
- Be happy with this kind of probabilistic analysis
- Remember, E.M. can get stuck in local minima, and empirically it <u>DOES</u>
- EM is coordinate ascent
- General case for EM

Acknowledgements

- K-means & Gaussian mixture models presentation contains material from excellent tutorial by Andrew Moore:
  - □ http://www.autonlab.org/tutorials/
- K-means Applet:
  - □ http://www.elet.polimi.it/upload/matteucc/Clustering/tutorial_html/AppletKM.html
- Gaussian mixture models Applet:
  - □ http://www.neurosci.aist.go.jp/%7Eakaho/MixtureEM.html

EM for HMMs a.k.a. The Baum-Welch Algorithm

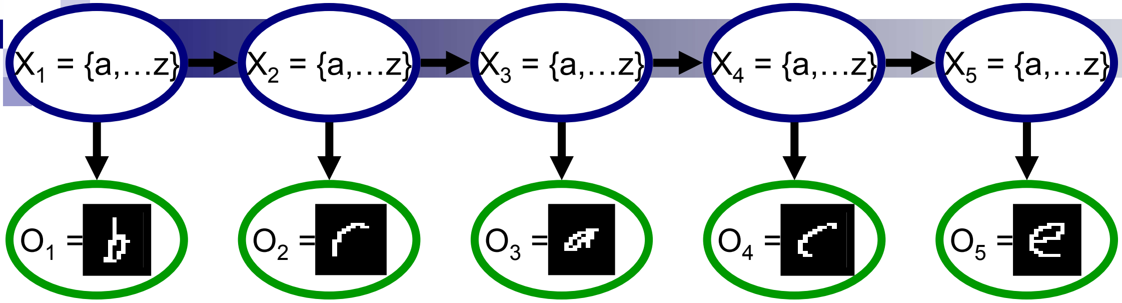
Machine Learning – 10701/15781

Carlos Guestrin

Carnegie Mellon University

April 9th, 2007

Learning HMMs from fully observable data is easy



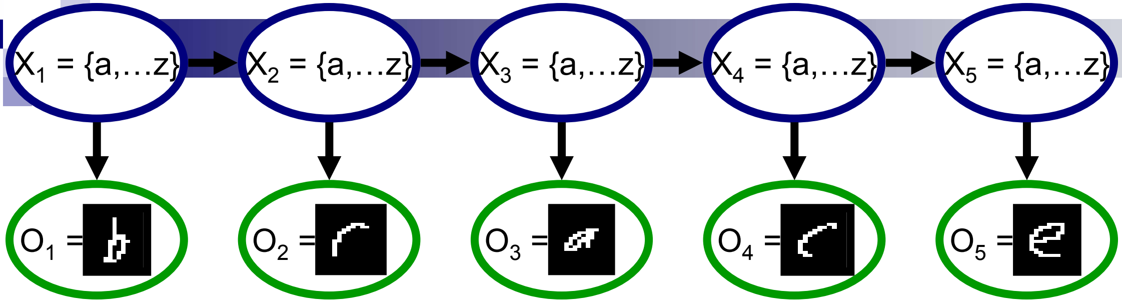
Learn 3 distributions:

$$P(X_1)$$

$$P(O_i \mid X_i)$$

$$P(X_i \mid X_{i-1})$$

Learning HMMs from fully observable data is easy



Learn 3 distributions:

$$P(X_1^{=a}) = \frac{\text{Count (\# first letter was a)}}{N = \text{dataset size}}$$

select training data where letter was a

$$P(O_i^{=\text{pixel 17 is white}} \mid X_i^{=a}) = \text{Count(pixel 17 was white, } X_i = a)$$

any … in

$$P(X_i^{=a} \mid X_{i-}^{=b}$$

)

What if **O** is observed, but **X** is hidden

©2005-2007 Carlos Guestrin

Log likelihood for HMMs when **X** is hidden

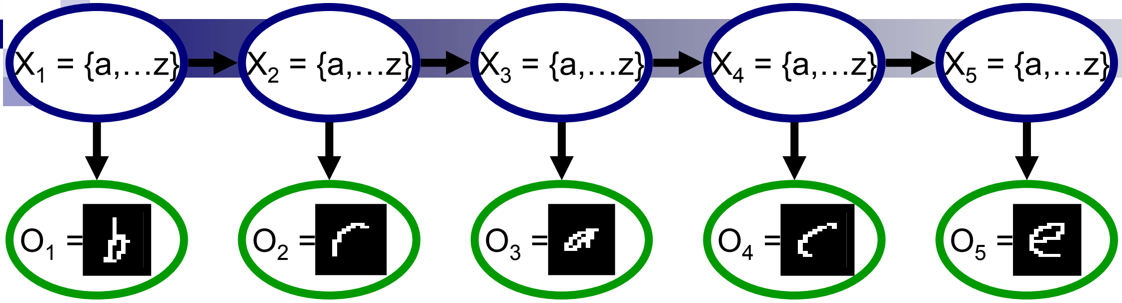
- Marginal likelihood – **O** is observed, **X** is missing
  - For simplicity of notation, training data consists of only one sequence:

$$\begin{aligned}\ell(\theta : \mathcal{D}) &= \log P(\mathbf{o} \mid \theta) \\ &= \log \sum_{\mathbf{x}} P(\mathbf{x}, \mathbf{o} \mid \theta)\end{aligned}$$

