EM Algorithm for Gaussian Mixture Model

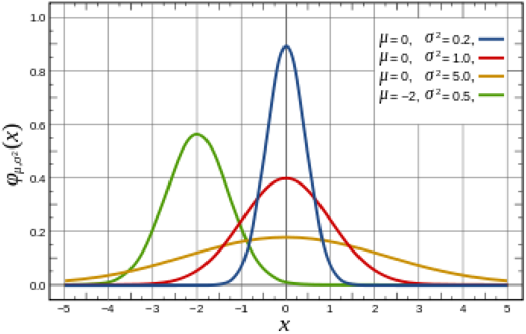
Kai Yu

Speech Lab
Department of Computer Science & Engineering
Shanghai Jiaotong University

April 2013

The Guassian distribution:

$$\mathcal{N}(x|\mu,\Sigma) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma|^{1/2}}exp\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\} \quad (1)$$

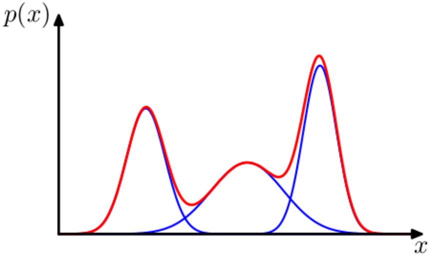


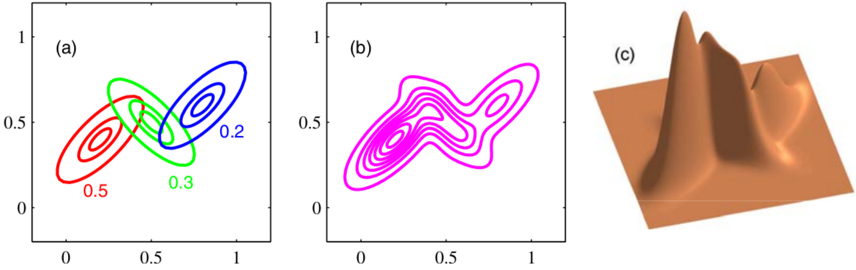
The Guassian Mixture distribution is a linear superposition of Guassians:

$$p(x) = \Sigma_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \tag{2}$$

Subject to:

$$\Sigma_{k=1}^{K} \pi_k = 1 \tag{3}$$



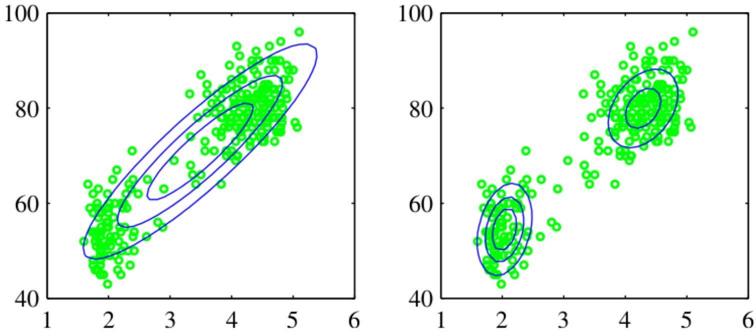


A 2-dimension example of GMM

Introduction of the Guassian Mixture Model

Now, for a Guassian Mixture Model, given the parameters:
k, the number of Guassian components
$\pi_1 ... \pi_k$, the mixture weights of the components
$\mu_1 ... \mu_k$, the mean of each component
$\Sigma_1 ... \Sigma_k$, the variance of each component
We can generate samples $s_1, s_2 ... s_n$ from the distribution.

Why do we need Guassian Mixture



In this example, we see that Guassian Mixture describes the data better a single Guassian.

The latent variable

Given a Guassian Mixture model, we introduce K-dimensional binary random variable z which only one element z_k is euqal to 1 and the others are all 0.

$$z = (0, 0, .., 1, 0, ..0) \tag{4}$$

So there are K possible states for z.And we let

$$p(z_k = 1) = \pi_k \tag{5}$$

That is to say,

$$p(z) = \prod_{k=1}^{K} {\pi_k}^{z_k} \tag{6}$$

The latent variable

Then, we define the conditional distribution of x given a particular z:

$$p(x|z_k = 1) = \mathcal{N}(x|\mu_k, \Sigma_k) \tag{7}$$

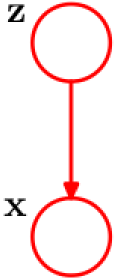
which can also be written as:

$$p(x|z) = \prod_{k=1}^{K} \mathcal{N}(x|\mu_k, \Sigma_k)^{z_k} \tag{8}$$

Now we can easily compute the marginal distribution of x

$$\begin{aligned} p(x) = \Sigma_z p(x|z)p(z) &= \Sigma_z \prod_{k=1}^{K} \mathcal{N}(x|\mu_k, \Sigma_k)^{z_k} \prod_{k=1}^{K} {\pi_k}^{z_k} \\ &= \Sigma_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \end{aligned}$$

The latent variable



Now, instead of working with $p(x)$ we can work with $p(x, z) = p(x|z)p(z)$, which will lead to significant simplification when we are introducing the EM algorithm.

