Mixture models and the EM algorithm

Iain Murray
http://iainmurray.net/
“Human brains are good at finding regularities in data. One way of expressing regularity is to put a set of objects into groups that are similar to each other. For example, biologists have found that most objects in the natural world fall into one of two categories: things that are brown and run away, and things that are green and don’t run away. The first group they call animals, and the second, plants.”

David MacKay, ITILA textbook p284
Mixture models motivation 1

Often see clusters in data.
Mixture models motivation 2

This lecture: density estimation when not Gaussian

[Figure from Neal 2001, Dirichlet Diffusion Trees]
But why density estimation?

Learning the joint distribution of variables $p(x)$: brute force approach to any question! Can get probability of any variable given any others.

**One practical example:** fit prior statistics of pixels, can denoise and deblur images using Bayes rule.

Zoran and Weiss, ICCV 2011

http://www.cs.huji.ac.il/~daniez/EPLLICCVCameraReady.pdf
Mixture models motivation 3

Mixture idea can be applied to supervised learning:

— Output/label comes from model or is occasionally junk

— E.g., $p(y \mid x, \theta) = \lambda \mathcal{N}(y; w^\top x, 0.1^2) + (1-\lambda) \mathcal{N}(y; 0, 100^2)$, \[ \lambda = 0.99 \]

— Makes model more robust to outliers

We’ll stick to the unsupervised case in this lecture.
Gaussian mixture model

\[ p(x_i \mid \theta) = \sum_{k=1}^{K} p(x_i, z_i = k \mid \theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i; \mu_k, \Sigma_k) \]

Point \( i \), picks component \( z_i \):

\[ P(z_i = k \mid \theta) = \pi_k \]

Gaussian given component:

\[ x_i \sim \mathcal{N}(\mu_k; \Sigma_k) \]

Parameters:

\[ \theta = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^{K} \]
Maximum likelihood fitting

If we knew cluster indicators $z_i$, would be easy:

Set ‘responsibilities’ $r_{ik} = I(z_i = k)$

$$\pi_k = \frac{1}{N} \sum_{i=1}^{N} r_{ik} = \frac{r_k}{N}$$

$$\mu_k = \frac{1}{r_k} \sum_{i=1}^{N} r_{ik} x_i$$

$$\Sigma_k = \frac{1}{r_k} \sum_{i=1}^{N} r_{ik} x_i x_i^\top - \mu_k \mu_k^\top$$
EM algorithm for Gaussian mixtures

1. **E-step:** Set ‘soft responsibilities’:
   
   \[ r_{ik} = P(z_i = k | \mathbf{x}_i, \theta) = \frac{\pi_k \mathcal{N}(\mathbf{x}_i; \mu_k, \Sigma_k)}{\sum_{k'} \pi_{k'} \mathcal{N}(\mathbf{x}_i; \mu_{k'}, \Sigma_{k'})} \]  
   \[ (\text{Bayes' rule}) \]

2. **M-step:**
   
   Update \( \theta = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^K \) as on previous slide

3. Go to 1. (if not converged and max iters not exceeded)

Constraints satisfied: \( 0 < \pi_k < 1, \sum_k \pi_k = 1, \mathbf{z}^\top \Sigma_k \mathbf{z} \geq 0 \quad \forall \mathbf{z} \)

Can prove finds local maximum of likelihood
EM algorithm for Gaussian mixtures

Bishop Figure 9.8, or see Murphy p353
Bound-based optimization idea

1. Form lower bound, tight at current $\theta$

2. Maximize bound \textit{wrt} $\theta$. Go to 1.

Bishop Figure 9.14, or Murphy p365
EM algorithm details

Want to maximize log probability

$$\log P(X | \theta) = \sum_i \log \sum_{z_i} P(x_i, z_i | \theta)$$

Sum inside log means no closed form solution. Can use gradient methods, although need to do work to satisfy constraints.

$$\log P(X | \theta) = \sum_i \log \sum_{z_i} q(z_i) \frac{P(x_i, z_i | \theta)}{q(z_i)}$$ \hspace{1cm} \text{by Jensen’s inequality:}

$$\geq \mathcal{L}(q, \theta) = \sum_i \sum_{z_i} q(z_i) \log P(x_i, z_i | \theta) - \sum_i \sum_{z_i} q(z_i) \log q(z_i)$$

E-step: Lower bound tight at $\theta = \theta^{old}$ if $q(z_i) = P(z_i | x_i, \theta^{old})$

M-step: First term is maximized by simple weighted averages under $q$. 
Summary

Mixture models
  — Find clusters
  — Fit complicated densities

EM algorithm
  — Find responsibility of each cluster. . .
    . . . the posterior probabilities $p(z_k \mid x, \theta)$
  — Parameters set to statistics of points in cluster
    simple averages weighted by responsibilities
  — EM is a bound-based optimizer