  - If there were m sequences:

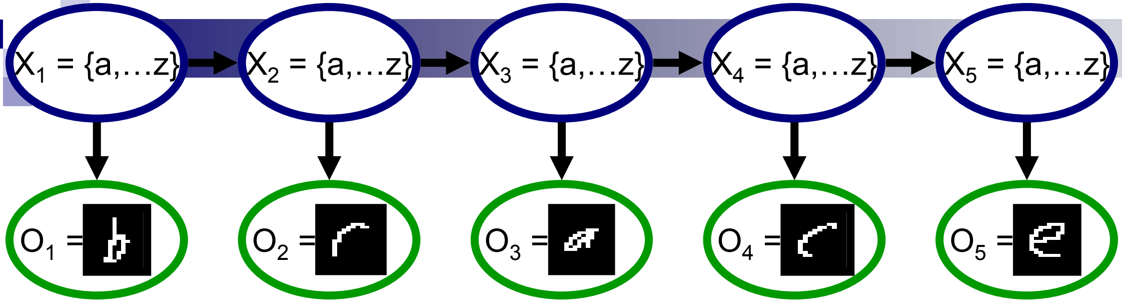
$$\ell(\theta : \mathcal{D}) = \sum_{j=1}^{m} \log \sum_{\mathbf{x}} P(\mathbf{x}, \mathbf{o}^{(j)} \mid \theta)$$

Computing Log likelihood for HMMs when **X** is hidden



$$\begin{aligned}\ell(\theta : \mathcal{D}) &= \log P(\mathbf{o} \mid \theta) \\ &= \log \sum_{\mathbf{x}} P(\mathbf{x}, \mathbf{o} \mid \theta)\end{aligned}$$

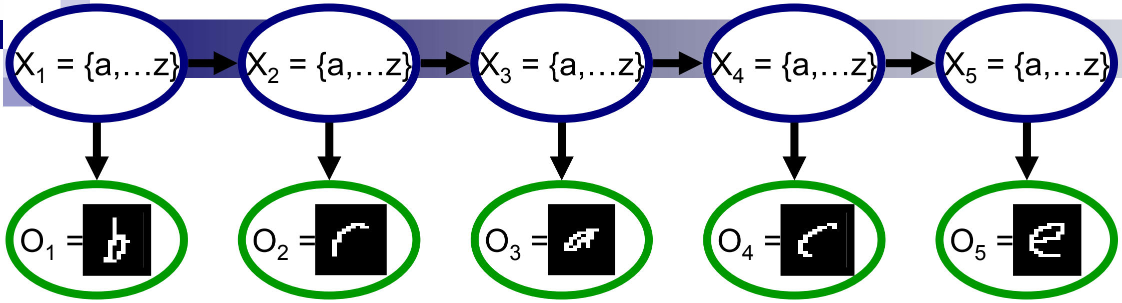
Computing Log likelihood for HMMs when **X** is hidden – variable elimination



- Can compute efficiently with variable elimination:

$$\begin{aligned}\ell(\theta : \mathcal{D}) &= \log P(\mathbf{o} \mid \theta) \\ &= \log \sum_{\mathbf{x}} P(\mathbf{x}, \mathbf{o} \mid \theta)\end{aligned}$$

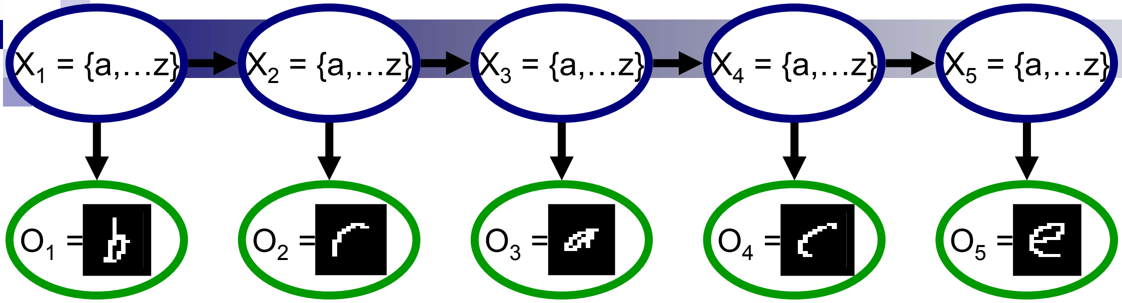
EM for HMMs when **X** is hidden



- E-step: Use inference (forwards-backwards algorithm)

- M-step: Recompute parameters with weighted data

E-step

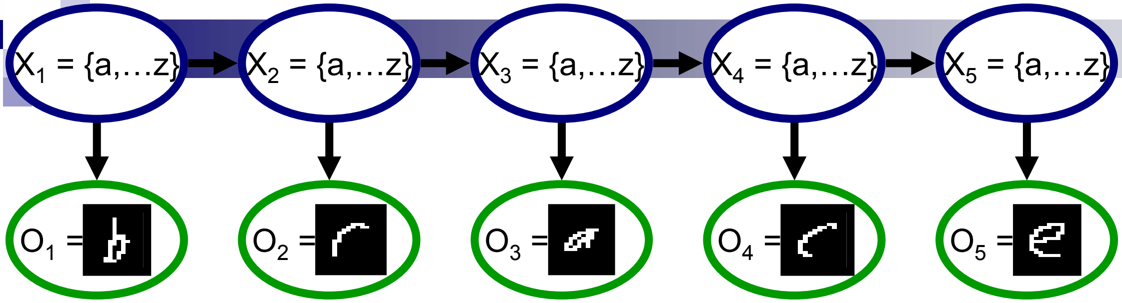


- E-step computes probability of hidden vars **x** given **o**

$$Q^{(t+1)}(\mathbf{x} \mid \mathbf{o}) \;=\; P(\mathbf{x} \mid \mathbf{o}, \theta^{(t)})$$

- Will correspond to inference
  - □ use forward-backward algorithm!

