Gaussian Mixture Models (GMMs)

Kaiwen Zhou

Contents

1 Topic: Gaussian Mixture Models

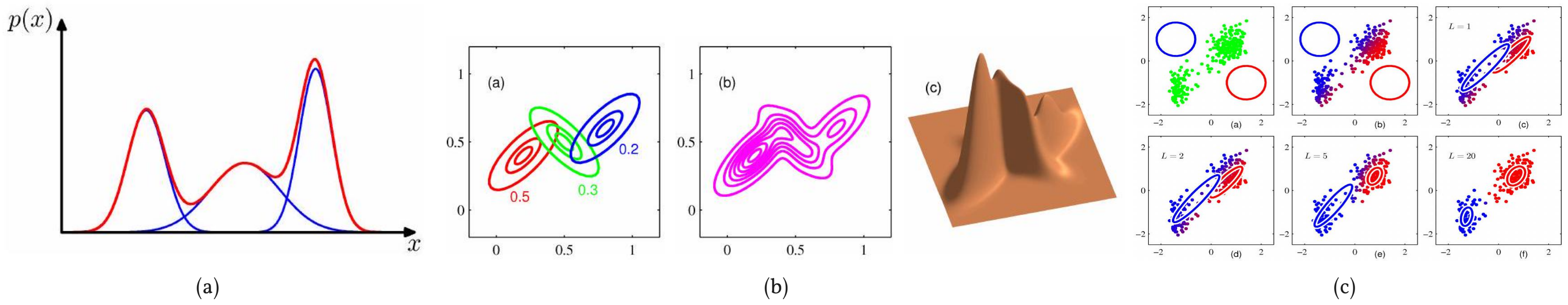

Figure 1: Gaussian Mixture Models
(a) One-dimensional Gaussian mixture distribution $p(x)$ formed from the sum of three Gaussians.
(b) Mixture of three Gaussians in a two dimensions. (b1) Contours of constant density for each mixture component with the values of the mixing coefficients. (b2) Contours of the marginal probability density $p(\mathrm{x})$ of the mixture distribution. (b3) Surface plot of the distribution $p(\mathbf{x})$.
(c) The EM algorithm applied to the Old Faithful dataset.

Mixture Models:

The idea of Mixture Models comes from the fact that almost any continuous density can be approximated to arbitrary accuracy by using a sufficient number of simple distributions (e.g. Gaussians); see Figure 1. Here, we will focus mainly on Gaussian components. But we note that General mixture models can built from linear combinations of other distributions.

Gaussians Mixture Model:

The Gaussians mixture model (GMM), or more Specifically, K-mixture of Gaussians, is a linear combination of K Gaussian densities of the form

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \tag{1}$$

where each Gaussian density $\mathcal{N}\left(\mathrm{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)$ is called a **component of the mixture** and has its own mean $\boldsymbol{\mu}_k$ and covariance $\boldsymbol{\Sigma}_k$. The parameters π_k in are called **mixing coefficients.**

To make $p(\mathrm{x})$ a proper density, we must have

$$\int p(\mathbf{x}) d\mathbf{x} = 1,\ \text{and}\ p(\mathbf{x}) \geqslant 0, \forall \mathbf{x} \quad \xrightarrow[\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\geqslant 0,\ \forall k]{\int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)d\mathbf{x}=1} \quad \sum_{k=1}^{K} \pi_k = 1,\ \text{and}\ 0 \leqslant \pi_k \leqslant 1, \forall k$$

On the other hand, the joint distribution $p(\mathbf{x}, k)$ is given by

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(\mathbf{x}, k) = \sum_{k=1}^{K} p(k) p(\mathbf{x} \mid k)$$

We observe that this is equivalent to Equation (1) where

- $p(k) \equiv \pi_k$ is the prior probability of choosing the k-th component, and
- $p(\mathbf{x} \mid k) \equiv \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)$ is the probability of x conditioned on k.

