Unsupervised Machine Learning with Python

Derivation of Expectation Maximization for the Gaussian Mixture Model

Section 1: Introduction

This document presents a derivation of the expectation maximization approach for the Gaussian Mixture Model.

Assume data points $X_0,\ X_1,\ X_2, \ldots, X_{M-1}$ in d dimensions. Assume that there are K clusters, where cluster k denoted C_k, has mean $\mu_k,$ covariance Σ_k, and weight ϕ_k. Note that weights satisfy $\phi_0 + \cdots + \phi_{K-1} = 1$. The probability density function for the mixture of Gaussians is:

$$P(X) = \sum_{k=0}^{K-1} \phi_k N(X, \mu_k, \Sigma_k)$$

The probability picking X and that it is part of cluster k is given by:

$$P(X \cap C_k) = \phi_k N(X, \mu_k, \Sigma_k)$$

Using Bayes Theorem, the conditional probability that a data point is in cluster k given X is

$$P(C_k|X) = \frac{P(X \cap C_k)}{P(X)} = \frac{\phi_k N(X, \mu_k, \Sigma_k)}{\sum_{k=0}^{K-1} \phi_k N(X, \mu_k, \Sigma_k)}$$

The joint probability density function for $X_0, \ldots, X_{M-1}$ is given by likelihood function:

$$P(X_0, \ldots, X_{M-1}) = \prod_{i=0}^{M-1} P(X_i) = \prod_{i=0}^{M-1} \sum_{k=0}^{K-1} \phi_k N(X_i, \mu_k, \Sigma_k)$$

Maximum Likelihood Estimation attempts to find the parameters (values of means $\{\mu_k\},$ covariances $\{\Sigma_k\}$, and weights $\{\phi_k\}$) that has the maximum likelihood for generating the given data points. Following convention, we will do this by maximizing the log likelihood function:

$$L = \log P(X_0, \ldots, X_{M-1}) = \sum_{i=0}^{M-1} \log \left[\sum_{k=0}^{K-1} \phi_k N(X_i, \mu_k, \Sigma_k) \right]$$

subject to the constraint $\phi_0 + \cdots + \phi_{K-1} = 1$. Expectation Maximization is an iterative approach for finding the parameters.

Section 2: One-Dimensional Case

For the 1-dimensional case, one must find the maximum of

$$L = \log P(X_0, \ldots, X_{M-1}) = \sum_{i=0}^{M-1} \log \left[\sum_{k=0}^{K-1} \phi_k N(X_i, \mu_k, \sigma_k^2) \right]$$

where σ_k^2 is the variance for the k'th Gaussian in the mixture.

Using the Lagrange multipliers approach: one maximizes

$$L' = \sum_{i=0}^{M-1} \log\left[\sum_{k=0}^{K-1} \phi_k N\left(X_i, \mu_k, \sigma_k^2\right)\right] + \lambda(\phi_0 + \cdots + \phi_{K-1} - 1)$$

where λ, is the Lagrange multiplier. Differentiating with respect to all the variables ($\{\mu_k\}$, $\{\sigma_k\}$, $\{\phi_k\}$, λ) and setting to 0, one gets:

$$\frac{\partial L'}{\partial \mu_k} = 0 \quad k = 0, \ldots, K-1$$

$$\frac{\partial L'}{\partial \sigma_k} = 0 \quad k = 0, \ldots, K-1$$

$$\frac{\partial L'}{\partial \phi_k} = 0 \quad k = 0, \ldots, K-1$$

$$\frac{\partial L'}{\partial \lambda} = 0$$

First start with derivatives of the normal distribution:

$$N(X, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(X-\mu)^2}{\sigma^2}}$$

The derivative with respect to μ is:

$$\frac{\partial N}{\partial \mu} = \frac{(X-u)^2}{\sigma^2} N$$

The derivative with respect σ is:

$$\frac{\partial N}{\partial \sigma} = \left(-\frac{1}{\sigma} + \frac{(X-u)^2}{\sigma^3}\right) N$$

Now returning to the derivatives of L', one has:

The derivative with respect to μ_k is:

$$\frac{\partial L'}{\partial \mu_k} = \sum_{i=0}^{M-1} \frac{\phi_k \frac{\partial N}{\partial \mu_k}}{\sum_{k=0}^{K-1} \phi_k N(X_i, \mu_k, \sigma_k^2)} = \sum_{i=0}^{M-1} \frac{\phi_k \frac{(X_i - \mu_k)}{\sigma^2} N(X_i, \mu_k, \sigma_k^2)}{\sum_{k=0}^{K-1} \phi_k N(X_i, \mu_k, \sigma_k^2)}$$

