Expectation Maximization and Gaussian Mixture Models

Emir Esenov

Uppsala University

2023

EM Algorithm review

- Recall the EM algorithm, an iterative process based on the "embedding principle"
- Use surrogate function $Q(\theta \,|\, \theta') = \mathbb{E}\left[\ln f(\mathbf{X} \,|\, \theta) \mid T(\mathbf{X}), \theta'\right]$
- θ is a free parameter in the likelihood function, and θ' is the parameter belonging to the underlying distribution.
- We are interested in finding

$$\widehat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta} Q(\theta \,|\, \theta_{true})$$

EM Algorithm review

- **E-Step:** Given the current state $\theta^{(k)}$, calculate $Q(\theta \mid \theta^{(k)})$.
- **M-Step:**
$$\theta^{(k+1)} = \arg\max_{\theta} Q(\theta \mid \theta^{(k)}).$$
- Want to converge to a good estimate

Gaussian Mixture Model (GMM) — motivating example

- Simple modeling assumption: data comes from Gaussian distribution

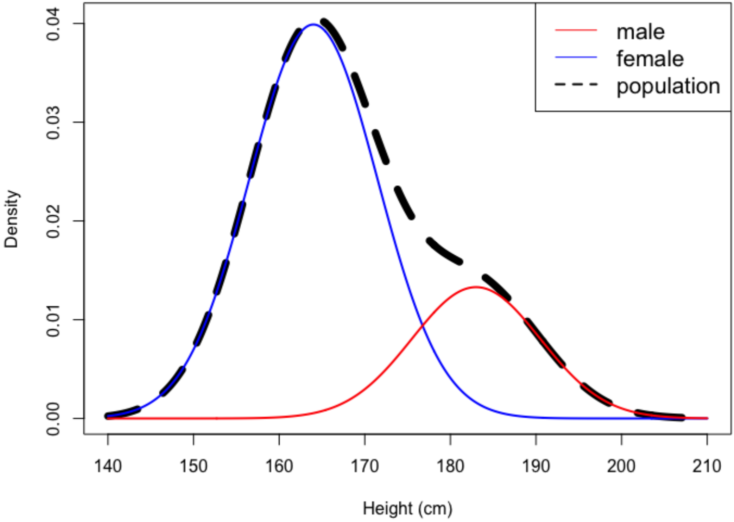
$$
\begin{aligned}
L(\mu) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x_i-\mu)^2}{2\sigma^2}\right) \\
\Rightarrow l(\mu) &= \left[\sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(x_i-\mu)^2}{2\sigma^2}\right] \\
\Rightarrow \frac{d}{d\mu} l(\mu) &= \sum_{i=1}^{n} \frac{x_i-\mu}{\sigma^2}.
\end{aligned}
$$

- Setting this equal to zero and solving for $\mu \Rightarrow \mu_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Similarly, we can derive $\sigma^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$

Gaussian Mixture Model (GMM) — motivating example

- So, we can find MLE of the parameters in closed form for simple Gaussian distribution case
- What about more complex cases?
- Consider the case where we measure the height of a given population made up of two sub-populations where 75% are female with height distribution $\mathcal{N}(164, 7.5^2)$, and 25% are male with height $\mathcal{N}(183, 7.5^2)$

Gaussian Mixture Model (GMM) — motivating example



Gaussian Mixture Model (GMM) — motivating example

- Assume we observe $X_1, \ldots, X_n$ and that each X_i is sampled from one of K mixture components
- In the example above, the mixture components are $\{$male,female$\}$
- Introduce a latent variable associated with each r.v. X_i, $Z_i \in \{1, \ldots, K\}$ which indicates which component X_i came from

Gaussian Mixture Model (GMM) — motivating example

- From the law of total probability, we know that the marginal probability of X_i is:

$$\begin{aligned} P(X_i = x) &= \sum_{k=1}^{K} \underbrace{P(Z_i = k)}_{\pi_k} P(X_i = x | Z_i = k) \\ &= \sum_{k=1}^{K} \pi_k P(X_i = x | Z_i = k) \\ &= \sum_{k=1}^{K} \pi_k N(x; \mu_k, \sigma_k^2) \end{aligned}$$

- Similarly, the joint probability of observations $X_1, \ldots, X_n$ is therefore:

$$P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k N(x_i; \mu_k, \sigma_k^2)$$

Gaussian Mixture Model (GMM) — motivating example

- Now we attempt the same strategy for deriving the MLE of our GMM. Our unknown parameters are $\theta = \{\mu_1, \ldots, \mu_K, \sigma_1, \ldots, \sigma_K, \pi_1, \ldots, \pi_K\}$, and so our likelihood is:

$$L(\theta | X_1, \ldots, X_n) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k N(x_i; \mu_k, \sigma_k^2)$$

And our log-likelihood is:

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

Gaussian Mixture Model (GMM) — motivating example

- If we try follow the same steps to derive the MLE of μ_k as before, we end up with the expression

$$\sum_{i=1}^{n} \frac{1}{\sum_{k=1}^{K} \pi_k N(x_i; \mu_k, \sigma_k)} \pi_k N(x_i; \mu_k, \sigma_k) \frac{(x_i - \mu_k)}{\sigma_k^2} = 0$$

