Case Study 2: Document Retrieval

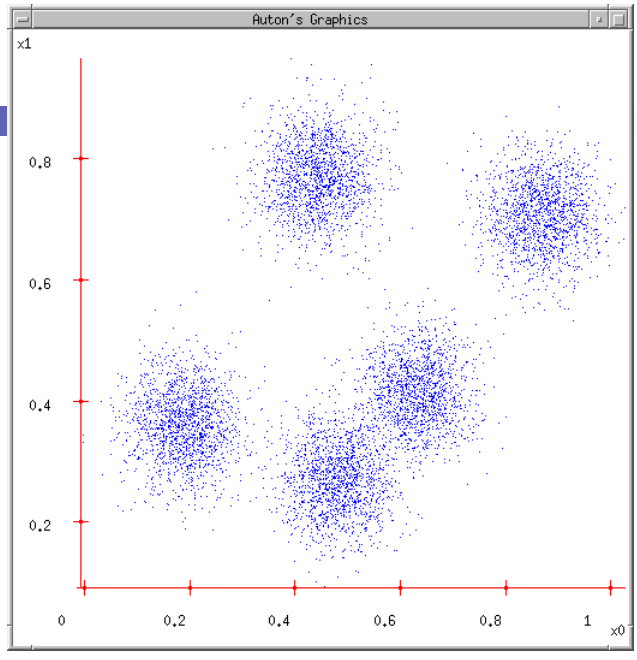
Review: Mixtures of Gaussians

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

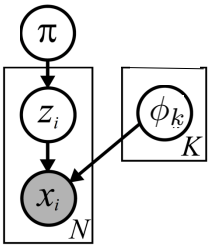
Emily Fox

January 28th, 2014

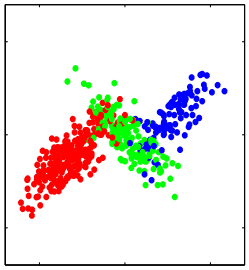
Some Data



Gaussian Mixture Model

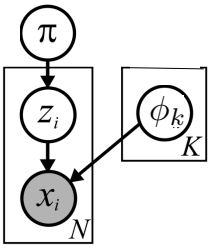


- Most commonly used mixture model
- Observations:
- Parameters:

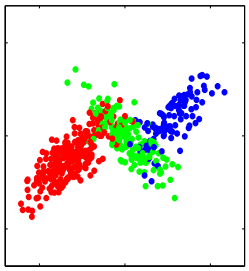


- Cluster indicator:
- Per-cluster likelihood:
- Ex. z^i = country of origin, x^i = height of i[th] person
  - *k*[th] mixture component = distribution of heights in country *k*

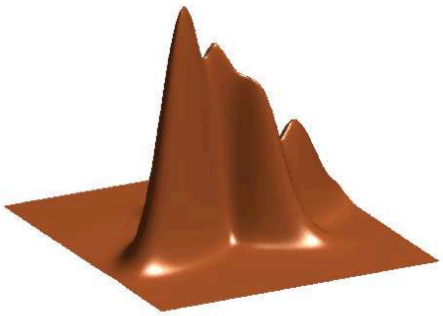
Generative Model



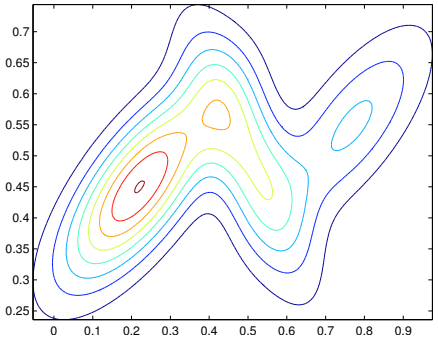
- We can think of *sampling* observations from the model
- For each observation *i*,
  - Sample a cluster assignment
  - Sample the observation from the selected Gaussian



