Gaussian Mixture Model

Damaris Stephanie Ndjebayi
Lionel Nanguep Komen
Omer Elhussien

African Master's in Machine Intelligence, AIMS-Senegal

April 6, 2023

Overview

Overview

Model assumptions

Let us suppose we are given a training set $\{x^{(1)}, \ldots, x^{(n)}\}$, which are *i.i.d.* We are interested in fitting the data using the below model,

$$\mathbb{P}(x^{(i)}, z^{(i)}) = \mathbb{P}(x^{(i)}|z^{(i)})\mathbb{P}(z^{(i)}),$$

where:

- $z^{(i)} \sim multinomial(\phi)$, $\phi_j \geq 0$, $\sum_{j=1}^{k} \phi_j = 1$, and $\phi_j = \mathbb{P}(z^{(i)} = j)$.
- $x^{(i)}|z^{(i)} = j \sim \mathcal{N}(\mu_j, \Sigma_j)$.

In the above model, we assume that each $x^{(i)}$ was generated by randomly choosing $z^{(i)} \in \{1, \ldots, k\}$, then $x^{(i)}$ was drawn from one of k Gaussian depending on $z^{(i)}$. The above model is called Gaussian Mixture Model. Further, $z^{(i)}$'s are called hidden (latent) variables.

The parameters of the above model are $\phi, \mu,$ and Σ. The likelihood of our data is given as,

$$\mathcal{L}(\phi, \mu, \Sigma) = \prod_{i=1}^{n} \mathbb{P}(x^{(i)}; \phi, \mu\Sigma) \iff$$

$$Log\left[\mathcal{L}(\phi, \mu, \Sigma)\right] = \sum_{i=1}^{n} log\left[\mathbb{P}(x^{(i)}; \phi, \mu, \Sigma)\right]$$

$$= \sum_{i=1}^{n} log \sum_{z^{(i)}=1}^{k} \mathbb{P}(x^{(i)}|z^{(i)}; \mu, \Sigma)\ \mathbb{P}(z^{(i)}; \phi)$$

The above likelihood could be easily solved with $z^{(i)}$'s are known. The likelihood will be,

$$log\,[\mathcal{L}] = \sum_{i=1}^{n} log\mathbb{P}(x^{(i)}|z^{(i)}; \phi, \mu, \Sigma) + log\mathbb{P}(z^{(i)}; \phi)$$

our estimated parameters are:

- $\phi_j = \dfrac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{z^{(i)} = j\},$
- $\mu_j = \dfrac{\sum_{i=1}^{n} \mathbb{I}\{z^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{n} \mathbb{I}\{z^{(i)} = j\}},$
- $\Sigma_j = \dfrac{\sum_{i=1}^{n} \mathbb{I}\{z^{(i)} = j\}(x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i=1}^{n} \mathbb{I}\{z^{(i)} = j\}}.$

Overview

Expectation Maximization algorithm

In the current problem, our $z^{(i)}$'s are unknown. We need to introduce the EM algorithm to solve the problem at hand.
EM is an iterative algorithm that has two main steps:

- E-step: It tries to guess the values of $z^{(i)}$'s.
- M-step: It updates the parameters of the model based on our guesses.

Algorithm 1: EM algorithm

repeat until convergence:

{

(E-step) for each i, j, set: $w_j^{(i)} = \mathbb{P}(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma)$.

(M-step) update the parameters:

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

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

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

}

In the E-step, we calculate

$$\mathbb{P}(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma) = \frac{\mathbb{P}(x^{(i)} | z^{(i)} = j; \mu, \Sigma)\, \mathbb{P}(z^{(i)} = j; \phi)}{\sum_{l=1}^{k} \mathbb{P}(x^{(i)} | z^{(i)} = l; \mu, \Sigma)\, \mathbb{P}(z^{(i)} = l; \phi)}.$$

- EM algorithm is vulnerable to local optima, so re-initializing at several different initial parameters will get us more accurate results.

Jensen's inequality

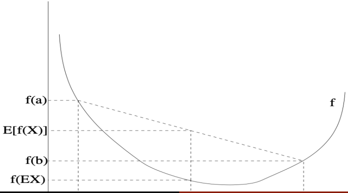
Theorem

Let f be a convex function, and let X be a random variable. Then,

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

Moreover, if f is strictly convex, then

$$\mathbb{E}\left[f(X)\right] = f\left[\mathbb{E}X\right] \ \textit{iff}\ X = \mathbb{E}(X) \textit{ with probability } 1.$$



The evidence lower bound

Let $\{x^{(1)}, \ldots, x^{(n)}\}$ be a training set which are *i.id*. Let $\mathbb{P}(x, z; \theta)$ be our model with z being the latent variable. Then,

$$\mathbb{P}(x; \theta) = \sum_{z} \mathbb{P}(x, z; \theta).$$

The likelihood is given by,

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} \mathbb{P}(x^{(i)}; \theta) \iff log\,[L] = \sum_{i=1}^{n} log\left[\mathbb{P}(x^{(i)}; \theta)\right]$$

$$= \sum_{i=1}^{n} log\left[\sum_{z^{(i)}} \mathbb{P}(x^{(i)}, z^{(i)}; \theta)\right]$$

The EM-algorithm maximizes $\mathcal{L}$ using two steps:

E-step: construct a lower bound on $\mathcal{L}$.

M-step: Optimize that lower-bound.

