Estimating the parameters of a Mixture of Gaussians model

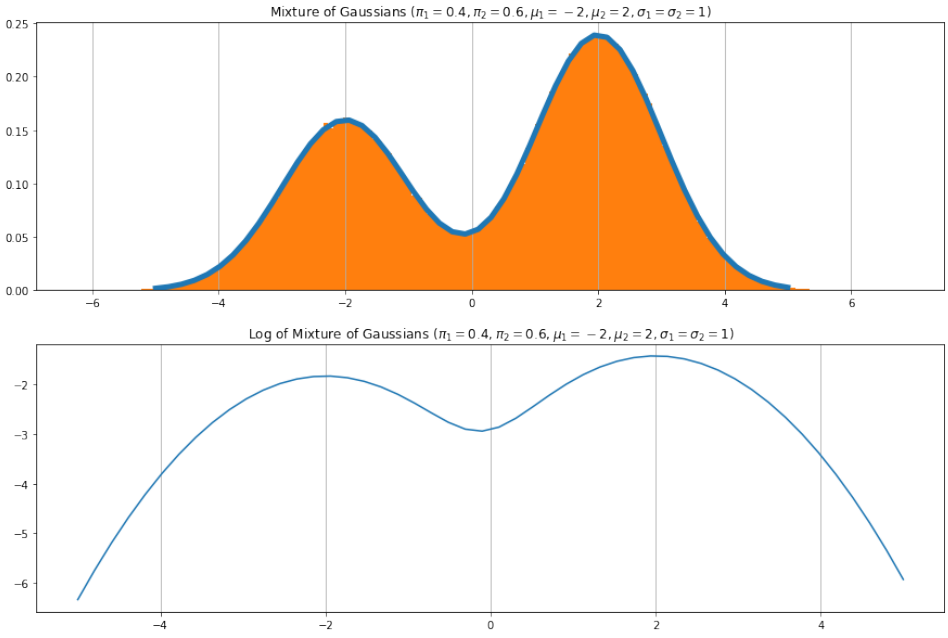
Mohsen Kiskani

August 16, 2019

1 Problem definition

We have a Mixture of Gaussians (MoG) model with parameters $\theta = (\pi_1, \mu_1, \sigma_1, \pi_2, \mu_2, \sigma_2)$. We want to estimate the parameter vector θ using the data $y_{1:N}$. We assume that each data point is gathered independently according to the following distribution

$$P(y \mid \theta) = \pi_1 \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left(-\frac{(y-\mu_1)^2}{2\sigma_1^2}\right) + \pi_2 \frac{1}{\sigma_2\sqrt{2\pi}} \exp\left(-\frac{(y-\mu_2)^2}{2\sigma_2^2}\right) \quad (1)$$



2 Frequentist methods

2.1 MAP

In this case, we are interested in finding the solution to the following problem,

$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta \mid y_{1:N}) = \arg\max_{\theta} P(y_{1:N} \mid \theta) P(\theta). \tag{2}$$

The problem with this approach is that we do not know $P(\theta)$.

2.2 MLE

This is a simpler problem in which we have

$$\hat{\theta}_{MLE} = \arg\max_{\theta} P(y_{1:N} \mid \theta). \tag{3}$$

Due to the independence of the data points we can write

$$\begin{aligned} \hat{\theta}_{MLE} &= \arg\max_{\theta} P(y_{1:N} \mid \theta) = \arg\max_{\theta} \prod_{i=1}^{N} P(y_i \mid \theta) \\ &= \arg\max_{\theta} \sum_{i=1}^{N} \log\left(P(y_i \mid \theta)\right). \end{aligned} \tag{4}$$

Finding the maximizing θ in this problem can be challenging. Trying to analytically solve the problem by taking the derivative is not possible since analytic solution does not exist. However, we can use gradient descent to estimate the parameter vector. In gradient descent estimation we want to minimize

$$\begin{aligned} l(\theta) &= -\sum_{i=1}^{N} \log\left(P(y_i \mid \theta)\right) \\ &= -\sum_{i=1}^{N} \log\left(\pi_1 \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left(-\frac{(y_i-\mu_1)^2}{2\sigma_1^2}\right) + \pi_2 \frac{1}{\sigma_2\sqrt{2\pi}} \exp\left(-\frac{(y_i-\mu_2)^2}{2\sigma_2^2}\right)\right) \end{aligned}$$

with respect to parameters $\theta = (\pi_1, \mu_1, \sigma_1, \pi_2, \mu_2, \sigma_2)$. We initialize by some random choice for these parameters and then iterate. The problem with this method is that sometimes, gradient is hard to compute.

