Introduction to EM: Gaussian Mixture Model (GMM)

Matthew Stephens and Heejung Shim

Learning goals

- Understand the purpose of the EM algorithm.
- Understand the EM algorithm
- Be able to derive and implement the EM algorithm

Expectation-Maximization (EM) algorithm

The **expectation-maximization** (EM) algorithm aims to obtain the maximum likelihood estimates (MLEs) of parameters.

We will first introduce the EM algorithm in the context of Gaussian Mixture Model (GMM).

If we observe independent samples $X_1, \ldots, X_n$ from GMM, the likelihood function of the parameters, $\theta = (\pi_1, \ldots \pi_{K-1}, \mu_1, \sigma_1^2, \ldots, \mu_K, \sigma_K^2)$, is

$$L(\theta) = \prod_{i=1}^{n} \left[\sum_{k=1}^{K} f(x_i; \mu_k, \sigma_k^2) \pi_k \right],$$

where $f(x_i; \mu_k, \sigma_k^2)$ is the probability density function of a normal (gaussian) distribution with mean $= \mu_k$ and variance σ_k^2.

We will describes the EM algorithm which aims to obtain the MLE of $\theta = (\pi_1, \ldots, \pi_{K-1}, \mu_1, \sigma_1^2, \ldots, \mu_K, \sigma_K^2)$ given observations $\{X_1, \ldots, X_n\}$.

Review: MLE of Normal Distribution

Suppose we have independent observations $X_1, \ldots, X_n$ from a Gaussian distribution with unknown mean μ and known variance σ^2. To find MLE for μ, we find the log-likelihood $\ell(\mu)$, take the derivative with respect to μ, set it equal zero, and solve for μ:

$p(x_1, \cdots x_n) = \prod_{i=1}^{n} p(x_i)$

$$L(\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \frac{(x_i - \mu)^2}{2\sigma^2} \tag{1}$$

$$\Rightarrow \ell(\mu) = \sum_{i=1}^{n} \left[\log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) - \frac{(x_i - \mu)^2}{2\sigma^2} \right] \tag{2}$$

$$\Rightarrow \frac{d}{d\mu}\ell(\mu) = \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma^2} \tag{3}$$

Review: MLE of Normal Distribution

$$\Rightarrow \frac{d}{d\mu}\ell(\mu) = \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma^2} \tag{4}$$

Setting this equal to zero and solving for μ, we get that $\hat{\mu}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Note that applying the log function to the likelihood helped us decompose the product and removed the exponential function so that we could easily solve for the MLE.

MLE of GMM

Now we attempt the same strategy for deriving the MLE of the GMM. Our parameters are $\theta = (\pi_1, \ldots, \pi_{K-1}, \mu_1, \sigma_1^2, \ldots, \mu_K, \sigma_K^2)$ and our log-likelihood is:

$$\ell(\theta) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k f(x_i; \mu_k, \sigma_k^2) \right).$$

Taking a look at the expression above, we already see a difference between this scenario and the simple setup in MLE for the normal distribution in the previous slides. We see that the summation over the K components "blocks" our log function from being applied to the normal densities.

$$\ell(\theta) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k f(x_i; \mu_k, \sigma_k^2)\right).$$

If we were to follow the same steps as above and differentiate with respect to μ_k and set the expression equal to zero, we would get:

$$\sum_{i=1}^{n} \frac{1}{\sum_{k'=1}^{K} \pi_{k'} f(x_i; \mu_{k'}, \sigma_{k'}^2)} \pi_k f(x_i; \mu_k, \sigma_k^2) \frac{(x_i - \mu_k)}{\sigma_k^2} = 0. \qquad (1)$$

Now we're stuck because we can't analytically solve for μ_k.

However, we make one important observation which provides intuition for what is to come: if we knew the latent variables Z_i, then we could simply gather all our samples X_i with $Z_i = k$ and simply modify the estimate from the normal distribution to estimate μ_k.