- No closed form solution, we are stuck
- However, if we knew the latent variables Z_i then we could simply gather all X_i such that $Z_i = k$ which would give us a closed form solution

Gaussian Mixture Model (GMM) — motivating example

- So, we want to know the latent variables Z_i. First, compute the posterior distribution

$$
\begin{aligned}
P(Z_i = k | X_i) &= \frac{P(X_i | Z_i = k) P(Z_i = k)}{P(X_i)} \\
&= \frac{\pi_k N(\mu_k, \sigma_k^2)}{\sum_{k=1}^{K} \pi_k N(\mu_k, \sigma_k^2)} \\
&= \gamma_{Z_i}(k)
\end{aligned}
$$

A vicious cycle:

1. If we knew the parameters, we could compute the posterior probabilities $\gamma_{Z_i}(k)$
2. If we knew the posteriors $\gamma_{Z_i}(k)$, we could easily compute the parameters

EM in GMM

- Solution: Expectation Maximization!
- Rewrite the former expression where we took the log-likelihood w.r.t μ_k as follows

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

- Trick: even though $\gamma_{Z_i}(k)$ depends on μ_k, pretend that it doesn't and solve for μ_k in this equation to get:

$$\hat{\mu_k} = \frac{\sum_{i=1}^{n} \gamma_{z_i}(k) x_i}{\sum_{i=1}^{n} \gamma_{z_i}(k)} = \frac{1}{N_k} \sum_{i=1}^{n} \gamma_{z_i}(k) x_i,$$

where we set $N_k = \sum_{i=1}^{n} \gamma_{z_i}(k)$. We can think of N_k as the effective number of points assigned to component k.

EM in GMM

- Similarly, if we apply a similar method to finding $\hat{\sigma_k^2}$ and $\hat{\pi_k}$, we get that:

$$\hat{\sigma_k^2} = \frac{1}{N_k} \sum_{i=1}^{n} \gamma_{z_i}(k)(x_i - \mu_k)^2$$

$$\hat{\pi_k} = \frac{N_k}{n}$$

EM in GMM

The EM algorithm now goes as follows:

1. Initialize the μ_k's, σ_k's and π_k's and evaluate the log-likelihood with these parameters
2. **E-step** Evaluate the posterior probabilities $\gamma_{Z_i}(k)$ using the current values of the μ_k's and σ_k's with equation
3. **M-step** Estimate new parameters $\hat{\mu_k}$, $\hat{\sigma_k^2}$ and $\hat{\pi_k}$ with the current values of $\gamma_{Z_i}(k)$
4. Evaluate the log-likelihood, continue iterating from step 2 until convergence

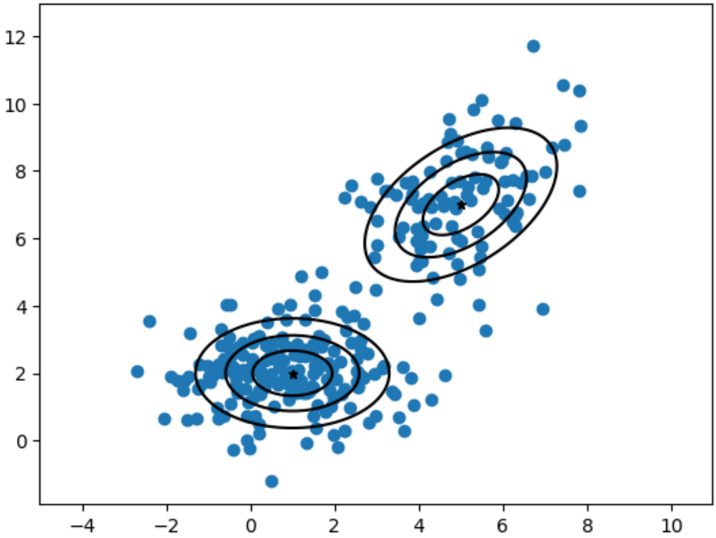
GMM in action — sample model

Consider a data set $X \in \mathbb{R}^{300\times 2}$ of 300 samples that are generated from a two-dimensional Gaussian mixture model with mixture weights $\pi_1 = 0.7$ and $\pi_2 = 0.3$ and mixture components with parameters

$$\mu_1 = \begin{bmatrix} 5 \\ 7 \end{bmatrix}, \quad \Sigma_1 = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, \quad \mu_2 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}.$$

Let's try to estimate these parameters from generated samples

GMM in action — sample model



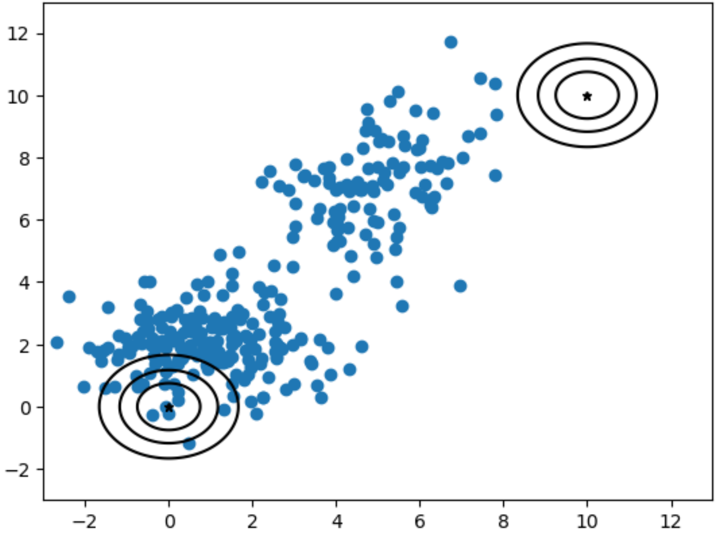
GMM in action — initialization

Start off with an initial guess

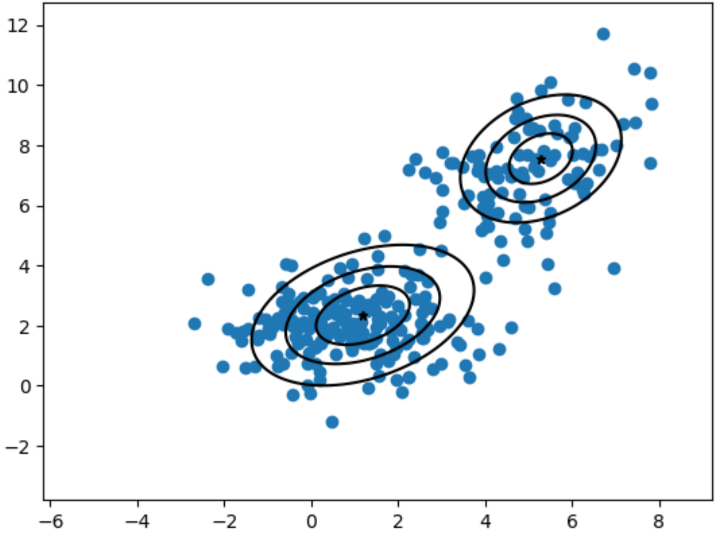
$$\widehat{\pi}_1 = 0.5, \widehat{\pi}_2 = 0.5$$

$$\widehat{\mu}_1 = \begin{bmatrix} 10 \\ 10 \end{bmatrix}, \quad \widehat{\Sigma}_1 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}, \quad \widehat{\mu}_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \widehat{\Sigma}_2 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}.$$

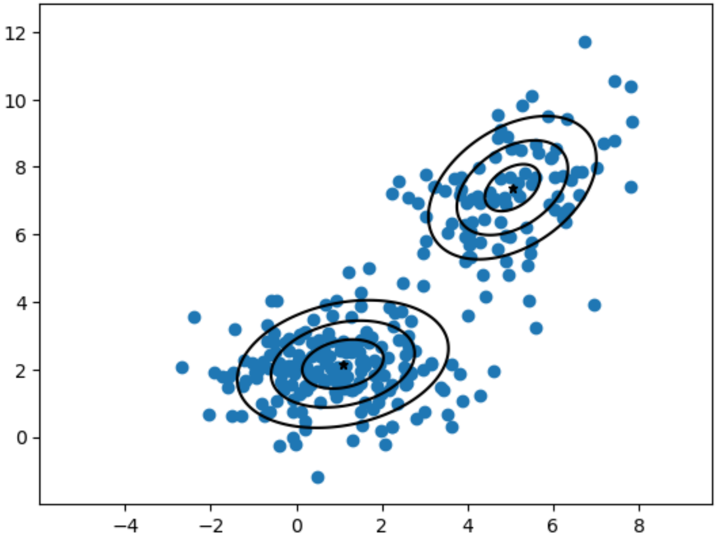
GMM in action — iteration 1



GMM in action — iteration 2



GMM in action — iteration 3



GMM in action — result

After three iterations, we end up with the following estimates for the results

$$\widehat{\pi}_1 = 0.68, \widehat{\pi}_2 = 0.32$$

$$\widehat{\mu}_1 = \begin{bmatrix} 4.9 \\ 7.3 \end{bmatrix}, \ \widehat{\Sigma}_1 = \begin{bmatrix} 1.6 & 0.8 \\ 0.8 & 1.9 \end{bmatrix}, \ \widehat{\mu}_2 = \begin{bmatrix} 1.0 \\ 2.1 \end{bmatrix}, \ \widehat{\Sigma}_2 = \begin{bmatrix} 2.2 & 0.2 \\ 0.2 & 1.1 \end{bmatrix}.$$

Compare to original distribution

$$\pi_1 = 0.7, \pi_2 = 0.3$$

$$\mu_1 = \begin{bmatrix} 5 \\ 7 \end{bmatrix}, \ \Sigma_1 = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, \ \mu_2 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \ \Sigma_2 = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}.$$