Defining

$$\gamma_{ki} = \frac{\phi_k N(X_i, \mu_k, \sigma_k^2)}{\sum_{k=0}^{K-1} \phi_k N(X_i, \mu_k, \sigma_k^2)} \; and \; M_k = \sum_{i=0}^{M-1} \gamma_{ki}$$

and setting

$$\frac{\partial L'}{\partial \mu_k} = \sum_{i=0}^{M-1} \gamma_{ki} \frac{(X_i - \mu_k)}{\sigma^2} = 0$$

implies

$$\mu_k = \frac{1}{M_k} \sum_{i=0}^{M-1} \gamma_{ki} X_i$$

The derivative with respect to σ_k is:

$$\frac{\partial L'}{\partial \sigma_k} = \sum_{i=0}^{M-1} \frac{\phi_k \frac{\partial N}{\partial \sigma_k}}{\sum_{k=0}^{K-1} \phi_k N(X_i, \mu_k, \sigma_k^2)} = \sum_{i=0}^{M-1} \frac{\phi_k \left(-\frac{1}{\sigma_k} + \frac{(X_i - \mu_k)^2}{\sigma_k^3} \right) N(X_i, \mu_k, \sigma_k^2)}{\sum_{k=0}^{K-1} \phi_k N(X_i, \mu_k, \sigma_k^2)}$$

Using the expression for γ_{ki} and setting to 0, we have

$$\frac{\partial L'}{\partial \sigma_k} = \sum_{i=0}^{M-1} \gamma_{ki} \left(-\frac{1}{\sigma_k} + \frac{(X_i - \mu_k)^2}{\sigma_k^3} \right) = 0$$

which implies:

$$\sigma_k^2 = \frac{1}{M_k} \sum_{i=0}^{M-1} \gamma_{ki} (X_i - \mu_k)^2$$

The derivative with respect to ϕ_k is:

$$\frac{\partial L'}{\partial \phi_k} = \sum_{i=0}^{M-1} \frac{N(X_i, \mu_k, \sigma_k^2)}{\sum_{k=0}^{K-1} \phi_k N(X_i, \mu_k, \sigma_k^2)} = \frac{1}{\phi_k} \sum_{i=0}^{M-1} \frac{\phi_k N(X_i, \mu_k, \sigma_k^2)}{\sum_{k=0}^{K-1} \phi_k N(X_i, \mu_k, \sigma_k^2)} + \lambda$$

Using the expression for γ_{ki} and setting to 0, we have

$$\frac{\partial L'}{\partial \sigma_k} = \frac{1}{\phi_k} \sum_{i=0}^{M-1} \gamma_{ki} + \lambda = 0$$

which implies:

$$\phi_k = -\frac{1}{\lambda} \sum_{i=0}^{M-1} \gamma_{ki}$$

Setting $\phi_0 + \cdots + \phi_{K-1} = 1$, we have

$$-\frac{1}{\lambda} \sum_{k=0}^{K-1} \sum_{i=0}^{M-1} \gamma_{ki} = 1$$

Based on the definition, is relatively straightforward to show that

$$\sum_{k=0}^{K-1}\sum_{i=0}^{M-1}\gamma_{ki} = M$$

Hence, λ=-M and

$$\phi_k = \frac{1}{M}\sum_{i=0}^{M-1}\gamma_{ki} = \frac{M_k}{M}$$

Summarizing, we have:

$$\gamma_{ki} = \frac{\phi_k N(X_i, \mu_k, \sigma_k^2)}{\sum_{k=0}^{K-1}\phi_k N(X_i, \mu_k, \sigma_k^2)} \; and \; M_k = \sum_{i=0}^{M-1}\gamma_{ki}$$

$$\phi_k = \frac{M_k}{M}$$

$$\mu_k = \frac{1}{M_k}\sum_{i=0}^{M-1}\gamma_{ki}X_i$$

$$\sigma_k^2 = \frac{1}{M_k}\sum_{i=0}^{M-1}\gamma_{ki}(X_i - \mu_k)^2$$

We cannot solve the equations directly for μ_k, σ_k^2, and ϕ_k because these variables appear on both sides of the equations. The Expectation Maximization algorithm attempts to solve the above equations using an iterative approach:

(1) Initialize: means $\{\mu_k\}$, variances $\{\sigma_k^2\}$, and weights $\{\phi_k\}$.
(2) Expectation Step: given for $\{\mu_k\}$, $\{\sigma_k^2\}$, and $\{\phi_k\}$, compute
$$\gamma_{ki} = \frac{\phi_k N(X_i, \mu_k, \sigma_k^2)}{\sum_{k=0}^{K-1}\phi_k N(X_i, \mu_k, \sigma_k^2)} \; and \; M_k = \sum_{i=0}^{M-1}\gamma_{ki}$$
(3) Maximization Step: given $\{\gamma_{ki}\}$ determined in the Expectation Step, compute means $\{\mu_k\}$, variances $\{\sigma_k^2\}$, and weights $\{\phi_k\}$.

Repeat steps (2) and (3) until convergence.

This approach converges to a local maximum of the log likelihood function. Typically, one continues steps (2) and (3) until the difference in the means from the current and previous iterations is less than a given tolerance.

Section 3: Multi-Dimensional Case

For the multi-dimensional case, one must maximize:

$$L' = \sum_{i=0}^{M-1} \log\left[\sum_{k=0}^{K-1} \phi_k N(X_i, \mu_k, \Sigma_k)\right] + \lambda(\phi_0 + \cdots + \phi_{K-1} - 1)$$

where λ is the Lagrange multiplier.

The multi-dimensional normal distribution probability density function in d dimensions is:

$$N(X, \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} e^{-\frac{1}{2}(X-\mu)^T \Sigma^{-1}(X-\mu)}$$

where X and μ are d-dimensional column vectors and the covariance Σ is dxd dimensional matrix. Note that |Σ| is the determinant.

Here are basic assumptions:

(A) The covariance matrix is symmetric
(B) The covariance matrix is invertible (determinant is non-zero)

Some basic matrix results for matrices:

(C) $(A^T)^{-1}=(A^{-1})^T = \mathrm{A}^{-T}$

Here are some results for derivatives (gradients) of functions of vectors and matrices. These results are taken from the Matrix Cookbook (see citation in the References section). Consider:

$$F = (X-\mu)^T \Sigma^{-1}(X-\mu)$$

First let's compute the gradient with respect to μ (see formula 86 in Matrix Cookbook):

$$\nabla_\mu F = -2\Sigma^{-1}(X-\mu)$$

Let's now compute the gradient with respect to the Σ (see formula 61 in Matrix Cookbook):

$$\nabla_\Sigma F = -\Sigma^{-T}(X-\mu)(X-\mu)^T\Sigma^{-T}$$

Let us define $G = |\Sigma|$ (determinant). The gradient is (see formula 49 in Matrix Cookbook):

$$\nabla_\Sigma G = |\Sigma|\Sigma^{-T}$$

Now we can compute the appropriate gradients of the normal pdf:

$$\nabla_\mu N(X, \mu, \Sigma) = \Sigma^{-1}(X-\mu)N(X, \mu, \Sigma)$$

$$\nabla_{\Sigma} N(X,\mu,\Sigma) = -\frac{1}{2}\Sigma^{-T} N(X,\mu,\Sigma) + \frac{1}{2} N(X,\mu,\Sigma)\Sigma^{-T}(X-\mu)(X-\mu)^{T}\Sigma^{-T}$$

We now have the tools to compute the gradients of L' and set them to 0. The equations are:

$$\nabla_{\mu_k} L' = 0 \quad k = 0, \dots, K-1$$

$$\nabla_{\Sigma_k} L' = 0 \quad k = 0, \dots, K-1$$

$$\frac{\partial L'}{\partial \phi_k} = 0 \quad k = 0, \dots, K-1$$

$$\frac{\partial L'}{\partial \lambda} = 0$$

Let's start with he gradient with respect to the means:

$$\nabla_{\mu_k} L' = \sum_{i=0}^{M-1} \frac{\phi_k \nabla_{\mu_k} N}{\sum_{k=0}^{K-1} \phi_k N(X_i,\mu_k,\Sigma_k)} = -\sum_{i=0}^{M-1} \frac{\phi_k \Sigma^{-1}(X_i-\mu_k) N(X_i,\mu_k,\Sigma_k)}{\sum_{k=0}^{K-1} \phi_k N(X_i,\mu_k,\Sigma_k)} \quad k = 0, \dots, K-1$$