The latent variable

Another quantity $p(z|x)$ will also be very important. We use $\gamma(z_k)$ to denote $p(z_k = 1|x)$, and we can use Bayes' theorem to deride its value.

$$\begin{aligned}\gamma(z_k) = p(z_k = 1|x) &= \frac{p(x|z_k = 1)p(z_k = 1)}{p(x)} \\ &= \frac{\mathcal{N}(x|\mu_k, \Sigma_k)}{\Sigma_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)}\end{aligned} \tag{9}$$

We usually say $\gamma(z_{nk})$ is the responsibility of component k for x_n.

Maximum likelihood

Suppose we have a data set of observations $\{x_1, ..., x_N\}$. And we wish to model this data set using Guassian Mixture model. We could represent this data set as an N*D matrix **X**, where N is the number of data vectors and D is the dimension of the vector. Then the log likelihood function is given by

$$lnp(X|\pi, \mu, \Sigma) = \Sigma_{n=1}^{N} ln\{\Sigma_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)\} \tag{10}$$

E-M algorithm for Gaussian mixtures

The expectation-maximization algorithm is an elegant and powerful method for finding maximum likelihood solutions for models with latent variables.
First, we set the derivatives of $lnp(X|\pi, \mu, \Sigma)$ in equation 10 with repect to μ_k to zero.

$$
\begin{aligned}
0 &= -\Sigma_{n=1}^{N} \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\Sigma_{j=1}^{K} \pi_j \mathcal{N}(x_n|\mu_j, \Sigma_j)} \Sigma_k (x_n - \mu_k) \\
&= -\Sigma_{n=1}^{N} \gamma(z_{nk}) \Sigma_k (x_n - \mu_k)
\end{aligned}
\tag{11}
$$

If we assume Σ_k to be nonsingular, we obtain

$$
\mu_k = \frac{\Sigma_{n=1}^{N} \gamma(z_{nk}) x_n}{\Sigma_{n=1}^{N} \gamma(z_{nk})} \tag{12}
$$

E-M algorithm for Gaussian mixtures

We set $N_k = \Sigma_{n=1}^{N} \gamma(z_{nk})$, as the effective number of points assigned to cluster k.
If we set the derivative of $lnp(X|\pi, \mu, \Sigma)$ with respect to Σ_k to zero,we get

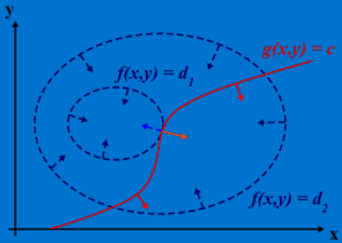
$$\Sigma_k = \frac{1}{N_k} \gamma_k(z_{nk})(x_n - \mu_k)(x_n - \mu_k)^T \tag{13}$$

E-M algorithm for Gaussian mixtures

Finally, when we maximize the log likelihood with respect to π, we need to take the constraint $\Sigma_{k=1}^{K}\pi_k = 1$ into consideration. This is done by using the Lagrange multiplier.

The Lagrange Multiplier

We want to maximize $f(x, y)$ subject $g(x, y) = c$.
Let $\Lambda(x, y\lambda) = f(x, y) + \lambda(g(x, y) - c)$.Then if $(x_0, y_0)$is a maximum of the original f, there exists (x_0, y_0, λ_0) is a stationary point for the Λ function.



The contour lines of f and g touch when the tangent vectors of the contour lines are parallel. Since the gradient of a function is perpendicular to the contour lines, this is the same as saying that the gradients of f and g are parallel.

The Lagrange Multiplier

So $\nabla_{x,y} f = -\lambda \nabla_{x,y} g$.
Combining with the constraint, we get $\nabla_{x,y,\lambda} \Lambda = 0$

Now, we apply the Lagrange Multiplier to maximize with respect to π. We will be maximizing

$$lnp(x|\pi, \mu, \Sigma) + \lambda(\Sigma_{k=1}^{K} \pi_k - 1) \tag{14}$$

By maximizing it we will get

$$\pi_k = \frac{N_k}{N} \tag{15}$$

We have to note that the solutions 12, 13, 15 are not closed. Because the responsibilities $\gamma(z_{nk})$ depend on the parameters. However, they suggest a iterative scheme for finding a solution to the maximum likelihood problem.

