Expectation - Maximization Algorithm

We have $X_1 \dots X_n$ iid (random sample)

Let we know that X_i comes from the mixture of two distributions and we know one of the components:

$$X_i \sim (1-\varepsilon) N(3,1) + \varepsilon N(\mu, 1)$$

We need to estimate unknown parameters ε, μ.

$\theta = (\varepsilon, \mu)$ - uknown parameters.

Standard way: likelihood function:

$$L(\theta/X) = \prod_{i=1}^{n} [(1-\varepsilon)\varphi(X_i - 3) + \varepsilon \varphi(X_i - \mu)],$$

where $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ - pdf of $N(0,1)$.

$\Rightarrow$ calculate $\hat{\theta}_{MLE} = \arg\max L(\theta/X)$.

However, analytical formula for $\hat{\theta}_{MLE}$ is very complicated.

The natural solution to this problem is to use Expectation - Maximization Algorithm (EM Algorithm)

I Let $z_1 \dots z_n$ be iid random variables from Bernoulli distribution:

$$P(z_i = 1) = \varepsilon = 1 - P(z_i = 0).$$

(latent indicators: variables that have impact on the data, but we don't know them)

$$\varphi(X_i / z_i) = \begin{cases} \varphi(X_i - 3), & \text{if } z_i = 0 \\ \varphi(X_i - \mu), & \text{if } z_i = 1 \end{cases}$$

assume, that we know z_i. Then

$$\hat{\varepsilon} = \frac{1}{n}\sum_{i=1}^{n} z_i ; \quad \hat{\mu} = \frac{\sum X_i z_i}{\sum z_i}$$

Γ Complete likelihood function:

$L(X, z/\theta) = L(X/z, \theta)\, L(z/\theta)$

$L(X/z, \theta) \overset{\text{we know } z}{=} \prod_{i=1}^{n} [\varphi(X_i - 3)]^{1-z_i} [\varphi(X_i - \mu)]^{z_i}$

Then:

$\log L(X/z, \theta) = \sum_{i=1}^{n} [(1-z_i) \log \varphi(X_i - 3) + z_i \log \varphi(X_i - \mu)] =$

$= / \log \varphi(X_i - 3) = \log \frac{1}{\sqrt{2\pi}} - \frac{(X_i - 3)^2}{2} / =$

$= \sum_{i=1}^{n} \left[\log \frac{1}{\sqrt{2\pi}} - (1-z_i) \frac{(X_i-3)^2}{2} - z_i \frac{(X_i - \mu)^2}{2} \right]$

$\log L(X, z/\theta) = \log(X/z, \theta) - \log L(z, \theta)$ ⇐ does not depend on μ

$\frac{\partial \log L(X, z/\theta)}{\partial \mu} = \sum_{i=1}^{n} (z_i X_i - z_i \mu) = 0$

$\Rightarrow \hat{\mu} = \frac{\sum_{i=1}^{n} z_i X_i}{\sum_{i=1}^{n} z_i}$

II EM Algorith

Complete likelihood function:

$L(X, z/\theta) = \prod_{i=1}^{n} f(X_i/z_i)\, P(z_i) =$

$= \prod_{i=1}^{n} [\varphi(X_i - 3)]^{1-z_i} [\varphi(X_i - \mu)]^{z_i}\, \varepsilon^{z_i} (1-\varepsilon)^{1-z_i}$

$\Rightarrow$

$\log L(X, z/\theta) = \sum_{i=1}^{n} [(1-z_i) \log \varphi(X_i - 3) + z_i \log \varphi(X_i - \mu) + z_i \log \varepsilon + (1-z_i) \log(1-\varepsilon)]$

In fact, we don't know z_i ⇒ we estimate z_i by mean (expectation step) ⇒ maximize likelihood function (maximization step).

<u>Itterations: (k+1)th Step</u>

Let $\theta_k = (\varepsilon_k, \mu_k)$ be the values of ε, μ at kth step

Expectation step.

We will replace z_i with $\pi_i^k = E(z_i / x, \theta_k)$

calculated on previous step.

$$\pi_i^k = E(z_i / x, \theta_k) = P(z_i = 1 / x, \theta_k) = /\text{Bayes' formula}/ =$$

$$= \frac{\varphi(x_i / z_i = 1, \theta_k)\, P_{\theta_k}(z_i = 1)}{\varphi(x_i / z_i = 1, \theta_k)\, P_{\theta_k}(z_i = 1) + \varphi(x_i / z_i = 0, \theta_k)\, P_{\theta_k}(z_i = 0)} =$$

$$= \frac{\varphi(x_i - \mu)\, \varepsilon_k}{\varphi(x_i - \mu)\, \varepsilon_k + \varphi(x_i - 3)(1 - \varepsilon_k)}$$

$$\Rightarrow Q(\theta / \theta_k) = E_{z / x, \theta_k} \log L(x, z / \theta) =$$

$$= \sum_{i=1}^{n} \Big[(1 - \pi_i^k) \log \varphi(x_i - 3) + \pi_i^k \log \varphi(x_i - \mu) + \pi_i^k \log \varepsilon + (1 - \pi_i^k) \log(1 - \varepsilon) \Big]$$