The M-step



- Maximization step:

$$\theta^{(t+1)} \leftarrow \arg\max_{\theta} \sum_{\mathbf{x}} Q^{(t+1)}(\mathbf{x} \mid \mathbf{o}) \log P(\mathbf{x}, \mathbf{o} \mid \theta)$$

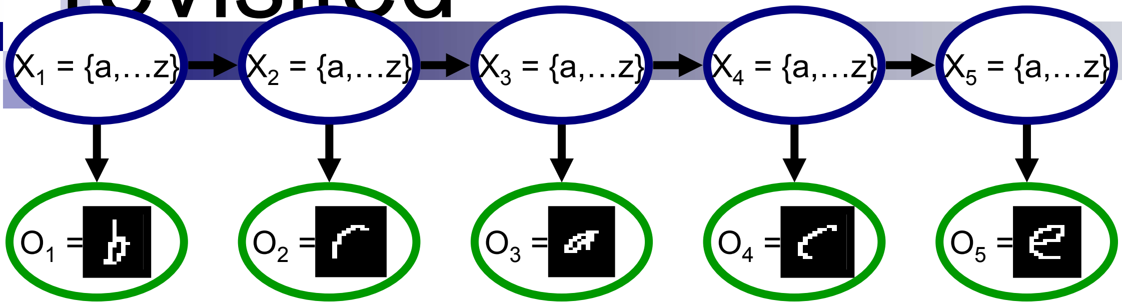
- Use expected counts instead of counts:
  - If learning requires Count(**x**,**o**)
  - Use $E_{Q(t+1)}$[Count(**x**,**o**)]

Decomposition of likelihood revisited

$$P(X_1)$$

$$P(O_i \mid X_i)$$

$$P(X_i \mid X_{i-1})$$



- Likelihood optimization decomposes:

$$\max_\theta \sum_{\mathbf{x}} Q(\mathbf{x} \mid \mathbf{o}) \log P(\mathbf{x}, \mathbf{o} \mid \theta) =$$

$$\max_\theta \sum_{\mathbf{x}} Q(\mathbf{x} \mid \mathbf{o}) \log P(x_1 \mid \theta_{X_1}) P(o_1 \mid x_1, \theta_{O|X}) \prod_{t=2}^{n} P(x_t \mid x_{t-1}, \theta_{X_t|X_{t-1}}) P(o_t \mid x_t, \theta_{O|X})$$

Starting state probability $P(X_1)$

- Using expected counts
  - $P(X_1{=}a) = \theta_{X1=a}$

$$\max_{\theta_{X_1}} \sum_{\mathbf{x}} Q(\mathbf{x} \mid \mathbf{o}) \log P(x_1 \mid \theta_{X_1})$$

$$\boxed{\theta_{X_1=a} = \frac{\sum_{j=1}^{m} Q(X_1 = a \mid \mathbf{o}^{(j)})}{m}}$$

Transition probability $P(X_t|X_{t-1})$

- Using expected counts
  - □ $P(X_t=a|X_{t-1}=b) = \theta_{Xt=a|Xt-1=b}$

$$\max_{\theta_{X_t|X_{t-1}}} \sum_{\mathbf{x}} Q(\mathbf{x} \mid \mathbf{o}) \log \prod_{t=2}^{n} P(x_t \mid x_{t-1}, \theta_{X_t|X_{t-1}})$$

$$\boxed{\theta_{X_t=a|X_{t-1}=b} \quad = \quad \frac{\sum_{j=1}^{m} \sum_{t=2}^{n} Q(X_t = a, X_{t-1} = b \mid \mathbf{o}^{(j)})}{\sum_{j=1}^{m} \sum_{t=2}^{n} \sum_{i=1}^{k} Q(X_t = i, X_{t-1} = b \mid \mathbf{o}^{(j)})}}$$

Observation probability $P(O_t|X_t)$

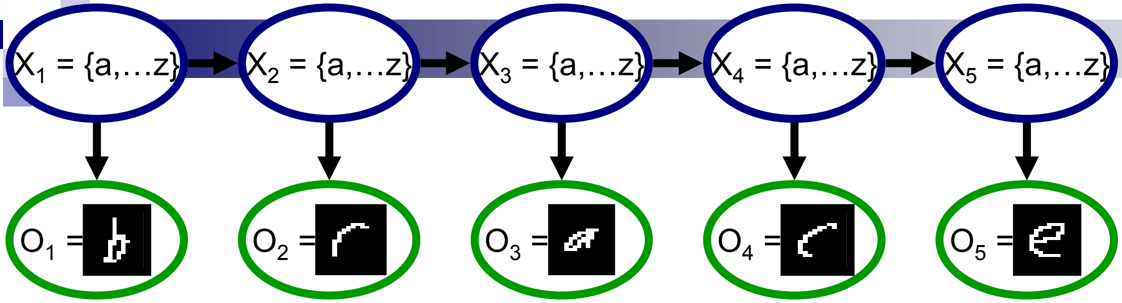
- Using expected counts
  - □ $P(O_t=a|X_t=b) = \theta_{Ot=a|Xt=b}$

$$\max_{\theta_{O|X}} \sum_{\mathbf{x}} Q(\mathbf{x} \mid \mathbf{o}) \log \prod_{t=1}^{n} P(o_t \mid x_t, \theta_{O|X})$$

$$\boxed{\theta_{O_t=a|X_t=b} \quad = \quad \frac{\sum_{j=1}^{m} \sum_{t=1}^{n} \delta(\mathbf{o}_t^{(j)} = a) Q(X_t = b \mid \mathbf{o}^{(j)})}{\sum_{j=1}^{m} \sum_{t=1}^{n} Q(X_t = b \mid \mathbf{o}^{(j)})}}$$

E-step revisited

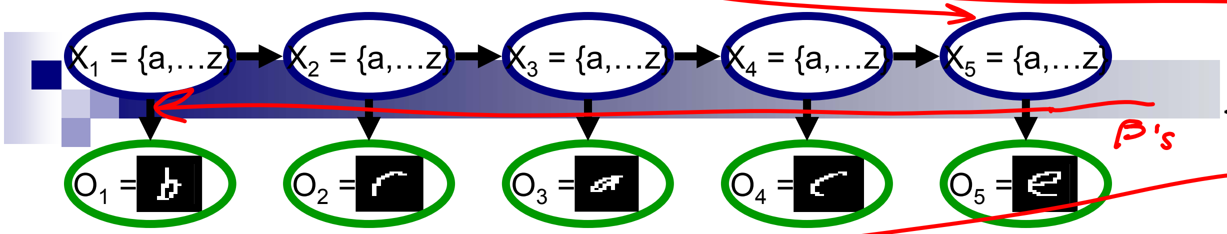
$$Q^{(t+1)}(\mathbf{x} \mid \mathbf{o}) \;=\; P(\mathbf{x} \mid \mathbf{o}, \theta^{(t)})$$



- E-step computes probability of hidden vars **x** given **o**
- Must compute:
  - □ Q(x_t=a|**o**) – marginal probability of each position
  - □ Q(x_{t+1}=a,x_t=b|**o**) – joint distribution between pairs of positions

α's

The forwards-backwards algorithm



$$P(X_i \mid o_{1..n})$$

drop O_1 in notation

- Initialization: $\underline{\alpha_1(X_1)} = \underline{P(X_1)}\,\underline{P(o_1 \mid X_1)}$
- For i = 2 to n
  - □ Generate a forwards factor by eliminating X_{i-1}

sum out previous var

$$\underline{\alpha_i(X_i)} = \sum_{x_{i-1}} \overbrace{P(o_i \mid X_i)}^{\text{prob obs}}\,\overbrace{P(X_i \mid X_{i-1} = x_{i-1})}^{\text{transition prob}}\,\overbrace{\alpha_{i-1}(x_{i-1})}$$

$\alpha_5(a)$
$\alpha_5(b)$
$\vdots$
$\alpha_5(z)$

$\alpha_n(X_n)$ normalized $= P(X_n \mid O_{1:n})$

$\beta_1(X_1)\alpha_1(X_1)$ normalized $= P(X_1 \mid O_{1:n})$

- Initialization: $\underline{\beta_n(X_n) = 1}$
- For i = n-1 to 1
  - □ Generate a backwards factor by eliminating X_{i+1}

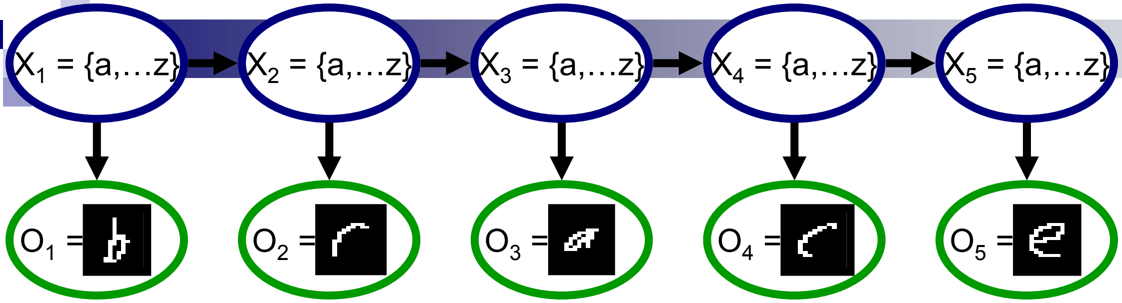
$\forall x_i$

$$\beta_i(\overset{=x_i}{X_i}) = \sum_{x_{i+1}} P(o_{i+1} \mid x_{i+1})\,P(x_{i+1} \mid \overset{=x_i}{X_i})\,\beta_{i+1}(x_{i+1})$$

- ∀ i, probability is: $\boxed{P(X_i \mid o_{1..n}) \propto \alpha_i(X_i)\beta_i(X_i)}$

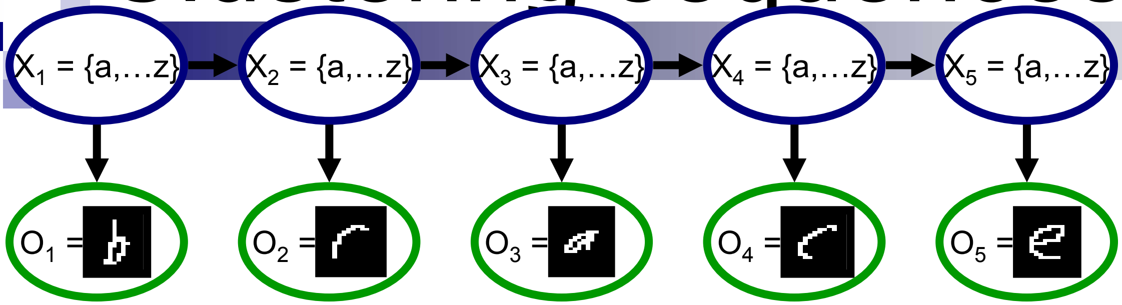
E-step revisited

$$Q^{(t+1)}(\mathbf{x} \mid \mathbf{o}) \;=\; P(\mathbf{x} \mid \mathbf{o}, \theta^{(t)})$$



- E-step computes probability of hidden vars **x** given **o**
- Must compute:
  - □ Q(x_t=a|**o**) – marginal probability of each position
    - Just forwards-backwards!
  - □ Q(x_{t+1}=a,x_t=b|**o**) – joint distribution between pairs of positions
    - Homework! ☺

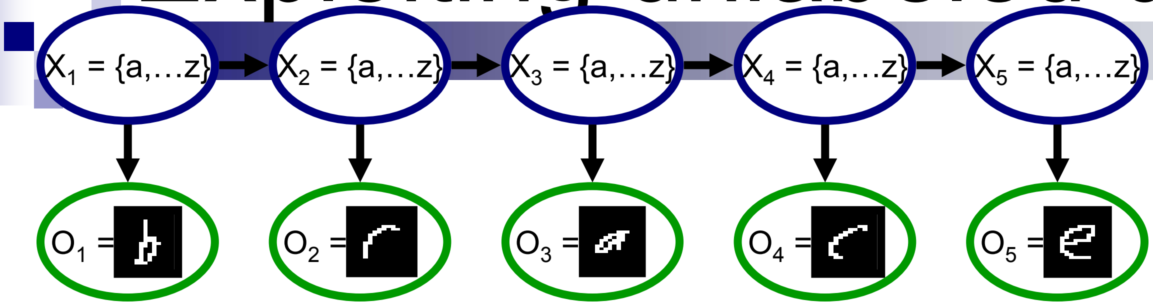
What can you do with EM for HMMs? 1 – Clustering sequences



Independent clustering:

Sequence clustering:

What can you do with EM for HMMs? 2 – Exploiting unlabeled data



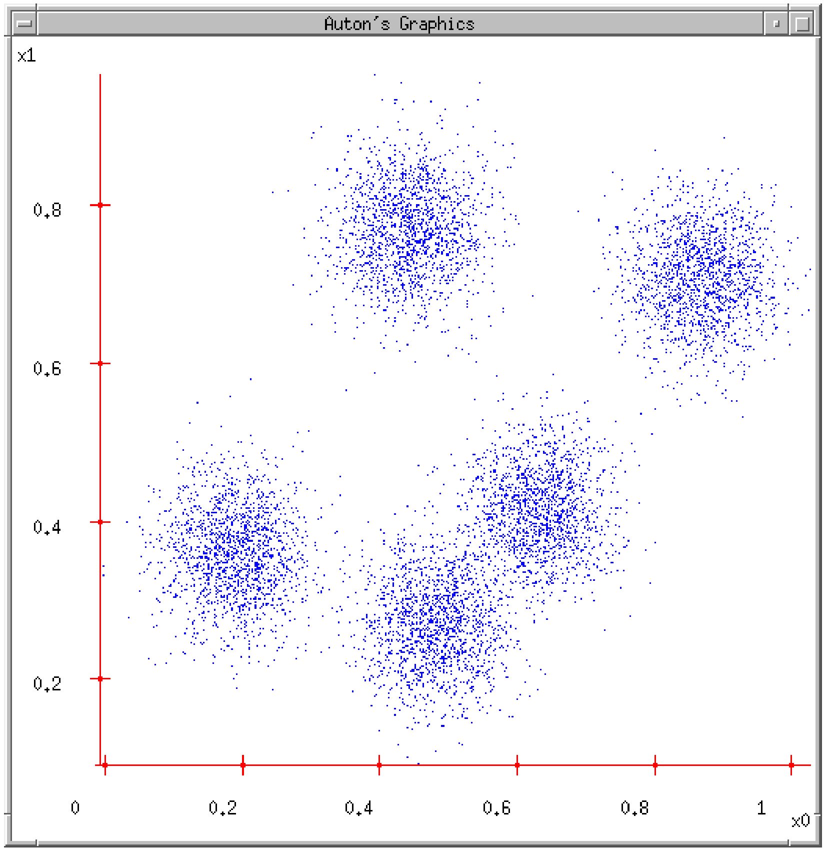
$X_1 = \{a,\dots,z\}$ $X_2 = \{a,\dots,z\}$ $X_3 = \{a,\dots,z\}$ $X_4 = \{a,\dots,z\}$ $X_5 = \{a,\dots,z\}$

