GMM Study Notes

Zhengyang Chen{chenzhengyang117@gmail.com}

December 24, 2018

Abstract

In this note,I write the points which I think important in the process of getting familiar with GMM.

Contents

1 Knowledge about Probability

1.1 The rules of Probability

Sum Rule:

$$p(X) = \sum_{Y} p(X, Y)$$

Product Rule:

$$p(X, Y) = p(X|Y)p(Y)$$

Independence:

$$p(X, Y) = p(X)p(Y)$$

1.2 Bayes' Theorem

$$p(X) = \sum_{Y} p(X|Y)p(Y)$$

$$p(Y|X) = \frac{p(X|Y)p(Y)}{(X)}$$

$$posterior \propto likelihood * prior$$

1.3 Concept comparision

Posterior $p(Y|X)$ v.s. conditional $p(X|Y)$
Marginal $p(X)$ v.s. prior $p(Y)$
Joint probability $p(X, Y)$

2 Knowledge about GMM

2.1 Definition

A Gaussian Mixture Model (GMM) is a parametric probability density function represented as a **weighted sum** of Gaussian component densities. GMMs are commonly used as a parametric model of the probability distribution of continuous measurements or features in a biometric system, such as vocal-tract related spectral features in a speaker recognition system. GMM parameters are estimated from training data using the iterative **Expectation-Maximization** (EM) algorithm or **Maximum A Posteriori** (MAP) estimation from a well-trained prior model.

2.2 Mathematical formula

A Gaussian mixture model is a weighted sum of M component Gaussian densities as given by the equation,

$$p(\mathbf{x}|\theta) = \sum_{m=1}^{M} c_m * \mathcal{N}(\mathbf{x}|\mu_{\mathbf{m}}, \mathbf{\Sigma_m}) \tag{1}$$

where $\mathbf{x}$ is a D-dimensional continuous-valued data vector (i.e. measurement or features),c_m , i = 1, . . . , M , are the mixture weights, and $\mathcal{N}(\mathbf{x}; \mu, \mathbf{\Sigma})$is the component Gaussian densities. Each component density is a D-variate Gaussian function of the form,

$$\mathcal{N}(\mathbf{x}; \mu, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}|^{1/2}} exp(-\frac{1}{2}(\mathbf{x} - \mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x} - \mu)) \tag{2}$$

with mean vector $\mu_{\mathbf{m}}$ and covariance matrix $\mathbf{\Sigma_m}$. The mixture weights satisfy the constraint that $\sum_{m=1}^{M} c_m = 1$.

The complete Gaussian mixture model is parameterized by the **mean vectors**, **covariance matrices** and **mixture weights** from all component densities. These parameters are collectively represented by the notation.

$$\theta = \{c_m, \mu_{\mathbf{m}}, \mathbf{\Sigma_m}\}, m = 1, ..., M$$

3 Expectation Maximization

3.1 Maximum Likelihood

There are several techniques available for estimating the parameters of a GMM. By far the most popular and well-established method is maximum likelihood (ML) estimation.

The aim of ML estimation is to find the model parameters which maximize the likelihood of the GMM given the training data. For a sequence of N training vectors $\mathbf{X} = \mathbf{x}_1, ..., \mathbf{x}_N$, the GMM likelihood, assuming independence between the vectors , can be written as,

$$p(\mathbf{X}|\theta) = \prod_{n=1}^{N} p(\mathbf{x}_n|\theta)$$

However,the equation is difficult to differentiate,we use the log on both sides

$$\mathcal{L}(\theta) = log(p(\mathbf{X}|\theta)) = \sum_{n=1}^{N} log(\sum_{m=1}^{M} c_m \mathcal{N}(\mathbf{x}_n; \mu_m, \mathbf{\Sigma}_m)) \tag{3}$$

3.2 Auxiliary function

The equation (3) is still difficult to differentiate.To ease this problem,we will build our auxiliary function.
We have known that if F is a upper convex function like log function,we have

$$F(\sum_{i=1}^{N} \lambda_i x_i) \geq \sum_{i=1}^{N} \lambda_i F(x_i) \tag{4}$$

Using the rule above,we build our auxiliary function:

$$\begin{aligned}
\mathcal{L}(\theta) = log(p(\mathbf{X}|\theta)) = \sum_{n=1}^{N} log(\sum_{m=1}^{M} p(\mathbf{x}_n, m|\theta) \\
&= \sum_{n=1}^{N} log \sum_{m=1}^{M} p(m|\mathbf{x}_n, \hat{\theta}) \frac{p(\mathbf{x}_n, m|\theta)}{p(m|\mathbf{x}_n, \hat{\theta})} \\
&\geq \sum_{n=1}^{N} \sum_{m=1}^{M} p(m|\mathbf{x}_n, \hat{\theta}) * log \frac{p(\mathbf{x}_n, m|\theta)}{p(m|\mathbf{x}_n, \hat{\theta})} \\
&= \sum_{n=1}^{N} H(p(m|\mathbf{x}_n, \hat{\theta})) + \sum_{n=1}^{N} \sum_{m=1}^{M} p(m|\mathbf{x}_n, \hat{\theta}) * log(p(\mathbf{x}_n, m|\theta)) \\
&= \Phi(\theta, \hat{\theta})
\end{aligned}$$

Suppose:

$$Q(\theta, \hat{\theta}) = \sum_{n=1}^{N} \sum_{m=1}^{M} p(m|\mathbf{x}_n, \hat{\theta}) * log(p(\mathbf{x}_n, m|\theta))$$

And:

$$\begin{aligned}
\mathcal{L}(\hat{\theta}) &= \sum_{n=1}^{N} log \sum_{m=1}^{M} p(m|\mathbf{x}_n, \hat{\theta}) \frac{p(\mathbf{x}_n, m|\hat{\theta})}{p(m|\mathbf{x}_n, \hat{\theta})} \\
&= \sum_{n=1}^{N} log \sum_{m=1}^{M} p(m|\mathbf{x}_n, \hat{\theta}) \frac{p(\mathbf{x}_n, m|\hat{\theta})}{p(\mathbf{x}_n, m|\hat{\theta})/p(\mathbf{x}_n|\hat{\theta})} \\
&= \sum_{n=1}^{N} log \sum_{m=1}^{M} p(m|\mathbf{x}_n, \hat{\theta}) p(\mathbf{x}_n|\hat{\theta}) \\
&= \sum_{n=1}^{N} p(\mathbf{x}_n|\hat{\theta})
\end{aligned}$$

The same reason:

$$\Phi(\hat{\theta}, \hat{\theta}) = \sum_{n=1}^{N} p(\mathbf{x}_n|\hat{\theta})$$

Namely:

$$\mathcal{L}(\hat{\theta}) = \Phi(\hat{\theta}, \hat{\theta})$$

According to $\mathcal{L}(\theta) \geq \Phi(\theta, \hat{\theta})$, we'll get

$$\mathcal{L}(\theta) - \mathcal{L}(\hat{\theta}) \geq \Phi(\theta, \hat{\theta}) - \Phi(\hat{\theta}, \hat{\theta})$$

which means when we find a better θ for $\Phi(\theta, \hat{\theta})$ or $Q(\theta, \hat{\theta})$,we find a better θ for $\mathcal{L}(\theta)$ at the same time.

3.3 Get θ when maximize $Q(\theta, \hat{\theta})$

Now we want get:

$$\theta_{new} = argmax_{\theta} Q(\theta, \hat{\theta})$$

And we can get θ_{new} by setting $\dfrac{\partial Q}{\partial \theta} = 0$, suppose:

$$\gamma_m(n) = p(m|\mathbf{x}_n, \hat{\theta}) = \frac{p(\mathbf{x}_n|m, \hat{\theta})p(m|\hat{\theta})}{\sum_{k=1}^{M} p(\mathbf{x}_n|k, \hat{\theta})p(k|\hat{\theta})}$$

The finally expression and constraint:

$$\begin{aligned}
Q(\theta, \hat{\theta}) &= \sum_{n=1}^{N} \sum_{m=1}^{M} p(m|\mathbf{x}_n, \hat{\theta}) * log(p(\mathbf{x}_n, m|\theta)) \\
&= \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_m(n) * log(c_m * p(\mathbf{x}_n|m, \mathbf{\Sigma}_m, \mu_m)) \\
&= \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_m(n) * log(p(\mathbf{x}_n|m, \mathbf{\Sigma}_m, \mu_m)) + \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_m(n) * log(c_m)
\end{aligned}$$

$$s.t. \sum_{m=1}^{M} c_m = 1$$

To maximize this expression, we can maximize the term containing c_m and the term containing $(\mathbf{\Sigma}_m, \mu_m)$ independently since they are not related.

3.3.1 Maximize w.r.t c_m

To find the expression for c_m, we introduce the Lagrange multiplier λ with the constraint that $\sum_{m=1}^{M} c_m = 1$, and solve the following equation:

$$\frac{\partial}{\partial c_m} \left[\sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_m(n) * log(c_m) + \lambda(\sum_{m=1}^{M} c_m - 1) \right] = 0$$

or

$$\sum_{n=1}^{N} \frac{\gamma_m(n)}{c_m} + \lambda = 0$$

Summing both sizes over m, we get that $\lambda = -N$ resulting in:

$$c_m = \frac{1}{N}\sum_{n=1}^{N}\gamma_m(n)$$

3.3.2 Maximize w.r.t $(\mathbf{\Sigma}_m, \mu_m)$

In this situation,we have

$$p(\mathbf{x}_n|m, \mathbf{\Sigma}_m, \mu_m) = \frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}_m|^{1/2}} exp(-\frac{1}{2}(\mathbf{x}_n - \mu_m)^T\mathbf{\Sigma}_m^{-1}(\mathbf{x}_n - \mu_m)) \quad (5)$$

To derive the update equations for this distribution, we need to recall some results from matrix algebra.

The trace of a square matrix $tr(A)$ is equal to the sum of A's diagonal elements. The trace of a scalar equals that scalar. Also, $tr(A + B) = tr(A) + tr(B)$, and $tr(AB) = tr(BA)$ which implies $\Sigma_i x_i^T A x_i = tr(AB)$ where $B = \Sigma_i x_i x_i^T$. Also note that $|A|$ indicates the determinant of a matrix, and that $|A|^{-1} = 1/|A|$.

Useful formulas of matrix calculus:

$$\frac{\partial tr(AX)}{\partial X} = A^T \quad (6)$$

$$\frac{\partial |X|}{\partial X} = |X|(X^{-1})^T, \frac{\partial ln|X|}{\partial X} = (X^{-1})^T \quad (7)$$

$$\frac{\partial (X^T A X)}{\partial X} = X^T(A + A^T) \quad (8)$$

We use the Equation 5 to replace the specific part of the $Q(\theta, \hat{\theta})$ which has $(\mathbf{\Sigma}_m, \mu_m)$,

$$\sum_{n=1}^{N}\sum_{m=1}^{M}\gamma_m(n) * log(p(\mathbf{x}_n|m, \mathbf{\Sigma}_m, \mu_m)) \quad (9)$$

$$= C + \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{M}\gamma_m(n) * [log(|\mathbf{\Sigma}_m|) + (\mathbf{x}_n - \mu_m)^T\mathbf{\Sigma}_m^{-1}(\mathbf{x}_n - \mu_m)] \quad (10)$$

Taking the derivative of Equation 10 with respect to μ_m and setting it equal to zero, we get:

$$\sum_{n=1}^{N}(\mathbf{x}_n - \mu_m)^T * (\mathbf{\Sigma}_m^{-1} + (\mathbf{\Sigma}_m^{-1})^T) * \gamma_m(n) = 0$$

$$\sum_{n=1}^{N}(\mathbf{x}_n - \mu_m)^T * 2 * \mathbf{\Sigma}_m^{-1} * \gamma_m(n) = 0$$

$$\mu_m^T \sum_{n=1}^{N} \gamma_m(n) = \sum_{n=1}^{N} \mathbf{x}_n * \gamma_m(n)$$

$$\mu_m^T = \frac{\sum_{n=1}^{N} \mathbf{x}_n * \gamma_m(n)}{\sum_{n=1}^{N} \gamma_m(n)} \tag{11}$$

Suppose: $A_{mn} = (\mathbf{x}_n - \mu_m)(\mathbf{x}_n - \mu_m)^T$, we can get

$$(\mathbf{x}_n - \mu_m)^T \mathbf{\Sigma}_m^{-1}(\mathbf{x}_n - \mu_m) = tr(\mathbf{\Sigma}_m^{-1} A_{mn}) \tag{12}$$

Taking the derivative of Equation 7 with respect to $\mathbf{\Sigma}_m^{-1}$ (not the $\mathbf{\Sigma}_m$) and setting it equal to zero, we get:

$$\sum_{n=1}^{N}[\mathbf{\Sigma}_m - A_{mn}^T] * \gamma_m(n) = 0$$

$$\mathbf{\Sigma}_m \sum_{n=1}^{N} \gamma_m(n) = \sum_{n=1}^{N} A_{mn} * \gamma_m(n)$$

$$\mathbf{\Sigma}_m = \frac{\sum_{n=1}^{N} \gamma_m(n)(\mathbf{x}_n - \mu_m)(\mathbf{x}_n - \mu_m)^T}{\sum_{n=1}^{N} \gamma_m(n)} \tag{13}$$

3.4 EM steps

- Expectation(E-step): Calculate posterior

$$\gamma_m(n) = p(m|\mathbf{x}_n, \hat{\theta}) = \frac{p(\mathbf{x}_n|m, \hat{\theta})p(m|\hat{\theta})}{\sum_{k=1}^{M} p(\mathbf{x}_n|k, \hat{\theta})p(k|\hat{\theta})}$$

- Maximization (M-step):Find parameters which maximize the auxiliary function $Q(\theta, \hat{\theta})$

$$\gamma_m = \sum_{n=1}^{N} \gamma_m(n) \tag{14}$$

$$\mu_m^T = \frac{\sum_{n=1}^{N} \mathbf{x}_n * \gamma_m(n)}{\sum_{n=1}^{N} \gamma_m(n)} = \frac{\sum_{n=1}^{N} \mathbf{x}_n * \gamma_m(n)}{\gamma_m} \tag{15}$$

$$\mathbf{\Sigma}_m = \frac{\sum_{n=1}^{N} \gamma_m(n)(\mathbf{x}_n - \mu_m)(\mathbf{x}_n - \mu_m)^T}{\sum_{n=1}^{N} \gamma_m(n)} \tag{16}$$

$$= \frac{\sum_{n=1}^{N} \gamma_m(n)(\mathbf{x}_n - \mu_m)(\mathbf{x}_n - \mu_m)^T}{\gamma_m} \tag{17}$$

$$c_m = \frac{1}{N} \sum_{n=1}^{N} \gamma_m(n) = \frac{\gamma_m}{\sum_{m=1}^{M} \gamma_m} \tag{18}$$