Also Useful for Density Estimation



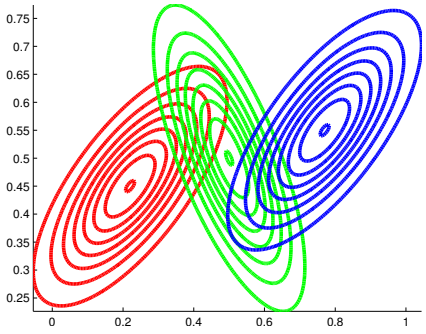
Contour Plot of Joint Density



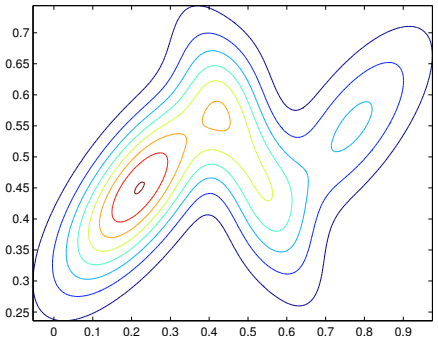
Density as Mixture of Gaussians

- Approximate density with a mixture of Gaussians

Mixture of 3 Gaussians



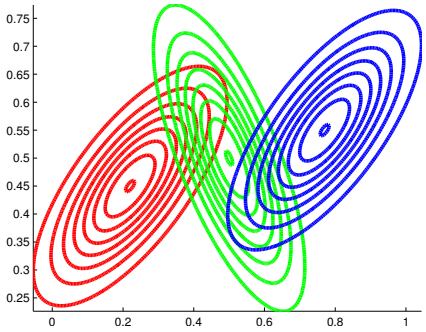
Contour Plot of Joint Density



Density as Mixture of Gaussians

- Approximate density with a mixture of Gaussians

Mixture of 3 Gaussians



$$p(x^i \mid \pi, \mu, \Sigma) =$$

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

January 28th, 2014

Task 2: Cluster Documents

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

Document Representation

- Bag of words model

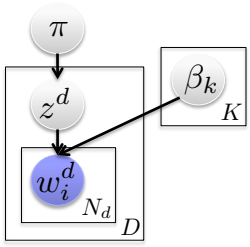
document *d*

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:

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

Case Study 2: Document Retrieval

Review: EM Algorithm

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

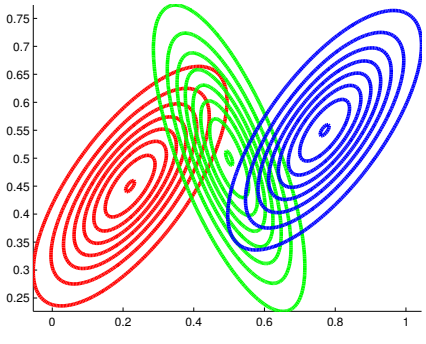
Emily Fox

January 28th, 2014

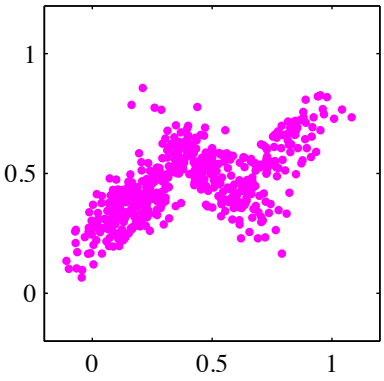
Learning Model Parameters

- Want to learn model parameters

Mixture of 3 Gaussians



Our actual observations



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

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

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

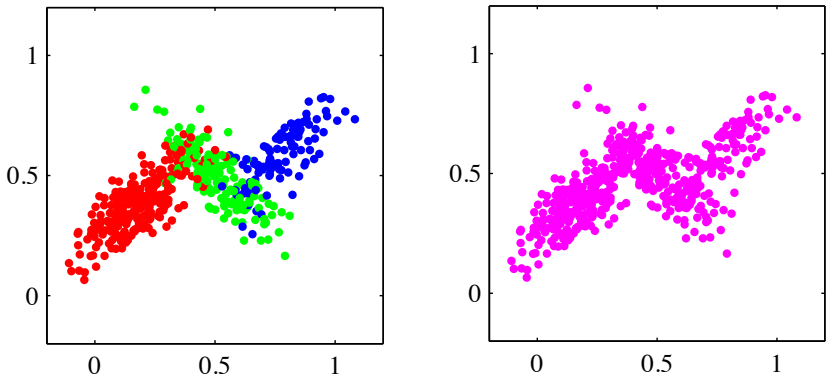
$$L_x(\theta) =$$

- Neither convex nor concave and local optima

Complete Data

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

Our actual observations



Complete data labeled by true cluster assignments

C. Bishop, Pattern Recognition & Machine Learning

If "complete" data were observed…

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

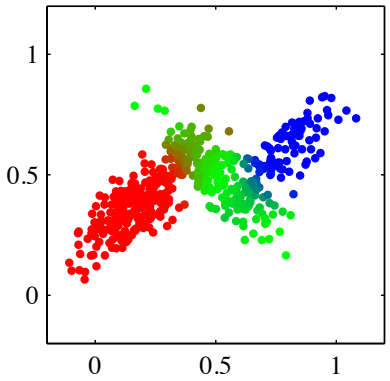
$$L_{x,z}(\theta) = \sum_i \log p(x^i, z^i \mid \theta)$$

