University of Illinois Fall 2018

ECE 544NA: Pattern Recognition
Lecture 17: Gaussian Mixture Models

Lecturer: Alexander Schwing Scribe: Haoran Qi

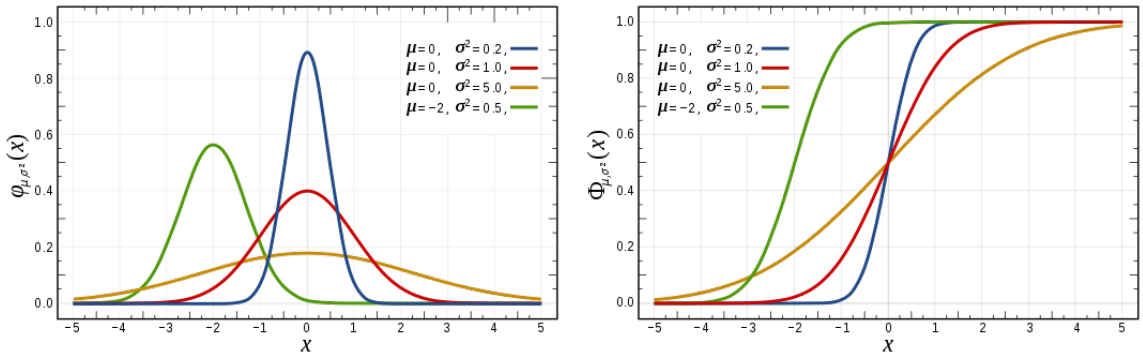
1 Overview

Lecture 17 introduces Gaussian Mixture Models, which is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. The goal of this lecture includes:

- Understanding Gaussian mixture models
- Getting to know more details about generative modeling
- Learning the relationship between Gaussian mixture models and k-Means

1.1 Introduction to Gaussian Distribution

In probability theory, the Gaussian (or normal) distribution is a common continuous probability distribution. Gaussian distributions are important in statistics and are often used in the natural and social sciences to represent real-valued random variables whose distributions are not known. A random variable with a Gaussian distribution is said to be normally distributed and is called a normal deviate [1].



(a) PDF of Gaussian Distribution (b) CDF of Gaussian Distribution

Figure 1: PDF and CDF of Gaussian Distribution

With Gaussian distribution, statistics becomes much easier, and more feasible. Central Limit Theorem [**?**], which is that when independent random variables are added, their properly normalized sum tends toward a normal distribution (informally a "bell curve") even if the original variables themselves are not normally distributed. It implies we can use Gaussian distribution to simplify the problem we want to solve.

1.2 Model the distribution of the data $X^{(i)}$

Our problem is that given a dataset $\mathcal{D} = \{X^{(i)}\}$, find $\theta = (\mu, \sigma)$ of distribution

$$p(X^{(i)}|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{1}{2\sigma^2}(X^{(i)} - \mu)^2) \tag{1}$$

Minimize negative log-likelihood to solve it, then

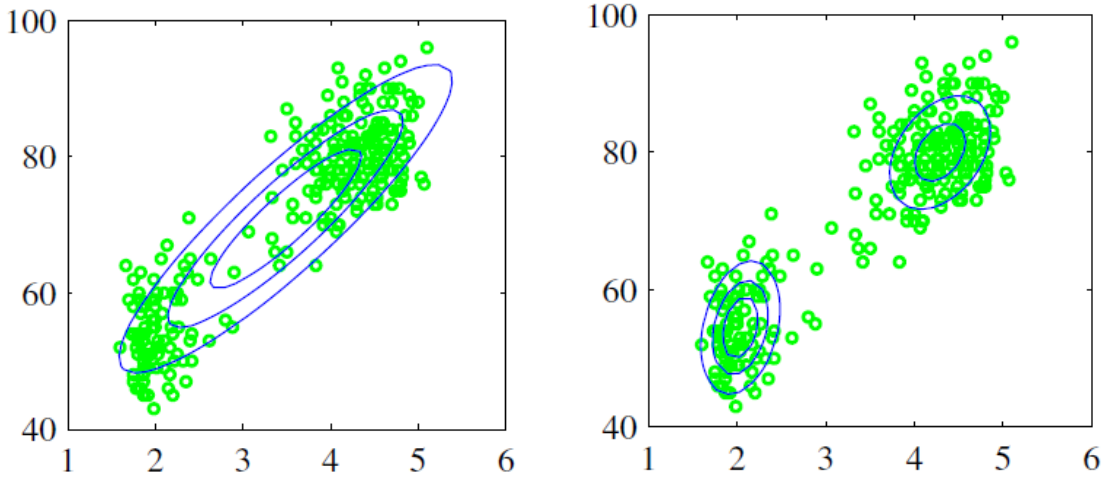
$$min - log \prod_{i\in\mathcal{D}} p(X^{(i)}|\mu,\sigma)) := \sum_{i\in\mathcal{D}} -\frac{1}{2\sigma^2}(X^{(i)} - \mu)^2 + \frac{N}{2} log(2\pi\sigma^2)) \tag{2}$$

Set the partial derivative for μ and σ to zero respectively, then

$$\frac{\partial}{\partial\mu} = \frac{1}{2\sigma^2}\sum_{i\in\mathcal{D}}(X^{(i)} - \mu) = 0 \quad \Rightarrow \quad \mu = \frac{1}{N}\sum_{i\in\mathcal{D}} X^{(i)} \tag{3}$$

$$\frac{\partial}{\partial\sigma} = -\frac{1}{\sigma^3}\sum_{i\in\mathcal{D}}(X^{(i)} - \mu)^2 + \frac{N}{\sigma} = 0 \quad \Rightarrow \quad \sigma^2 = \frac{1}{N}\sum_{i\in\mathcal{D}}(X^{(i)} - \mu)^2 \tag{4}$$

So the distribution can be determined using equation (3) and (4). However, single Gaussian model is not flexible enough to handle the problem.



(a) Single Gaussian (b) Two Gaussians

Figure 2: Single Gaussian Model is not flexible enough

2 Latent Variable Models

Model with hidden variables are called latent variable models or LVMs. Such models are harder to fit than models with no latent variables. However, they can have significant advantages for two main reasons: [1]

- LVMs often have fewer parameters than models that directly represent correlation in the visible space, so it's easier to compute
- The hidden variables in an LVM can serve as a bottleneck, which computes a compressed representation of the data, which forms the basis of unsupervised learning

3 Gaussians Mixture Model

3.1 Mixture of Gaussians

Gaussian mixture models are used commonly when the underlying populations can be explained by a normal distribution and there are many heterogeneous populations.

To add more flexibility, we use linear superposition of Gaussians. In this model, each base distribution in the mixture is a multivariate Gaussian with mean μ_k and covariance matrix σ_k. Thus the model has the form

$$p(X^{(i)}|\pi,\mu,\sigma) = \sum_{k=1}^{K} \pi_k \mathcal{N}(X^{(i)}|\mu_k,\sigma_k), \text{ s.t.} \sum_{k=1}^{K} \pi_k = 1, \pi_k \geq 0 \tag{5}$$

