Expectation Maximization and Gaussian Mixture Models

Emir Esenov
Uppsala University
emiresenov96@gmail.com

***Abstract*—We outline the Expectation Maximization (EM) algorithm and its application in Gaussian Mixture Models (GMM). Finally, we implement a GMM classifier to analyze the Iris flower dataset. The classifier was trained for 20 iterations with the EM algorithm and achieved 97.4% accuracy.**

I. The Expectation Maximization algorithm

The expectation-maximization (EM) algorithm [1] is an algorithm based on the embedding principle, where we have data from a complicated model that we embed in larger model with desirable properties to which we can apply straightforward methods. In EM this allows us to find maximum likelihood estimates (MLE) of parameters in models with latent variables, and the embedding principle finds expression as follows: suppose we have observations

$$\mathbf{Y} = (Y_1, \ldots, Y_m)$$

from the observed model with density $g(\mathbf{y}|\theta)$, and we want to find the MLE of θ. We interpret the observations as a transformation

$$\mathbf{Y} = T(\mathbf{X}), \tag{1}$$

where

$$\mathbf{X} = (X_1, \ldots, X_n).$$

The trick is to use the conditional expectation of the log-likelihood function $\ln f(\mathbf{x}|\theta)$ as an estimation criterion. The conditional mean projects the likelihood function from the complete model on to the observed model — we call this the surrogate function and define it as

$$Q(\theta \,|\, \theta^{'}) = \mathbb{E}\left[\ln f(\mathbf{X} \,|\, \theta) \,|\, T(\mathbf{X}), \theta^{'}\right]. \tag{2}$$

Here we have that θ is a free parameter in the likelihood function, and $\theta^{'}$ is the parameter belonging to the underlying distribution. We are interested in finding

$$\widehat{\theta}_{\text{MLE}} = \arg\max_{\theta} Q(\theta \,|\, \theta_{true}) \tag{3}$$

in cases where T is an insufficient statistic. Note that when T is insufficient, the conditional expectation of $\mathbf{X}$ given $T(\mathbf{X})$ depends on the underlying parameter θ_{true}. Conversely, in cases where T is sufficient, minimizing the surrogate function equates to minimizing the likelihood function $l(\theta) = \ln g(\mathbf{y}|\theta)$, so this is an uninteresting case.

For problems where it is difficult to solve (3), we can make a guess about the complete data $\mathbf{X}$ and solve for the θ that maximizes (2) for our guess. With this new θ, we can make a better guess about the complete data $\mathbf{X}$, and continue to iterate to try and converge to a good estimate. This is EM as outlined in Algorithm 1.

Algorithm 1: EM Algorithm

E-step: Given the current state $\theta^{(k)}$, calculate $Q(\theta \,|\, \theta^{(k)})$.

M-step:

$$\theta^{(k+1)} = \arg\max_{\theta} Q(\theta \,|\, \theta^{(k)}).$$

Note that there is no guarantee that EM finds the global optimum likelihood; the only guarantee is that the EM estimate never gets worse. In practice, this means that we can either find global maxima, local maxima, or get stuck on saddle points. This property is called the monotonicity of the EM algorithm and follows from Theorem 1.

Theorem 1. *Let $L(\theta) = \ln f(\mathbf{x} \,|\, \theta)$ be the log-likelihood function. If $Q(\theta \,|\, \theta^{(k)}) \geq Q(\theta^{(k)} \,|\, \theta^{(k)})$, then $L(\theta) \geq L(\theta^{(k)})$.*

For the EM algorithm, the M-step ensures that

$$\theta^{(k+1)} = \arg\max_{\theta} Q(\theta \,|\, \theta^{(k)}),$$

hence it must be that

$$Q(\theta^{(k+1)} \,|\, \theta^{(k)}) \geq Q(\theta^{(k)} \,|\, \theta^{(k)}).$$

Therefore, we can apply Theorem 1 and conclude that

$$L(\theta^{(k+1)}) \geq L(\theta^{(k)}).$$

We omit the proof for this theorem and defer to [2]. Despite lacking global convergence guarantees, EM still proves useful in many settings.

II. Gaussian Mixture Models

A. Review of the Gaussian distribution and a motivating example

We can sometimes make simple modeling assumptions about our data, such as assuming that each observation comes from a Gaussian distribution. Recall that in such cases, we

can derive an analytical solution for the MLE of the model parameters: suppose we have n observations $X_1, \ldots, X_n$ from a Gaussian distribution with unknown mean and variance σ^2, then we have that

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

Setting this equal to zero and solving for μ, we get that $\mu_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} x_i$, the average of our samples —— note that the estimate does not depend on the variance σ^2. Similarly, we can use the log-likelihood of σ^2 and derive that $\sigma^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$ for which we would plug in $\mu = \mu_{\text{MLE}}$.

