Case Study 2: Document Retrieval

Clustering Documents, Mixture Models, Expectation Maximization

Machine Learning/Statistics for Big Data
CSE599C1/STAT592, University of Washington

Emily Fox

January 29th, 2013

Document Retrieval

- **Goal:** Retrieve documents of interest
- **Challenges:**
  - Tons of articles out there
  - How should we measure similarity?

Task 1: Find Similar Documents

- **So far…**
  - **Input:** Query article
  - **Output:** Set of k similar articles

Task 2: Cluster Documents

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

Document Representation

- Bag of words model

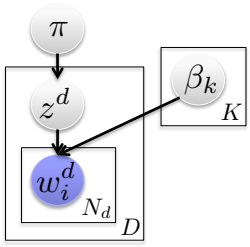
document *d*

A Generative Model

- Documents:
- Associated topics:
- Parameters: $\theta = \{\pi, \beta\}$

A Generative Model

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



Form of Likelihood

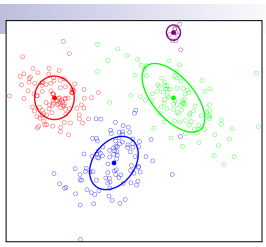
- Conditioned on topic...

$$p(x^d \mid z^d, \beta) =$$

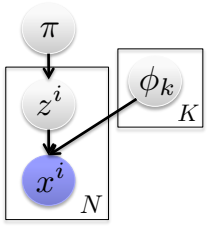
- Marginalizing latent topic assignment:

$$p(x^d \mid \beta, \pi) =$$

Gaussian Mixture Model



- Most commonly used mixture model
- Observations:
- Parameters:

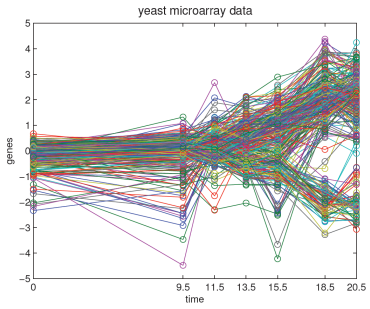


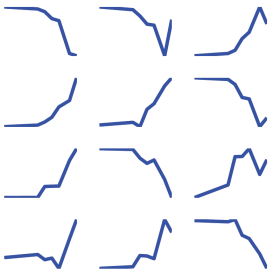
- Likelihood:
- Ex. z^i = country of origin, x^i = height of i[th] person
  - *k*[th] mixture component = distribution of heights in country *k*

Another Example

(Taken from Kevin Murphy's ML textbook)

- Data: gene expression levels
- Goal: cluster genes with similar expression trajectories





Mixture models are useful for…

- Density estimation
  - Allows for multimodal density
- Clustering
  - Want membership information for each observation
    - e.g., topic of current document
  - Soft clustering:

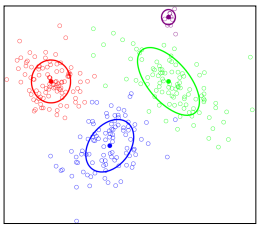
$$p(z^i = k \mid x^i, \theta) =$$

  - Hard clustering:

$$z^{i*} = \arg\max_k p(z^i = k \mid x^i, \theta) =$$

Issues

- Label switching
  - Color = label does not matter
  - Can switch labels and likelihood is unchanged



- Log likelihood is not convex in the parameters
  - Problem is simpler for "complete data likelihood"

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

$$\hat{\theta}^{ML} =$$

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

$$L_x(\theta) =$$

- Neither convex nor concave and local optima

If "complete" data were observed…

- Assume class labels z^i were observed in addition to x^i

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

- Compute ML estimates
  - Separates over clusters *k*!

- Example: mixture of Gaussians (MoG) $\theta = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^K$

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

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

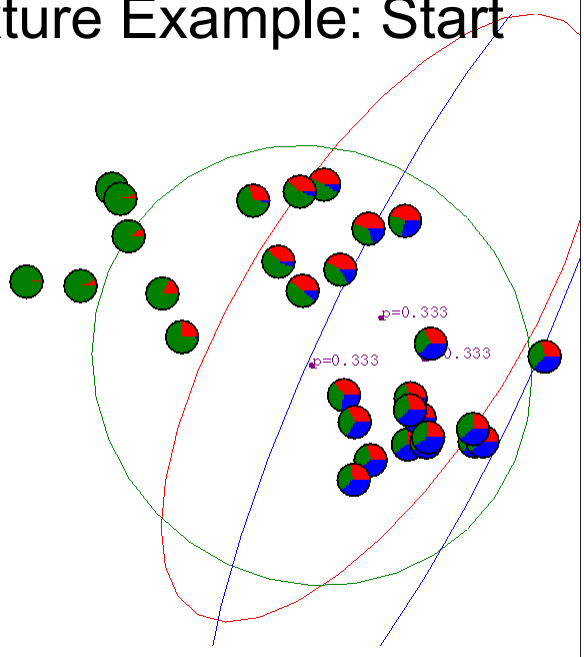
  $r_{ik} = p(z^i = k \mid x^i, \hat{\theta}^{(t-1)}) =$

