STA 5106
Computational Methods in Statistics I

Department of Statistics
Florida State University

Class 11
October 1, 2019

About Homework

- independently finish your own (but discussion is allowed).
- include all Matlab (and Python) code
- label each figure
- provide detailed procedure, not just the final result

About Python

- This class focuses on computational methods and we illustrate their practical use with Matlab. Python is optional.
- I give Python example code, which provides most information for the homework.
- Students who want to program well should learn more by themselves.
- Tons of tutorials can be found online: e.g.
  - https://docs.python.org/3/tutorial/index.html
  - https://www.youtube.com/watch?v=HBxCHonP6Ro&list=PL6gx4Cwl9DGAcbMi1sH6oAMk4JHw91mC_

3.7 Expectation Maximization (EM) Algorithm

Likelihood Maximization

- So far we have looked at maximizing likelihood for parameter θ given an observation x (= $\{x_1, \ldots, x_n\}$) and a likelihood function $f(x|\theta)$.

- In such cases it is assumed that the maximizer exists and the techniques described earlier can solve of it.

- However, in some problems the full observation may not be provided and that makes the maximization a difficult problem.

- Let the data vector x be made up of two components $x = \{x_o, x_m\}$ where x_o denotes the observed part of x and x_m stands for the missing part.

Likelihood Maximization

- Assume that given the full data x, it is possible to solve for the maximum likelihood estimate of θ but goal now is to solve for the maximizer:

$$\hat{\theta} = \arg\max_{\theta} f(x_o \mid \theta)$$

- Expectation Maximization (EM) algorithm is beneficial mostly where $f(x|\theta)$ is easier to maximize compared to maximizing $f(x_o|\theta)$.

Example

- **Example 4 (Mixture of Gaussians)**

  Let Y be a real-valued random variable such that $Y = Y^{(j)}$ with probability α_j, where

  $$Y^{(j)} \sim N(\mu_j, \sigma_j^2) \text{ and } \sum_j \alpha_j = 1.$$

  The goal is to use independent observations of Y, say $Y_1, Y_2, \ldots, Y_n$, to estimate the parameters α_j, μ_j, σ_j^2, for all j.

  To simplify this discussion let the number of densities in the mixture be two, and Y becomes a mixture of two Gaussian random variables.

  We seek a maximum likelihood estimate of the unknowns

  $$\theta = (\mu_1, \sigma_1^2, \alpha_1, \mu_2, \sigma_2^2, \alpha_2)$$

Example

- The likelihood function is given by:

$$f(Y|\theta) = \prod_{i=1}^{n} f(Y_i|\theta) = \prod_{i=1}^{n} [\alpha_1 f_1(Y_i|\mu_1, \sigma_1^2) + \alpha_2 f_2(Y_i|\mu_2, \sigma_2^2)]$$

  where f_1 and f_2 are the two normal density functions with appropriate parameters.

- The log-likelihood function is given by:

$$\log f(Y|\theta) = \sum_{i=1}^{n} \log[\alpha_1 f_1(Y_i|\mu_1, \sigma_1^2) + \alpha_2 f_2(Y_i|\mu_2, \sigma_2^2)]$$

- Solving for $\hat{\theta} = \arg\max_{\theta} \log f(Y|\theta)$ is difficult because of the summation inside the log function.

Example

- Consider a different situation: in addition to Y_i' s one also observes a label l_i that equals one if $Y_i \sim f_1$ and two if $Y_i \sim f_2$.

- In other words, l_i tells us what density Y_i came from. Form a larger observation (Y, l), where $l = (l_1, l_2, \ldots, l_n)$, and derive the log-likelihood function:

$$\log f(Y, l \mid \theta) = \sum_{i=1}^{n} \log f(Y_i, l_i \mid \theta)$$

$$= \sum_{i=1}^{n} \log[f(Y_i \mid l_i, \theta) P(l_i \mid \theta)]$$

$$= \sum_{i;l_i=1} \log[\alpha_1 f_1(Y_i \mid \mu_1, \sigma_1^2)] + \sum_{i;l_i=2} \log[\alpha_2 f_2(Y_i \mid \mu_2, \sigma_2^2)]$$

Example

- Now, the data breaks into two groups: one from f_1 and the other from f_2.