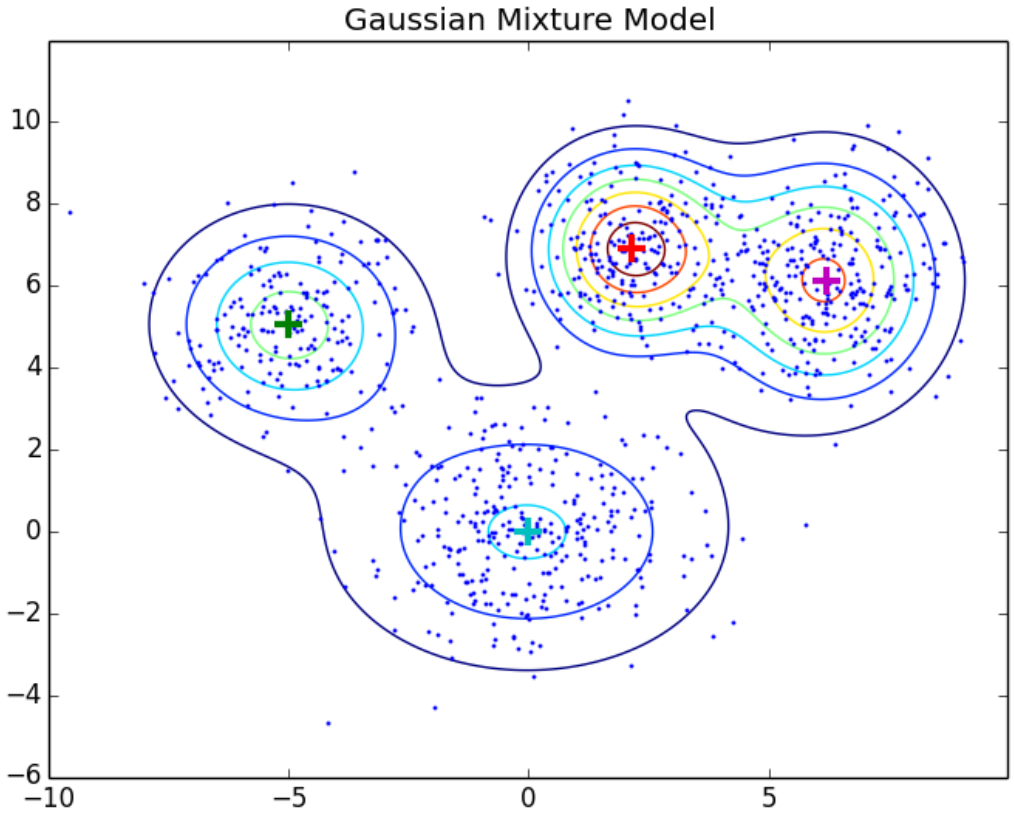


Figure 3: Gaussian Mixture Model

To find π, μ, σ, minimize negative log-likelihood:

$$min - log \prod_{i \in \mathcal{D}} p(X^{(i)}|\pi, \mu, \sigma)) := - \sum_{i \in \mathcal{D}} log \sum_{k=1}^{K} \pi_k \mathcal{N}(X^{(i)}|\mu_k, \sigma_k) \tag{6}$$

3.2 Optimization of GMM

Use latent variable $Z_{ik} \in \{0, 1\}$, with $\sum_{k=1}^{K} Z_{ik} = 1, \forall i$.

The marginal distribution for Z_{ik} is:

$$p(Z_{ik} = 1) = \pi_k \quad \Rightarrow \quad p(\mathbf{Z}_i) = \prod_{k=1}^{K} \pi_k^{Z_{ik}}, \text{ where } \mathbf{Z}_i = [Z_{i1}, ..., Z_{ik}]^T \tag{7}$$

And the conditional distribution for Z_{ik} is:

$$p(X^{(i)}|Z_{ik} = 1) = \mathcal{N}(X^{(i)}|\mu_k, \sigma_k) \tag{8}$$

Then the marginal distribution for $X_{(i)}$ is:

$$\begin{aligned} p(X^{(i)}|\pi, \mu, \sigma) &= \sum_{\mathbf{Z}_i} p(X^{(i)}|\mathbf{Z}_i) p(\mathbf{Z}_i) \\ &= \sum_{\mathbf{Z}_i} \prod_{k=1}^{K} \pi_k^{Z_{ik}} \mathcal{N}(X^{(i)}|\mu_k, \sigma_k)^{Z_{ik}} \\ &= \sum_{k=1}^{K} \pi_k \mathcal{N}(X^{(i)}|\mu_k, \sigma_k) \end{aligned} \tag{9}$$

Use Bayes' theorem, calculate the posterior:

$$\begin{aligned} r_{ik} &= p(Z_{ik} = 1|X^{(i)}) \\ &= \frac{p(Z_{ik} = 1) p(X^{(i)}|Z_{ik} = 1)}{\sum_{\widehat{k}=1}^{K} p(Z_{i\widehat{k}} = 1) p(X^{(i)}|Z_{i\widehat{k}} = 1)} \\ &= \frac{\pi_k \mathcal{N}(X^{(i)}|\mu_k, \sigma_k)}{\sum_{\widehat{k}=1}^{K} \pi_{\widehat{k}} \mathcal{N}(X^{(i)}|\mu_{\widehat{k}}, \sigma_{\widehat{k}})} \end{aligned} \tag{10}$$

With equation (7), (8), (9), (10), equation (6) now is

$$min_{\pi,\mu,\sigma} - log \prod_{i \in \mathcal{D}} p(X^{(i)}|\pi, \mu, \sigma)) := - \sum_{i \in \mathcal{D}} log \sum_{k=1}^{K} \pi_k \mathcal{N}(X^{(i)}|\mu_k, \sigma_k), \quad \text{s.t.} \sum_{k=1}^{K} \pi_k = 1 \tag{11}$$

Set the partial derivative for μ and σ to zero respectively, and define per cluster weight $N_k = \sum_{i \in \mathcal{D}} r_{ik}$, then

$$\frac{\partial}{\partial \mu_k} = \frac{1}{2\sigma_k^2} \sum_{i \in \mathcal{D}} r_{ik}(X^{(i)} - \mu) = 0 \quad \Rightarrow \quad \mu_k = \frac{1}{N_k} \sum_{i \in \mathcal{D}} r_{ik} X^{(i)} \tag{12}$$

$$\frac{\partial}{\partial \sigma_k} = \sum_{i \in \mathcal{D}} r_{ik}\left(\frac{1}{\sigma_k} - \frac{1}{\sigma_k^2}(X^{(i)} - \mu_k)^2\right) = 0 \quad \Rightarrow \quad \sigma_k^2 = \frac{1}{N}\sum_{i \in \mathcal{D}}(X^{(i)} - \mu_k)^2 \tag{13}$$

For π_k, we introduce the Lagrange multiplier

$$\frac{\partial}{\partial \pi_k}: \quad \sum_{i \in \mathcal{D}} \frac{\mathcal{N}(X^{(i)}|\mu_k, \sigma_k)}{\sum_{\widehat{k}=1}^{K} \pi_{\widehat{k}} \mathcal{N}(X^{(i)}|\mu_{\widehat{k}}, \sigma_{\widehat{k}})} + \lambda = 0 \tag{14}$$

For equation(14), multiply both sides with π_k and sum over k, then right side becomes $-\lambda$ and left side becomes N:

$$\sum_k \pi_k \sum_{i \in \mathcal{D}} \frac{\mathcal{N}(X^{(i)}|\mu_k, \sigma_k)}{\sum_{\widehat{k}=1}^{K} \pi_{\widehat{k}} \mathcal{N}(X^{(i)}|\mu_{\widehat{k}}, \sigma_{\widehat{k}})} = -\lambda \sum_k \pi_k \qquad \Rightarrow \quad \lambda = -N \tag{15}$$

With equation(10), (14), and (15), we can get the way of calculating π_k:

$$\pi_k = \frac{N_k}{N} \tag{16}$$

3.3 Gaussian Mixture Model Algorithm:

Input: $X^{(i)}$

Output: μ, σ, π

Initialize μ, σ, π randomly

Iterate:

E-step: Update

$$r_{ik} = \frac{\pi_k \mathcal{N}(X^{(i)}|\mu_k, \sigma_k)}{\sum_{\widehat{k}=1}^{K} \pi_{\widehat{k}} \mathcal{N}(X^{(i)}|\mu_{\widehat{k}}, \sigma_{\widehat{k}})}$$

S-step: Update

$$\mu_k = \frac{1}{N_k}\sum_{i \in \mathcal{D}} r_{ik} X^{(i)}$$

$$\sigma_k^2 = \frac{1}{N}\sum_{i \in \mathcal{D}}(X^{(i)} - \mu_k)^2$$

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

return $con(r_i)$;

Algorithm 1: Gaussian Mixture Model algorithm

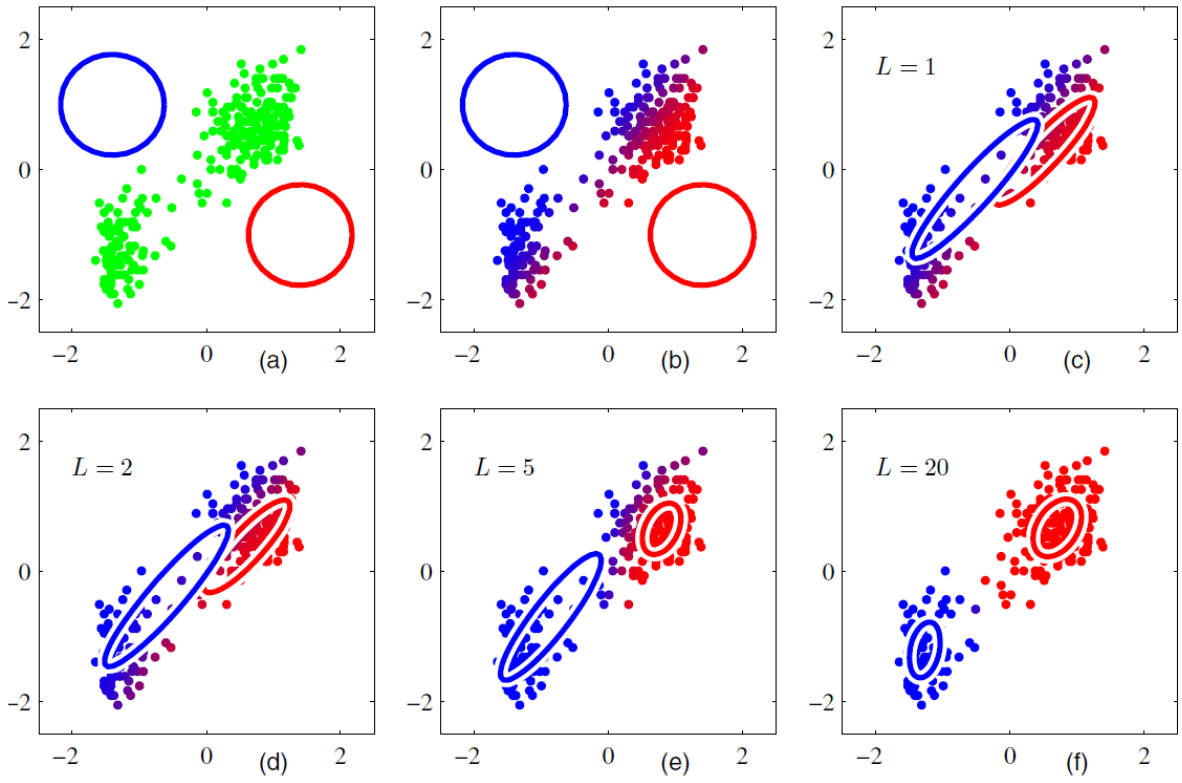


Figure 4: The convergence of Gaussian Mixture Model

4 GMM and K-Means

4.1 Similarity

1. r_{ik} is an assignment of sample i to cluster k
2. μ_k are the cluster centers

Let $\sigma_k^2 = \epsilon,\ \forall k$, then equation (10) becomes

$$r_{ik} = \frac{\pi_k \exp(-\frac{1}{2\epsilon}(X^{(i)} - \mu_k)^2)}{\sum_{\hat{k}=1}^{K} \pi_{\hat{k}} \exp(-\frac{1}{2\epsilon}(X^{(i)} - \mu_{\hat{k}})^2)} \tag{17}$$

Discussion: When $\epsilon \to 0$, all terms of the denominator will go to zero except the one for which $(X^{(i)} - \mu_{\hat{k}})^2$ is smallest, which will go to unity, responsibilities will go from soft assignments to hard assignments and cost function will be identical. At this time, the GMM looks like the same as K-means

4.2 Comparison

4.2.1 K-Means

K-Means is an algorithm, which classifies samples based on attributes/features into K number of clusters. [?, ?, ?, ?, ?] Clustering of samples is done by minimizing the distance between sample

and the center. i.e. Assign the center and optimize the center based on the distances from the points to it. This is called as Hard Assignment i.e. We are certain that particular points belong to particular center and then based on the least squares distance method, we will optimize the place of the center.

Advantages of K-Means:

- Better for high dimensional data
- Easy to interpret and Implement

Disadvantages of K-Means:

- Sensitive to outliers: Outliers will also be assigned to specific clusters, which distorts the corresponding cluster centers
- Hard Assignment might lead to mis-grouping
- Need to choose the number of clusters K
- Can get stuck in local minima: K-Means are only guaranteed to converge to local minimum, and the result can be arbitrarily bad/wrong

4.2.2 Guassian Mixture Model

Guassian Mixture Model is used if we are uncertain about the data points where they belong or to which group, instead of hard assigning data points to a cluster. It uses probability of a sample to determine the feasibility of it belonging to a cluster [?, ?, ?, ?, ?, ?].

Advantages:

- Does not assume clusters to be of any geometry. Works well with non-linear geometric distributions as well
- Does not bias the cluster sizes to have specific structures as does by K-Means (Circular)

Disadvantages:

- Uses all the components it has access to, so initialization of clusters will be difficult when dimensionality of data is high
- Difficult to interpret

5 Summary

In this lecture, we introduced Gaussian Mixture Model, which is a probabilistic model that each base distribution in the mixture is a multivariate Gaussian.

Quiz1: The maximum likelihood solution of fitting the mean and variance of a Gaussian?

Solution:

$$\mu = \frac{1}{N} \sum_{i \in \mathcal{D}} X^{(i)}$$

$$\sigma^2 = \frac{1}{N}\sum_{i\in\mathcal{D}}(X^{(i)} - \mu)^2$$

Quiz2: Why do we consider mixtures of Gaussians?

Solution: Single Gaussian model is not flexible enough.

Quiz3: How do we find the means, variances and responsibilities of the Gaussian mixture model?

Solution: r_{ik} correspends to the responsibility of kth Gaussian in the mixture model, and we have:

$$r_{ik} = \frac{\pi_k \mathcal{N}(X^{(i)}|\mu_k, \sigma_k)}{\sum_{\widehat{k}=1}^{K} \pi_{\widehat{k}} \mathcal{N}(X^{(i)}|\mu_{\widehat{k}}, \sigma_{\widehat{k}})}$$

For the means and variances, we have:

$$\mu_k = \frac{1}{N_k}\sum_{i\in\mathcal{D}} r_{ik} X^{(i)}$$

$$\sigma_k^2 = \frac{1}{N}\sum_{i\in\mathcal{D}}(X^{(i)} - \mu_k)^2$$

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

References

[1] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.