Responsibilities $\gamma_k(\mathrm{x})$:

Define the responsibility that component k takes for 'explaining' the observation x, $\gamma_k(\mathrm{x})$ to be the posterior $p(k \mid \mathrm{x})$ given by the Bayes' theorem

$$\gamma_k(\mathbf{x}) := p(k \mid \mathbf{x}) = \frac{p(k) p(\mathbf{x} \mid k)}{\sum_l p(l) p(\mathbf{x} \mid l)} = \frac{\pi_k \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}{\sum_l \pi_l \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l\right)}$$

Sample from GMMs:

To sample from a Gaussian mixture, for each data point we use the following steps:

1. Pick a component $k \in \{1, \ldots, K\}$ with probabilities $\{\pi_1, \ldots, \pi_K\}$
2. Draw a sample $\mathbf{x}_n \sim p(\mathbf{x} \mid k) \equiv \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$

Maximum Likelihood for GMMs:

Given a data set $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$, we use maximum likelihood estimator to estimate the parameters $\boldsymbol{\pi} := \{\pi_1, \ldots, \pi_K\}$, $\boldsymbol{\mu} := \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\}$ and $\boldsymbol{\Sigma} := \{\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K\}$ in the Gaussian mixture distribution $p(\mathbf{x}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.

The log-likelihood function is given by

$$\log p(\mathbf{x} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^N \log \left\{ \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

Due to the presence of the summation over k inside the logarithm, the maximum likelihood solution for the parameters do not have a closed-form solution. One can maximize the likelihood function with gradient based techniques; however, the caveat is that it might not converge. And an alternative is to use expectation maximization (EM) algorithm.

2 Topic: Formulating GMMs Using Discrete Latent Variables

Now, we will find an equivalent formulation of the Gaussian mixture distribution using explicit, discrete latent variables $\mathbf{z}$. The latent variable representation will then help us formulate the expectation-maximization (EM) algorithm.

Let us consider the Gaussian mixture distribution

$$p(\mathbf{x}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

1-of-K representation of $\mathbf{z}$:

Define $\mathbf{z} \in \mathbb{R}^K$ as a binary random vector where only one element z_k is equal to 1 and all other elements are equal to 0. (Think standard basis of dimension K, and we use $z_k = 1$ to denote $\mathbf{z}$ being the k-th standard basis.)

Discrete Latent Variables $\mathbf{z}$:

Define the marginal distribution $p(\mathbf{z})$ to be extended-Bernoulli and the conditional distribution $p(\mathbf{x} \mid \mathbf{z})$ to be Gaussian respectively as

$$\begin{cases} p(z_k = 1) & := \pi_k \\ p(\mathbf{x} \mid z_k = 1) & := \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \end{cases} \implies \begin{cases} p(\mathbf{z}) & := \prod_{k=1}^K \pi_k^{z_k} \\ p(\mathbf{x} \mid \mathbf{z}) & := \prod_{k=1}^K \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k} \end{cases} \tag{2}$$

where $\sum_{k=1}^K \pi_k = 1$, $0 \leqslant \pi_k \leqslant 1$.

Formulate GMM using $\mathbf{z}$:

Since $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{z}) p(\mathbf{x} \mid \mathbf{z})$, we can always define a joint distribution $p(\mathbf{x}, \mathbf{z})$ in terms of its marginal and conditional distributions. Therefore, the marginal distribution $p(\mathbf{x})$ by summing the joint distribution over all possible states of $\mathbf{z}$, that is

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z}) p(\mathbf{x} \mid \mathbf{z}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

We see that this marginal distribution is a GMM. In particular, for every observed data point $\mathbf{x}_n$ in a set of observations $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, we can pick a corresponding latent variable $\mathbf{z}_n$ as in a way described in Equation (2).

The responsibility that component k takes in explaining an observation $\mathbf{x}$, $\gamma(z_k)$,is then

$$\gamma(z_k) \equiv p(z_k = 1 \mid \mathbf{x}) = \frac{p(z_k = 1) p(\mathbf{x} \mid z_k = 1)}{\sum_{j=1}^K p(z_j = 1) p(\mathbf{x} \mid z_j = 1)} = \frac{\pi_k \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

Interpretation: π_k is the prior probability of $z_k = 1$, and $\gamma(z_k)$ is the posterior probability once we have observed $\mathbf{x}$.

Remark 2.1. To see this clearly, we have

$$\mathbf{z} \sim \begin{cases} (1, 0, \ldots, 0), & \mathbb{P} = \pi_1 \\ (0, 1, \ldots, 0), & \mathbb{P} = \pi_2 \\ & \vdots \\ (0, 0, \ldots, 1), & \mathbb{P} = \pi_K \end{cases} \iff \begin{cases} p(z_1 = 1) & = \pi_1 \\ p(z_2 = 1) & = \pi_2 \\ & \vdots \\ p(z_K = 1) & = \pi_K \end{cases} \iff p(\mathbf{z}) := \prod_{k=1}^K \pi_k^{z_k}$$

△

3 Topic: EM Algorithm in GMMs

To estimate the GMM, we will want to maximize the log-likelihood function

$$\log p(\mathbf{x} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^N \log \left\{ \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\} \tag{3}$$

First-order condition for $\boldsymbol{\mu}_k$:

Taking the partial derivative of Equation (3) with respect to $\boldsymbol{\mu}_k$, we obtain the first order condition

$$0 = \sum_{n=1}^{N} \underbrace{\frac{\pi_k \mathcal{N}\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}{\sum_j \pi_j \mathcal{N}\left(\mathbf{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)}}_{\gamma(z_{nk})} \boldsymbol{\Sigma}_k^{-1}\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right) \quad \Longrightarrow \quad \boldsymbol{\Sigma}_k^{-1} \cdot \left(\sum_{n=1}^{N} \gamma\left(z_{nk}\right) \mathbf{x}_n\right) = \boldsymbol{\Sigma}_k^{-1} \cdot \left(\sum_{n=1}^{N} \gamma\left(z_{nk}\right)\right) \boldsymbol{\mu}_k \quad \Longrightarrow \quad \boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma\left(z_{nk}\right) \mathbf{x}_n$$

where $N_k := \sum_{n=1}^{N} \gamma\left(z_{nk}\right)$.

Interpretation:

- N_k is the "effective number of points" assigned to the k-th cluster; and

- The mean $\boldsymbol{\mu}_k$ of the k-th Gaussian component is a weighted mean of all of the points in the data set for which the weighting factor for the data point, $\mathbf{x}_n$, is given by the posterior probability $\gamma\left(z_{nk}\right)$ that component k was responsible for generating $\mathbf{x}_n$.

First-order condition for $\boldsymbol{\Sigma}_k$:

Taking the partial derivative of Equation (3) with respect to $\boldsymbol{\Sigma}_k$, it is easy to see that we obtain the first order condition

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma\left(z_{nk}\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)^{\top}$$

Note that this equation is of the same form as that of a single Gaussian trained on the data set, but with (a) each data point weighted by the corresponding posterior probability and (b) the denominator given by the effective number of points associated with the corresponding component.

First-order condition for π_k:

Let us maximize the log likelihood function with respect to π_k, subject to the constraint $\sum_{k=1}^{K} \pi_k = 1$. The resulting Lagrangian takes the form

$$L := \log p(\mathbf{x} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \lambda\left(\sum_{k=1}^{K} \pi_k - 1\right)$$

where λ is a Lagrange multiplier. Taking the partial derivative of (14) with respect to π_k, we obtain the first order condition

$$0 = \frac{\partial L}{\partial \pi_k} = \lambda + \sum_{n=1}^{N} \frac{\mathcal{N}\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}{\sum_j \pi_j \mathcal{N}\left(\mathbf{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)} = \lambda + \frac{1}{\pi_k} \sum_{n=1}^{N} \gamma\left(z_{nk}\right) \tag{4}$$

Multiply both sides of Equation (4) by π_k and then summing over k, we obtain

$$0 = \sum_{k=1}^{K}\left(\lambda \pi_k + \pi_k \cdot \frac{1}{\pi_k} \sum_{n=1}^{N} \gamma\left(z_{nk}\right)\right) = \lambda\left(\sum_{k=1}^{K} \pi_k\right) + \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma\left(z_{nk}\right) \overset{\sum_{k=1}^{K} \pi_k=1}{\underset{\sum_{k=1}^{K} z_{nk}=1}{=\!=\!=\!=\!=}} \lambda + \sum_{n=1}^{N} 1 \quad \Longrightarrow \quad \lambda = -N$$

Inserting $\lambda = -N$ into Equation (4), we conclude that

$$\pi_k = \frac{N_k}{N}$$

Interpretation: The mixing coefficient for the k-th component is given by the average of the responsibilities which that component takes for explaining the data points.

