Intuition of expectation maximization (EM)

EM can be understood intuitively by tossing different types of biased coins, as shown in link below. http://labs.seas.wustl.edu/bme/raman/Lectures/Lecture16_Clustering.pdf (http://labs.seas.wustl.edu/bme/raman/Lectures/Lecture16_Clustering.pdf). Another intuitive picture is from k-means, as it is a special case of EM. EM soft-assigns data to a class, while k-means hard assigns data. However, the main idea is same.

EM is essentially a maximum likelihood estimation (MLE) approach. While typical MLE method is a single-step iterative algorithm for finding optimal parameters, EM is a two-step iterative algorithm due to the existence of latent variables. When applying MLE, the optimized parameters are usually in the form of sample data average. This is similarly done in the M-step of EM algorithm. In the two intuitive pictures of either tossing coins or K-means, we also have the procedure of summing and taking average.

Gaussian Discriminant Analysis (GDA) and Mixtures of Gaussians

http://cs229.stanford.edu/notes/cs229-notes7b.pdf (http://cs229.stanford.edu/notes/cs229-notes7b.pdf)

The log likelihood of mixtures of Gaussians can be written as

$$l(\phi, \mu, \Sigma) = \sum_{i=1}^{m} \log p(x^{(i)}; \phi, \mu, \Sigma) = \sum_{i=1}^{m} \log \sum_{z^{(i)}=1}^{k} p(x^{(i)} \mid z^{(i)}; \mu, \Sigma) p(z^{(i)}; \phi)$$

If go move forward with usual maximization way and set to zero the derivatives of the log likelihood, then it is impossible to find the parameters in closed form. The key is that the random variable $z^{(i)}$ (indicating which of the k Gaussians each $x^{(i)}$ had come from) is not known.

If we know $z^{(i)}$, then the log likelihood can be written as

$$l(\phi, \mu, \Sigma) = \sum_{i=1}^{m} \log p(x^{(i)} \mid z^{(i)}; \mu, \Sigma) + \log p(z^{(i)}; \phi)$$

Maximizing the parameters with the usual way we can obtain

$$\phi_j = \frac{1}{m}\sum_{i=1}^{m} 1\{z^{(i)} = j\}$$

$$\mu_j = \frac{\sum_{i=1}^{m} 1\{z^{(i)} = j\}x^{(i)}}{\sum_{i=1}^{m} 1\{z^{(i)} = j\}}$$

$$\Sigma_j = \frac{\sum_{i=1}^{m} 1\{z^{(i)} = j\}(x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i=1}^{m} 1\{z^{(i)} = j\}}$$

In fact, if $z^{(i)}$ are known, then the above approach is very similar to GDA except that we let the variance matrix of each Gaussian varies and that we may have more than two classes.

Then what happens if we go back to the original problem where $z^{(i)}$ are unknown?

Handling of Gaussian mixtures with EM

When $z^{(i)}$ are unknown, EM will handle the problem with following two steps:

E-step: For each i, j, set

$$w_j^{(i)} = p(z^{(i)} = j \mid x^{(i)}; \phi, \mu, \Sigma) = \frac{p(x^{(i)} \mid z^{(i)} = j; \mu, \Sigma)p(z^{(i)} = j; \phi)}{\sum_{l=1}^{k} p(x^{(i)} \mid z^{(i)} = l; \mu, \Sigma)p(z^{(i)} = l; \phi)}$$

M-step: Update the parameters

$$\phi_j = \frac{1}{m}\sum_{i=1}^{m} w_j^{(i)}$$

$$\mu_j = \frac{\sum_{i=1}^{m} w_j^{(i)} x^{(i)}}{\sum_{i=1}^{m} w_j^{(i)}}$$

$$\Sigma_j = \frac{\sum_{i=1}^{m} w_j^{(i)}(x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i=1}^{m} w_j^{(i)}}$$

- The fundamental point of EM is that if we directly maximize the log likelihood it would be intractable. However, with E and M steps, then each step is tractable as we usually with closed-form expressions, as shown in example above.
- In the E step, we calculate posterior probability $w_j^{(i)}$ using Bayes formula with the initial guess or calculated parameters. This is similar to the $p(y^i = 1|x)$ in GDA except there y is not latent.

