Mixture densities & EM algorithm.

In general, densities can be multi-modal.

A single parametric density model might not be a good choice for class-conditional densities.

Solution: use convex-combination of multiple parametric densities.

One example: Gaussian mixture models.

$$f(x_i) = \sum_{j=1}^{K} \alpha_j f_j(x_i)$$

Each $f_j \sim N(\mu_j, \sigma_j)$.

$$l(\alpha, \mu, \sigma) = \sum_{i=1}^{n} \log \sum_{j=1}^{K} \alpha_j f_j(x_i).$$

Likelihood cannot be maximized directly (we have log of sum).

Wish to model these kinds of densities using another latent variable & specifying the joint distribution of the data & the latent variable.

Specifically,

Define a new 'un-observed' RV Z

s.t., $z^i \sim \text{Multinomial}(\alpha) \in \{1, 2, \dots, K\}$.

$$\alpha_j \geq 0 \quad \sum_{j=1}^{K} \alpha_j = 1.$$

$$\alpha_j = P(z^i = j).$$

Model: Each x_i was generated by 'selecting' one of K values for z_i & drawing x_i from that Gaussian.

Now, we have

$$l(\alpha, \mu, \sigma) = \sum_{i=1}^{n} \log f(x_i ; \alpha, \mu, \sigma).$$

$$= \sum_{i=1}^{n} \log \sum_{z_i=1}^{K} P(x_i | z_i ; \mu, \sigma) P(z_i ; \alpha)$$

Now suppose z_i's were known, then optimizing the above would have been easy.

$\because$ z_i was fixed for every i.

Thus, $$l = \sum_{i=1}^{n} \log \left(f(x_i | z_i ; \sigma, \mu) \right) + \log P(z_i ; \alpha)$$

Now, $$\frac{\partial l}{\partial \alpha_j} = \frac{1}{n} \sum_{i=1}^{n} I_{[z_i = j]}$$

$$\frac{\partial l}{\partial \mu_j} = \frac{\sum_{i=1}^{n} I_{[z_i=j]} x_i}{\sum_{i=1}^{n} I_{[z_i=j]}}$$

$I \rightarrow$ Indicator of $z_i = j$

$$\sigma_j = \frac{\sum_{i=1}^{m} I_{[z_i=j]} (x_i - \mu_j)^2}{\sum_{i=1}^{n} I_{[z_i=j]}}.$$

Basically, the estimates are similar to MLE for a single gaussian but they are weighted by the proportion of occurance of each component.

<u>But the problem</u>: we do not have any information about z_i.

Question: How to estimate MLE for GMMs in that case?

<u>Answer</u>. use an iterative algo:

i) Guess a value for z_i

ii) update the parameters with that guess

iii) Repeat (1) & (2) until convergence.

<u>Procedure:</u> For every x_i, Calculate

i) $$\omega_{ij}^{t} = P\left[z_i = j \,\middle|\, x_i ; \alpha^{t}, \mu^{t}, \sigma^{t} \right]$$

(Guess for z_i)

ii) $$\alpha_j^{t+1} = \frac{1}{n} \sum_{i=1}^{m} \omega_{ij}^{t}$$

$$\mu_j^{t+1} = \frac{\sum_{i=1}^{m} \omega_{ij}^{t} x_i}{\sum_{i=1}^{n} \omega_{ij}}$$

$$\sigma_j^{t+1} = \frac{\sum_{i=1}^{n} \omega_{ij}^{t} (x_i - \mu_j)^2}{\sum_{i=1}^{n} \omega_{ij}^{t}}$$

[update the parameters].

How to find ω_{ij}^{t} ?

Answer: use Bayes rule.

$$w_{ij} = P\left(z_i = j \mid x_i ; \alpha^t, \mu^t, \sigma^t\right)$$

$$= \frac{P\left(x_i \mid z_i = j ; \mu^t, \sigma^t\right) P\left(z_i = j ; \alpha^t\right)}{\sum_{l=1}^{K} P\left(x_i \mid z_i = l\right) P\left(z_i = l ; \alpha^t\right)}.$$

We can compute all the above terms.

Thus, EM algorithm looks like giving a 'weighted' averages with 'weights' being 'soft' in the sense that they are probabilistic.
[This accounts for the fact that z_i's are only guesses]

Question: Why should this procedurce work?

