The EM Algorithm for Maximum Likelihood Estimation of Gaussian Mixture Models.

Gaussian Mixture Models

LDA and QDA assume a Gaussian model for $X|Y=0$ and $X|Y=1$, estimate the model parameters via MLE, and plug in to the Bayes classifier.

Advantages

- Simple
- Mahalanobis distance interpretation

Disadvantages

- Data is often non Gaussian.

A Gaussian mixture model (GMM) extends the basic Gaussian model and affords a more flexible representation:

$$g(x|\theta) = \sum_{k=1}^{K} w_k \phi(x|\mu_k, \Sigma_k),$$

where $\theta = (w_1, \ldots, w_K, \mu_1, \Sigma_1, \ldots, \mu_K, \Sigma_K)$, $\Sigma w_k = 1$, and ϕ denotes the Gaussian pdf.

A classifier is obtained by estimating a GMM for each class, $\hat{\theta}_0$ and $\hat{\theta}_1$, and plugging in to the Bayes classifier:

$$\frac{g(x \mid \hat{\theta}_1)}{g(x \mid \hat{\theta}_0)} \underset{0}{\overset{1}{\gtrless}} \frac{\hat{\pi}_0}{\hat{\pi}_1}$$

In this lecture we will focus on an algorithm, called the EM algorithm, for computing the MLE of a GMM.

GMMs: A generative view

Suppose

$$X \sim g(x \mid \theta) = \sum_{k=1}^{K} w_k \phi(x \mid \mu_k, \Sigma_k)$$

and θ is given.

How could you generate a realization of X using basic random number generators (e.g., uniform and Gaussian)?

Let $S \in \{1, 2, \ldots, K\}$ be a discrete RV such that

$$\Pr\{S = k\} = w_k.$$

Generate X as follows:

1. Generate a realization s of S.
2. Generate $X \sim \phi(x \mid \mu_s, \Sigma_s)$.

Then the density of X generated in this way is

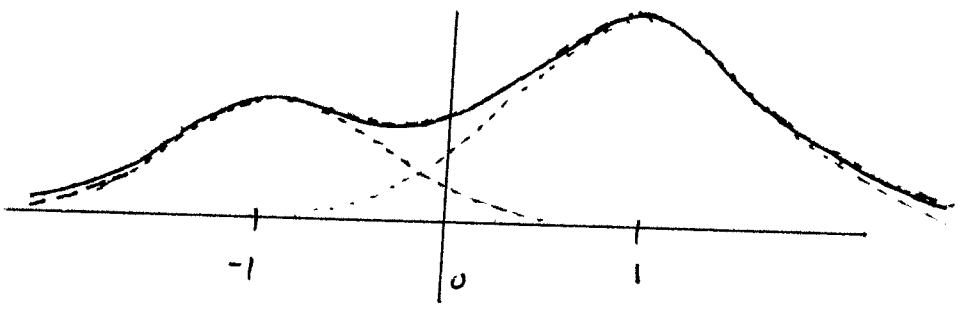
$$\begin{aligned} q(x) &= \sum_{k=1}^{K} q(x \mid S = k) \cdot \Pr\{S = k\} \\ &= \sum_{k=1}^{K} w_k \, \phi(x \mid \mu_k, \Sigma_k) \end{aligned}$$

as desired.

The variable S is called a (hidden) <u>state</u> variable. We will imagine that every realization from a GMM is associated with a specific realization of a state variable.

Example

$$X \sim \frac{1}{3}\phi(x \mid -1, 1) + \frac{2}{3}\phi(x \mid 1, 1)$$



↑ simulate this $\frac{1}{3}$ of the time

↑ simulate this $\frac{2}{3}$ of the time

Maximum Likelihood Estimation

Consider the likelihood function

$$\ell(\theta \mid \underline{x}) = \prod_{i=1}^{n} g(x_i \mid \theta)$$

based on a random sample $\underline{x} = (x_1, \ldots, x_n)$.

Recall each $x_i \in \mathbb{R}^d$ is a vector.