E-M algorithm for Gaussian mixtures

Initialize Initialize the parameters μ_k, Σ_k, and π_k

E-step Evaluate the responsibilities using the current parameter values.

$$\gamma(z_{nk}) = \frac{\mathcal{N}(x_n|\mu_k, \Sigma_k)}{\Sigma_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)} \tag{16}$$

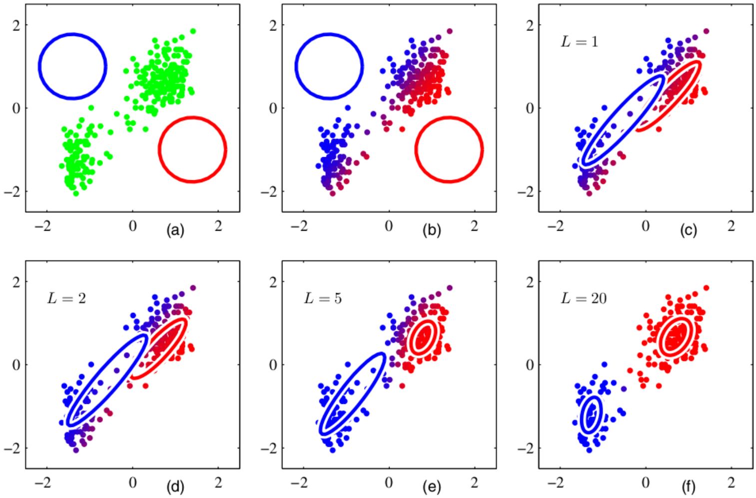
E-M algorithm for Gaussian mixtures

M-step Re-estimate the parameters using the current responsibilities.

$$
\begin{aligned}
\mu_k^{new} &= \frac{1}{N_k} \Sigma_{n=1}^{N} \gamma(z_{nk}) x_n \\
\Sigma_k^{new} &= \frac{1}{N_k} \gamma_k(z_{nk})(x_n - \mu_k)(x_n - \mu_k)^T \\
\pi_k^{new} &= \frac{N_k}{N}
\end{aligned}
\tag{17}
$$

Likelihood Recalculate the log likelihood function to see if it converges, if not, go to step 1 again.

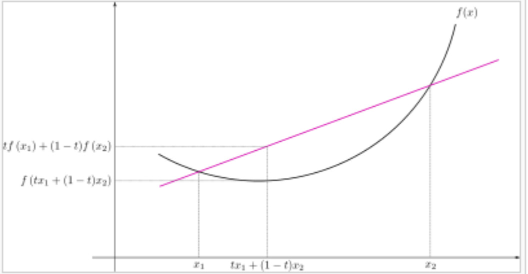
An Example



Now we've know the inituition of the EM algorithm, and we are going to proof its property. We will need the Jensen Inequality.

Jensen's Inequality

If X is a random variable and φ is a convex function. Then $\varphi(E[X]) \leq E[\varphi(X)]$.



Jensen's inequality generalizes the statement that a secant line of a convex function lies above the graph.

The General EM

Theorem

In each iteration, the E-M algorithm gives a solution that gives higher likelihood.

*Proof:*For simplicity, we denote all parameters as θ. Remembering the latent variable Z, we can now write the log likelihood as

$$lnp(X|\theta) = ln\{\Sigma_z p(X, z|\theta)\} \tag{18}$$

In the E-step, we are are actually forming an ancillary function(go back to 12, 13, 15 and check it):

$$\mathcal{Q}(\theta, \theta^{old}) = \Sigma_z p(z|X, \theta^{old}) lnp(X, z|\theta) \tag{19}$$

Then, in the M-step, we are maximizing $\mathcal{Q}(\theta, \theta^{old})$, and get the θ^{new}.

The General EM

Now we use Jensen's inequality to transform the log likelihood function.

$$\begin{aligned} lnp(X|\theta) &= ln\{\Sigma_z \frac{q(z)p(X,z|\theta)}{q(z)}\} \\ &\geq \Sigma_z q(z) ln(\frac{p(X,z|\theta)}{q(z)}) \\ &= \Sigma_z q(z) ln(p(X,z|\theta)) - \Sigma_z q(z) ln(q(z)) \end{aligned} \tag{20}$$

To continue we need to prove a lemma first.

lemma

If $q(z) = p(z|\theta, X)$, equation 20 becomes an equality.
Proof:

$$\begin{aligned} & lnp(X|\theta) - \Sigma_z q(z) ln(\frac{p(X,z|\theta)}{q(z)}) \\ = & \Sigma_z q(z)\{ln(p(X|\theta)) - ln(\frac{p(X,z|\theta)}{q(z)})\} \\ = & \Sigma_z q(z)\{ln(\frac{q(z)}{p(z|\theta,X)})\} \\ = & \Sigma_z q(z) ln(1) = 0 \end{aligned} \quad (21)$$

Now we let $q(z) = p(z|X, \theta^{old})$, then we can complete our proof.

$$
\begin{aligned}
ln(p(X|\theta^{new}) &\geq \Sigma_z q(z) ln(p(X, z|\theta^{new})) - \Sigma_z q(z) ln(q(z)) \\
&= \mathcal{Q}(\theta^{new}, \theta^{old}) - \Sigma_z q(z) ln(q(z)) \\
&\geq \mathcal{Q}(\theta^{old}, \theta^{old}) - \Sigma_z q(z) ln(q(z)) \\
&= ln(p(X|\theta^{old})
\end{aligned} \tag{22}
$$

Bibliography

- Some pictures are from Wiki or PRML