STA6171: Statistical Computing for DS 1

EM Algorithm

Ick Hoon Jin

Yonsei University, Department of Statistics and Data Science

2020.10.07

1. Introduction
2. EM Algorithm
3. EM Variants

Introduction

- Develop for handling missing outcomes.
- Use a lot in optimization

- The expectation-maximization (EM) algorithm is an iterative optimization strategy motivated by a notion of missingness and by consideration of the conditional distribution of what is missing given what is observed.
- Popularity of the EM algorithm
  - Simple to implement
  - Reliable to find the global optimum.

Introduction

Observed Data : X
Missing Data : Z) Complete Data : $Y=(X,Z)$

- Frequentist Setting
  - Observed data from X along with missing data from Z.
  - Complete data $Y = (X, Z)$. We want to maximize $L(\theta|X)$
  - Given observed data x, maximize a likelihood $L(\theta|x)$. Difficult to work with this likelihood

    Difficult to work with this likelihood
    (Not able to apply Netwon method)

    Use EM algorithm
  - A easier way is working with the density $Y|\theta$ and $Z|x, \theta$.

    E $f(Y|\theta)$ $f(Z|X,\theta)$ → maximize $L(\theta|X)$
- Bayesian Setting: Interest Often focuses on estimating the mode of a posterior distribution $f(\theta|x)$.
- Missing data may not truly be missing: they may be only a conceptual ploy that simplifies the problem. In this case, Z is often referred to as *latent*.

  Sometimes we want to maximize $L(\theta|X)$

  Z : latent variable $Z|X,\theta$

  $(Y|\theta)$

Missing Data and Marginalization - Frequentist

observed

$Y = (X, Z)$

missing

- In the presence of missing data, only a function of the complete-data y is observed.

$f(y|\theta) = f(x,z|\theta) = f_x(x|\theta)\, f_{z|x}(z|x,\theta)$

$l_y(\theta) = l_x(\theta) + \log f_{z|x}(z|x,\theta)$

- Assume that the missing data are random, so that

$$f(y|\theta) = f(x,z|\theta) = f_X(x|\theta) \cdot f_{Z|X}(z|x,\theta).$$

Thus, it follows that

$l_x(\theta) = \boxed{l_y(\theta) - \log f_{z|x}(z|x,\theta)}$

maximization diff.

↳ maximization is straightforward.

$$l_X(\theta) = l_Y(\theta) - \log f_{Z|X}(Z|X,\theta).$$

- Useful when maximizing $l_X(\theta)$ can be difficult but maximizing the complete log-likelihood l is simple.

Missing Data and Marginalization - Bayesian

Complete data likelihood $L(\theta|y) = L(\theta|x,z)$

$L(\theta|x)$: marginalization of $L(\theta|y)$

- View the likelihood $L(\theta|x)$ as a marginalization of the complete-data likelihood $L(\theta|y) = L(\theta|x, z)$.
- Consider there to be ~~missing parameter~~ ψ, whose inclusion simplifies Bayesian calculations even though ψ is of no interest itself. Since Z and ψ are both missing random quantities, it matters little whether we use notation that suggests the missing variables to be unobserved data or parameters.

nuisance parameter (latent)

EM Algorithm

- EM algorithm iteratively seeks to maximize $L(\theta|x)$ with respect to θ.

estimate.

- Let $\boxed{\theta^{(t)}}$ denote the estimated maximizer at iteration t, for $t = 0, 1, \cdots$.
- Define $Q(\theta|\theta^{(t)})$ to be the expectation of the joint log-likelihood for the complete data, conditional on the observed data $X = x$.

$$Q\left(\theta|\theta^{(t)}\right) = E\left\{\log L(\theta|Y)\Big|x, \theta^{(t)}\right\}.$$

$Q(\theta|\theta^{(t)}) = E\{\log L(\theta|Y)|x,\theta^{(t)}\}$

- Then, $Q\left(\theta|\theta^{(t)}\right)$ is maximized w.r.t θ, that is $\theta^{(t+1)}$ is found such that

$$Q\left(\theta^{(t+1)}|\theta^{(t)}\right) \geq Q\left(\theta|\theta^{(t)}\right)$$

for all $\theta \in \Theta$.

Example: Simple Exponential Distribution

$f(y|\theta) = \prod_{i=1}^{2} \theta e^{-\theta y_i} = \theta^2 e^{-\theta \sum_{i=1}^{2} y_i}$

- Suppose $Y_1, Y_2 \overset{iid}{\sim} exp(\theta)$ and $y_1 = 5$ is observed but the value y_2 is missing.

$Y = (\boxed{Y_1}, Y_2)$ — Y_1 → observed, Y_2 → missing.

$l_y(\theta) = 2\log\theta - \theta y_1 - \theta \boxed{y_2}$ → missing

- The complete-data log likelihood function is

$$\log L(\theta|y) = 2\log\theta - \theta y_1 - \theta y_2.$$

make $\boxed{y_2}$ using conditional expectation

$E(Y_2|y_1,\theta^{(t)}) = E(Y_2|\theta^{(t)}) = 1/\theta^{(t)}$

- Because

$$E(Y_2|y_1,\theta^{(t)}) = E(Y_2|\theta^{(t)}) = \frac{1}{\theta^{(t)}},$$

the conditional expectation of $\log L(\theta|Y)$ yields

$Q(\theta|\theta^{(t)}) = E[l_y(\theta)|y_1,\theta^{(t)}] = E[2\log\theta - \theta y_1 - \theta y_2 | y_1, \theta^{(t)}]$

$$Q(\theta|\theta^{(t)}) = 2\log\theta - 5\theta - \theta/\theta^{(t)}.$$

$= 2\log\theta - 5\theta - \theta E[y_2|y_1,\theta^{(t)}] = 2\log\theta - 5\theta - \theta/\theta^{(t)}$

- The maximizer of $Q(\theta|\theta^{(t)})$ with respect to θ is easily to found to be the root of $2/\theta - 5 - 1/\theta^{(t)} = 0$.

$\frac{2}{\theta} = \frac{5\theta^{(t)}+1}{\theta^{(t)}}$ $\quad \theta^{(t+1)} = \frac{2\theta^{(t)}}{5\theta^{(t)}+1}$