Machine Learning

Expectation Maximization

We've seen the update eqs. of GMM, but

- How are they derived?
- What is the general algorithm of EM?

General Settings for EM

- Given data $X = (x_1, \ldots, x_n)$, where $x_i \sim p(x; \theta)$.
- Want to maximize log-likelihood $\log p(X; \theta)$.
- $p(X; \theta)$ is difficult to maximize because it involves some latent variables Z.
- But maximizing the complete data log-likelihood $\log p(X, Z; \theta)$ would be easy (if we observed Z). We think of Z as missing data.
- In this scenario, EM gives an efficient method to maximize likelihood $p(X; \theta)$ to some local maximum.

Basic Idea of EM

- Since maximizing data log-likelihood $\log p(X;\theta)$ is hard but maximizing complete data log-likelihood $\log p(X,Z;\theta)$ is easy (if we observed Z), out bet is to maximize the latter and hopefully it also increases the former.

- (E step) Since we didn't observe Z, we cannot maximize $\log p(X,Z;\theta)$ directly. We will consider its expected value under the posterior dist. of Z, using old parameter.

- (M step) We then update parameter θ to maximize the expected complete data log-likelihood.

EM Algorithm in General

- 1. Initialize parameter θ^{old}.
- 2. E step: evaluate posterior dist. of latent variables $p(Z|X;\theta^{\text{old}})$, using old parameter. Then the expected complete data log-likelihood, under this dist. would be

$$Q(\theta;\theta^{\text{old}}) = \sum_{Z} p(Z|X;\theta^{\text{old}}) \log p(X,Z;\theta)$$

A function of θ

Taking expectation

Complete data log-likelihood

- 3. M step: update parameters to maximize the expected complete data log-likelihood.

$$\theta^{\text{new}} = \arg\max_{\theta} Q(\theta;\theta^{\text{old}})$$

- 4. Check convergence criterion. Return to step 2 if not satisfied.

GMM Revisited – E step

- Evaluate posterior dist. of latent variables $p\left(Z|X;\theta^{\text{old}}\right)$, using old parameters.
  - Since data are i.i.d., we evaluate for each i.

$$q_i^{(j)} \equiv p(z_i = j|x_i) = \frac{p(z_i = j, x_i)}{p(x_i)}$$

$$= \frac{p(x_i|z_i = j)p(z_i = j)}{\sum_{l=1}^{K} p(x_i|z_i = l)p(z_i = l)}$$

$p(x_i|z_i = j)$ ← $\frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{(x_i-\mu_j)^2}{2\sigma^2{}_j}}$

$p(z_i = j)$ ← w_j

GMM Revisited – E step

- Then the expected complete data log-likelihood, under this dist. would be

$$Q(\theta;\theta^{\text{old}}) = \sum_{i=1}^{N}\sum_{j=1}^{K} p(z_i = j|x_i)\log p(x_i, z_i;\theta^{\text{old}})$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{K} q_i{}^{(j)}\log p(x_i, z_i;\theta^{\text{old}})$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{K} q_i{}^{(j)}\log\{p(z_i = j)p(x_i|z_i = j)\}$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{K} q_i{}^{(j)}\log\left\{w_j \cdot \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{(x_i-\mu_j)^2}{2\sigma^2{}_j}}\right\}$$

GMM Revisited – M step (for μ_j)

- Update parameters to maximize the expected complete data log-likelihood.

$$Q(\theta;\theta^{\text{old}}) = \sum_{i=1}^{N}\sum_{j=1}^{K} {q_i}^{(j)} \log\left\{ w_j \cdot \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{(x_i-\mu_j)^2}{2{\sigma^2}_j}} \right\}$$

- For μ_j, set derivative to 0.

$$\frac{\partial Q(\theta;\theta^{\text{old}})}{\partial \mu_j} = \sum_{i=1}^{N} {q_i}^{(j)} \frac{x_i - \mu_j}{{\sigma^2}_j} = 0$$

- We get update equation for means:

$$\mu_j = \frac{\sum_{i=1}^{N} {q_i}^{(j)} x_i}{\sum_{i=1}^{N} {q_i}^{(j)}}.$$

GMM Revisited – M step (for σ_j)

- The derivation is similar to the derivation for μ_j
- Try it yourself…

- We get update equation for variances:

$$\sigma^2{}_j = \frac{\sum_{i=1}^{N} q_i{}^{(j)} \left(x_i - \mu_j\right)^2}{\sum_{i=1}^{N} q_i{}^{(j)}}$$

GMM Revisited – M step (for w_j)

- For w_j, recall there is a constraint $\sum_{j=1}^{K} w_j = 1$.
- To maximize $Q\left(\theta; \theta^{\text{old}}\right)$ w.r.t. w_j, we construct the Lagrangian

$$Q'\left(\theta; \theta^{\text{old}}\right) = Q\left(\theta; \theta^{\text{old}}\right) + \beta\left(\sum_{j=1}^{K} w_j - 1\right)$$

- Set derivative to 0

$$\frac{\partial Q'\left(\theta; \theta^{\text{old}}\right)}{\partial w_j} = \sum_{i=1}^{N} \frac{q_i{}^{(j)}}{w_j} + \beta = 0$$

GMM Revisited – M step (for w_j)

- We get $w_j = \frac{\sum_{i=1}^{N} {q_i}^{(j)}}{-\beta}$
- Sum over j, and using $\sum_{j=1}^{K} w_j = 1$, we get

$$-\beta = \sum_{j=1}^{K} \sum_{i=1}^{N} {q_i}^{(j)}$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{K} {q_i}^{(j)} = N$$

- So we get update equation for weights:

$$w_j = \frac{1}{N} \sum_{i=1}^{N} {q_i}^{(j)}$$

More Theoretical Questions

- We've seen how the update equations of GMM are derived from the EM algorithm.

- But… do these equations really work?
  - Will the data likelihood be maximized (at least to some local maximum)?
  - Will the algorithm converge?

Answers

- We will show that the data log-likelihood never decrease in each iteration.
- We also know that log-likelihood (which is log of probability) is bounded above by 0.
- Therefore, EM algorithm always converges!

Expected data log-likelihood increases

- Recall that in M step, we maximize the expected complete log-likelihood

$$\theta^{\text{new}} = \arg\max_{\theta} Q\left(\theta; \theta^{\text{old}}\right)$$

$$= \arg\max_{\theta} \sum_{Z} p\left(Z|X; \theta^{\text{old}}\right) \log p(X, Z; \theta)$$

- Therefore

$$\sum_{Z} p\left(Z|X; \theta^{\text{old}}\right) \log p(X, Z; \theta^{\text{new}})$$

$$\geq \sum_{Z} p\left(Z|X; \theta^{\text{old}}\right) \log p\left(X, Z; \theta^{\text{old}}\right)$$

Therefore…

- From previous slide

$$\sum_Z p\left(Z|X;\theta^{\text{old}}\right)\log p(X,Z;\theta^{\text{new}}) \geq \sum_Z p\left(Z|X;\theta^{\text{old}}\right)\log p\left(X,Z;\theta^{\text{old}}\right)$$

- So we also have

$$\sum_Z p\left(Z|X;\theta^{\text{old}}\right)\log\frac{p(X,Z;\theta^{\text{new}})}{p(Z|X;\theta^{\text{old}})}$$

$$\geq \sum_Z p\left(Z|X;\theta^{\text{old}}\right)\log\frac{p\left(X,Z;\theta^{\text{old}}\right)}{p(Z|X;\theta^{\text{old}})} \quad (1)$$

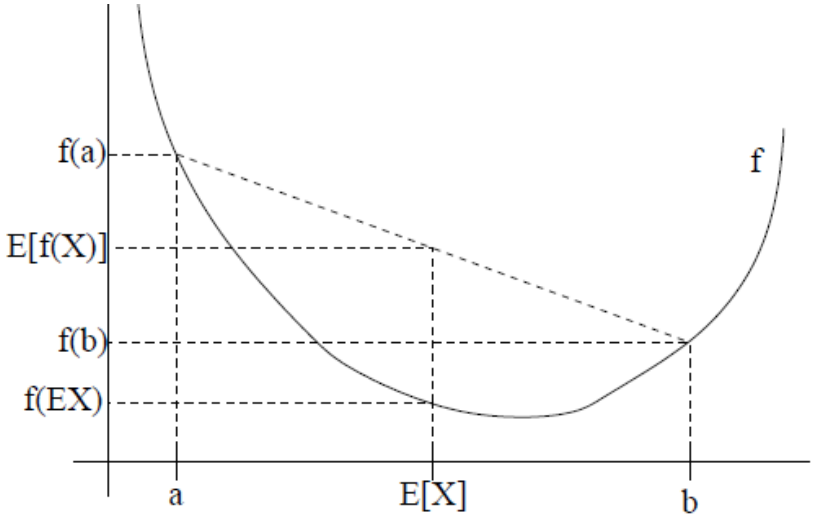
$$= \sum_Z p\left(Z|X;\theta^{\text{old}}\right)\log p\left(X;\theta^{\text{old}}\right) = \log p\left(X;\theta^{\text{old}}\right)$$

Data log-likelihood using old parameter

Jensen's Inequality

- Let f be a convex function, X be a random variable, then

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}X)$$



- Since $\log()$ is a concave function, we have

Expectation — Concave function — Random variable

$$\sum_Z p(Z|X;\theta^{\text{old}}) \log \frac{p(X,Z;\theta^{\text{new}})}{p(Z|X;\theta^{\text{old}})}$$

$$\leq \log \sum_Z p(Z|X;\theta^{\text{old}}) \frac{p(X,Z;\theta^{\text{new}})}{p(X,Z;\theta^{\text{old}})}$$

Continue…

$$\sum_Z p\left(Z|X;\theta^{\text{old}}\right)\log\frac{p(X,Z;\theta^{\text{new}})}{p(Z|X;\theta^{\text{old}})}$$

$$\leq \log\sum_Z p\left(Z|X;\theta^{\text{old}}\right)\frac{p(X,Z;\theta^{\text{new}})}{p(X,Z;\theta^{\text{old}})} \qquad (2)$$

$$= \log\sum_Z p(X,Z;\theta^{\text{new}}) = \boxed{\log p(X;\theta^{\text{new}})}$$

Data log-likelihood using new parameter

Finally…

- Putting (1) and (2) together, we get

$$\log p\left(X;\theta^{\text{old}}\right)$$
$$\leq \sum_{Z} p\left(Z|X;\theta^{\text{old}}\right)\log\frac{p(X,Z;\theta^{\text{new}})}{p(Z|X;\theta^{\text{old}})}$$
$$\leq \log p(X;\theta^{\text{new}})$$

- Data log-likelihood monotonically increases!

You Should Know

- EM algorithm is an efficient way to do maximum likelihood estimation, when there are latent variables or missing data.
- The general algorithm of EM
  - E step: calculate posterior dist. of latent variables.
  - M step: update parameters by maximizing the expected complete data log-likelihood.
- How to derive EM update equations of GMM?
  - Can you derive EM update equations for parameter estimation of a mixture of categorical distributions?
- Why does EM always converge (to some local optimum)?