3. EM Algorithm

In "Gaussian Mixture Model", we give out the "Maximum Likelihood" view to see E-M algorithm, but we cannot prove why algorithm will converge. In this section, we will look at EM algorithm from the information theorem's perspect.

1. Kullback-Leibler Divergence.

We will use KL Divergence in the analysis of EM.

Suppose we have two distributions $f_0(x)$, $f_1(x)$, we want to measure the "distance" between these two, Divergence is a way.

$$D(f_0 \| f_1) = \int_x \log_2 \left(\frac{f_0(x)}{f_1(x)} \right) f_0(x)\, dx$$

It also has an entropy like interpretation

Use $H(f_0)$ to represent the entropy for R.V satisfing $f_0(x)$.

$$H(f_0) = -\sum_x \left(\log_2 f_0(x)\right) f_0(x)$$

↑ general summation, for both Cont. & DT R.V.

$H(f_0, f_1)$ is called cross entropy.

$$H(f_0, f_1) = -\sum f_0(x) \log f_1(x)$$

Then it's easy to check that

$$D(f_0 \| f_1) = H(f_0, f_1) - H(f_0)$$

Note that, $H(f_0, f_1) \geq H(f_0)$, equality can be achieved when $f_0 = f_1$

Some useful properties of KL-Divergence.

① None-negativity

$\because H(f_0, f_1) \geq H(f_0)$ (proof ignored here)

$\therefore D(f_0, f_1) = H(f_0, f_1) - H(f_0) \geq 0$

② $D(f_0, f_1) = 0 \iff f_0 = f_1$

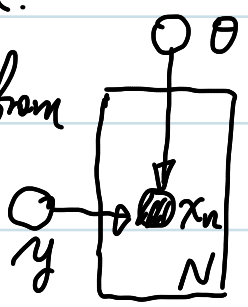
③ $D(f_0 \| f_1) \neq D(f_1 \| f_0)$

Our model is x, y, θ

x: set of samples from the mixture model.

y: latent variable, determine which model x_n is from

θ: parameters for each model.



For the EM algorithm

$$\ln f(x, y|\theta) = \ln\left(f(y|x,\theta) \cdot f(x|\theta)\right)$$
$$= \underbrace{\ln f(y|x,\theta)}_{\text{log posterior}} + \ln f(x|\theta)$$

Now, we start with a initial guess θ_0, and try to improve it.

Suppose currently, θ_0 turns to θ_t

the difference of current model based on θ_t and real model is

(We are talking about likelihood)

$$\log f(x, y; \theta) - \log f(x, y | \theta_t)$$

$$= \log f(y|x,\theta) + \log f(x|\theta) - \log f(y|x,\theta_t) - \log f(x|\theta_t)$$

(Reorder)

$$= \log f(x|\theta) - \log f(x|\theta_t) + \underbrace{\log \frac{f(y|x;\theta_t)}{f(y|x;\theta)}}$$

↑ we want to turn this to KL Divergence.

Now, take expectation to $y \sim f(y|x;\theta_t)$,

$$\mathbb{E}_y[\log f(x,y|\theta)] - \mathbb{E}_y[\log f(x,y|\theta_t)]$$

$$= \log f(x|\theta) - \log f(x|\theta_t) - \underbrace{D(f(y|x;\theta_t) \| f(y|x;\theta))}_{\geq 0}$$

we now have

$$\underbrace{\log f(x|\theta)}_{l(\theta)} - \underbrace{\log f(x|\theta_t)}_{l(\theta_t)} \geq \underbrace{\mathbb{E}_y[\log f(x,y|\theta)]}_{Q(\theta,\theta_t)} - \underbrace{\mathbb{E}_y[\log f(x,y|\theta_t)]}_{Q(\theta_t,\theta_t)}$$

we have

$$l(\theta) - l(\theta_t) \geq Q(\theta,\theta_t) - Q(\theta_t,\theta_t)$$

We can choose θ s.t

$$\theta_{t+1} = \arg\max_{\theta} Q(\theta,\theta_t)$$

$Q(\theta, \theta_t)$ is $\int_y \log f(x, y|\theta) \cdot f(y|x, \theta_t)\, dy = E_y[\log f(x, y|\theta)]$
is the expected log likelihood.

Now the algorithm turns to

* E-step.

Compute $Q(\theta, \theta_t)$

$Q(\theta, \theta_t) = \int_y \log f(x, y|\theta) \cdot f(y|x, \theta_t)\, dy$

* M-step

$\theta_{t+1} = \underset{\theta}{\operatorname{argmax}}\ Q(\theta, \theta_t)$

Geometric Interpretation

Our ultimate target is to find θ to maximize $\log(X|\theta)$

In each step, we can use current θ_t to give a best guess of posterior $f(y|x, \theta_t)$, and try to maximize $Q(\theta, \theta_{t+1})$ based on that.