Expectation Maximization [3] for Fitting Mixture Models

Qi Zhao

August 13, 2018

1. Introduction

To learn the MoG parameters [4] $\theta = \{\lambda_k, \mu_k, \Sigma_k\}_{k=1}^K$ from training data $\{x_i\}_{i=1}^I$ it applies the EM algorithm [5]. And it initializes the parameters randomly and alternate between performing the E- and M-steps.

In the E-step, it maximizes the bound with respect to the distributions $q_i\ (h_i)$ by finding the posterior probability distribution Pr($h_i\ |x_i$) of each hidden variable h_i given the observation x_i and the current parameter settings in Equation 1.

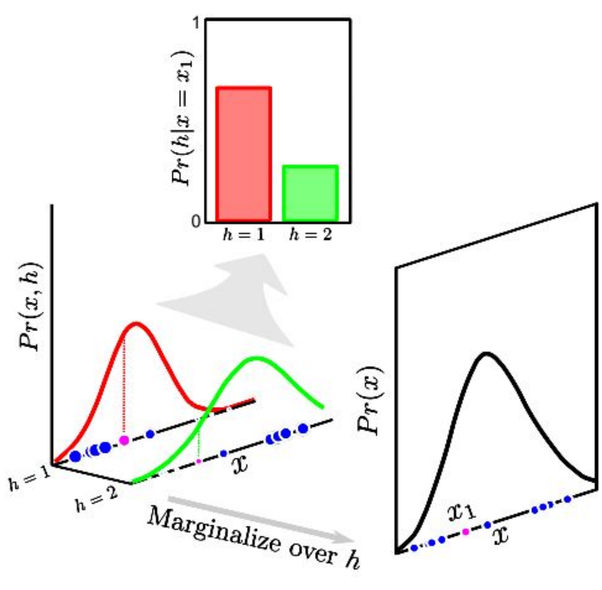

Figure 1. E-step for fitting the mixture of Gaussians model. For each of the I data points $x_{1...I}$, it calculates the posterior distribution Pr($h_i|x_i$) over the hidden variable h_i.The posterior probability [2] Pr(h_i = k$|x_i$) that h i takes value k can be understood as the responsibility of normal distribution k for data point x_i. For example, for data point x_1 (magenta circle), component 1 (red curve) is more than twice as likely to be responsible than component 2 (green curve). Note that in the joint distribution (left), the size of the projected data point indicates the responsibility.

$$\begin{aligned}
q_i(h_i) &= Pr(h_i = k|x_i, \theta^{[t]}) \\
&= \frac{Pr(x_i|h_i = k, \theta^{[t]})Pr(h_i = k, \theta^{[t]})}{\sum_{j=1}^K Pr(x_i|h_i = j, \theta^{[t]})Pr(h_i = j, \theta^{[t]})} \\
&= \frac{\lambda_k Norm_{x_i}[\mu_k, \Sigma_k]}{\sum_{j=1}^K \lambda_j Norm_{x_i}[\mu_j, \Sigma_j]} \\
&= r_{ik}
\end{aligned} \tag{1}$$

In other words it computes the probability Pr(h_i = k$|x_i$, $\theta^{[t]}$) that the k^{th} normal distribution was responsible for the i^{th} data point (Figure 1). It denotes this responsibility by r_{ik} for short. In the M-step, it maximizes the bound with respect to the parameters $\theta = \{\lambda_k, \mu_k, \Sigma_k\}_{k=1}^K$ in Equation 2.

$$\begin{aligned}
\hat{\theta}^{[t+1]} &= argmax_\theta[\sum_{i=1}^I \sum_{k=1}^K \hat{q}_i(h_i = k)log[Pr(x_i, h_i = k|\theta)]] \\
&= argmax_\theta[\sum_{i=1}^I \sum_{k=1}^K r_{ik}log[\lambda_k Norm_{x_i}[\mu_k, \Sigma_k]]]
\end{aligned} \tag{2}$$

2. Conclusions

This maximization can be performed by taking the derivative of the expression with respect to the parameters, equating the result to zero and rearranging, taking care to enforce the constraint $\sum_k \lambda_k$ = 1 using Lagrange multipliers [1].

References

[1] F. B. Belgacem. The mortar finite element method with Lagrange multipliers. *Numerische Mathematik*, 84(2):173–197, 1999. 1

[2] P. Erixon, B. Svennblad, T. Britton, and B. Oxelman. Reliability of bayesian posterior probabilities and bootstrap frequencies in phylogenetics. *Systematic Biology*, 52(5):665–673, 2003. 1

[3] T. K. Moon. *The expectation maximization algorithm*. Springer, 2000. 1

[4] C. W. Roy, M. Seed, J. F. van Amerom, B. Al Nafisi, L. Grosse-Wortmann, S.-J. Yoo, and C. K. Macgowan. Dynamic imaging of the fetal heart using metric optimized gating. *Magnetic Resonance in Medicine*, 70(6):1598–1607, 2013. 1

[5] C. F. J. Wu. On the convergence properties of the EM algorithm. *Annals of Statistics*, 11(1):95–103, 1983. 1