MIXTURE MODELS & EM ALGORITHM

General set-up:

Description of distribution:

- There are K component distributions specified by densities
  $f_j(x;\theta_j) \qquad j=1,2,\dots,K$
- A "selecting" random variable Z takes values in $\{1,2,\dots,K\}$, with probability distribution π, so
  $$\mathbb{P}(Z=j)=\pi_j,\ j=1,2,\dots,K$$

Sometimes we will represent Z by a K-dimensional vector $Z=(Z_1,\dots,Z_K)$ with

$$Z_j=\begin{cases}1 & \text{if } j \text{ is selected}\\ 0 & \text{otherwise}\end{cases}$$

So, e.g., if 3 is selected, $Z=(0,0,1,0,\dots,0)$

- Given that $Z=j$, the random variable X is sampled from $f_j(\cdot;\theta_j)$ above

Thus the overall density of X is, by conditioning on Z, given by

$$f_X(x;\theta_1,\dots,\theta_j,\pi)=\sum_{j=1}^{K} f_j(x;\theta_j)\mathbb{P}(Z=j)$$

$$f_X(x;\theta_1,\dots,\theta_j,\pi) = \sum_{j=1}^{k} f_j(x;\theta_j)\,\pi_j$$

The pair (X,Z) has <u>joint</u> density

$$f_{X,Z}(x,z;\theta_1,\dots,\theta_j,\pi) = \prod_{j=1}^{k}\left[\pi_j f_j(x;\theta_j)\right]^{z_j}$$

because if $z = (0,0,\dots,\underset{j_0^{th}}{1},\dots,0)$ in its j_0^{th} component vector representation, then

$$f_{X,Z}(x,(0,\dots,1,\dots,0);\theta_1,\dots,\theta,\pi) = \pi_j f_j(x;\theta_j)$$

since $z_j = 1$ only for $j = j_0$

Now, suppose $(X_1,Z_1),(X_2,Z_2),\dots,(X_n,Z_n)$ are sampled from above distribution.

<u>NOTE</u> Now Z_i in its vector representation is

$$Z_i = (Z_{i,1},\dots,Z_{i,k}) \text{ with}$$

$$Z_{i,j} = \begin{cases} 1 & \text{if } X_i \text{ is sampled from } f_j(\cdot;\theta_j) \\ 0 & \text{otherwise} \end{cases}$$

Note: The Z_i's are HIDDEN, or LATENT VARIABLES

We have already seen an example where each $f_j(x;\theta_j)$ is $N(\mu_j, \Sigma_j)$ a Normal (or GAUSSIAN) density

The goal is to estimate $\theta_1, \ldots, \theta_K$ and π using only $X_1, \ldots, X_n$, the observed data. For each X_i, we do not know which distribution f_j generated X_i. (The Z_i variable would tell us this information)

The E-M algorithm provides method for maximizing the likelihood function for the joint observations $(X_1, \ldots, X_n)$

$$L(\theta_1, \ldots, \theta_K, \pi; x_1, \ldots, x_n)$$

$$= f(x_1, \ldots, x_n; \theta_1, \ldots, \theta_K, \pi)$$

← JOINT density of $(x_1, \ldots, x_n)$

$$= \prod_{i=1}^{n} f_X(x_i; \theta_1, \ldots, \theta_K, \pi) \quad \text{by INDEPENDENCE}$$

Here $f_X(x_i; \theta_1, \ldots, \theta_K, \pi) = \sum_{j=1}^{K} \pi_j f_j(x_i; \theta_j)$

as above

The E-M algorithm starts by writing down the complete data log-likelihood, that is,

$$\log f_{X,Z}(x_1,\ldots,x_n,z_1,\ldots,z_n;\theta_1,\ldots,\theta_K,\pi)$$

$$f_{X,Z}(x_1,\ldots,x_n,z_1,\ldots,z_n;\theta)$$

$(x_1,\ldots,x_n)$ $(z_1,\ldots,z_n)$ write $\theta=(\theta_1,\ldots,\theta_K,\pi)$

$$= \prod_{i=1}^{n} f_{X,Z}(x_i,z_i;\theta)$$

$$= \prod_{i=1}^{n}\prod_{j=1}^{K}\left[f_j(x_i;\theta)\pi_j\right]^{z_{ij}}$$

since each pair (x_i,z_i) has density $\prod_{j=1}^{K}\left[f_j(x_i;\theta)\pi_j\right]^{z_{ij}}$

Taking logarithms shows

$$\log f_{X,Z}(x_1,\ldots,x_n,z_1,\ldots,z_n;\theta) = \sum_{i=1}^{n}\sum_{j=1}^{K} z_{ij}\log\pi_j + \sum_{i=1}^{n}\sum_{j=1}^{K} z_{ij}\log f_j(x_i;\theta_j)$$

In the E-step of the algorithm, the "Q-function" is calculated. This is defined as

$$Q(\theta,\theta_0) = \mathbb{E}\left[\log f_{X,Z}(x_1,\ldots,x_n,z_1,\ldots,z_n;\theta)\right]$$

where the expectation $\mathbb{E}$ is calculated using

The distribution $p_{Z|X}(z_1,\dots,z_n \mid x_1,\dots,x_n;\theta_0)$

That is, the $x_1,\dots,x_n$ are held constant,

and the function $\log f_{X,Z}(x_1,\dots,x_n,z_1,\dots,z_n;\theta)$ is averaged with respect to the probability distribution over $(z_1,\dots,z_n)$ given by

$$p_{Z|X}(z_1,\dots,z_n \mid x_1,\dots,x_n;\theta_0)$$

So using our formula for $\log f_{X,Z}(\underset{\sim}{x},\underset{\sim}{z};\theta)$ above

($\underset{\sim}{x}=(x_1,\dots,x_n)$, $\underset{\sim}{z}=(z_1,\dots,z_n)$)

$$Q(\theta,\theta_0) = \sum_{i=1}^{n}\sum_{j=1}^{k} E[z_{i,j} \mid x_1,\dots,x_n;\theta_0)\log\pi_j + \sum_{i=1}^{n}\sum_{j=1}^{k} E[z_{i,j} \mid x_1,\dots,x_n;\theta_0)\log f_j(x_i;\theta_j)$$

So all we need is

$$E[z_{i,j} \mid x_1,\dots,x_n;\theta_0] = E[z_{i,j} \mid x_i;\theta_0]$$

(z_i only depends on x_i)

$$= P(z_{i,j}=1 \mid x_i;\theta_0) \overset{\text{DEF}}{=} \gamma^{(i)}(j)$$

This is the conditional prob., given observation x_i, that this observation came from distribution j

Bayes Theorem allows calculation

$$\gamma^{(i)}(j) = \frac{\pi_j f_j(x_i; \theta_j)}{\sum_{\ell=1}^{K} \pi_\ell f_\ell(x_i; \theta_\ell)}$$

Note that $\sum_{j=1}^{K} \gamma^{(i)}(j) = 1$

The distribution $(\gamma^{(i)}(1), \ldots, \gamma^{(i)}(K))$ gives the conditional prob.'s that X_i came from each of the K distributions $f_j(\ ; \theta_j)$

To conclude,

$$Q(\theta, \theta_0) = \sum_{j=1}^{K} \sum_{i=1}^{n} \gamma^{(i)}(j) \log \pi_j + \sum_{j=1}^{K} \sum_{i=1}^{n} \gamma^{(i)}(j) \log f_j(x_i; \theta_j)$$

Let $N_j = \sum_{i=1}^{n} \gamma^{(i)}(j) = \mathbb{E}[\#X_i\text{'s from } j^{\text{th}} \text{ component}]$

$$Q(\theta, \theta_0) = \underbrace{\sum_{j=1}^{K} N_j \log \pi_j}_{T_1} + \underbrace{\sum_{j=1}^{K} \sum_{i=1}^{n} \gamma^{(i)}(j) \log f_j(x_i; \theta_j)}_{T_2}$$

We are done with the E-step; this amounts to calculation of γ's as above

In the M-step, θ_0 is held fixed and $Q(\theta, \theta^0)$ is maximized over θ

The maximization of $T_1 = \sum_{j=1}^{k} N_j \log \pi_j$ is always the same, it does not depend on the form of f_j's

The solution is always

$$\pi_j = \frac{N_j}{n}$$

The maximization of T_2 depends on form of f_j's

Let us try when f_j is poisson (λ_j)

$$f_j(x) = e^{-\lambda_j} \lambda_j^x / x! \qquad x = 0, 1, \ldots$$

$$\log f_j(x_i) = -\lambda_j + x_i \log \lambda_j - \log x_i!$$

Thus $$T_2 = \sum_{j=1}^{k} \sum_{i=1}^{n} \gamma^{(i)}(j) \left[-\lambda_j + x_i \log \lambda_j - \log x_i! \right]$$

we can maximize each inner sum independently

$$\frac{\partial}{\partial \lambda_j} \sum_{i=1}^{n} \gamma^{(i)}(j) \left[-\lambda_j + x_i \log \lambda_j - \log x_i! \right]$$

$$= \sum_{i=1}^{n} \gamma^{(i)}(j) \left[-1 + x_i / \lambda_j \right]$$

$$= -\sum_i \gamma^{(i)}(j) + \frac{1}{\lambda_j}\sum_{i=1}^n \gamma^{(i)}(j)\, x_i$$

$$= -N_j + \frac{1}{\lambda_j}\sum_{i=1}^n \gamma^{(i)}(j)\, x_i$$

Solving for λ_j after setting to 0 yields

$$\lambda_j = \frac{1}{N_j}\sum_{i=1}^n \gamma^{(i)}(j)\, x_i$$

Thus given the posterior prob. $\gamma^{(i)}(j)$ that x_i came from component j, the estimate of λ_j weights the data point x_i by $\gamma^{(i)}(j)$, and the weighted average is the estimate of the mean λ_j of the j^{th} component

So for Poisson, the "M-step" amounts to writing down

$$\lambda_j = \frac{1}{N_j}\sum_{i=1}^n \gamma^{(i)}(j)\, x_i$$

The EM-algorithm then iterates; once $\lambda_{j,n}$ is calculated as above, then and the π_j's

The γ's are recomputed

This converges to a local max of the likelihood function.

Above can be applied where each f_j is a BERNOULLI vector

eg digit recognition

First we describe the "single-variable" dist'n.

Let $X = (X_1, \ldots, X_D)$ be a collection of independent BITS, where

$$\mathbb{P}(X_\ell = 1) = \mu_\ell$$
$$\mathbb{P}(X_\ell = 0) = 1 - \mu_\ell$$

Thus $\mathbb{P}(X_1 = x_1, \ldots, X_D = x_D) = \prod_{\ell=1}^{D} \mu_\ell^{x_\ell} (1-\mu_\ell)^{1-x_\ell}$

For $x_\ell = 0$ or 1 $\quad \ell = 1, \ldots, D$

This distribution has parameters $(\mu_1, \ldots, \mu_D)$

Now suppose we have K such distributions

The j^{th} distribution has parameters

$$(\mu_{j,1}, \ldots, \mu_{j,D})$$

Now we mix these with mixing probabilities

$$\pi_1, \ldots, \pi_j$$

So the set-up is as above, but now

$$f_j\left(\underset{\uparrow}{x_i}\,;\,\mu_j=(\mu_{j,1},\ldots,\mu_{j,D})\right)$$

$$(x_{i,1},\ldots,x_{i,D}) \qquad = \prod_{l=1}^{D} \mu_{j,l}^{x_{i,l}} (1-\mu_{j,l})^{1-x_{i,l}}$$

i = "which data point"

j = "which component distribution"

l = "which bit out of D bits"

T_1 term in $Q(\theta,\theta_0)$ does not change

T_2 term is

$$\sum_{j=1}^{K} \sum_{i=1}^{n} \gamma^{(i)}(j) \left[\sum_{l=1}^{D} x_{i,l} \log \mu_{j,l} + (1-x_{i,l}) \log(1-\mu_{j,l}) \right]$$

Fixing j, we can differentiate with respect to μ_{j,l_0} on inside sum

$$\frac{\partial}{\partial \mu_{j,l_0}} = \sum_{i=1}^{n} \gamma^{(i)}(j) \left[x_{i,l_0} \frac{1}{\mu_{j,l_0}} - (1-x_{i,l_0}) \frac{1}{1-\mu_{j,l_0}} \right]$$

Setting to 0 and solving (multiply both sides by $\mu_{j,l_0}(1-\mu_{j,l_0})$

$$0 = \sum_{i=1}^{n} \gamma^{(i)}(j)\, x_{i,l_0} (1-\mu_{j,l_0}) - \mu_{j,l_0} \Big(N_j - \underbrace{\sum_{i=1}^{n} \gamma^{(i)}(j)}_{\uparrow} x_{i,l_0} \Big)$$

$$0 = s_{l_0}(j) - \mu_{j,l_0} s_{l_0}(j) - \mu_{j,l_0} N_j + \mu_{j,l_0} s_{l_0}(j)$$

$$\mu_{j,l_0} = s_{l_0}(j) / N_j$$

In vector notation

$$\mu_j = \frac{s(j)}{N_j} \leftarrow \quad s(j) = (s_1(j), \ldots, s_D(j))$$

$\uparrow$

$(\mu_{j,1}, \ldots, \mu_{j,D})$

Thus to apply E-M in this case, during M-step, set

$$\mu_j = \frac{s(j)}{N_j}, \quad s_l(j) = \sum_{i=1}^{n} \gamma^{(i)}(j) x_{i,l}$$

HIGH-LEVEL SUMMARY

For mixtures, the E-step is always the same: compute γ^i's

The M-step depends on form of f_j's