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

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

Cluster Responsibilities

- We must infer the cluster assignments from the observations



Soft assignments to clusters

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

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

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

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

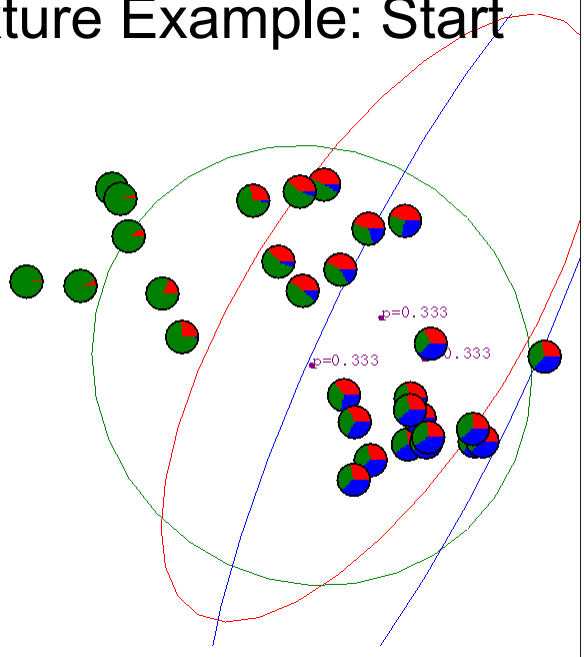
  $r_{ik} = p(z^i = k \mid x^i, \hat{\theta}^{(t-1)}) =$