Maximization step:

$$\theta_{k+1} = \underset{\theta}{\operatorname{argmax}}\, Q(\theta / \theta_k)$$

$$\frac{\partial Q(\theta, \theta_k)}{\partial \varepsilon} = \sum_{i=1}^{n} \left(\frac{\pi_i^k}{\varepsilon} - \frac{(1 - \pi_i^k)}{1 - \varepsilon} \right) = 0.$$

$$\Rightarrow (1 - \hat{\varepsilon}) \sum_{i=1}^{n} \pi_i^k = \left(n - \sum_{i=1}^{n} \pi_i^k\right) \hat{\varepsilon} \Rightarrow$$

$$\Rightarrow \hat{\varepsilon} = \frac{\sum_{i=1}^{n} \pi_i^k}{n} =: \varepsilon_{k+1}$$

$$\frac{\partial Q(\theta, \theta_k)}{\partial \mu} = \sum_{i=1}^{n} \pi_i^k \frac{\partial \log \varphi(x_i - \mu)}{\partial \mu} = \left| \begin{array}{l} \log(x_i - \mu) = \\ = \log \frac{1}{\sqrt{2\pi}} - \frac{(x_i - \mu)^2}{2} \end{array} \right| =$$

$$= \sum_{i=1}^{n} \pi_i^k (x_i - \mu) = 0 \Rightarrow$$

$$\Rightarrow \hat{\mu} = \frac{\sum_{i=1}^{n} \pi_i^k x_i}{\sum_{i=1}^{n} \pi_i^k} =: \mu_{k+1}$$

$$\Rightarrow \theta_{k+1} = (\varepsilon_{k+1}, \mu_{k+1})$$

$\Rightarrow$ go back to the Expectation step.

(III) Let us show that EM Algorithm leads to the improvement of the likelihood function:

$$L(X/\theta_{k+1}) \geq L(X/\theta_k).$$

Γ $L(z, x/\theta) = L(z/x, \theta)\, L(x, \theta)$

$\log L(z, x/\theta) = \log L(z/x, \theta) + \log L(x, \theta)$ $\Big/ E_{z/x,\theta_k} \ldots$

does not depend on z

$\Rightarrow \log L(x, \theta) = \underbrace{E_{z/x,\theta_k} \log L(z, x/\theta)}_{Q(\theta/\theta_k)} -$

$- E_{z/x,\theta_k} \log L(z/x, \theta) =$

$= Q(\theta/\theta_k) + H(\theta/\theta_k),$

where $H(\theta/\theta_k) = - E_{z/\theta_k, x} \log L(z/x, \theta)$.

$\Rightarrow L(x/\theta_{k+1}) - L(x, \theta_k) = \underbrace{\left(Q(\theta_{k+1}/\theta_k) - Q(\theta_k/\theta_k)\right)}_{1} +$

$+ \underbrace{\left(H(\theta_{k+1}/\theta_k) - H(\theta_k/\theta_k)\right)}_{2}$

1: $Q(\theta_{k+1}/\theta_k) \geq Q(\theta_k/\theta_k)$, because

$\theta_{k+1} = \arg\max Q(\theta/\theta_k)$

2: $H(\theta_{k+1}/\theta_k) - H(\theta_k/\theta_k)$

Γ $H(\theta/\theta_k) = - E_{z/\theta_k, x} \log L(z/x, \theta) =$

$= - \int \log L(z/x, \theta)\, L(z/x, \theta_k)\, dz$

pdf of z

$\Rightarrow H(\theta_{k+1}/\theta_k) - H(\theta_k/\theta_k) = - \int \log L(z/x, \theta_{k+1})\, L(z/x, \theta_k)\, dz$

$+ \int \log L(z/x, \theta_k)\, L(z/x, \theta_k)\, dz =$

$= - \int \log \dfrac{L(z/x, \theta_{k+1})}{L(z/x, \theta_k)}\, L(z/x, \theta_k)\, dz.$

this function is the partial case of

Kullback - Leibler divergence (relative entropy)

$D_{KL}(P \| Q)$ shows how distribution Q

probability measures

is different from the distribution P. In the case when $dP = p\,dx$; $dQ = q\,dx$

$$D_{KL}(P \| Q) = -\int \log\left(\frac{q(x)}{p(x)}\right) p(x)\,dx.$$

Property: $D_{KL}(P \| Q) \geq 0$

$\ulcorner$ $D_{KL}(P \| Q) = -\int \log \frac{q}{p}\, p\,dx = \Big/ \begin{matrix} \log x \leq x-1, \\ \text{for } x > 0 \end{matrix} \Big/ \geq$

$\geq -\int \left(\frac{q}{p} - 1\right) p\,dx = -\int q\,dx + \int p\,dx = 0$ $\lrcorner$

$\Rightarrow H(\theta_{k+1} / \theta_k) - H(\theta_k / \theta_k) \geq 0$ $\lrcorner$ $\lrcorner$