In all, we have the first order necessary conditions are

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma\left(z_{nk}\right) \mathbf{x}_n, \quad \boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma\left(z_{nk}\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)^{\top}, \quad \pi_k = \frac{N_k}{N},$$

We note that this is not a closed-form solution for the parameters of the GMM. (Why?)

However, we can try to solve the problem using an iterative scheme as follows. This scheme is a version of the Expectation-Maximization (EM) algorithm applied to the GMM.

Algorithm 1 EM Algorithm for Gaussian Mixtures

```
procedure EM-GMMs( $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k, \mathbf{x}_n$)
    $\delta, tol = 1, 0.005$
    $\boldsymbol{\mu}_k \leftarrow \boldsymbol{\mu}_k^0$
    $\boldsymbol{\Sigma}_k \leftarrow \boldsymbol{\Sigma}_k^0$
    $\pi_k \leftarrow \pi_k^0$
    $L \leftarrow \sum_{n=1}^N \log \left\{ \sum_{k=1}^K \pi_k^0 \mathcal{N}\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k^0, \boldsymbol{\Sigma}_k^0\right) \right\}$
    while $\delta \geq tol$ do
        E step: Evaluate the responsibilities using the current parameter values
        $\gamma\left(z_{nk}\right) \leftarrow \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)}$
        M step: Re-estimate the parameters using the current responsibilities
        $N_k \leftarrow \sum_{n=1}^N \gamma\left(z_{nk}\right)$
        $\boldsymbol{\mu}_k^{\text{new}} \leftarrow \frac{1}{N_k} \sum_{n=1}^N \gamma\left(z_{nk}\right) \mathbf{x}_n$
        $\boldsymbol{\Sigma}_k^{\text{new}} \leftarrow \frac{1}{N_k} \sum_{n=1}^N \gamma\left(z_{nk}\right) \left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right) \left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right)^\top$
        $\pi_k^{\text{new}} \leftarrow \frac{N_k}{N}$
        $L^{\text{new}} \leftarrow \sum_{n=1}^N \log \left\{ \sum_{k=1}^K \pi_k^{\text{new}} \mathcal{N}\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k^{\text{new}}, \boldsymbol{\Sigma}_k^{\text{new}}\right) \right\}$
        $\delta \leftarrow |L^{\text{new}} - L|$
        $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k, L \leftarrow \boldsymbol{\mu}_k^{\text{new}}, \boldsymbol{\Sigma}_k^{\text{new}}, \pi_k^{\text{new}}, L^{\text{new}}$
    return $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k$
```

In general, The log likelihood function of GMMs are not convex, and hence there will be multiple local maxima. While EM is not guaranteed to find the largest of these, it is guaranteed to converge monotonically to a local optima. See Bishop (2006, Section 9.4) for a proof.

EM Algorithm in GMMs v.s. K-Means:

Remark 3.1.

1. In comparison, the EM algorithm for GMMs takes many more iterations to converge than standard K-means clustering. Each iteration of the EM algorithm is also significantly more expensive. To speed up convergence it is common to first perform K-means clustering to find suitable starting points, and then perform the EM algorithm from these.
2. Note that (a) the covariance matrices can be initialized to the sample covariances of the clusters found by K-means, and (b) the mixing coefficients can be set to the fractions of data points assigned to the respective clusters.

△

References

[1] Bishop, Christopher M. (2006). Pattern Recognition And Machine Learning. Springer. URL: https://www.microsoft.com/enus/research/uploads/prod/2006/01/Bishop-PatternRecognition-and-Machine-Learning-2006.pdf.