As in the one-dimensional case, let us define:

$$\gamma_{ki} = \frac{\phi_k N(X_i,\mu_k,\Sigma_k)}{\sum_{k=0}^{K-1} \phi_k N(X_i,\mu_k,\Sigma_k)} \; and \; M_k = \sum_{i=0}^{M-1} \gamma_{ki}$$

Rewriting the gradient and setting to 0:

$$\nabla_{\mu_k} L' = -\sum_{i=0}^{M-1} \gamma_{ki}\, \Sigma_k^{-1}(X_i-\mu_k) = 0 \;\; k = 0, \dots, K-1$$

Multiplying by Σ_k and solving for μ_k, we have

$$\mu_k = \frac{\sum_{i=0}^{M-1} \gamma_{ki} X_i}{\sum_{i=0}^{M-1} \gamma_{ki}} = \frac{1}{M_k} \sum_{i=0}^{M-1} \gamma_{ki} X_i \quad k = 0, \dots, K-1$$

For the gradients with respect to the covariance matrices, we have:

$$\nabla_{\Sigma_k} L' = \sum_{i=0}^{M-1} \frac{\phi_k \nabla_{\Sigma_k} N}{\sum_{k=0}^{K-1} \phi_k N(X_i,\mu_k,\Sigma_k)} = \sum_{i=0}^{M-1} \gamma_{ki} \left(-\frac{1}{2}\Sigma_k^{-T} N + \frac{1}{2}\Sigma_k^{-T}(X_i-\mu_k)(X_i-u_k)^{T}\Sigma_k^{-T} N\right)$$

Setting the gradient to 0, and dividing out 2N and multiplying by Σ_k^{T}, we have:

$$\sum_{i=0}^{M-1} \gamma_{ki} \left(-1 + (X_i-\mu_k)(X_i-u_k)^{T}\Sigma_k^{-T}\right) = 0 \quad k = 0, \dots, K-1$$

Since the covariance matrix is symmetric $\Sigma_k^{-T} = \Sigma_k^{-1}$. Solving this last equation for Σ_k, we have:

$$\Sigma_k = \frac{1}{M_k} \sum_{i=0}^{M-1} \gamma_{ki} (X_i-\mu_k)(X_i-\mu_k)^{T} \quad k = 0, \dots, K-1$$

The derivative with respect to ϕ_k is:

$$\frac{\partial L'}{\partial \phi_k} = \sum_{i=0}^{M-1} \frac{N(X_i, \mu_k, \Sigma_k)}{\sum_{k=0}^{K-1} \phi_k N(X_i, \mu_k, \Sigma_k)} = \frac{1}{\phi_k} \sum_{i=0}^{M-1} \frac{\phi_k N(X_i, \mu_k, \Sigma_k)}{\sum_{k=0}^{K-1} \phi_k N(X_i, \mu_k, \Sigma_k)} + \lambda$$

Using the same approach as in the one-dimensional case, one can show

$$\phi_k = \frac{1}{M} \sum_{i=0}^{M-1} \gamma_{ki} = \frac{M_k}{M}$$

Summarizing, we have for k = 0,…,K-1:

$$\gamma_{ki} = \frac{\phi_k N(X_i, \mu_k, \Sigma_k)}{\sum_{k=0}^{K-1} \phi_k N(X_i, \mu_k, \Sigma_k)} \; and \; M_k = \sum_{i=0}^{M-1} \gamma_{ki}$$

$$\phi_k = \frac{M_k}{M}$$

$$\mu_k = \frac{1}{M_k} \sum_{i=0}^{M-1} \gamma_{ki} X_i$$

$$\Sigma_k = \frac{1}{M_k} \sum_{i=0}^{M-1} \gamma_{ki} (X_i - \mu_k)(X_i - \mu_k)^T$$

We cannot solve the equations directly for μ_k, Σ_k and ϕ_k because these variables appear on both sides of the equation. The Expectation Maximization algorithm attempts to solve the above equations using an iterative approach:

(1) Initialize: means $\{\mu_k\}$, variance matrices $\{\Sigma_k\}$, and weights $\{\phi_k\}$.
(2) Expectation Step: compute $\{\gamma_{ki}\}$ and $\{M_k\}$
(3) Maximization Step: given γ_{ki} determined in the Expectation Step, compute means $\{\mu_k\}$, variance matrices $\{\Sigma_k\}$, and weights $\{\phi_k\}$.

Section 4: References

Kaare Brandt Petersen and Michael Syskind Pedersen, *The Matrix Cookbook*, Version November 15, 2012. https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf