AMS 597: Statistical Computing

Pei-Fen Kuan

Applied Math and Stats, Stony Brook University

EM Algorithm

- Expectation-maximization (EM) method is an iterative method for maximizing difficult likelihood problems.
- It was first introduced by Dempster et al. (J. Roy. Statist. Soc. 1997).
- Suppose we have a random sample $X_1, X_2, \ldots, X_n$ iid from $f(x_i|\theta)$, then the likelihood function $L(\theta) = \prod_{i=1}^n f(x_i|\theta)$
- A maximum likelihood estimate of θ is a value $\hat{\theta}$ that maximizes $L(\theta)$.

EM Algorithm

- In other words, we wish to find the maximum likelihood estimator

$$\hat{\theta} = \arg\ \max_\theta \prod_{i=1}^{n} f(x_i|\theta) = \arg\ \max_\theta \sum_{i=1}^{n} \log f(x_i|\theta)$$

- If θ is a scalar, the parameter space Θ is an open interval, and $L(\theta)$ is differentiable and assumes a maximum on Θ, then $\hat{\theta}$ is a solution of $\frac{\mathrm{d}L(\theta)}{\mathrm{d}\theta}$ or $\frac{\mathrm{d}\log L(\theta)}{\mathrm{d}\theta}$.
- However, sometimes maximizing $\log L(\theta)$ is difficult. We will look at one motivating example in the following page.

EM Algorithm

- Suppose $X_1, X_2, \ldots, X_n$ are iid from a mixture of two normal distribution, i.e., $f_X(x_i) = (1-p_1)N(x_i|\mu_0, \sigma_0^2) + p_1 N(x_i|\mu_1, \sigma_1^2)$.
- The the log likelihood is $l(\theta) = \log L(\theta) = \sum_{i=1}^n \log((1-p_1)N(x_i|\mu_0, \sigma_0^2) + p_1 N(x_i|\mu_1, \sigma_1^2))$. Here $\theta = (\mu_0, \sigma_0^2, \mu_1, \sigma_1^2, p_1)$
- Direct maximization of $l(\theta)$ is quite difficult numerically, because we have the sum of terms inside the logarithm function.

EM Algorithm

- However, if we pretend that we "know" for each X_i which normal component it is generated from, then the maximization problem is simplified. Let $Z_i = 0$ if X_i is generated from the first component $N(x_i|\mu_1, \sigma_1^2)$ and $Z_i = 1$ if X_i is generated from the second component $N(x_i|\mu_2, \sigma_2^2)$. That is, assuming that we have the "complete data" $(X_1, Z_1), (X_2, Z_2), \ldots, (X_n, Z_n)$. In this case, the "complete" likelihood function becomes

$$L(\theta) = \prod_{i=1}^{n}((1-p_1)N(x_i|\mu_0, \sigma_0^2))^{(1-z_i)}(p_1 N(x_i|\mu_1, \sigma_1^2))^{z_i}$$

$$l(\theta) = \sum_{i=1}^{n}[(1-z_i)\log N(x_i|\mu_0, \sigma_0^2) + z_i \log N(x_i|\mu_1, \sigma_1^2)]$$

$$+ \sum_{i=1}^{n}[(1-z_i)\log(1-p_1) + z_i \log p_1]$$

EM Algorithm

- Since the values of Z_i's are actually unknown, we will substitute it with its expected value $\tau_i(\theta) = E(Z_i|\theta, X) = p(Z_i = 1|\theta, X)$.
- Initialization: Take initial guesses for the parameters $\hat{\mu_0}, \hat{\sigma_0^2}, \hat{\mu_1}, \hat{\sigma_1^2}, \hat{p_1}$
- E-step: compute

$$\hat{\tau}_i(\theta) = \frac{(1-p_1)N(x_i|\mu_0,\sigma_0^2)}{(1-p_1)N(x_i|\mu_0,\sigma_0^2) + p_1 N(x_i|\mu_1,\sigma_1^2)}, i = 1, 2, \ldots, n$$

EM Algorithm

- M-step: Get updated estimates $\hat{\mu_0}, \hat{\sigma_0^2}, \hat{\mu_1}, \hat{\sigma_1^2}, \hat{p_1}$:

$$\hat{\mu_0} = \frac{\sum_{i=1}^n (1 - \hat{\tau}_i(\theta)) x_i}{\sum_{i=1}^n (1 - \hat{\tau}_i(\theta))}$$

$$\hat{\mu_1} = \frac{\sum_{i=1}^n \hat{\tau}_i(\theta) x_i}{\sum_{i=1}^n \hat{\tau}_i(\theta)}$$

$$\hat{\sigma_0^2} = \frac{\sum_{i=1}^n (1 - \hat{\tau}_i(\theta))(x_i - \hat{\mu_0})^2}{\sum_{i=1}^n (1 - \hat{\tau}_i(\theta))}$$

$$\hat{\sigma_1^2} = \frac{\sum_{i=1}^n \hat{\tau}_i(\theta)(x_i - \hat{\mu_1})^2}{\sum_{i=1}^n \hat{\tau}_i(\theta)}$$

$$\hat{p_1} = \sum_{i=1}^n \hat{\tau}_i(\theta)/n$$

- Iterate E-step and M-step until convergence

Hidden Markov Model (HMM)

- In the mixture of normal problem Z_i's are also known as the latent/hidden states.
- Suppose you observed $(X_1, X_2, \ldots, X_n)$ and assuming there is also a vector of latent/hidden states $(Z_1, Z_2, \ldots, Z_n)$.
- In HMM, the "Markov" is to model the relationship between Z_i's. Specifically, in first order HMM, we assume Z_{i+1} depends on Z_i and that $P(Z_{i+1}|Z_i)$ is the transition probability.
- Let's take an example, suppose there are 2 hidden states Z_i is the health status, where $Z_i =$ "sick" or "healthy" for day i, and X_i's are the temperature read from thermometer for day i. You are interested to figure out the health status of a person for day 1 to 100, given that you are only told the temperature of this person for each day.

Hidden Markov Model (HMM)

- A transition probability matrix for this problem will take the form:

$$\begin{array}{cc} & \begin{array}{cc} \text{Healthy} & \text{Sick} \end{array} \\ \begin{array}{c} \text{Healthy} \\ \text{Sick} \end{array} & \begin{pmatrix} a_{HH} & a_{HS} \\ a_{SH} & a_{SS} \end{pmatrix} \end{array}$$

  where $a_{HH} = P(Z_{i+1} = \text{Healthy}|Z_i = \text{Healthy})$ and $a_{HS} = P(Z_{i+1} = \text{Sick}|Z_i = \text{Healthy})$. Thus we have $a_{HH} + a_{HS} = 1$ and $a_{SH} + a_{SS} = 1$.
- We then assume that $P(X_i|Z_i) = N(x_i|\mu_{Z_i}, \sigma^2_{Z_i})$. Thus, you can derive the "complete" likelihood function.

Hidden Markov Model (HMM)

- In the E-step: Compute $P(Z_{i+1}, Z_i|X, \theta)$ using forward backward algorithm (also known as Baum-Welch algorithm, a dynamic programming method)
- In the M-step: Based on $P(Z_{i+1}, Z_i|X, \theta)$ from E-step, find updated estimates of the transition probabilities a_{ij}'s, μ_{Z_i}, $\sigma^2_{Z_i}$) as well as the initial state distribution (i.e., the distribution of Z_i's of day 1 to start the Markov chain).
- Iterate E-step and M-step until convergence.