- This example illustrates a situation where the log-likelihood function for the observed data $\log(f(Y|\theta))$ is rather difficult to maximize, while the same function for a complete data, assuming additional data in form of the labels l, is much easier to maximize.

- The EM algorithm is applied in such situations where the additional data, called the missing data, is not available but could have greatly simplified the optimization problem.

Derivation of EM Algorithm

- Our goal is to construct a sequence $\{\theta_k\}$ such that: (i) $f(x_o|\theta_{k+1}) \geq f(x_o|\theta_k)$, and (ii) with some additional conditions this sequence converges to the estimator $\hat{\theta}$.

- Rearrange the equation: $f(x|\theta) = f(x_o|\theta)\, f(x_m|x_o, \theta)$ to write:

$$f(x_o \mid \theta) = \frac{f(x \mid \theta)}{f(x_m \mid x_o, \theta)}$$

- Taking log on both sides, we get

$$\log f(x_o \mid \theta) = \log f(x \mid \theta) - \log f(x_m \mid x_o, \theta)$$

- Next, take expectation on both sides with respect to the density function $f(x_m|x_o, \theta_k)$, for some θ_k.

Derivation of EM Algorithm

- Given x_o, the left hand side is a constant and remains same.

- Then
$$\log f(x_o \mid \theta)$$
$$= E[\log f(x \mid \theta) \mid \theta_k, x_o] - E[\log f(x_m \mid \theta, x_o) \mid \theta_k, x_o]$$
$$= Q(\theta \mid \theta_k, x_o) - H(\theta \mid \theta_k, x_o)$$
where Q and H are defined by above equations.

- Considering the second term first, we focus on the difference:
$$H(\theta \mid \theta_k, x_o) - H(\theta_k \mid \theta_k, x_o)$$
$$= E[\log f(x_m \mid \theta, x_o) \mid \theta_k, x_o] - E[\log f(x_m \mid \theta_k, x_o) \mid \theta_k, x_o]$$
$$= E[\log \frac{f(x_m \mid \theta, x_o)}{f(x_m \mid \theta_k, x_o)} \mid \theta_k, x_o]$$

Derivation of EM Algorithm

$$\leq \log E[\frac{f(x_m | \theta, x_o)}{f(x_m | \theta_k, x_o)} | \theta_k, x_o]$$

$$= \log[\int \frac{f(x_m | \theta, x_o)}{f(x_m | \theta_k, x_o)} f(x_m | \theta_k, x_o) dx_m] = 0$$

- The inequality comes from the Jensen's inequality which says that $E[\log(Y)] \leq \log(E[Y])$ since log is a concave function.
- This implies that $H(\theta|\theta_k, x_o) \leq H(\theta_k|\theta_k, x_o)$ for any θ.

Derivation of EM Algorithm

- Therefore,

$$\log f(x_o \mid \theta_{k+1}) - \log f(x_o \mid \theta_k)$$
$$= [Q(\theta_{k+1} \mid \theta_k, x_o) - H(\theta_{k+1} \mid \theta_k, x_o)] - [Q(\theta_k \mid \theta_k, x_o) - H(\theta_k \mid \theta_k, x_o)]$$
$$= [Q(\theta_{k+1} \mid \theta_k, x_o) - Q(\theta_k \mid \theta_k, x_o)] - [H(\theta_{k+1} \mid \theta_k, x_o) - H(\theta_k \mid \theta_k, x_o)]$$

- We set

$$\theta_{k+1} = \arg\max_{\theta} Q(\theta \mid \theta_k, x_o)$$

- Then, the first term by definition is non-negative, and we have already shown that the second term is non-positive. Together, these two conditions imply that

$$\log f(x_o \mid \theta_{k+1}) \geq \log f(x_o \mid \theta_k)$$

Algorithm

- **Algorithm 30 (EM Algorithm)**

Choose an initial value for θ_0 and set $k = 0$.

1. **Expectation Step:** Compute

$$Q(\theta \mid \theta_k, x_o) = E[\log f(x_o, x_m \mid \theta) \mid \theta_k, x_o].$$

2. **Maximization Step:** Set

$$\theta_{k+1} = \arg\max_{\theta} Q(\theta \mid \theta_k, x_o).$$

3. Check convergence. If not converged, set $k = k + 1$ and go to Step 1.