We will consider our optimization using a single data point, then after we bring the summation back.

$$log\mathbb{P}(x;\theta) = log\left[\sum_z \mathbb{P}(x,z;\theta)\right].$$

Let the possible values of z follow a distribution $\mathbb{Q}$, $\sum_z Q(z) = 1$, and $Q(z) \geq 0$.

$$log\mathbb{P}(x;\theta) = log\left[\sum_z \mathbb{P}(x,z;\theta)\right] = log\left[\sum_z \mathbb{Q}(z)\frac{\mathbb{P}(x,z;\theta)}{\mathbb{Q}(z)}\right]$$

Using Jensen's inequality, we get

$$\geq \sum_z \mathbb{Q}(z) log\left[\frac{\mathbb{P}(x,z;\theta)}{\mathbb{Q}(z)}\right].$$

So for any distribution $\mathbb{Q}$, the above formula gives a lower-bound on $log\mathbb{P}(x;\theta)$.

The above inequality could be tighten to equality by taking,

$$\frac{\mathbb{P}(x, z; \theta)}{\mathbb{Q}(z)} = c,$$

c is a constant does not depend on z.
So taking,

$$\mathbb{Q}(z) \propto \mathbb{P}(x, z; \theta)$$

will give us the above equality. Further, we know that $\sum_z \mathbb{Q}(z) = 1$.

So, we know that

$$\mathbb{Q}(z) = \frac{\mathbb{P}(x, z; \theta)}{\sum_z \mathbb{P}(x, z; \theta)} = \frac{\mathbb{P}(x, z; \theta)}{\mathbb{P}(x; \theta)} = \mathbb{P}(z|x; \theta).$$

With the above value of $\mathbb{Q}(z)$ we get,

$$\begin{aligned}\sum_z \mathbb{Q}(z) log \left[\frac{\mathbb{P}(x, z; \theta)}{\mathbb{Q}(z)}\right] &= \sum_z \mathbb{P}(z|x; \theta) log \frac{\mathbb{P}(z|x; \theta)\mathbb{P}(x; \theta)}{\mathbb{P}(z|x; \theta)} \\ &= log \left[\mathbb{P}(x; \theta)\right].\end{aligned}$$

The evidence lower bound

$$ELBO(x; \mathbb{Q}, \theta) = \sum_z \mathbb{Q}(z) log \left[\frac{\mathbb{P}(x, z; \theta)}{\mathbb{Q}(z)} \right].$$

So,

$$\forall \mathbb{Q}, \theta, x, \quad log \mathbb{P}(x; \theta) \geq ELBO(x; \mathbb{Q}, \theta).$$

EM algorithm updates $\mathbb{Q}$ and θ by:

a. setting $\mathbb{Q}(z) = \mathbb{P}(z|x; \theta)$.

b. maximizing $ELBO(x; \mathbb{Q}, \theta)$ w.r.t θ while fixing the choice of $\mathbb{Q}$.

EM convergence

Now, we are interesting on answering the question, will the EM algorithm converge?
Let $\theta^{(t)}$ and $\theta^{(t+1)}$ are the parameters from two successive iterations of EM. We are interested in showing that EM always monotonically improves the log-likelihood.

$$log\left[\mathcal{L}(\theta^{(t)})\right] = \sum_{i=1}^{n} ELBO(x^{(i)}, \mathbb{Q}_i^{(t)}, \theta^{(t)})$$

$$log\left[\mathcal{L}(\theta^{(t+1)})\right] \geq \sum_{i=1}^{n} ELBO(x^{(i)}; \mathbb{Q}_i^{(t)}, \theta^{(t+1)})$$

$$\geq \sum_{i=1}^{n} ELBO(x^{(i)}; \mathbb{Q}_i^{(t)}, \theta^{(t)}) = log\left[\mathcal{L}(\theta^{(t)})\right]$$

We have seen ELBO given as,

$$ELBO(x; Q, \theta) = \sum_z Q(z) \log \frac{p(x, z; \theta)}{Q(z)}$$

.

But, there are several other forms of ELBO:

$$\begin{aligned} ELBO(x; Q, \theta) &= E_{z \sim Q}[\log p(x, z; \theta)] - E_{z \sim Q}[\log Q(z)] \\ &= E_{z \sim Q}[\log p(x|z; \theta)] - D_{KL}(Q||p_z), \end{aligned}$$

where:

- D_{KL} is the KL divergence given by:

$$D_{KL}(Q||p_z) = \sum_z Q(z) \log \frac{Q(z)}{p(z)}.$$

Also,

$$ELBO(x; Q, \theta) = \log p(x) - D_{KL}(Q||p_{z|x}).$$

Now, writing the log-likelihood for our Gaussian Mixture model, we will have:

$$log\left[\mathcal{L}(\phi, \mu, \Sigma)\right] = \sum_{i=1}^{n} \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \phi, \Sigma, \mu)}{Q_i(z^{(i)})}$$

$$= \sum_{i=1}^{n} \sum_{j}^{k} Q_i(z^{(i)} = j) \log \frac{p(x^{(i)}|z^{(i)} = j; \mu, \Sigma) p(z^{(i)} = j; \phi)}{Q_i(z^{(i)} = j)}$$

$$= \sum_{i=1}^{n} \sum_{j}^{k} w_j^{(i)} \log \frac{\frac{1}{(2\pi)^{(d/2)}|\Sigma_j|^{(1/2)}} \exp\left(-\frac{1}{2}(x^{(i)} - \mu_j)^T \Sigma_j^{(-1)} (x^{(i)} - \mu_j)\right).\phi_j}{w_j^{(i)}}.$$

Appendix

For the proves, please reach the below link:
https://drive.google.com/file/d/1cL_g0Xttf0aaBb2dV1IrwrkVxd0zQb1M/view?usp=share_link

For the GitHub implementation:
https://github.com/ndams55/Gaussian_Mixture_Model_AMMI_2023

References

Brian S. Everitt.
Cluster Analysis.
Wiley Press, 2001.

Andrew Ng.
CS229 Lecture notes.