The Generalized EM algorithm.

The general setting:

$D = \{x_1, x_2, \dots x_n\}$ are sampled IID

from a GMM. [seen data]

z_i indicates the Component from which x_i

was drawn. [latent|unseen|hidden variable]

Goal: Find a procedurce to obtain the MLE

for such a model.

$$l(\theta) = \sum_{i=1}^{n} \log f(x_i ; \theta).$$

$$= \sum_{i=1}^{n} \log \sum_{z} f(x_i, z_i ; \theta).$$

General strategy of EM:

repeatedly

Construct a lower bound on $l(\theta)$ &

repeatedly optimize the lower bound.

$\forall i$, Let ϕ_i be a distribution over zs.

$\because \sum_z \phi_i(z) = 1, \; \phi_i(z) \geq 0.$

$$l = \sum_i \log f(x_i; \theta)$$

$$= \sum_i \log \sum_{z_i} f(x_i, z_i; \theta)$$

$$= \sum_i \log \sum_{z_i} \phi_i(z_i) \frac{f(x_i, z_i; \theta)}{\phi_i(z_i)} \quad -(2).$$

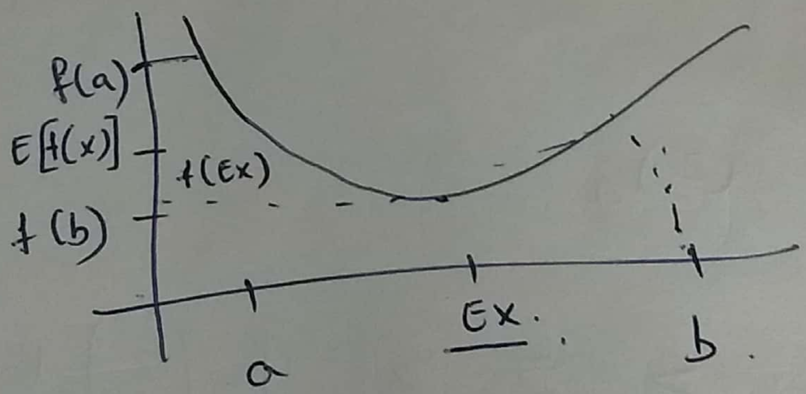
Jensen's inequality :

f be a real-valued convex function

i.e, $f''(x) \geq 0 \;\; \forall \; x \in R$.

If X is a RV, then

$$E\left[f(x)\right] \geq f(EX).$$



$X \in \{a, b\}$

$P(X=a) = 1/2$

For concave functions, $E[f(x)] \leq f(EX)$.

In (2), since log is a concave fn,

$$\sum_{z_i} \phi_i(z_i) \left[\frac{f(x_i, z_i; \theta)}{\phi_i(z_i)} \right] =$$

$$(\quad A \quad)$$

$E_{z_i \sim \phi_i}(A)$ [Hence the name EM].

Thus, $f\left(E_{z_i}(A)\right) \geq E_{z_i \sim \phi_i}[A]$

Thus in (2),

$$l \geq \sum_i \sum_{z_i} \phi_i(z_i) \log \frac{f(x_i, z_i; \theta)}{\phi_i(z_i)}.$$

Any ϕ_i would give a lower bound on l. But we want a tight-one

Jensen's inequality would be tight if $E[X]$ is a constant.

in our case,

$$\frac{f(x_i, z_i)}{\phi_i(z_i)} = c. \quad [\text{not depending upon } z_i].$$

$$\Rightarrow \phi_i(z_i) \propto f(x_i, z_i; \theta).$$

Also, $\sum_{z} \phi_i(z_i) = 1 \quad \because$ its a distributio

Thus, $$\phi_i(z_i) = \frac{f(x_i, z_i; \theta)}{\sum_{z} f(x_i, z; \theta)}$$

$$= \frac{f(x_i, z_i; \theta)}{f(x_i; \theta)}$$

$$= f(z_i | x_i; \theta).$$

$\phi_i(z_i) \sim$ posterior of z_i given x_i

[~~kimi~~ same as our previous guess].

Now, EM:

for each i, set

$$\phi_i(z_i) = P[z_i | x_i; \theta^t]. \quad [\text{E-step}]$$

$$\text{M-step}: \theta^{t+1} = \underset{\theta}{\operatorname{argmax}} \sum_i \sum_{z_i} \phi_i(z_i) \log \frac{f(x_i, z_i; \theta^t)}{\phi_i(z_i)}.$$

Question: Does this procedure converge?

consider: E-step. Ensures that

$$\ell(\theta^t) = \sum_i \sum_{z^i} \phi_i^t(z_i) \log \frac{f(x_i, z_i; \theta^t)}{\phi_i^t(z_i)}.$$

θ^{t+1} is obtained by maximizing the above.

$$\therefore \ell(\theta^{t+1}) \geq \sum_i \sum_{z_i} \phi_i^t(z_i) \log \frac{f(x_i, z_i; \theta^{t+1})}{\phi_i^t(z_i)}.$$

[Jensen's inequality].

$$\geq \sum_i \sum_{z_i} \phi_i^{(t)}(z_i) \log \frac{f(x_i, z_i; \theta^t)}{\phi_i^t(z_i)},$$

[$\because \theta^{t+1}$ is explicitly obtained to maximize $\ell(\theta^t)$].

$$= \ell(\theta^{(t)})$$

[by choice]

$\therefore \quad l(\theta^{t+1}) \geq l(\theta^{t})$.

Note that global convergence is not gaurenteed.

Thus, EM gaurentees the increase in likelihood for each iteration.

EM for the GMM case:

As before, lets define E-step at t^{th} iteration

$$\phi_i^t(z_i=j) = P\left[z_i=j \mid x_i ; \alpha^t, \mu^t, \sigma^t\right].$$

Lets call it w_{ij}^t

M-Step:

$$l = \sum_{i=1}^{n} \sum_{i} \phi_i^t(z_i) \log \frac{f(x_i, z_i; \alpha^t, \mu^t, \sigma^t)}{\phi_i^t(z_i)}$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{k} \phi_i^t(z_i = j) \log \frac{f(x_i | z_i = j, \mu^t, \sigma^t) \, P(z_i = j; \alpha^t)}{\phi_i^t(z_i = j)}.$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{k} \omega_{ij}^t \log \frac{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-\frac{1}{2}(x_i - \mu_j^t)^2}{\sigma_j^t}\right) \cdot \alpha_j^t}{\omega_{ij}^t}$$

$$\frac{\partial l}{\partial \mu_j^t} = 0 \Rightarrow \mu_j^{t+1} = \frac{\sum_{i=1}^{n} \omega_{ij}^t x_i}{\sum_{i=1}^{n} \omega_{ij}}$$

Similarly for $\sigma_l^{t+1} = \frac{\sum_{i=1}^{n} \omega_{ij}^t (x_i - \mu_j)^2}{\sum_{i=1}^{n} \omega_{ij}^t}$

(69).

We have to estimate α_i^s too.

in l, the terms that depend upon α

~~out~~ is $l' = \sum_{i=1}^{n} \sum_{j=1}^{K} w_{ij}^{t} \log \alpha_j^{t}$

This again is a constrained optimization problem $\because \sum_{j=1}^{K} \alpha_j = 1$. $\because \alpha_j = P(z_i = j)$.

$$\therefore L(\alpha) = \sum_{i=1}^{n} \sum_{j=1}^{K} w_{ij}^{t} \log \alpha_j^{t} + \lambda \left(\sum_{j=1}^{K} \alpha_j^{t} - 1 \right).$$

$$\Rightarrow \alpha_j^{t} = \frac{\sum_{i=1}^{n} w_{ij}^{t}}{\lambda}$$

since $\sum_{j=1}^{K} \alpha_j^{t} = 1$,

$\cancel{-\lambda = \sum_{i=1}^{n} \sum_{j=1}^{K} \alpha_j^{t}}$

$$\sum_{j=1}^{K} \sum_{i=1}^{n} \frac{w_{ij}^{t}}{\lambda} = 1.$$

$$\frac{n}{\lambda} = 1 \Rightarrow \lambda = n$$

Thus, $\boxed{\alpha_j^{t} = \frac{1}{n} \sum_{i=1}^{n} w_{ij}^{t}}$.