HistogramGMM Mathematical Formulation

thomasvf

March 2024

1 Introduction

This code adapts the EM for Gaussian Mixture Models found in Bishop's 2006 book to work with histograms.

Essentially, we work with bins instead of examples, and each bin is associated with a certain number of examples indicated by the histogram. That is, $\mathbf{x}_b$ gives the $M-$ dimensional position of bin b, and h_b is the number of examples in that bin (that is, that have this same position). We then only need to change the summations so that they emulate the iteration over all the examples in a bin. Since all bins have the same position, this simply means multiplying the summed term by the appropriate value in h_b.

2 E-Step

The E-step evaluates the responsibilities r_{nk} of each cluster $k = 1, ..., K$ and point $\mathbf{x}_n$. r_{nk} is also equivalent to $p(k|\mathbf{x}_n)$, that is, the probability of a certain cluster k given example x_n. In our case, we deal with the positions of each bin b. Since all examples in a bin have the same position, we only need to compute the responsibilities r_{bk} for each bin position $\mathbf{x}_b$.

The E-step becomes:

$$r_{bk} = \frac{\pi_k \mathcal{N}(\mathbf{x}_b|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_b|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}.$$

Note that we have simply changed the index from n to b.

3 M-Step

We adapt the M-step to emulate the iteration over all the examples in all bins. Let h_b be the number of examples in bin b for $b = 1, ..., B$.

The total number of examples and the effective number of examples in a cluster become, respectively:

$$N = \sum_{b=1}^{B} h_b$$

and

$$N_k = \sum_{b=1}^{B} h_b r_{bk}.$$

The estimations of the Gaussian parameters are similarly modified:

$$\boldsymbol{\mu}_k^{new} = \frac{1}{N_k} \sum_{b=1}^{B} h_b \, r_{bk} \, \mathbf{x}_b$$

$$\boldsymbol{\Sigma}_k^{new} = \frac{1}{N_k} \sum_{b=1}^{B} h_b \, r_{bk} \, (\mathbf{x}_b - \boldsymbol{\mu}_k^{new})(\mathbf{x}_b - \boldsymbol{\mu}_k^{new})^{\mathsf{T}}$$

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