The MLE of the parameters can thus be derived in closed form when the modeling assumption is that all samples are drawn from a single Gaussian distribution. However, we may encounter situations where samples come from different distributions and there are groupings in the data. In such cases, we would like to make more complex modeling assumptions. As an example, consider the results in figure 1, where the height of a given population is made up of two sub-populations, where 75% are female with a height distribution of $\mathcal{N}(164, 7.5^2)$, and 25% are male with a height distribution of $\mathcal{N}(183, 7.5^2)$. In this case, the distribution is not normally distributed, and we would instead like to model the measurements as a linear combination (a mixture) of two components, each with a Gaussian distribution. This is a Gaussian Mixture Model (GMM).

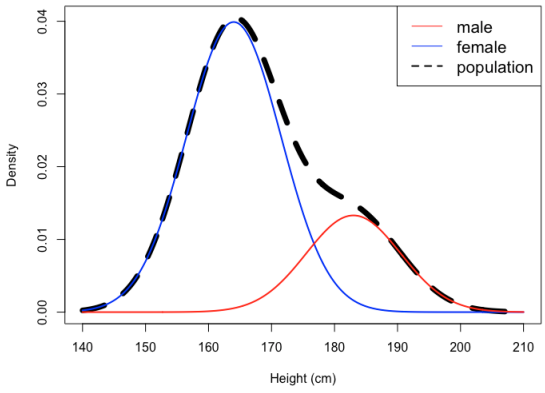

Fig. 1. Height measurements from a sample population.

The above example gives rise to the general notion of a mixture model, which assumes that each observation is generated from one of K mixture components. Formally [3]–[5], suppose again that we have n observations $X_1, \ldots, X_n$ for each person's height, and that each X_i is sampled from one of K mixture components. In our example from Figure 1, the mixture components were {male, female}. Now, we introduce a latent variable $Z_i \in \{1, \ldots, K\}$ which indicates which component X_i came from. We have 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:

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

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:

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

If we try follow the same steps as before and differentiate with respect to μ_k, we obtain the expression

$$\sum_{i=1}^{n} \frac{\pi_k N(x_i; \mu_k, \sigma_k)}{\sum_{k=1}^{K} \pi_k N(x_i; \mu_k, \sigma_k)} \frac{(x_i - \mu_k)}{\sigma_k^2} = 0 \tag{4}$$

which contains a mix of ratios of exponentials and linear terms and has no closed-form solution like the one we obtained for the simple Gaussian distribution. However, note that 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. Consider the posterior distribution for the latent variable Z_i :

$$\begin{aligned} P(Z_i = k | X_i) &= \frac{P(X_i | Z_i = k) P(Z_i = k)}{P(X_i)} && (5) \\ &= \frac{\pi_k N(\mu_k, \sigma_k^2)}{\sum_{k=1}^{K} \pi_k N(\mu_k, \sigma_k^2)} && (6) \\ &= \gamma_{Z_i}(k). && (7) \end{aligned}$$

We can use this to rewrite (4) as

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

Now we have a cycle: if we knew the parameters, we could compute the posterior probablities $\gamma_{Z_i}(k)$, and if we knew the posteriors $\gamma_{Z_i}(k)$, then we could compute the parameters θ. We deploy the EM algorithm to tackle this circular problem.

B. EM algorithm in GMM

For the GMM, the likelihood is [4]:

$$P(X, Z|\mu, \sigma, \pi) = \prod_{i=1}^{n} \prod_{k=1}^{K} \pi_k^{I(Z_i=k)} N(x_i|\mu_k, \sigma_k)^{I(Z_i=k)}$$

which gives us the log-likelihood

$$\sum_{i=1}^{n} \sum_{k=1}^{K} I(Z_i = k) \left(\log(\pi_k) + \log(N(x_i|\mu_k, \sigma_k))\right).$$

Here, the latent variables are unobserved, so we consider the expectation of the log-likelihood with respect to the posterior of the latent variables. The expected value of the log-likelihood is thus:

$$E_{Z|X}[\log(P(X, Z|\mu, \sigma, \pi))] \tag{8}$$

$$= E_{Z|X}\left[\sum_{i=1}^{n} \sum_{k=1}^{K} I(Z_i = k) \left(\log(\pi_k) + \log(N(x_i|\mu_k, \sigma_k))\right)\right] \tag{9}$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} E_{Z|X}[I(Z_i = k)] \left(\log(\pi_k) + \log(N(x_i|\mu_k, \sigma_k))\right) \tag{10}$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_{Z_i}(k) \left(\log(\pi_k) + \log(N(x_i|\mu_k, \sigma_k))\right) \tag{11}$$

EM proceeds as follows: given values for μ, σ, π (initialized with a rough approximation for the first iteration), use these in the E-step to evaluate the $\gamma_{Z_i}(k)$. Then, with $\gamma_{Z_i}(k)$ fixed, perform the M-step by maximizing the expected log-likelihood (11) with respect to μ_k, σ_k and π_k. This leads to the closed-form solutions:

$$\widehat{\mu}_k = \frac{1}{N_k} \sum_{i=1}^{n} \gamma_{z_i}(k) x_i, \quad \widehat{\sigma}_k^2 = \frac{1}{N_k} \sum_{i=1}^{n} \gamma_{z_i}(k)(x_i - \mu_k)^2$$

$$\widehat{\pi}_k = \frac{N_k}{n}.$$

III. Experiment

We implement a GMM to perform cluster analysis and classification on the Iris flower dataset [1].

