

Expectation Maximization

Hossein Falaki
Computer Science Department,
University of California, Los Angeles
falaki@cs.ucla.edu

1 Introduction

This note is an introduction to the *Expectation Maximization* method for estimation. We will demonstrate this method through an example: fitting data to a mixture model.

We assume X_1, X_2, \dots, X_n are i.i.d. samples from a mixture model:

$$\lambda N(\mu_1, \sigma_1^2) + (1 - \lambda)N(\mu_0, \sigma_0^2) \quad (1)$$

We wish to find the unknowns: $\Theta = (\lambda, \mu_0, \sigma_0^2, \mu_1, \sigma_1^2)$.

2 Maximum Likelihood

Before starting with the EM algorithm, we need to know the *Maximum Likelihood* method. This method finds the most plausible Θ to explain the data with the given model. To do so, it maximizes the likelihood function:

$$likelihood(\Theta) = Pr(X_1, \dots, X_n | \Theta) \quad (2)$$

For our example the likelihood function is:

$$L(\Theta) = \prod_{i=1}^n (\lambda f_1(X_i | \mu_1, \sigma_1^2) + (1 - \lambda) f_0(X_i | \mu_0, \sigma_0^2))$$

where:

$$f_0(X | \mu_0, \sigma_0^2) = \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(X-\mu_0)^2}{2\sigma_0^2}}$$

and

$$f_1(X | \mu_1, \sigma_1^2) = \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(X-\mu_1)^2}{2\sigma_1^2}}$$

The logarithm of the likelihood function, log likelihood, is easier to deal with:

$$l(\Theta) = \log(L(\Theta)) = \sum_{i=1}^n (\lambda f_1(X_i|\mu_1\sigma_1^2) + (1-\lambda)f_0(X_i|\mu_0\sigma_0^2)) \quad (3)$$

By taking the derivative of $l(\Theta)$, we get five different equations that can be used to find the five values of Θ .

3 Expectation Maximization

EM is the most commonly used algorithm to find the maximum likelihood fit.

We can add hidden random variables, Z_i , to the data:

$$Z_1, Z_2, \dots, Z_n \sim \text{Bernoulli}(\lambda)$$

where

$$\begin{aligned} [X_i|Z_i = 1] &\sim f_1 \\ [X_i|Z_i = 0] &\sim f_0 \end{aligned}$$

With this *complete-data assumption* we can write the likelihood:

$$\begin{aligned} L_{\text{complete-data}}(\Theta) &= Pr((X_1, Z_1)(X_2, Z_2) \dots | \Theta) \\ &= \prod_{i=1}^n Pr(X_i, Z_i | \Theta) \\ &= \prod_{i=1}^n Pr(Z_i | \Theta) (Pr(X_i | Z_i, \Theta)) \\ &= \prod_{i=1}^n \lambda^{Z_i} (1-\lambda)^{(1-Z_i)} f_1(X_i|\mu_1\sigma_1^2)^{Z_i} f_0(X_i|\mu_0\sigma_0^2)^{(1-Z_i)} \end{aligned}$$

To maximize this function it is easier to consider the logarithm:

$$\begin{aligned} l_{\text{complete-data}}(\Theta) &= \log \lambda \sum_i Z_i + \log(1-\lambda) \sum_i (1-Z_i) \\ &\quad + \sum_i Z_i \log f_1(X_i|\mu_1, \sigma_1^2) + \sum_i (1-Z_i) \log f_0(X_i|\mu_0, \sigma_0^2) \end{aligned}$$

By taking the partial derivatives we can find the values for Θ :

$$\frac{\partial}{\partial \lambda} \implies \hat{\lambda} = \frac{\sum Z_i}{n}$$

similarly:

$$\begin{aligned}\hat{\mu}_1 &= \frac{\sum X_i Z_i}{\sum Z_i} \\ \hat{\sigma}_1^2 &= \frac{\sum (X_i - \hat{\mu}_1) Z_i}{\sum Z_i} \\ \hat{\mu}_0 &= \frac{\sum X_i (1 - Z_i)}{\sum (1 - Z_i)} \\ \hat{\sigma}_0^2 &= \frac{\sum (X_i - \hat{\mu}_1) (1 - Z_i)}{\sum (1 - Z_i)}\end{aligned}$$

This is the maximization step of EM. Now we need to impute Z_i or more precisely: $E_{[Z_i|X_i, \Theta]} l_{\text{complete-data}}(\Theta)$

$$\begin{aligned}E(Z_i|X_i, \Theta) &= 0 \times Pr(Z_i = 0|X_i, \Theta) + 1 \times Pr(Z_i = 1|X_i, \Theta) \\ &= \frac{\lambda f_1}{\lambda f_1 + (1 - \lambda) f_0}\end{aligned}$$

We continue this process and find a better Θ in each iteration.

4 Acknowledgment

This note is based on Prof. Ying Nian Wu's lectures on Theoretical Statistics at UCLA. Figures 1 and 2 are from the Internet.