$O_1 =$ $O_2 =$ $O_3 =$ $O_4 =$ $O_5 =$

- Labeling data is hard work → save (graduate student) time by using both labeled and unlabeled data
  - □ Labeled data:
    - <X="brace",O= >
  - □ Unlabeled data:
    - <X=?????,O= >

Exploiting unlabeled data in clustering

- A few data points are labeled
  - <x,o>
- Most points are unlabeled
  - <?,o>
- In the E-step of EM:
  - If i’th point is unlabeled:
    - compute $Q(X|o_i)$ as usual
  - If i’th point is labeled:
    - set $Q(X=x|o_i)=1$ and $Q(X\neq x|o_i)=0$
- M-step as usual



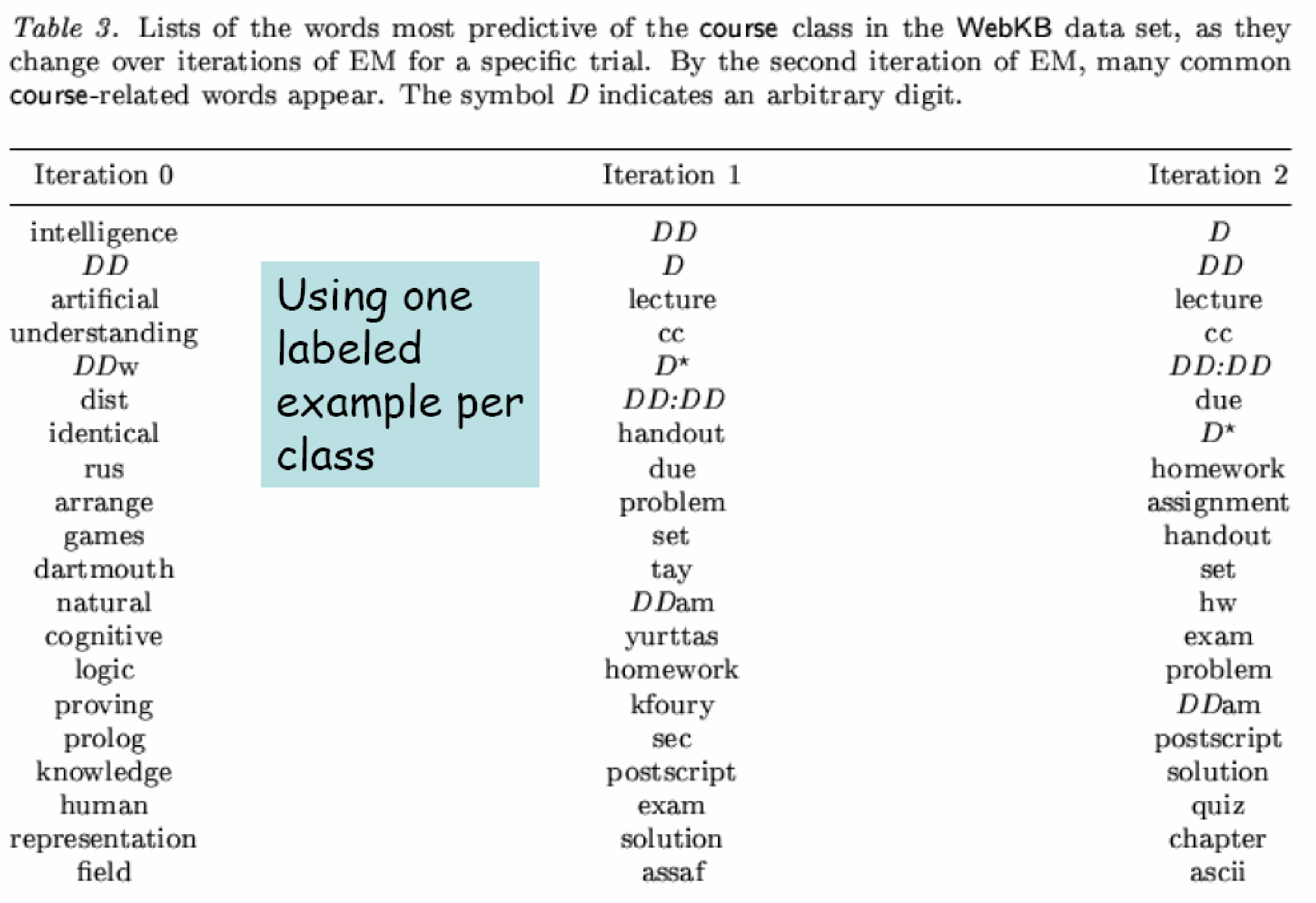
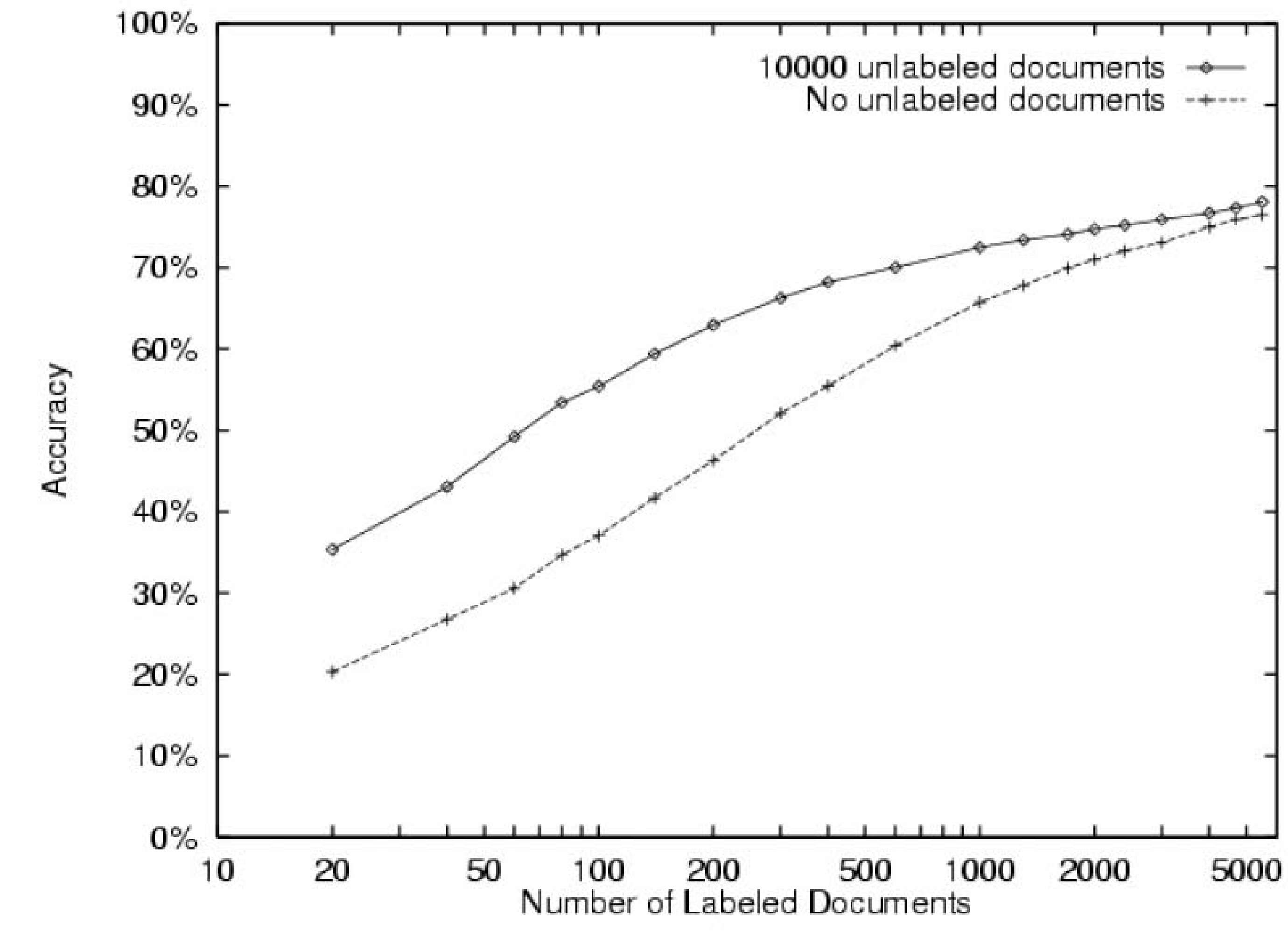


