Case Study 2: Document Retrieval

Review: Mixtures of Gaussians

Machine Learning for Big Data
CSE547/STAT548, University of Washington

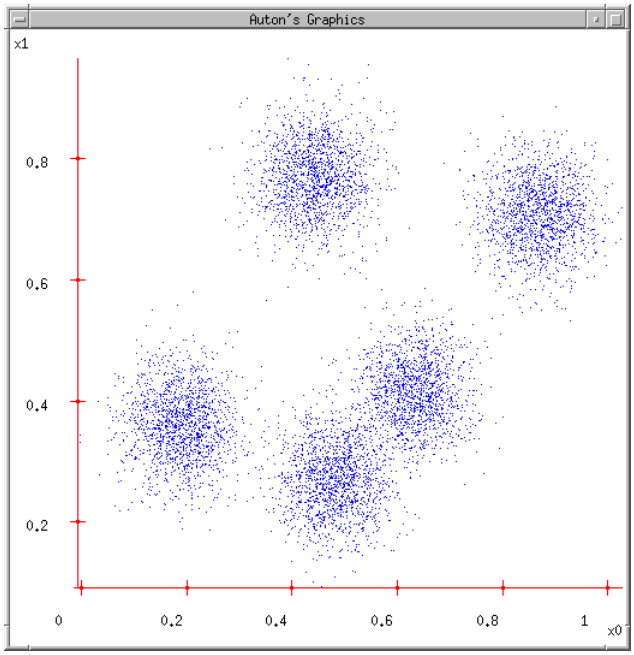
Emily Fox

April 21st, 2015

Some Data

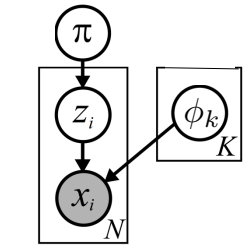
want to cluster

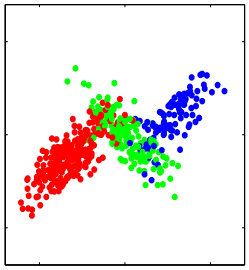
- unsup.
- generative approach



Gaussian Mixture Model

- Most commonly used mixture model
- Observations: $x^1, \dots, x^N$ $\quad x^i \in \mathbb{R}^d$
- Parameters: mix. weights
  $\pi = [\pi_1, \dots, \pi_K]$ # clusters
  $\phi = \{\phi_k\} = \{\mu_k, \Sigma_k\}$ params for cluster k
- Cluster indicator:
  $z^i \in \{1, \dots, K\}$ $\quad \Pr(z^i = k) = \pi_k$
- Per-cluster likelihood:
  $N(x^i \mid \mu_k, \Sigma_k, z^i = k)$
- Ex. z^i = country of origin, x^i = height of i[th] person
  - k^{th} mixture component = distribution of heights in country k





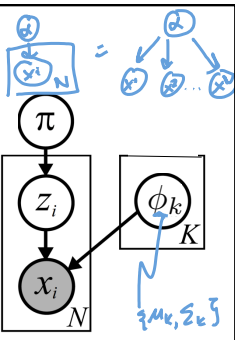
Generative Model

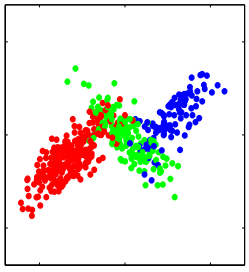
- We can think of *sampling* observations from the model
- For each observation *i*,
  - Sample a cluster assignment
    $z^i \sim \pi$ sample from or "drawn" from
  - Sample the observation from the selected Gaussian
    $x^i \mid z^i \sim N(x^i \mid \mu_{z^i}, \Sigma_{z^i})$
    ↑ given, "conditioned upon"

pmf $\pi_1, \pi_2, \pi_3, \dots, \pi_K$ over $1, 2, \dots, K$; $\sum_{i=1}^{K} \pi_i = 1$

$\{\mu_k, \Sigma_k\}$

can "generate" obs.