For the GMM we have

$$\ell(\theta \mid \underline{x}) = \prod_{i=1}^{n} \left(\sum_{k=1}^{K} w_k \phi(x_i \mid \mu_k, \Sigma_k) \right)$$

Unfortunately, solving

$$\frac{\partial \ell(\theta|\underline{x})}{\partial \theta} = 0 \quad \text{or} \quad \frac{\partial \log \ell(\theta|x)}{\partial \theta} = 0$$

is not straightforward.

Complete data

Let $\underline{s} = (s_1, \ldots, s_n)$ be the state variables corresponding to $\underline{x} = (x_1, \ldots, x_n)$. We call $\underline{s}$ the unobserved, or latent, or hidden data.

We call $\underline{z} = (\underline{x}, \underline{s})$ the complete data.

Why? Because with knowledge of $\underline{s}$, the MLE is easy to compute.

MLE of θ given complete data

Introduce the notation

$$I_k = \{ i : s_i = k \}$$

$$n_k = |I_k|.$$

Exercise

Find the likelihood of the complete data

$$\bar{l}(\theta \mid \underline{z}) =$$

Solution /

$$\bar{\ell}(\theta \mid \underline{z}) = \prod_{i=1}^{n} g(z_i \mid \theta)$$

$$= \prod_{i=1}^{n} g(x_i, s_i \mid \theta)$$

$$= \prod_{i=1}^{n} g(x_i \mid \theta, S_i = s_i) \cdot Pr(S_i = s_i)$$

$$= \prod_{i=1}^{n} w_{s_i} \cdot \prod_{i=1}^{n} \phi(x_i \mid \mu_{s_i}, \Sigma_{s_i})$$

$$= \prod_{k=1}^{K} w_k^{n_k} \cdot \prod_{k=1}^{K} \cdot \prod_{i \in I_k} \phi(x_i \mid \mu_k, \Sigma_k).$$

$\Rightarrow$ The likelihood factors into independent terms

$\Rightarrow$ Can maximize w.r.t. (μ_k, Σ_k) independently for each k.

$\Rightarrow$

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i \in I_k} x_i$$

$$\hat{\Sigma}_k = \frac{1}{n_k} \sum_{i \in I_k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$$

What about $\hat{w}_1, \ldots, \hat{w}_K$? Because of the constraint $\sum w_k = 1$ we cannot optimize each term separately. However, this is the MLE for a multinomial distribution:

$$\hat{w}_k = \frac{n_k}{n}.$$

Use Lagrange multipliers

In summary, the complete data makes the problem easy. Unfortunately, we don't observe $\underline{z}$.

The EM Algorithm

Introduce the indicator variable

$$\Delta_{i,k} = \begin{cases} 1 & \text{if } s_i = k \\ 0 & \text{if } s_i \neq k \end{cases}.$$

Observe that the complete log-likelihood may be written

$$\log \bar{\ell}(\theta | \underline{z}) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \Delta_{i,k} w_k \phi(x_i | \mu_k, \Sigma_k) \right)$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} \Delta_{i,k} \left[\log w_k + \log \phi(x_i | \mu_k, \Sigma_k) \right]$$