Table 3. Lists of the words most predictive of the course class in the WebKB data set, as they change over iterations of EM for a specific trial. By the second iteration of EM, many common course-related words appear. The symbol D indicates an arbitrary digit.

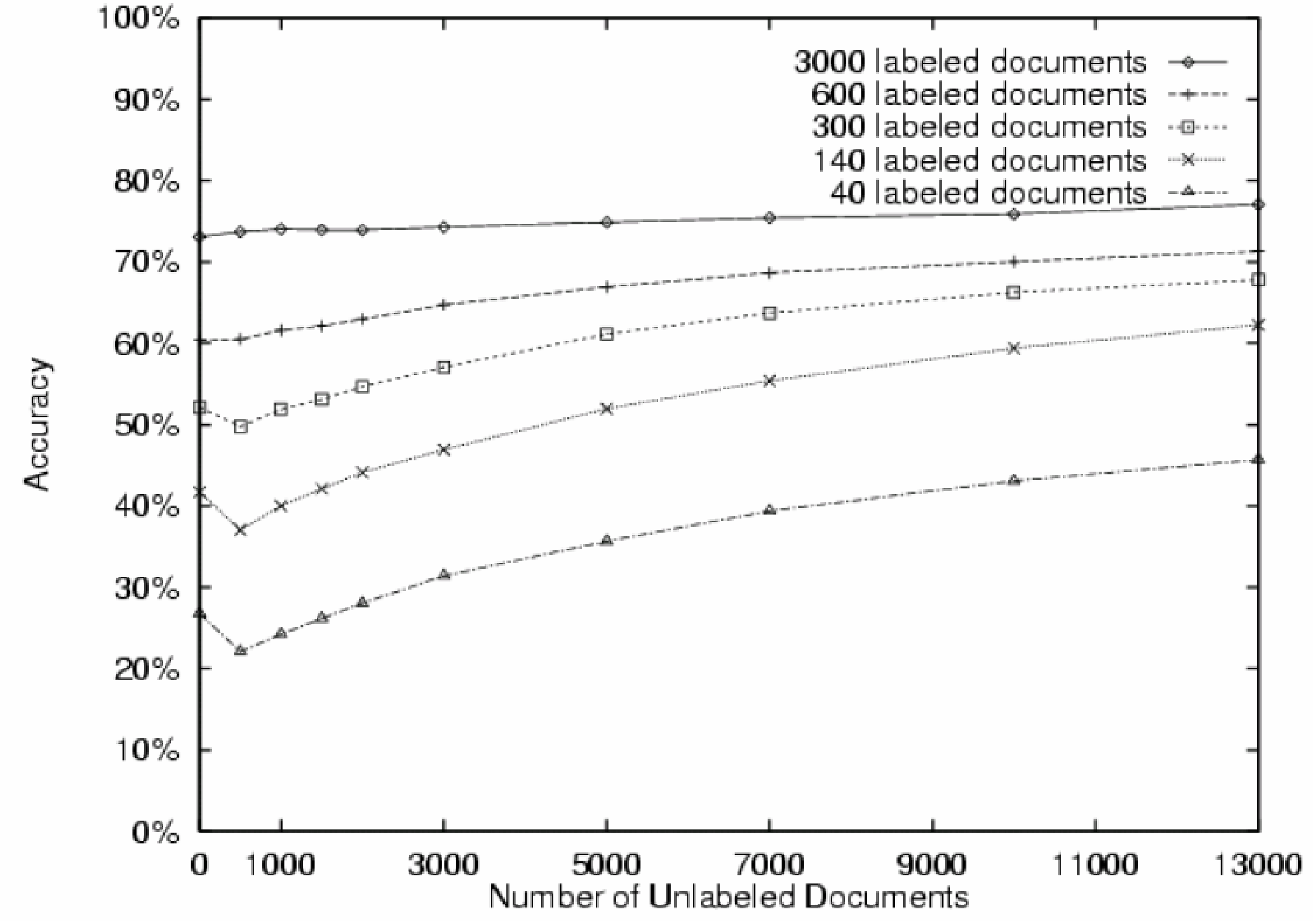
Iteration 0	Iteration 1	Iteration 2
intelligence	DD	D
DD	D	DD
artificial	lecture	lecture
understanding	cc	cc
DDw	$D^\star$	$DD{:}DD$
dist	$DD{:}DD$	due
identical	handout	$D^\star$
rus	due	homework
arrange	problem	assignment
games	set	handout
dartmouth	tay	set
natural	DDam	hw
cognitive	yurttas	exam
logic	homework	problem
proving	kfoury	DDam
prolog	sec	postscript
knowledge	postscript	solution
human	exam	quiz
representation	solution	chapter
field	assaf	ascii

Using one labeled example per class

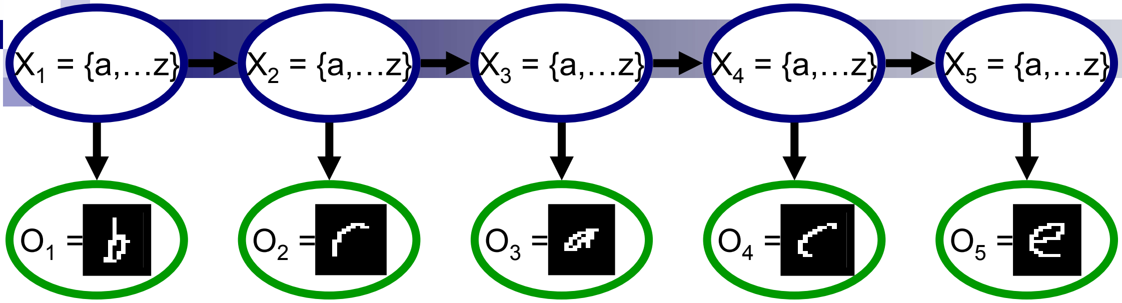
20 Newsgroups data – advantage of adding unlabeled data



20 Newsgroups data – Effect of additional unlabeled data



Exploiting unlabeled data in HMMs



- A few data points are labeled
  - □ <x,o>
- Most points are unlabeled
  - □ <?,o>
- In the E-step of EM:
  - □ If i'th point is unlabeled:
    - compute $Q(X|o_i)$ as usual
  - □ If i'th point is labeled:
    - set $Q(X=x|o_i)=1$ and $Q(X\neq x|o_i)=0$
- M-step as usual
  - □ Speed up by remembering counts for labeled data

What you need to know

- Baum-Welch = EM for HMMs
- E-step:
  - Inference using forwards-backwards
- M-step:
  - Use weighted counts
- Exploiting unlabeled data:
  - Some unlabeled data can help classification
  - Small change to EM algorithm
    - In E-step, only use inference for unlabeled data

Acknowledgements

- Experiments combining labeled and unlabeled data provided by Tom Mitchell