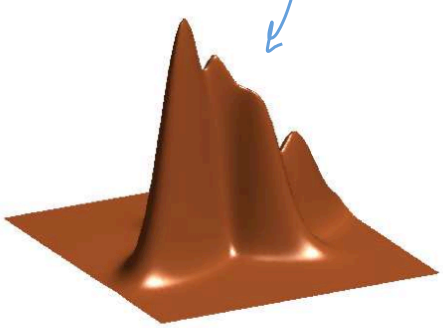


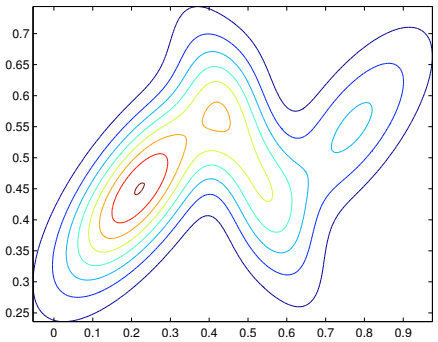


Also Useful for Density Estimation

$x^i \in \mathbb{R}^2$

$p(x^i \mid \theta)$



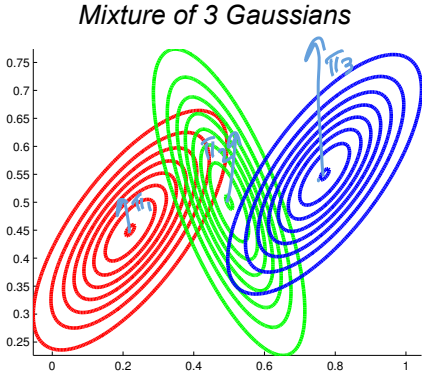


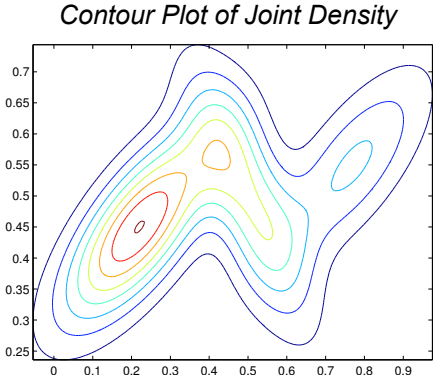
bird's eye view

Density as Mixture of Gaussians

- Approximate density with a mixture of Gaussians

$K =$



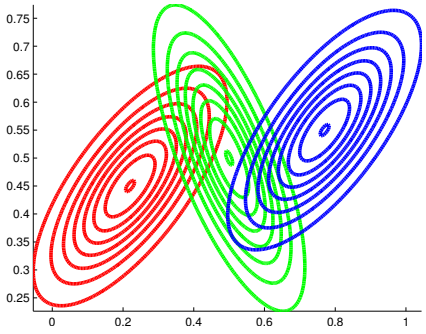


Each Gauss. has weight π_k $(\sum \pi_k = 1)$ and shape params $\{\mu_k, \Sigma_k\}$

Density as Mixture of Gaussians

- Approximate density with a mixture of Gaussians

Mixture of 3 Gaussians

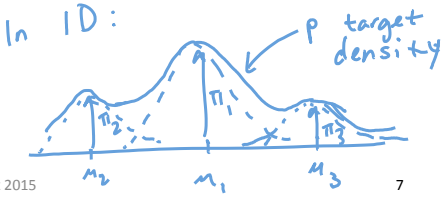


$$p(x^i \mid \pi, \mu, \Sigma) = \sum_{k=1}^{K} \pi_k N(x^i \mid \mu_k, \Sigma_k)$$

$\theta = [\pi_1, \ldots, \pi_K, \{\mu_k, \Sigma_k\}]$

$p(z^i = k)$ $\quad p(x^i \mid z^i, \phi)$

In 1D:



Summary of GMM Components

- Observations $x_i \in \mathbb{R}^d, \quad i = 1, 2, \ldots, N$
- Hidden cluster labels $z_i \in \{1, 2, \ldots, K\}, \quad i = 1, 2, \ldots, N$
- Hidden mixture means $\mu_k \in \mathbb{R}^d, \quad k = 1, 2, \ldots, K$
- Hidden mixture covariances $\Sigma_k \in \mathbb{R}^{d \times d}, \quad k = 1, 2, \ldots, K$
- Hidden mixture probabilities $\pi_k, \quad \sum_{k=1}^{K} \pi_k = 1$

Gaussian mixture marginal and conditional likelihood **:**

$$p(x_i \mid \pi, \mu, \Sigma) = \sum_{z_i=1}^{K} \pi_{z_i} \mathcal{N}(x_i \mid \mu_{z_i}, \Sigma_{z_i})$$

$$p(x_i \mid z_i, \pi, \mu, \Sigma) = \mathcal{N}(x_i \mid \mu_{z_i}, \Sigma_{z_i})$$

Case Study 2: Document Retrieval

Application to Document Modeling

Machine Learning for Big Data
CSE547/STAT548, University of Washington

Emily Fox

April 21st, 2015

 9

Task 2: Cluster Documents

- **Now:**
  - Cluster documents based on topic

"sports"

"world news"

 10

Document Representation

- Bag of words model

document *d*

previously ← vector fcn of word counts

$X = [\ ]$ (e.g. tf-idf)

performed operations on this vector

now:

$X = \{w_1, \ldots, w_N\}$ list

unordered ~~set~~ of N words in doc.

$w_i \in V$ (vocab)

A Generative Model

- Documents: $x^1, \ldots, x^D$ ← # of docs; with $x^d = \{w_1^d, \ldots, w_{N_d}^d\}$ ← word ID (not count)
- Associated topics: $z^1, \ldots, z^D$ with $z^d \in \{1, \ldots, K\}$ ← # of topics; # of words in doc d
- Parameters: $\theta = \{\pi, \beta\}$ ← mix weights, cluster params

$\pi = [\pi_1, \ldots, \pi_K]$ topic prob.

$Pr(z^i = k) = \pi_k$

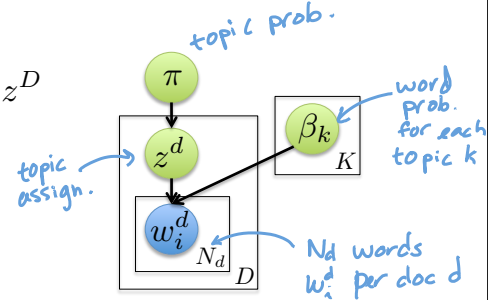
$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix}$ topics × words (1 … V)

$\sum_{v=1}^{V} \beta_{kv} = 1$

β_k: $\beta_{k1}, \beta_{k2}, \ldots, \beta_{kv}$ ← prob. of word 2 in topic k; pmf

A Generative Model

- Documents: $x^1, \ldots, x^D$
- Associated topics: $z^1, \ldots, z^D$
- Parameters: $\theta = \{\pi, \beta\}$
- Generative model:



Sample topic:
$z^d \sim \pi$

sample words:
$w_i^d \mid z^d \sim \beta_{z^d}$ $i = 1, \ldots, N_d$

Given topic $z^d = k$ for doc d, draw each word from β_k

Form of Likelihood

- Conditioned on topic...

N_d iid words from topic z^d

$$p(x^d \mid z^d, \beta) = \prod_{i=1}^{N_d} p(w_i^d \mid z^d, \beta) = \prod_{i=1}^{N_d} \beta_{z^d, w_i^d}$$

- Marginalizing latent topic assignment:

$$p(x^d \mid \beta, \pi) = \sum_{k=1}^{K} \pi_k \, p(x^d \mid z^d = k, \beta_k)$$