Intuition for EM algorithm

Intuitively, the latent variables Z_i should help us find the MLEs. We first attempt to compute the posterior distribution of Z_i given the observations:

$$P(Z_i = k|X_i) = \frac{P(X_i|Z_i = k)P(Z_i = k)}{P(X_i)} = \frac{\pi_k f(x_i; \mu_k, \sigma_k^2)}{\sum_{k'=1}^{K} \pi_{k'} f(x_i; \mu_{k'}, \sigma_{k'}^2)}. \quad (2)$$

Now we can rewrite equation (1), the derivative of the log-likelihood with respect to μ_k, as follows:

$$\sum_{i=1}^{n} P(Z_i = k|X_i) \frac{(x_i - \mu_k)}{\sigma_k^2} = 0.$$

Even though $P(Z_i = k|X_i)$ depends on μ_k, we can cheat a bit and pretend that it doesn't. Now we can solve for μ_k in this equation to get:

pretend $P(Z_i=k|x_i)$ doesn't depend on μ_k

$$\hat{\mu}_k = \frac{\sum_{i=1}^{n} P(Z_i = k|X_i)x_i}{\sum_{i=1}^{n} P(Z_i = k|X_i)}.$$

$$\hat{\mu}_k = \frac{\sum_{i=1}^n P(Z_i = k|X_i)x_i}{\sum_{i=1}^n P(Z_i = k|X_i)} = \frac{1}{N_k}\sum_{i=1}^n P(Z_i = k|X_i)x_i, \tag{3}$$

where we set $N_k = \sum_{i=1}^n P(Z_i = k|X_i)$. We can think of N_k as the effective number of points assigned to component k. We see that $\hat{\mu}_k$ is therefore a weighted average of the data with weights $P(Z_i = k|X_i)$.

"membership probability"

Similarly, if we apply a similar method to finding $\hat{\sigma}_k^2$ and $\hat{\pi}_k$, we find that:

$$\hat{\sigma}_k^2 = \frac{1}{N_k}\sum_{i=1}^n P(Z_i = k|X_i)(x_i - \mu_k)^2 \tag{4}$$

$$\hat{\pi}_k = \frac{N_k}{n} \tag{5}$$

Again, remember that $P(Z_i = k|X_i)$ depends on the unknown parameters, so these equations are not closed-form expressions.

We are now in the following situation:

- If we knew the parameter estimates, we could compute the posterior probabilities $P(Z_i = k|X_i)$.
- If we knew the posteriors $P(Z_i = k|X_i)$, we could easily compute the parameter estimates.

Let's iterate!

Motivated by the two observations above, the following algorithm can be considered:

1. Initialize the current values of $\theta = (\pi_1, \ldots, \pi_{K-1}, \mu_1, \sigma_1^2, \ldots, \mu_K, \sigma_K^2)$. Evaluate the log-likelihood with the current values.
2. Evaluate the posterior probabilities $P(Z_i = k|X_i)$ using the current values of θ with equation (2).
3. Compute new parameter estimates $\hat{\theta}$ with the current values of $P(Z_i = k|X_i)$ using equations (3), (4) and (5).
4. Evaluate the log-likelihood with the new parameter estimates. If the log-likelihood has changed by less than some small ϵ, stop. Otherwise, replace the current values with the new parameter estimates and go back to step 2.

stop criterion

Motivated by the two observations above, the following algorithm can be considered:

1. Initialize the current values of $\theta = (\pi_1, \ldots, \pi_{K-1}, \mu_1, \sigma_1^2, \ldots, \mu_K, \sigma_K^2)$. Evaluate the log-likelihood with the current values.
2. **E-step:** Evaluate the posterior probabilities $P(Z_i = k|X_i)$ using the current values of θ with equation (2).
3. **M-step:** Compute new parameter estimates $\hat{\theta}$ with the current values of $P(Z_i = k|X_i)$ using equations (3), (4) and (5).
4. Evaluate the log-likelihood with the new parameter estimates. If the log-likelihood has changed by less than some small ϵ, stop. Otherwise, replace the current values with the new parameter estimates and go back to step 2.

One property of the EM algorithm is that the log-likelihood increases at every step (why?).

greedy algorithm → doesn't guarantee for global maximum

The EM algorithm is sensitive to the initial values. In practice, we run EM algorithm multiple times with different initial values and pick estimators with the highest log-likelihood.

EM algorithm

eg mixture model, HMM

The EM algorithm attempts to find maximum likelihood estimates for models with latent variables. Now, we describe a more abstract view of EM algorithm which can be extended to other latent variable models.

$= (x_1, \cdots x_n)$ $(z_1, \cdots z_m)$

Let X be the entire set of observed variables and Z the entire set of latent variables. The log-likelihood is therefore:

$l(\theta) = \log(P(X|\theta))$ → important for EM

$$\log\left(P(X|\theta)\right) = \log\left(\sum_{Z} P(X, Z|\theta)\right), \quad = \log\left(\sum_{z} P(X|z,\theta)P(z|\theta)\right)$$

where we've simply marginalized Z out of the joint distribution.

As we noted above, the existence of the sum inside the logarithm prevents us from applying the log to the densities which results in a complicated expression for the MLE. As we noted previously, if we knew Z, the maximization would be easy.

$$\log\left(P(X|\theta)\right) = \log\left(\sum_Z P(X, Z|\theta)\right).$$

We typically don't know Z, but the information we do have about Z is contained in the posterior $P(Z|X, \theta)$. Since we don't know $\log P(X, Z|\theta)$ (which is called the complete log-likelihood), we consider its expectation under the posterior distribution of the latent variables. We maximize this expectation to find a new estimate for the parameters.

Note: $\log P(X|\theta)$ is called the incomplete log-likelihood.

EM algorithm is to find a value to maximize incomplete log-likelihood

EM algorithm

1. Initialize the current values of θ. Let θ^0 denote the current values. Evaluate the incomplete log-likelihood with θ^0.
2. **E-step:** Using θ^0, compute the posterior distribution of the latent variables given by $P(Z|X, \theta^0)$.
3. **M-step:** Obtain new parameter estimates $\hat{\theta}$ which maximise the expectation of the complete log-likelihood with respect to this posterior $P(Z|X, \theta^0)$. Specifically,

$$\hat{\theta} = \text{argmax}_\theta Q(\theta, \theta^0), \text{ where}$$

$$Q(\theta, \theta^0) = E_{Z|X,\theta^0}[\log(P(X, Z|\theta))]$$

4. Evaluate the incomplete log-likelihood with $\hat{\theta}$. If the incomplete log-likelihood has changed by less than some small ϵ, stop. Otherwise, replace the current values with the new parameter estimates ($\theta^0 \leftarrow \hat{\theta}$) and go back to step 2.

EM guarantee log $P(X|\theta)$ increases at every step.

EM algorithm

One property of the EM algorithm is that the incomplete log-likelihood increases at every step (why?).

The EM algorithm is sensitive to the initial values. In practice, we run EM algorithm multiple times with different initial values and pick estimators with the highest incomplete log-likelihood.

Formal derivation of the EM algorithm for GMM

Now we derive the EM steps for GMM and compare it to our "intuitive" derivation above.

Let $\theta = (\pi_1, \ldots, \pi_{K-1}, \mu_1, \sigma_1^2, \ldots, \mu_K, \sigma_K^2)$. The complete likelihood is:

$$\begin{aligned} P(X, Z|\theta) &= \prod_{i=1}^{n} \prod_{k=1}^{K} \left[P(Z_i = k|\theta) P(X|Z_i = k, \theta) \right]^{I(Z_i=k)} \\ &= \prod_{i=1}^{n} \prod_{k=1}^{K} \pi_k^{I(Z_i=k)} f(x_i; \mu_k, \sigma_k^2)^{I(Z_i=k)}. \end{aligned}$$

where $I(A)$ is an indicator function.

So the complete log-likelihood takes the form:

$$\log\left(P(X, Z|\theta)\right) = \sum_{i=1}^{n} \sum_{k=1}^{K} I(Z_i = k) \left(\log(\pi_k) + \log(f(x_i; \mu_k, \sigma_k^2)) \right).$$

Note that for the complete log-likelihood, the logarithm acts directly on the normal density which leads to a simpler solution for the MLE.

As we said, in practice, we do not observe the latent variables, so we consider the expectation of the complete log-likelihood with respect to the posterior of the latent variables. $P(X, Z|\theta^0)$

Let $\theta^0 = (\pi_1^0, \ldots, \pi_{K-1}^0, \mu_1^0, {\sigma_1^2}^0, \ldots, \mu_K^0, {\sigma_K^2}^0)$ denote the current value of the parameters. Then, the expected value of the complete log-likelihood is therefore:

$$\begin{aligned} Q(\theta, \theta^0) &= E_{Z|X,\theta^0}[\log(P(X, Z|\theta))] \\ &= E_{Z|X,\theta^0}\left[\sum_{i=1}^{n}\sum_{k=1}^{K} I(Z_i = k)\left(\log(\pi_k) + \log(f(x_i; \mu_k, \sigma_k^2))\right)\right] \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} E_{Z|X,\theta^0}[I(Z_i = k)]\left(\log(\pi_k) + \log(f(x_i; \mu_k, \sigma_k^2))\right), \end{aligned}$$

inner r.v is Z.