2.2.1 Expectation Maximization (EM)

The EM algorithm attempts to find maximum likelihood estimates for models with latent variables. In the following, we will give a brief overview of the EM algorithm.

We assume that the data y is coming from a latent random variable z. Hence,

$$\begin{aligned} f(\theta) = \log\left(P(y \mid \theta)\right) &= \log\left(\sum_z P(y, z \mid \theta)\right) \\ &= \log\left(\sum_z q(z \mid y, \theta)\frac{P(y, z \mid \theta)}{q(z \mid y, \theta)}\right) \\ &\geq \sum_z q(z \mid y, \theta)\log\left(\frac{P(y, z \mid \theta)}{q(z \mid y, \theta)}\right) \\ &= \sum_z q(z \mid y, \theta)\log\left(\frac{P(z \mid y, \theta)P(y \mid \theta)}{q(z \mid y, \theta)}\right) \\ &= -D_{KL}\left(q(z \mid y, \theta) \mid\mid P(z \mid y, \theta)\right) + \log\left(P(y \mid \theta)\right).v \end{aligned} \tag{5}$$

Where $q(z \mid y, \theta)$ is an arbitrary density over z and the inequality is valid due to Jensen's inequality and concavity of the log function. Now, instead of maximizing $f(\theta)$ directly, one can maximize the lower bound, i.e.

$$\begin{aligned} F(q, \theta) \triangleq \sum_z q(z \mid y, \theta)\log\left(\frac{P(z \mid y, \theta)P(y \mid \theta)}{q(z \mid y, \theta)}\right) \\ = -D_{KL}\left(q(z \mid y, \theta) \mid\mid P(z \mid y, \theta)\right) + \log\left(P(y \mid \theta)\right) \end{aligned} \tag{6}$$

via coordinate ascent algorithm. This will be the EM algorithm with the following steps:

$$\mathbf{E} - \mathbf{Step} : q^{(t+1)} = \arg\max_q F(q, \theta^{(t)}) \tag{7}$$

$$\mathbf{M} - \mathbf{Step} : \theta^{(t+1)} = \arg\max_\theta F(q^{(t+1)}, \theta). \tag{8}$$

Starting with some initial value of $\theta^{(0)}$, one can cycle between the E and M steps until $\theta^{(t)}$ converges to a local maximum. Computing the E step seem to be a pain in the ass since we are maximizing over the space of distributions. But notice that maximizing $F(q, \theta)$ over q is equivalent to minimizing the KL distance between $q(z \mid y, \theta)$ and $P(z \mid y, \theta)$. Hence,

$$q^{(t+1)} = P(z \mid y, \theta^{(t)}). \tag{9}$$

Notice that at this minimizing choice for KL distance, the lower bound becomes equal to the negative log likelihood function for $\theta^{(t)}$.

So at this point, we don't need to worry about finding the optimal q since we know it's exact form and we know that it is a distribution that depends on $\theta^{(t)}$. In M step, we

fix q and we note that

$$\begin{aligned} F(q,\theta) &= \sum_z q(z \mid y,\theta) \log\left(\frac{P(y,z \mid \theta)}{q(z \mid y,\theta)}\right) \\ &= \sum_z q(z \mid y,\theta) \log\left(P(y,z \mid \theta)\right) \\ &- \sum_z q(z \mid y,\theta) \log\left(q(z \mid y,\theta)\right) \\ &= \sum_z q(z \mid y,\theta) \log\left(P(y,z \mid \theta)\right) + H(q), \end{aligned} \tag{10}$$

where $H(q)$ is the entropy of q and is not changing by θ in the M step as q is fixed in the E step. Hence, maximizing $F(q,\theta)$ in the M step is equivalent to maximizing the expected complete log-likelihood function. In other words, if we define

$$\begin{aligned} Q(\theta \mid \theta^{(t)}) &\triangleq \sum_z q(z \mid y,\theta^{(t)}) \log\left(P(y,z \mid \theta)\right) \\ &= \mathbb{E}_{P(z|y,\theta^{(t)})}\left[\log\left(P(y,z \mid \theta)\right)\right] \end{aligned} \tag{11}$$

then using equation (9), we have

$$\arg\max_\theta F(q^{(t+1)},\theta) = \arg\max_\theta F\left(P(z \mid y,\theta^{(t)}),\theta\right) = \arg\max_\theta Q(\theta \mid \theta^{(t)}). \tag{12}$$