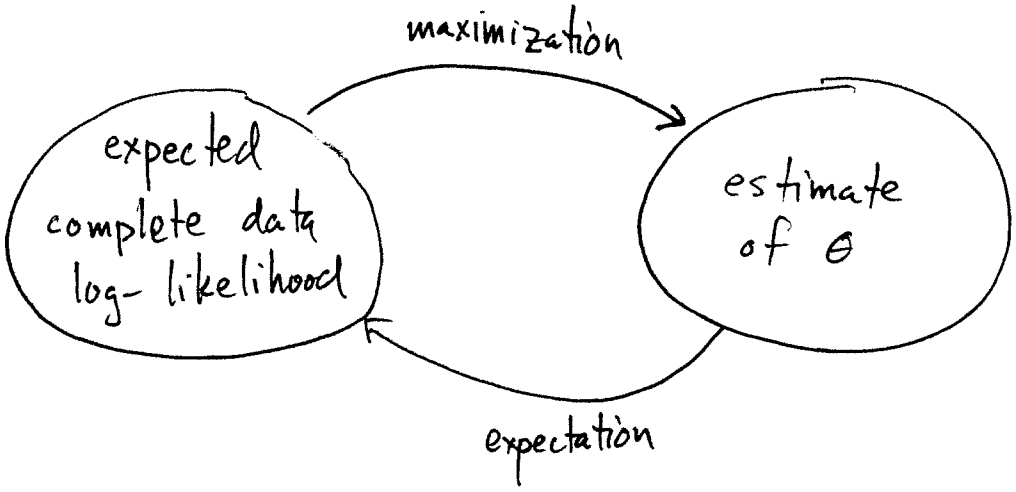
Big Idea #1

Replace the "hard" assignments $\Delta_{i,k}$ by the "soft" assignments $\gamma_{i,k} = E[\Delta_{i,k} \mid \underline{x}]$

$\uparrow$ w.r.t. $\leq$.

Big Idea #2

Since $\gamma_{i,k} = \gamma_{i,k}(\theta)$, proceed in an iterative fashion.



EM Algorithm

Initialize $\theta^{(0)}$

Repeat

E-Step : Compute

$$Q(\theta, \theta^{(j)}) = E\left[\log \bar{l}(\theta \mid \underline{z}) \mid \theta^{(j)}, \underline{x}\right]$$

M-Step :

$$\theta^{(j+1)} = \arg\max_{\theta} Q(\theta, \theta^{(j)})$$

Until termination criterion satisfied.

E-Step for GMM

$$Q(\theta, \theta^{(j)}) = \sum_{i=1}^{n} \sum_{k=1}^{k} \gamma_{i,k}(\theta^{(j)}) \cdot \Big[\log w_k - \frac{1}{2} \log |\Sigma| - \frac{d}{2} \log 2\pi - \frac{1}{2} (x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) \Big]$$

where

$$\gamma_{i,k}(\theta^{(j)}) = E\left[\Delta_{i,k} \mid x_i, \theta^{(j)} \right].$$

Exercise Express $\gamma_{i,k}(\theta^{(j)})$ in terms of $x_i, \theta^{(j)}$.

Solution

$$\gamma_{i,k}(\theta^{(j)}) = E[\Delta_{i,k} \mid x_i, \theta^{(j)}]$$

$$= Pr\{\Delta_{i,k} = 1 \mid x_i, \theta^{(j)}\}$$

$$= Pr\{s_i = k \mid x_i, \theta^{(j)}\}$$

Bayes rule →

$$= \frac{Pr\{s_i = k \mid \theta^{(j)}\} \cdot g(x_i \mid s_i = k, \theta^{(j)})}{g(x_i \mid \theta^{(j)})}$$

$$= \frac{w_k^{(j)} \cdot \phi(x_i \mid \mu_k^{(j)}, \Sigma_k^{(j)})}{\sum_{\ell=1}^{K} w_\ell^{(j)} \phi(x_i \mid \mu_\ell^{(j)}, \Sigma_\ell^{(j)})}$$

M-Step for GMM

$$Q(\theta, \theta^{(j)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_{i,k}^{(j)} \left[\log w_k - \frac{d}{2} \log 2\pi \right.$$

$$\left. - \frac{1}{2} \log |\Sigma| - \frac{1}{2}(x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) \right]$$

Maximizing w.r.t. θ yields ...

$$\mu_k^{(j+1)} = \frac{\sum_{i=1}^{n} \gamma_{i,k}^{(j)} \cdot x_i}{\sum_{i=1}^{n} \gamma_{i,k}^{(j)}}, \quad k = 1, \ldots, K$$

$$\Sigma_k^{(j+1)} = \frac{\sum_{i=1}^{n} \gamma_{i,k}^{(j)} \cdot (x_i - \mu_k^{(j+1)}) \cdot (x_i - \mu_k^{(j+1)})^T}{\sum_{i=1}^{n} \gamma_{i,k}^{(j)}}, \quad k = 1, \ldots, K$$

$$w_k^{(j+1)} = \frac{1}{n} \sum_{i=1}^{n} \gamma_{i,k}^{(j)}, \quad k = 1, \ldots, K.$$

The number $\gamma_{i,k}$ is called the <u>responsibility</u> of component k for observation i.

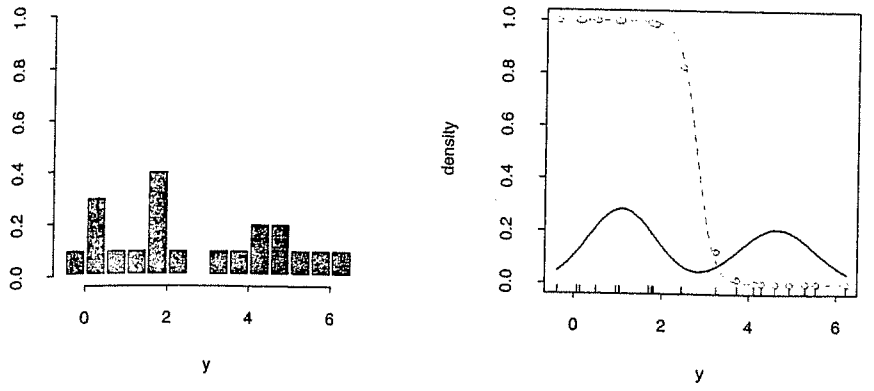


FIGURE 8.5. *Mixture example. Left panel: histogram of data. Right panel: maximum likelihood fit of Gaussian densities (solid red) and responsibility (dotted green) of the left component density for observation y, as a function of y.*

TABLE 8.1. *20 fictitious data points used in the two-component mixture example in Figure 8.5.*

-0.39	0.12	0.94	1.67	1.76	2.44	3.72	4.28	4.92	5.53
0.06	0.48	1.01	1.68	1.80	3.25	4.12	4.60	5.28	6.22

← Hastie, Tibshirani & Friedman, The Elements of Statistical Learning

Initialization

In general, the expected complete-data log-likelihood has several local maxima. Therefore initialization of the EM algorithm is critical.

In fact, a global maximum is obtained by putting $\mu_1 = x_i$ for some i, $\Sigma_1 = 0$, $w_1 = 1$, which is not a useful solution. Hence, we are actually seeking a local maxima.

A good initialization for the GMM is

$$\mu_k^{(0)} = \text{random } x_i \quad \text{(distinct)}$$

$$\Sigma_k^{(0)} = \text{sample covariance}$$

$$w_k^{(0)} = \frac{1}{K}$$

In practice, it may be beneficial to initialize the algorithm and run it several times, and select the final estimate with largest expected complete-data log-likelihood.

Termination

The EM algorithm can be terminated when

$$|Q(\theta^{(j+1)}, \theta^{(j)}) - Q(\theta^{(j)}, \theta^{(j-1)})| \leq \epsilon$$

for some user defined tolerance ϵ, or when the estimated parameters do not change by more than a predetermined amount.

Question | Why is the EM algorithm couched in terms of the log likelihood as opposed to just the likelihood?

The EM Algorithm in General

The EM algorithm is a general class of algorithms for maximum likelihood estimation. It is useful whenever MLE given the observed data $\underline{x}$ is intractable, but there exists some unobserved data $\underline{z}$ that makes the MLE tractable. Applications are numerous. The EM algorithm also applies to maximum a posteriori (MAP) estimation.

Convergence Properties

Theorem: For each $j = 1, 2, \ldots$

$$\ell(\theta^{(j)} \mid \underline{x}) \geq \ell(\theta^{(j-1)} \mid \underline{x})$$

where $\ell(\theta \mid \underline{x})$ is the likelihood of θ given the observed data $\underline{x}$.