Case Study 2: Document Retrieval

Review: EM Algorithm

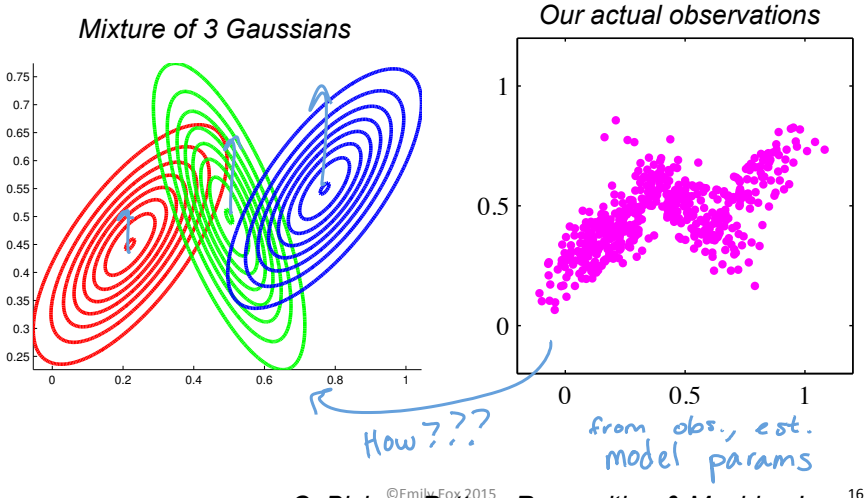
Machine Learning for Big Data
CSE547/STAT548, University of Washington

Emily Fox

April 21st, 2015

Learning Model Parameters

- Want to learn model parameters



C. Bishop, Pattern Recognition & Machine Learning

ML Estimate of Mixture Model Params

- Log likelihood

$$L_x(\theta) \triangleq \log p(\{x^i\} \mid \theta) = \sum_i \log \sum_{z^i} p(x^i, z^i \mid \theta)$$

- Want ML estimate

$$\hat{\theta}^{ML} = \arg\max_\theta L_x(\theta)$$

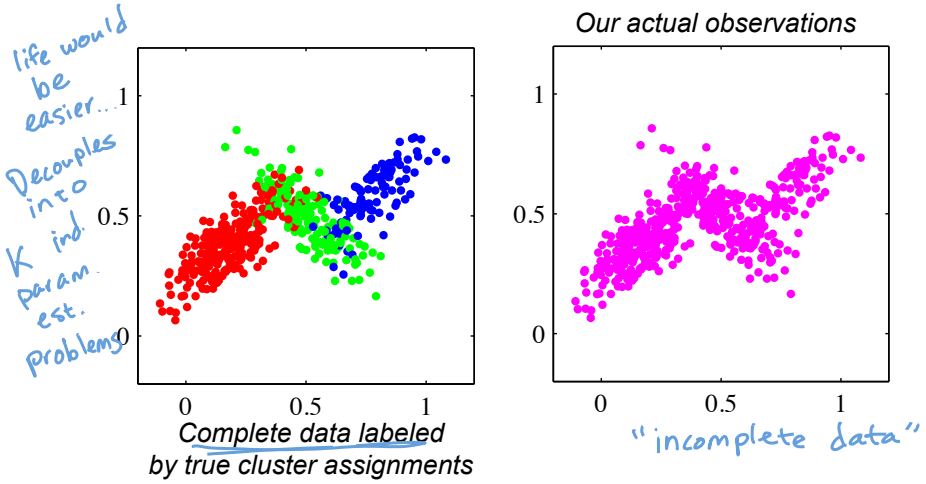
- Assume exponential family $p(x, z \mid \theta) = \frac{1}{Z(\theta)} e^{\theta' \phi(x,z)}$

$$L_x(\theta) = \sum_i \log \left(\sum_{z_i} e^{\theta^T \phi(z^i, x^i)} \right) - N \log Z(\theta)$$

- Neither convex nor concave and local optima

Complete Data

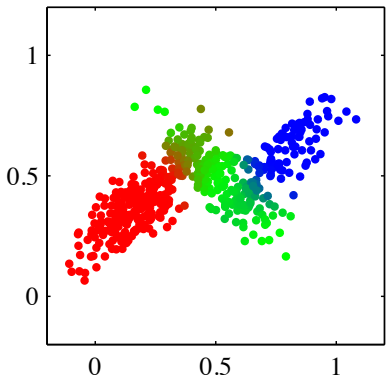
- Imagine we have an assignment of each x^i to a cluster



C. Bishop, Pattern Recognition & Machine Learning

Cluster Responsibilities

- We must infer the cluster assignments from the observations



Soft assignments to clusters

★ Motivates iterative alg ★

"responsibility" of cluster k for point i

- Posterior probabilities of assignments to each cluster *given* model parameters:

$$r_{ik} = p(z^i = k \mid x^i, \pi, \phi) =$$

(obs, params)

$$= \frac{\pi_k \, p(x^i \mid \phi_k)}{\sum_{j=1}^{K} \pi_j \, p(x^i \mid \phi_j)}$$

e.g. $N(x^i \mid \mu_j, \Sigma_j)$

C. Bishop, Pattern Recognition & Machine Learning

Iterative Algorithm

- Motivates a coordinate ascent-like algorithm:
  1. Infer missing values z^i given estimate of parameters $\hat{\theta}$
  2. Optimize parameters to produce new $\hat{\theta}$ given "filled in" data z^i
  3. Repeat

$z^i \xrightarrow{\text{est.}} \hat{\theta}$, $\hat{\theta} \xrightarrow{\max} z^i$

- Example: MoG (derivation soon… + HW)
  1. Infer "responsibilities"

$$r_{ik}^{(t)} = p(z^i = k \mid x^i, \hat{\theta}^{(t-1)}) = \frac{\pi_k^{(t-1)} \, p(x^i \mid \phi_k^{(t-1)})}{\sum_j \pi_j^{(t-1)} \, p(x^i \mid \phi_j^{(t-1)})}$$

  2. Optimize parameters

$\max$ w.r.t. π_k : $\pi_k^{(t)} = \frac{1}{N} \sum_i r_{ik}^{(t)} = \frac{r_k^{(t)}}{N}$ ← soft counts!

$\max$ w.r.t. ϕ_k :

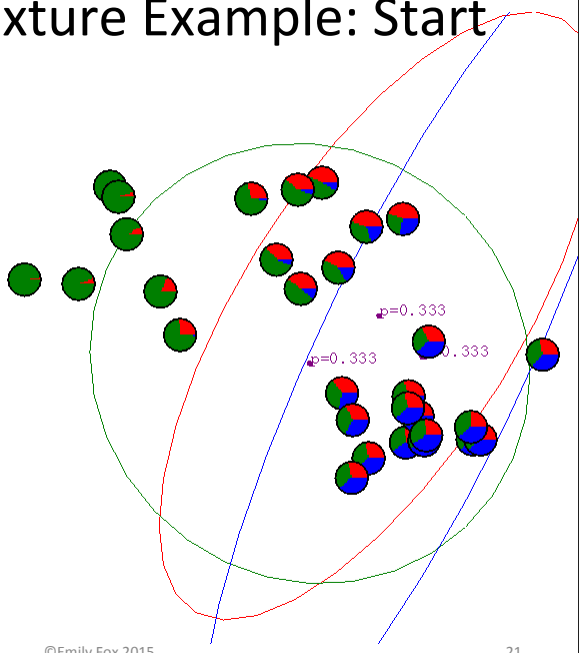
$\mu_k^{(t)} = \frac{\sum_i r_{ik}^{(t)} x^i}{r_k^{(t)}}$ ← weighted mean

$\Sigma_k^{(t)} = \frac{1}{r_k^{(t)}} \sum_i r_{ik}^{(t)} x^i x^{i\,T} - \mu_k^{(t)} \mu_k^{(t)\,T}$

Gaussian Mixture Example: Start

Initialize
$\pi^{(0)}, \phi^{(0)}$

→ compute
$r_{ik}^{(1)}$

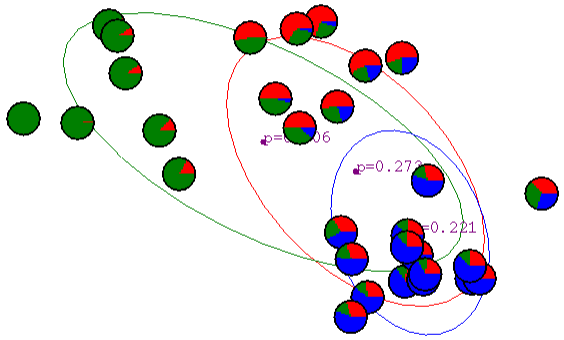


After first iteration

max like. given
soft counts

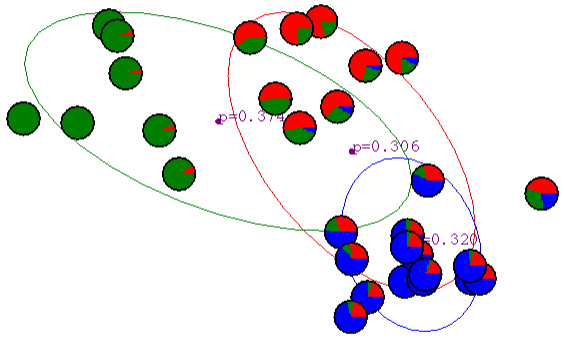
→ $\pi^{(1)}, \phi^{(1)}$

→ new $r_{ik}^{(2)}$

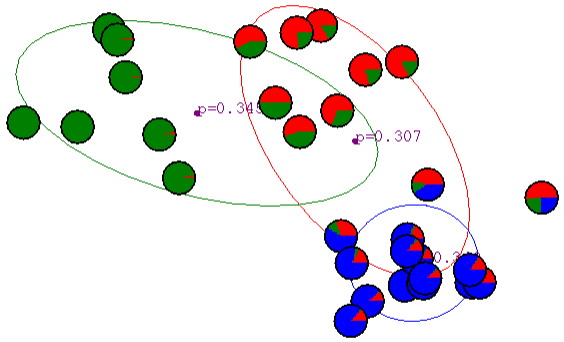


After 2nd iteration

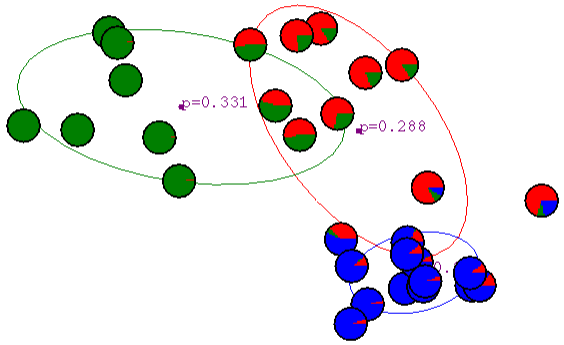
rinse + repeat



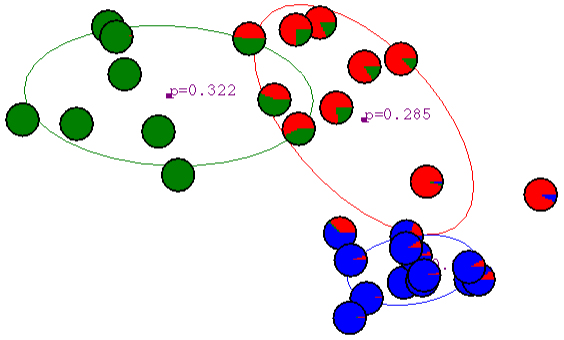
After 3rd iteration



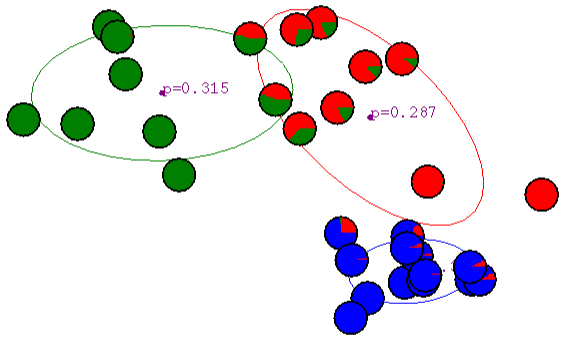
After 4th iteration



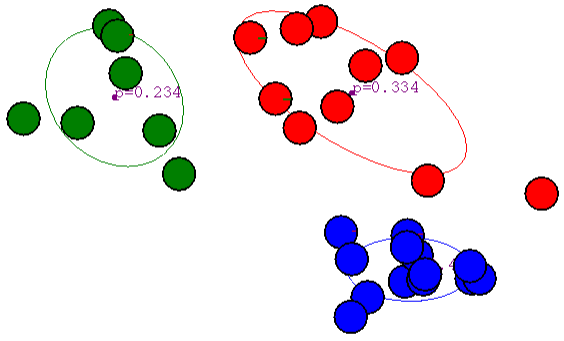
After 5th iteration



After 6th iteration



After 20th iteration



Expectation Maximization (EM) – Setup

- More broadly applicable than just to mixture models considered so far
- Model: x observable – *"incomplete" data* ← what we actually have
  y not (fully) observable – *"complete" data* ← what we wish we had
  θ parameters
- Interested in maximizing (wrt θ):

$$p(x \mid \theta) = \sum_y p(x, y \mid \theta)$$

introduce + marg. complete data

- Special case:

$$x = g(y)$$

deterministic fcn

e.g. $y = \begin{bmatrix} z \\ x \end{bmatrix}$ ← class labels ← obs.

in standard mix model

EM Algorithm

- Initial guess: $\hat{\theta}^{(0)}$
- Estimate at iteration *t*: $\hat{\theta}^{(t)}$
- **E-Step**

  Compute $U(\theta, \hat{\theta}^{(t)}) = E[\log p(y|\theta) \mid x, \hat{\theta}^{(t)}]$

  (complete data; obs. data)

- **M-Step**

  Compute $\hat{\theta}^{(t+1)} = \arg\max_\theta U(\theta, \hat{\theta}^{(t)})$

$\Rightarrow L_x(\hat{\theta}^{(t+1)}) \geq L_x(\hat{\theta}^{(t)})$

mild assump. $\Rightarrow$ $\hat{\theta}$ converges to a local mode

Example – Mixture Models

- **E-Step** Compute $U(\theta, \hat{\theta}^{(t)}) = E[\log p(y \mid \theta) \mid x, \hat{\theta}^{(t)}]$
- **M-Step** Compute $\hat{\theta}^{(t+1)} = \arg\max_{\theta} U(\theta, \hat{\theta}^{(t)})$

- Consider $y^i = \{z^i, x^i\}$ i.i.d.

$$p(x^i, z^i \mid \theta) = \pi_{z^i} p(x^i \mid \phi_{z^i}) = \prod_{k=1}^{K} (\pi_k p(x^i \mid \phi_k))^{I(z^i = k)}$$

$$E_{q_t}[\log p(y \mid \theta)] \overset{\text{i.i.d}}{=} \sum_i E_{q_t}[\log p(x^i, z^i \mid \theta)] =$$

$$= \sum_i \sum_k r_{ik} \log \pi_k + \sum_i \sum_k r_{ik} \log p(x^i \mid \phi_k)$$

$E[I(z^i = k)]$

M-step: max maximize w.r.t. π_k, ϕ_k

E-step: compute the r_{ik} based on $\hat{\theta}^{(t)}$

fcn of params

$E_{q_t}[I(z^i = k)] = p(z^i = k \mid x, \hat{\theta}^{(t)}) = r_{ik}$

Initialization

- In mixture model case where $y^i = \{z^i, x^i\}$, there are many ways to initialize the EM algorithm

- Examples:
  - Choose K observations at random to define each cluster. Assign other observations to the nearest "centriod" to form initial parameter estimates
  - Pick the centers sequentially to provide good coverage of data
  - Grow mixture model by splitting (and sometimes removing) clusters until K clusters are formed

- Can be quite important to convergence rates in practice + quality of local mode

What you need to know

- Mixture model formulation
  - Generative model
  - Likelihood
- Expectation Maximization (EM) Algorithm
  - Derivation
  - Concept of non-decreasing log likelihood
  - Application to standard mixture models

Case Study 2: Document Retrieval

Review: Connection to k-means

Machine Learning for Big Data
CSE547/STAT548, University of Washington

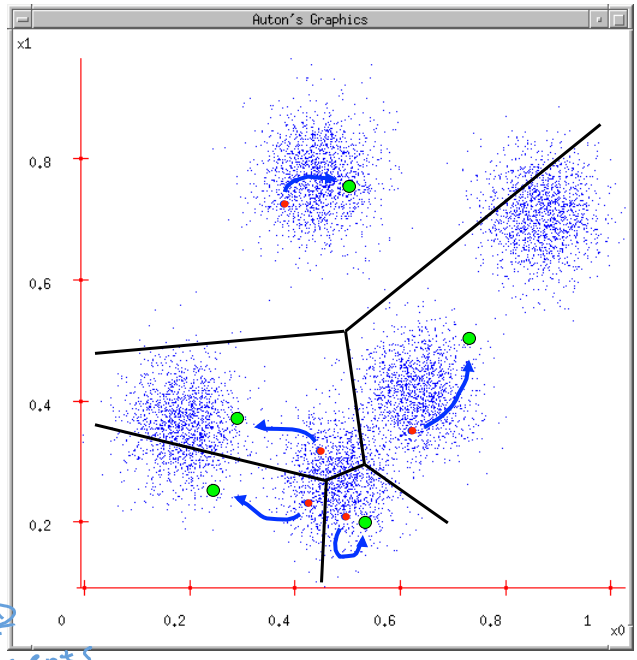
Emily Fox

April 21st, 2015

K-means

1. Ask user how many clusters they'd like. *(e.g. k=5)*
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to.
4. Each Center finds the centroid of the points it owns

iterative alg. making HARD assignments



K-means

- Randomly initialize *k* centers
  $\mu^{(0)} = \mu_1^{(0)},\ldots, \mu_k^{(0)}$
- **Classify**: Assign each point $j \in \{1,\ldots N\}$ to nearest center:

$$z^j \leftarrow \arg\min_i ||\mu_i - \mathbf{x}^j||_2^2$$

hard assign.

- **Recenter**: μ_i becomes centroid of its point:

$$\mu_i^{(t+1)} \leftarrow \arg\min_\mu \sum_{j:z^j=i} ||\mu - \mathbf{x}^j||_2^2$$

update params.

  – Equivalent to μ_i ← average of its points!

Special Case: Spherical Gaussians + hard assignments

$$P(z^i = k, \mathbf{x}^i) = \frac{1}{(2\pi)^{d/2} \| \Sigma_k \|^{1/2}} \exp\left[-\frac{1}{2} \left(\mathbf{x}^i - \mu_k \right)^T \Sigma_k^{-1} \left(\mathbf{x}^i - \mu_k \right) \right] P(z^i = k)$$

$P(x^i \mid z^i, \theta)$

- If P(X|z=k) is spherical, with same σ for all classes:

$$P(\mathbf{x}^i \mid z^i = k) \propto \exp\left[-\frac{1}{2\sigma^2} \left\| \mathbf{x}^i - \mu_k \right\|^2 \right]$$

assume: $\Sigma_k = \begin{pmatrix} \sigma^2 & & \\ & \ddots & \\ & & \sigma^2 \end{pmatrix} = \sigma^2 I$

- Then, compare EM objective with k-means:

EM: $\max_\theta \prod_i \sum_{z^i} p(x^i, z^i \mid \theta)$

maximizing marg. likelihood

k-means: $\max_{\{z^i\}, \theta} \prod_i p(x^i \mid z^i, \theta)$

OR if $\pi_k = \frac{1}{K}$ $\forall k$

$\max_{\{z^i\}, \theta} \prod_i p(x^i, z^i \mid \theta)$

$\max_\theta \prod_i \max_{z^i} p(x^i, z^i \mid \theta)$