Mixture of Gaussians [2]

Qi Zhao

August 9, 2018

The mixture of Gaussians (MoG) is a prototypical example of a model where learning is suited to the EM algorithm [6]. The data are described as a weighted sum of K normal distributions [1] in Equation 1.

$$Pr(x|\theta) = \sum_{k=1}^{K} \lambda_k Norm_x[\mu_k, \Sigma_k] \tag{1}$$

where $\mu_{1...K}$ and $\Sigma_{1...K}$ are the means and covariances of the normal distributions and $\lambda_{1...K}$ are positive valued weights that sum to one. The mixtures of Gaussians model describes complex multi-modal probability [5] densities by combining simpler constituent distributions (Figure 1).

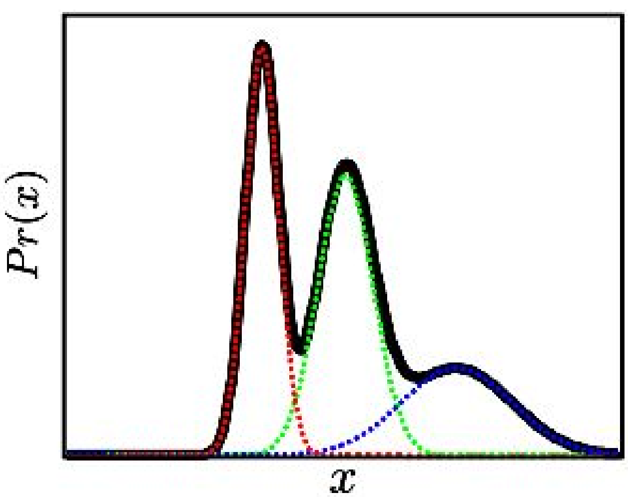

Figure 1. Mixture of Gaussians model in 1D. A complex multi-modal probability density function (black solid curve) is created by taking a weighted sum or mixture of several constituent normal distributions with different means and variances (red, green and blue dashed curves). To ensure that the final distribution is a valid density, the weights must be positive and sum to one.

To learn the parameters $\theta = \{\mu_k, \Sigma_k, \lambda_k\}_{k=1}^{K}$ from training data $\{x_i\}_{i=1}^{I}$ it could apply the straightforward maximum likelihood [3] approach in Equation 2.

$$\begin{aligned} \hat{\theta} &= argmax_x[\sum_{i=1}^{I} log[Pr(x_i|\theta)]] \\ &= argmax_x[\sum_{i=1}^{I} log[\sum_{k=1}^{K} \lambda_k Norm_x[\mu_k, \Sigma_k]] \end{aligned} \tag{2}$$

Unfortunately, if it takes the derivative with respect to the parameters θ and equate the resulting expression to zero, it is not possible to solve the resulting equations in closed form. The sticking point is the summation inside the logarithm, which precludes a simple solution. Of course, it could use a nonlinear optimization approach, but this would be complex as it would have to maintain the constraints on the parameters; the weights λ must sum to one and the covariances $\{\Sigma_k\}_{k=1}^{K}$ must be positive definite. For a simpler approach, it expresses the observed density as a marginalization [4] and use the EM algorithm to learn the parameters.

References

[1] D. F. Andrews and C. L. Mallows. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society*, 36(1):99–102, 1974. 1

[2] S. Dasgupta. Learning mixture of Gaussians. In *IEEE Symposium on Foundations of Computer Science*, pages 634–644, 1999. 1

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977. 1

[4] E. G. Larsson and J. Jalden. Fixed-complexity soft mimo detection via partial marginalization. *IEEE Transactions on Signal Processing*, 56(8):3397–3407, 2008. 1

[5] M. Nagode and M. Fajdiga. A general multi-modal probability density function suitable for the rainflow ranges of stationary random processes. *International Journal of Fatigue*, 20(3):211–223, 1998. 1

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