A. Data analysis

We analyze the data in Figure (2) and Table (I). We note that petal length and petal width are highly correlated and nicely grouped, and we use these two features for the GMM.

B. Implementation

We implement the GMM algorithm and train it for 20 iterations in an unsupervised setting. Figure (3) shows the EM algorithm at work during different stages of iteration in the training process. Next, we perform classification with the GMM and obtain 97.4% test accuracy on a hold-out set.

[1] https://en.wikipedia.org/wiki/Iris_flower_data_set

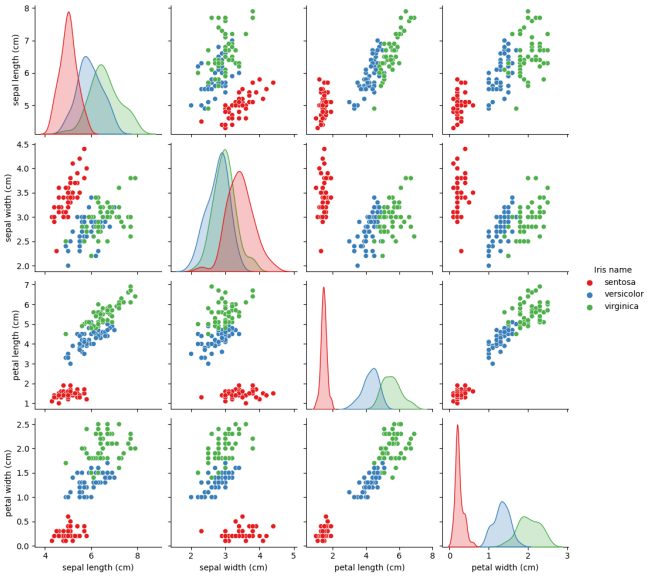

Fig. 2. Pairplot of Iris dataset features.

	sepal length	sepal width	petal length	petal width
sepal length	1	-0.12	0.87	0.82
sepal width	-0.12	1	-0.43	-0.37
petal length	0.87	-0.43	1	0.96
petal width	0.82	-0.37	0.96	1

TABLE I
FEATURE CORRELATION MATRIX.

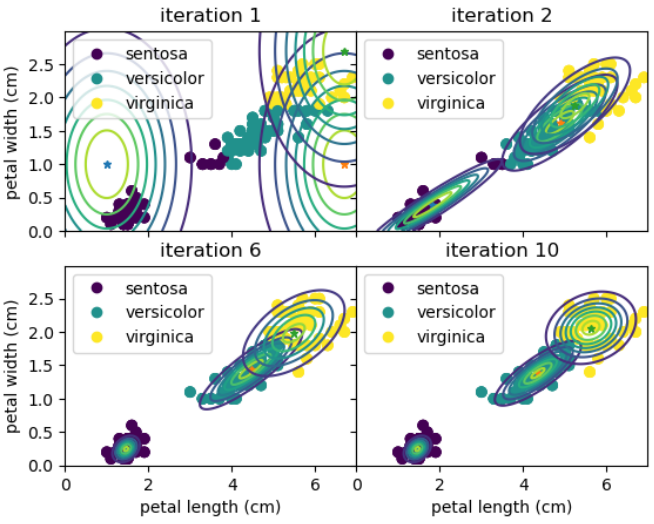

Fig. 3. Training of the GMM.

IV. Code

The code for this project can be found at:
https://github.com/emiresenov/CISDM-Project/tree/main.

REFERENCES

[1] S. Zwanzig and B. Mahjani, *Computer Intensive Methods in Statistics*. Taylor & Francis Group, 12 2019.

[2] Y. Chen and M. R. Gupta, "Em demystified: An expectation-maximization tutorial," 2010. [Online]. Available: https://api.semanticscholar.org/CorpusID:26878370

[3] "Introduction to mixture models," https://stephens999.github.io/fiveMinuteStats/intro_to_mixture_models.html, accessed: 2023-12-12.

[4] "Introduction to em: Gaussian mixture models," https://stephens999.github.io/fiveMinuteStats/intro_to_em.html, accessed: 2023-12-12.

[5] R. Sridharan, "Gaussian mixture models and the em algorithm," 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:16820177