Therefore, in summary the EM algorithm can be summarized as follows

$$\mathbf{E}-\mathbf{Step}: \text{Compute } Q(\theta \mid \theta^{(t)}) = \mathbb{E}_{P(z|y,\theta^{(t)})}\left[\log\left(P(y,z \mid \theta)\right)\right] \tag{13}$$

$$\mathbf{M}-\mathbf{Step}: \theta^{(t+1)} = \arg\max_\theta \mathbb{E}_{P(z|y,\theta^{(t)})}\left[\log\left(p(y,z \mid \theta)\right)\right]. \tag{14}$$

Now, back to our original problem, to generalize the problem to a mixture of K Gaussians, let z be a K-dimensional binary random variable which has a 1-of-K representation in which a particular element z^k is equal to 1 and all other elements are equal to 0. The values of z^k therefore satisfy $z^k \in \{0,1\}$ and $\sum_k z^k = 1$, and we see that there are K possible states for the vector z according to which element is nonzero. The marginal distribution over z is specified in terms of the mixing coefficients π_k, such that $P(z^k = 1) = \pi_k$ and $0 \leq \pi_k \leq 1$ and $\sum_k \pi_k = 1$. For the i^{th} sample, we can write

$$P(z_i \mid \theta) = \prod_{k=1}^{K} \pi_k^{z_i^k}. \tag{15}$$

Hence,

$$P(z_{1:N} \mid \theta) = \prod_{i=1}^{N}\prod_{k=1}^{K} \pi_k^{z_i^k}. \tag{16}$$

similarly

$$P(y_{1:N} \mid z_{1:N}, \theta) = \prod_{i=1}^{N} \prod_{k=1}^{K} \mathcal{N}(y_i; \mu_k, \sigma_k)^{z_i^k}. \tag{17}$$

and

$$P(y_{1:N}, z_{1:N} \mid \theta) = \prod_{i=1}^{N} \prod_{k=1}^{K} (\pi_k \mathcal{N}(y_i; \mu_k, \sigma_k))^{z_i^k}. \tag{18}$$

and

$$\log\left(P(y_{1:N}, z_{1:N} \mid \theta)\right) = \sum_{i=1}^{N} \sum_{k=1}^{K} z_i^k \log\left(\pi_k \mathcal{N}(y_i; \mu_k, \sigma_k)\right). \tag{19}$$

Using Bayes rule,

$$\begin{aligned} P(z_{1:N} \mid y_{1:N}, \theta) &= \frac{P(y_{1:N} \mid z_{1:N}, \theta) P(z_{1:N} \mid \theta)}{\sum_z P(y_{1:N} \mid z_{1:N}, \theta) P(z_{1:N} \mid \theta)} \\ &= \frac{\prod_{i=1}^{N} \prod_{k=1}^{K} (\pi_k \mathcal{N}(y_i; \mu_k, \sigma_k))^{z_i^k}}{\prod_{i=1}^{N} \sum_{j=1}^{K} \pi_j \mathcal{N}(y_i; \mu_j, \sigma_j)} \\ &= \prod_{i=1}^{N} \frac{\prod_{k=1}^{K} (\pi_k \mathcal{N}(y_i; \mu_k, \sigma_k))^{z_i^k}}{\sum_{j=1}^{K} \pi_j \mathcal{N}(y_i; \mu_j, \sigma_j)}. \end{aligned} \tag{20}$$

Denoting $\gamma(z_i^k; y_i, \theta) \triangleq P(z_i^k = 1 \mid y_i, \theta)$, we have

$$\gamma(z_i^k; y_i, \theta) = \frac{\pi_k \mathcal{N}(y_i; \mu_k, \sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(y_i; \mu_j, \sigma_j)}. \tag{21}$$

Since z can only take discrete values, the E-Step in equation (13) can be written as

$$\begin{aligned} Q(\theta \mid \theta^{(t)}) &= \mathbb{E}_{P(z_{1:N} \mid y_{1:N}, \theta^{(t)})} \left[\log\left(P(y_{1:N}, z_{1:N} \mid \theta)\right)\right] \\ &= \mathbb{E}_{P(z_{1:N} \mid y_{1:N}, \theta^{(t)})} \left[\sum_{i=1}^{N} \sum_{k=1}^{K} z_i^k \log\left(\pi_k \mathcal{N}(y_i; \mu_k, \sigma_k)\right)\right] \\ &= \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbb{E}_{P(z_{1:N} \mid y_{1:N}, \theta^{(t)})} \left[z_i^k \log\left(\pi_k \mathcal{N}(y_i; \mu_k, \sigma_k)\right)\right] \\ &= \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbb{E}_{P(z_{1:N} \mid y_{1:N}, \theta^{(t)})} \left[z_i^k\right] \log\left(\pi_k \mathcal{N}(y_i; \mu_k, \sigma_k)\right) \\ &= \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbb{E}_{P(z_i \mid y_i, \theta^{(t)})} \left[z_i^k\right] \log\left(\pi_k \mathcal{N}(y_i; \mu_k, \sigma_k)\right) \\ &= \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma(z_i^k; y_i, \theta^{(t)}) \log\left(\pi_k \mathcal{N}(y_i; \mu_k, \sigma_k)\right) \end{aligned} \tag{22}$$

and the M-Step in equation (14) involves maximizing the above $Q(\theta \mid \theta^{(t)})$ to find the parameters of the next iteration. Taking the derivative with respect to μ_k we have

$$\frac{\partial Q(\theta \mid \theta^{(t)})}{\partial \mu_k} = 0 \Rightarrow \sum_{i=1}^{N} \gamma(z_i^k; y_i, \theta^{(t)}) \left(\mu_k - y_i\right) = 0.$$

Therefore,

$$\mu_k^{(t+1)} = \frac{1}{\sum_{i=1}^{N} \gamma(z_i^k; y_i, \theta^{(t)})} \sum_{i=1}^{N} \gamma(z_i^k; y_i, \theta^{(t)}) y_i. \tag{23}$$

Denoting

$$N_k \triangleq \sum_{i=1}^{N} \gamma(z_i^k; y_i, \theta^{(t)}), \tag{24}$$

we have

$$\mu_k^{(t+1)} = \frac{1}{N_k} \sum_{i=1}^{N} \gamma(z_i^k; y_i, \theta^{(t)}) y_i \tag{25}$$

Similarly, taking the derivative with respect to σ_k, we have

$$\frac{\partial Q(\theta \mid \theta^{(t)})}{\partial \sigma_k} = 0 \Rightarrow \sum_{i=1}^{N} \gamma(z_i^k; y_i, \theta^{(t)}) \left(\frac{(y_i - \mu_k)^2}{\sigma_k^2} - 1 \right) = 0.$$

Plugging in $\mu_k = \mu_k^{(t+1)}$ we have

$$\sigma_k^{(t+1)} = \sqrt{\frac{1}{N_k} \sum_{i=1}^{N} \gamma(z_i^k; y_i, \theta^{(t)}) (y_i - \mu_k^{(t+1)})^2}. \tag{26}$$

It only remains to find $\pi_k^{(t+1)}$. To find $\pi_k^{(t+1)}$, we should notice that $\sum_{k=1}^{K} \pi_k = 1$. Therefore, to optimize for π_k's we need to add a Lagrange constraint to $Q(\theta \mid \theta^{(t)})$ when taking the derivative. Hence, we should maximize

$$L(\theta \mid \theta^{(t)}) \triangleq Q(\theta \mid \theta^{(t)}) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right)$$

Taking the derivative with respect to π_k's we have

$$\frac{\partial L(\theta \mid \theta^{(t)})}{\partial \pi_k} = \frac{\partial Q(\theta \mid \theta^{(t)})}{\partial \pi_k} + \lambda = \sum_{i=1}^{N} \frac{\gamma(z_i^k; y_i, \theta^{(t)})}{\pi_k} + \lambda = \frac{N_k}{\pi_k} + \lambda = 0 \Rightarrow \pi_k = \frac{N_k}{\lambda}$$

Since sum of all π_k's should be one, we have

$$\sum_{k=1}^{K} \frac{N_k}{\lambda} = 1 \Rightarrow \lambda = \sum_{k=1}^{K} N_k = \sum_{k=1}^{K} \sum_{i=1}^{N} \gamma(z_i^k; y_i, \theta^{(t)})$$
$$= \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma(z_i^k; y_i, \theta^{(t)}) = \sum_{i=1}^{N} 1 = N$$

Therefore,

$$\pi_k^{(t+1)} = \frac{N_k}{N}. \tag{27}$$

To summarize it all, the E-Step in mixture of Gaussians is equivalent to calculating $\gamma(z_i^k; y_i, \theta^{(t)})$ based on equation (21) and the M-Step is equivalent to updating the parameters based on equations (24), (25), (26) and (27). These are pretty consistent with equations in page 439 of [1].

References

[1] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.