  2. Optimize parameters

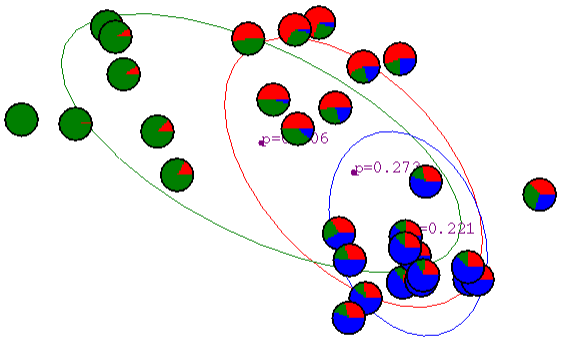
  $\max \text{ w.r.t. } \pi_k :$

  $\max \text{ w.r.t. } \phi_k :$

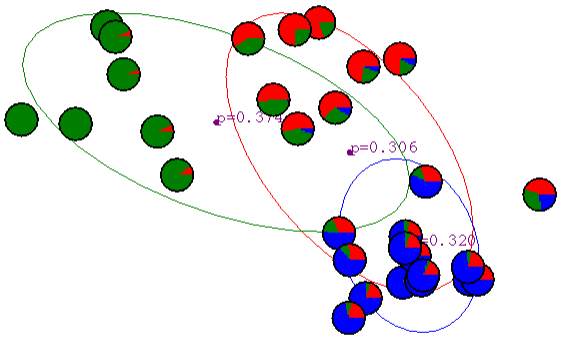
Gaussian Mixture Example: Start



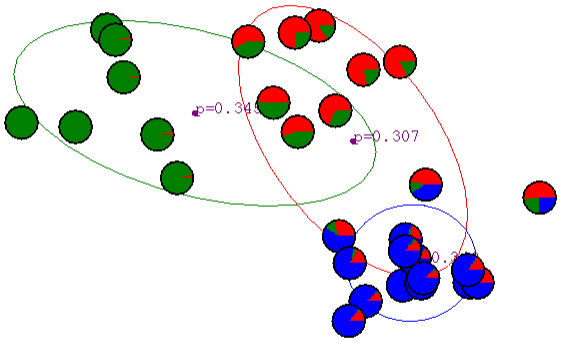
After first iteration



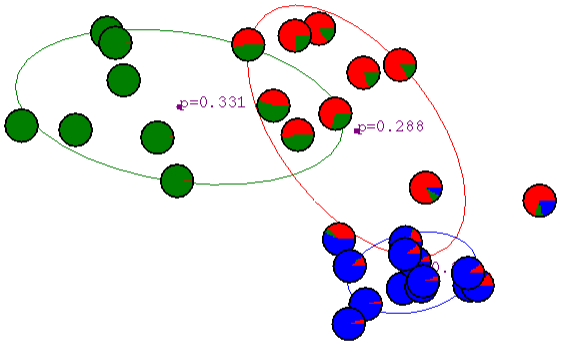
After 2nd iteration



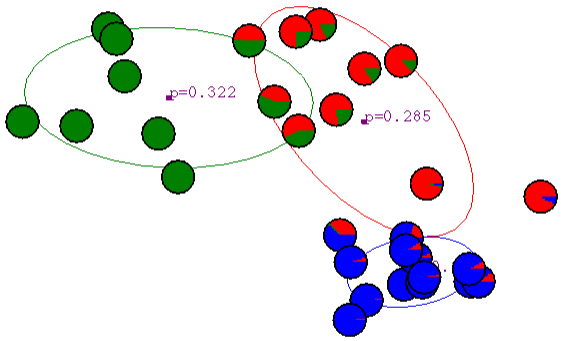
After 3rd iteration



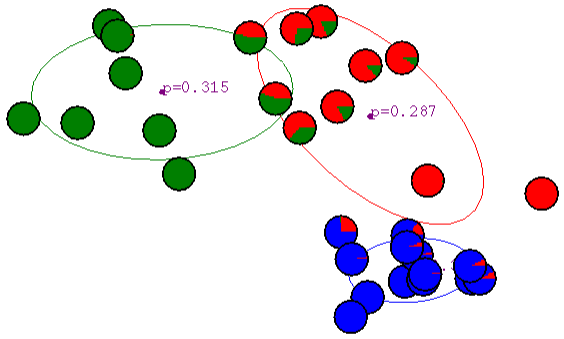
After 4th iteration



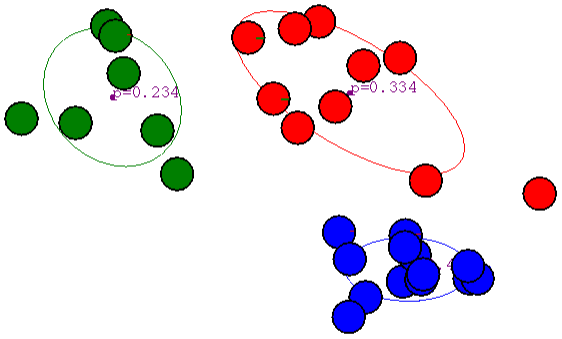
After 5th iteration



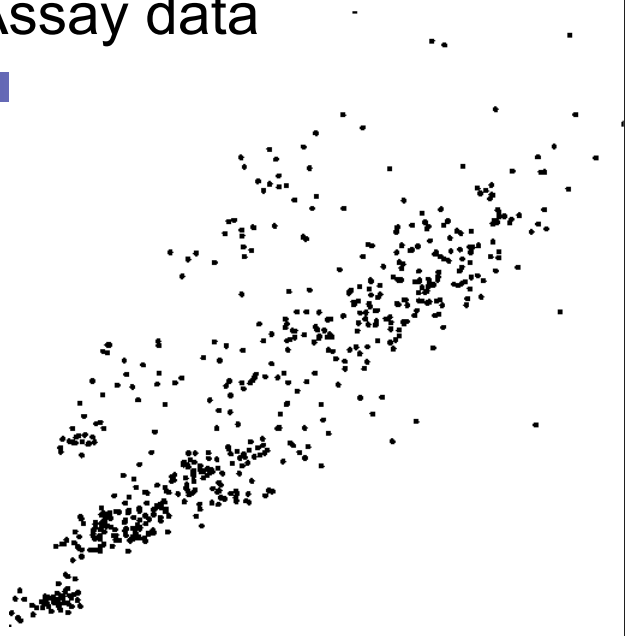
After 6th iteration



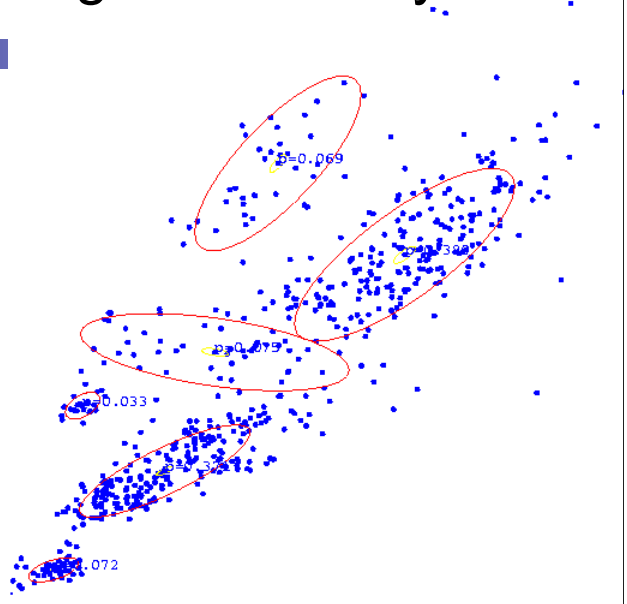
After 20th iteration



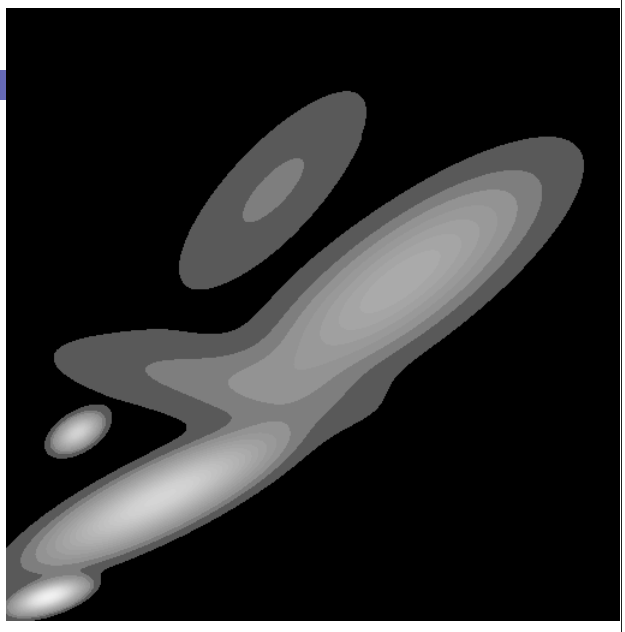
Some Bio Assay data



GMM clustering of the assay data



Resulting Density Estimator



Expectation Maximization (EM) – Setup

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

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

- Special case:

$$x = g(y)$$

Expectation Maximization (EM) – Derivation

- Step 1
  - Rewrite desired likelihood in terms of complete data terms

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

- Step 2
  - Assume estimate of parameters $\hat{\theta}$
  - Take expectation with respect to $p(y \mid x, \hat{\theta})$

Expectation Maximization (EM) – Derivation

- Step 3
  - Consider log likelihood of data at any θ relative to log likelihood at $\hat{\theta}$

$$L_x(\theta) - L_x(\hat{\theta})$$

- **Aside: Gibbs Inequality** $E_p[\log p(x)] \geq E_p[\log q(x)]$

Proof:

Expectation Maximization (EM) – Derivation

$$L_x(\theta) - L_x(\hat{\theta}) = [U(\theta, \hat{\theta}) - U(\hat{\theta}, \hat{\theta})] - [V(\theta, \hat{\theta}) - V(\hat{\theta}, \hat{\theta})]$$

- Step 4
  - Determine conditions under which log likelihood at θ exceeds that at $\hat{\theta}$

Using Gibbs inequality:

If

Then

$$L_x(\theta) \geq L_x(\hat{\theta})$$

Motivates EM Algorithm

- Initial guess:
- Estimate at iteration *t*:

- **E-Step**

  Compute

- **M-Step**

  Compute

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.

$p(x^i, z^i \mid \theta) = \pi_{z^i} p(x^i \mid \phi_{z^i}) =$

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

Coordinate Ascent Behavior

- Bound log likelihood:

$$L_x(\theta) = U(\theta, \hat{\theta}^{(t)}) + V(\theta, \hat{\theta}^{(t)})$$
$$\geq$$
$$L_x(\hat{\theta}^{(t)}) = U(\hat{\theta}^{(t)}, \hat{\theta}^{(t)}) + V(\hat{\theta}^{(t)}, \hat{\theta}^{(t)})$$

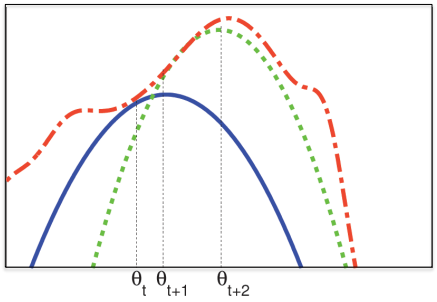


Figure from KM textbook

Comments on EM

- Since Gibbs inequality is satisfied with equality only if *p=q*, any step that changes θ should strictly **increase likelihood**
- In practice, can replace the **M-Step** with increasing *U* instead of maximizing it (**Generalized EM**)
- Under certain conditions (e.g., in exponential family), can show that EM **converges to a stationary point** of $L_x(\theta)$
- Often there is a **natural choice for *y*** … has physical meaning
- If you want to choose any *y*, not necessarily *x=g(y)*, replace $p(y \mid \theta)$ in *U* with $p(y, x \mid \theta)$

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

MAP Estimation

- Bayesian approach:
  - Place **prior** $p(\theta)$ on parameters
  - Infer **posterior** $p(\theta \mid x)$
- Many, many, many motivations and implications
  - For the sake of this class, simplest motivation is to think of this as akin to regularization

$$\hat{\theta}^{MAP} = \arg\max_{\theta} \log p(\theta \mid x)$$

  - Saw importance of regularization in logistic regression (ML estimate can overfit data and lead to poor generalization)

EM Algorithm – MAP Case

- Re-derive EM algorithm for $p(\theta \mid x)$
- Add $\log p(\theta)$ to $U(\theta, \hat{\theta}^{(t)})$
  - What must be computed in E-Step remains unchanged because this term does not depend on *y*.
  - M-Step becomes:

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

MAP EM Example – MoG

- For mixture of Gaussians, conjugate priors are:

$$\pi \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_K) \quad \{\mu_k, \Sigma_k\} \sim \mathrm{NIW}(m_0, \kappa_0, \nu_0, S_0)$$

- Results in following M-Step:

$$\hat{\mu}_k = \frac{r_k \bar{x}_k + \kappa_0 m_0}{r_k + \kappa_0} \qquad \hat{\pi}_k = \frac{r_k + \alpha_k - 1}{N + \sum_k \alpha_k - K}$$

$$\hat{\Sigma}_k = \frac{S_0 + r_k S_k + \frac{\kappa_0 r_k}{\kappa_0 + r_k}(\bar{x}_k - m_0)(\bar{x}_k - m_0)'}{\nu_0 + r_k + d + 2}$$

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

Course Announcements

- Homework 2 will be posted on Thursday
  - Due 2 weeks later (Feb 14)
- Project proposals:
  - Initial ideas now posted
  - Deadline extended to Tues, Feb 5
  - 1 page, 1-2 people
- Recitation on Thursday (Linda)
- Office hours as normal