$P(z_i = k|X, \theta^0)$ or $P_{z|x,\theta^0}(z_i = k)$

Since $E_{Z|X,\theta^0}[I(Z_i = k)] = P(Z_i = k|X, \theta^0)$, we see that this is simply $P(Z_i = k|X)$ which we computed using equation (2) with θ^0.

EM proceeds as follows: first initialize the current values of θ and use them in the E-step to compute $P(Z_i = k|X, \theta^0)$. Then, with $P(Z_i = k|X, \theta^0)$ fixed, find new parameter estimates which maximize the expected complete log-likelihood above.

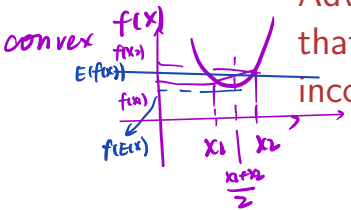
Exercise: Derive E-step and M-step of the EM algorithm for GMM.

- In the M-step, show that the new parameter estimates which maximize the expected complete log-likelihood above are equivalent to the closed form solutions in equations (3), (4), and (5).
- Therefore, we can see the "intuitive" derivation above is equivalent to the "formal" derivation here.

You should understand and use the "formal" derivation of EM-algorithm which can be applied to more general models.

Implementation of EM algorithm for GMM

Implement the EM algorithm for GMM: see EM.pdf.



Advanced topic: prove that the EM algorithm guarantees that parameter estimates at each iteration increase the incomplete log likelihood

concave

Jensen's inequality: Let f be a convex function, and let X be a random variable. Then $E[f(X)] \geq f(E[X])$. Moreover, if f is strictly convex, then $E[f(X)] = f(E[X])$ holds true if and only if $X = E[X]$ (i.e., if X is a constant).

Let $l(\theta)$ denote the incomplete log likelihood. We wish to to show $l(\hat{\theta}) \geq l(\theta^0)$, where θ^0 and $\hat{\theta}$ represent the current and new values of the parameters θ.

$\log p(x|\hat{\theta})$

$$l(\hat{\theta}) = \log \sum_Z \mathrm{P}(X, Z|\hat{\theta}) = \log \sum_Z Q(Z) \frac{\mathrm{P}(X, Z|\hat{\theta})}{Q(Z)}$$

$$= \log \mathrm{E}_{Q(Z)} \left[\frac{\mathrm{P}(X, Z|\hat{\theta})}{Q(Z)} \right] \quad \text{for probability mass function Q(Z)}$$

log is concave function

$$\geq \mathrm{E}_{Q(Z)} \left[\log \frac{\mathrm{P}(X, Z|\hat{\theta})}{Q(Z)} \right] \quad \text{by the Jensen's inequality}$$

$= E_{z|x,\theta^0}[\log p(x,z|\hat{\theta}) - \log p(x,z|\theta^0)]$

by def of $\hat{\theta}$ ⇒ to max $E_{z|x,\theta^0}[\log p(x,z|\theta)]$

Let $Q(Z) = \mathrm{P}(Z|X, \theta^0)$. Then by the definition of $\hat{\theta}$,

$$\geq \mathrm{E}_{Z|X,\theta^0} \left[\log \frac{\mathrm{P}(X, Z|\theta^0)}{\mathrm{P}(Z|X, \theta^0)} \right]$$

→ constant

$= E_{z|x,\theta^0}[\log p(x,z|\theta^0) - \log p(z|x,\theta^0)]$

Here, $\mathrm{P}(X, Z|\theta^0)/\mathrm{P}(Z|X, \theta^0) = \mathrm{P}(X|\theta^0)$ doesn't depend on Z.

Then, the equality holds in the Jensen's inequality, so

swap the order of log & E

$$= \log \mathrm{E}_{Z|X,\theta^0} \left[\frac{\mathrm{P}(X, Z|\theta^0)}{\mathrm{P}(Z|X, \theta^0)} \right] = \log \sum_Z \mathrm{P}(X, Z|\theta^0) = l(\theta^0)$$

$= \log p(x|\theta^0)$.

↳ $\log \sum_z p(z|x,\theta^0) \cdot \frac{p(x,z|\theta^0)}{p(z|x,\theta^0)}$

Chapter 9.4 of 'Pattern Recognition and Machine Learning, Christopher Bishop, 2006' provide a different interpretation of the EM algorithm (using a language in machine learning).

Learning goals

- Understand the purpose of the EM algorithm.
- Understand the EM algorithm
- Be able to derive and implement the EM algorithm