- In M step, we update the parameters with the guess or calculated $w_j^{(i)}$. The updating equations are from the usual maximization process by setting derivatives w.r.t. parameters to zero, and hence is called maximization step.
- After updating the parameters in M step, we plug these parameters in the $w_j^{(i)}$ of E-step to obtain a newly updated $w_j^{(i)}$. Then we use this new $w_j^{(i)}$ to further update the parameters in M-step. Repeat this process until convergence.

A side comments:
Both K-means and the more general EM approach can be straightforwardly understood as coordinate ascent. Define the log likelihood as $J(Q, \Theta)$. Then the E-step is maximizing w.r.t. Q and M-step is maximizing w.r.t. Θ.

Derivation of EM

The equations in the E and M steps for mixtures of Gaussians in the previous section can be rigorously derived from the general theory or EM. Below are some key points in the derivation of EM. http://cs229.stanford.edu/notes/cs229-notes8.pdf (http://cs229.stanford.edu/notes/cs229-notes8.pdf).

For a training set $\{x_{(1)}, \dots x_{(m)}\}$, we wish to fit the parameters of a model $p(x, z)$ to the data, where the likelihood is given by

$$l(\theta) = \sum_{i=1}^{m} \mathrm{log} p(x; \theta) = \sum_{i=1}^{m} \log \sum_{z} p(x, z; \theta)$$

For each i, let Q_i be some distribution over the z's, i.e., $\sum_z Q_i(z) = 1, Q_i(z) \geq 0$.

Consider the following (assume z is discrete random variable)

$$\begin{aligned} \sum_{i=1}^{m} \mathrm{log} p(x^{(i)}; \theta) &= \sum_{i=1}^{m} \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)}; \theta) \\ &= \sum_{i=1}^{m} \log \sum_{z^{(i)}} Q_i(z^{(i)}) \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)}} \\ &\geq \sum_{i=1}^{m} \sum_{z^{(i)}} Q_i(z^{(i)}) \mathrm{log} \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)}} \qquad (1) \end{aligned}$$

where the Jensen's inequality is used in the last step.

For any particular value of θ, we want make the lower bound function tight. That is, we will make the inequality above hold with equality at this θ. On the graph, this indicates the lower bound function 'connected' to the general log likelihood function at this θ value.

To make the bound tight for a particular value of θ, we need for the step involving Jensen's inequality in the derivation to hold with equality. It turns out we can achieve this goal by setting $Q_i(z^{(i)}) = p(z^{(i)} \mid x^{(i)}; \theta)$**. In other words, $Q_i(z^{(i)})$ is set to be the posterior distribution of the $z^{(i)}$ given $x^{(i)}$ and the parameter θ**

For the above choice of Q_i, Eq. (1) gives a lower-bound on the likelihood $l(\theta)$. This completes our E-step of EM algorithm. In the M-step, we plug the Q_i to RHS of Eq. (1) and try to obtain a new θ through a normal optimization process.

To summarize, the EM algorithm include the following two steps:

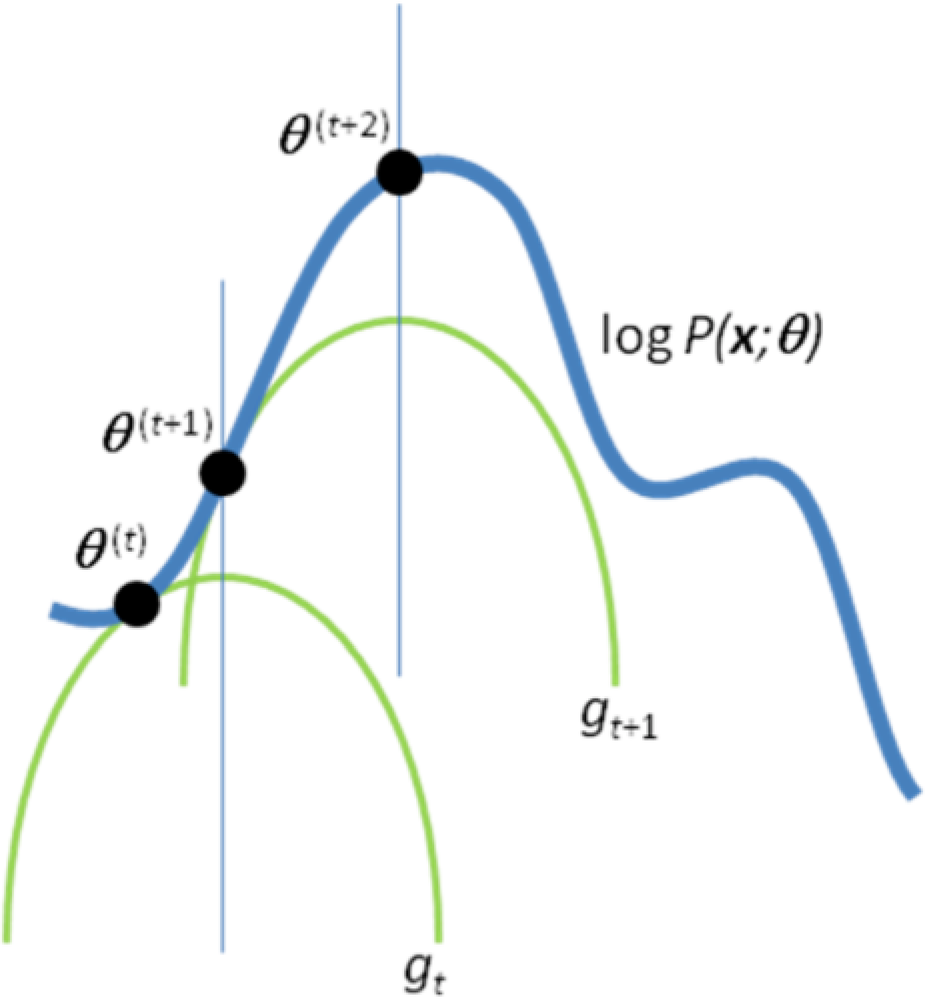
E-step

$$Q_i(z^{(i)}) = p(z^{(i)} \mid x^{(i)}; \theta)$$

M-step

$$\theta = \arg\max_{\theta} \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)}}$$

The following figure summarize the key steps of maximizing log likelihood by EM. [https://people.duke.edu/~ccc14/sta-663/EMAlgorithm.html (https://people.duke.edu/~ccc14/sta-663/EMAlgorithm.html)](https://people.duke.edu/~ccc14/sta-663/EMAlgorithm.html)



Supplementary Figure 1 Convergence of the EM algorithm. Starting from initial parameters $\theta^{(t)}$, the E-step of the EM algorithm constructs a function g_t that lower-bounds the objective function $\log P(x;\theta)$. In the M-step, $\theta^{(t+1)}$ is computed as the maximum of g_t. In the next E-step, a new lower-bound g_{t+1} is constructed; maximization of g_{t+1} in the next M-step gives $\theta^{(t+2)}$, etc.

Convergence and local maximum

- The convergence of EM algorithm is guaranteed, see detailed proof in http://cs229.stanford.edu/notes/cs229-notes8.pdf (http://cs229.stanford.edu/notes/cs229-notes8.pdf).
- However, we may converge to a local maximum because the log likelihood is not guaranteed to be convex. This can be understood from the cartoon picture, if we start from the right side of the picture, we may converge to the local maximum on the right side.

Application of EM

- Using the EM, we can derive the two steps of mixtures of Gaussian model introduced earlier.
- We can also obtain many generative algorithms such as mixtures of naive Bayes, etc.
- K-means is a special case of EM.
- EM can be used to derive factor analysis. http://cs229.stanford.edu/notes/cs229-notes9.pdf (http://cs229.stanford.edu/notes/cs229-notes9.pdf)