  2. Optimize parameters

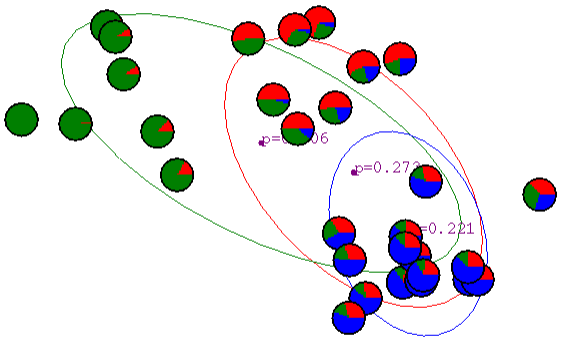
  $\max \text{ w.r.t. } \pi_k :$

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

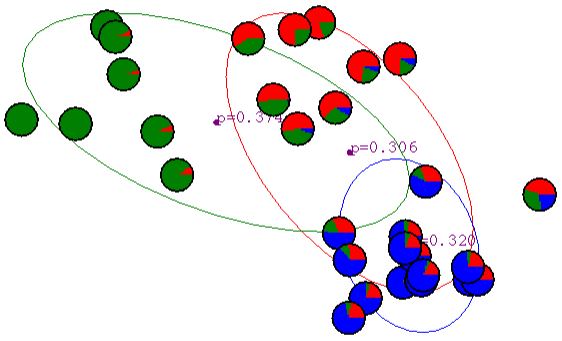
Gaussian Mixture Example: Start



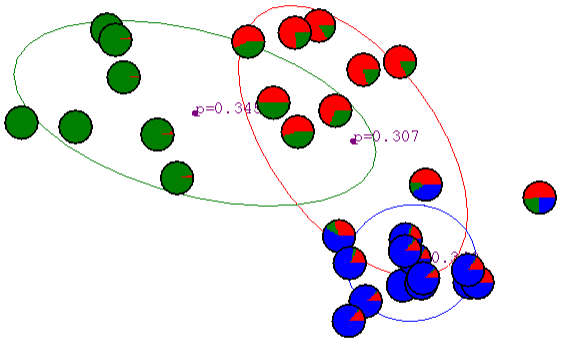
After first iteration



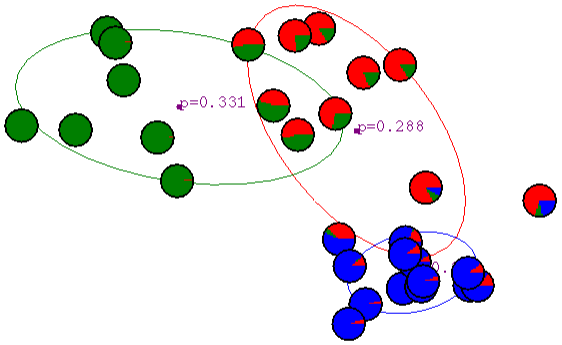
After 2nd iteration



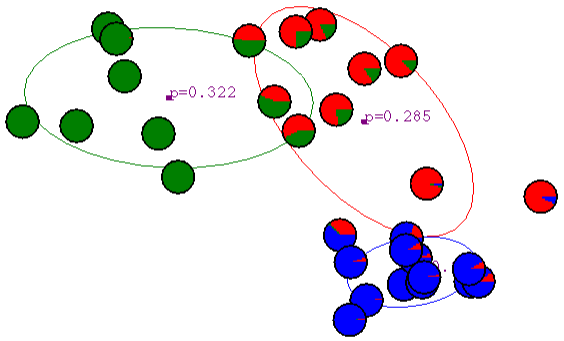
After 3rd iteration



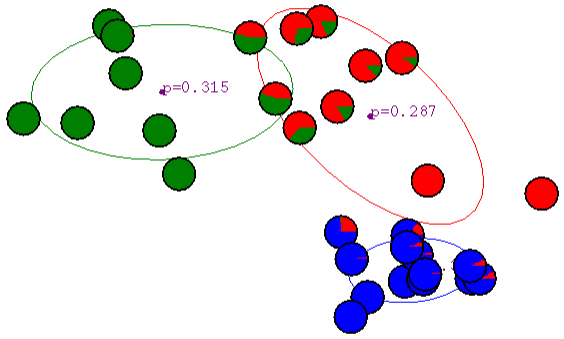
After 4th iteration



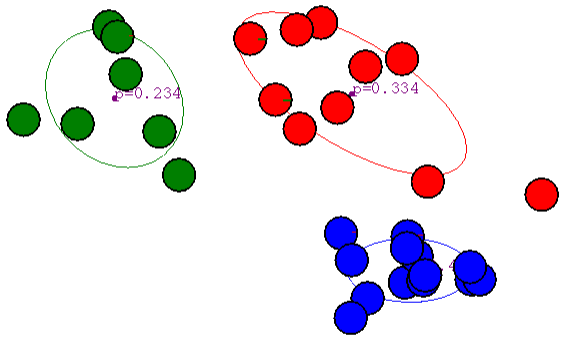
After 5th iteration



After 6th iteration



After 20th iteration



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:

$$x = g(y)$$

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.

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

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

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

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

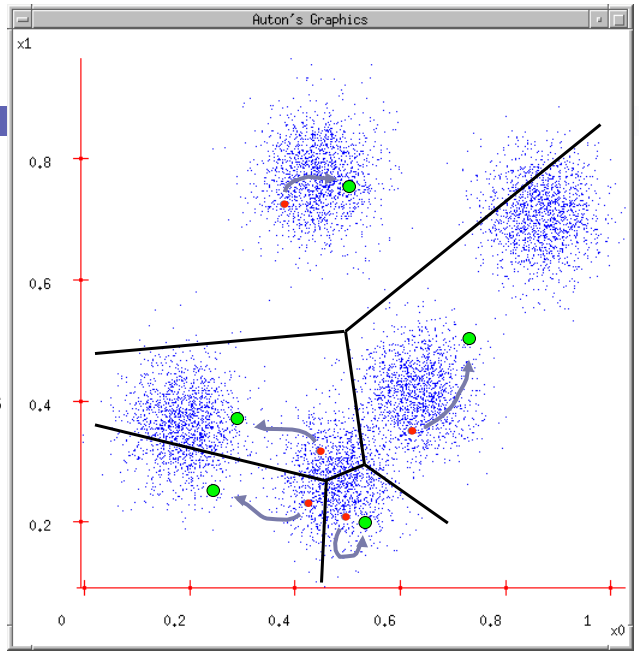
Machine Learning for Big Data
CSE547/STAT548, University of Washington

Emily Fox

January 28th, 2014

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



K-means

- Randomly initialize *k* centers
  - $\mu^{(0)} = \mu_1^{(0)}, \ldots, \mu_k^{(0)}$
- **Classify**: Assign each point j∈{1,…m} to nearest center:
  - $z^j \leftarrow \arg\min_i ||\mu_i - \mathbf{x}^j||_2^2$
- **Recenter**: μ_i becomes centroid of its point:
  - $\mu_i^{(t+1)} \leftarrow \arg\min_\mu \sum_{j:z^j=i} ||\mu - \mathbf{x}^j||_2^2$
  - Equivalent to μ_i ← average of its points!

Special case: spherical Gaussians and hard assignments

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

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

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

- Then, compare EM objective with k-means: