

Note Set 4: The EM Algorithm for Gaussian Mixtures

Probabilistic Learning: Theory and Algorithms, CS 274A

The Gaussian Mixture Model

We are given a data set $D = \{\underline{x}(1), \dots, \underline{x}(N)\}$ where $\underline{x}(i)$ is a d -dimensional vector measurement. Assume that the points are generated in an IID fashion from an underlying density $p(\underline{x})$. We further assume a Gaussian mixture model with K components for $p(\underline{x})$:

$$p(\underline{x}|\Theta) = \sum_{k=1}^K \alpha_k p_k(\underline{x}|\theta_k)$$

where $\sum_{k=1}^K \alpha_k = 1$ are the mixture weights and where

$$\Theta = \{\alpha_1, \dots, \alpha_K, \theta_1, \dots, \theta_K\}$$

and where each component is a multivariate Gaussian density

$$p_k(\underline{x}|\theta_k) = \frac{1}{(2\pi)^{d/2} |\Sigma_k|^{1/2}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu}_k)^t \Sigma_k^{-1} (\underline{x}-\underline{\mu}_k)}$$

with its own parameters $\theta_k = \{\underline{\mu}_k, \Sigma_k\}$.

We can compute the membership of data point $\underline{x}(i)$ in cluster k , given parameters Θ as

$$w_{ik} = p(C = k|\underline{x}(i), \Theta) = \frac{p_k(\underline{x}(i)|\theta_k) \cdot \alpha_k}{\sum_{m=1}^K p_m(\underline{x}(i)|\theta_m) \cdot \alpha_m}, \quad 1 \leq k \leq K, \quad 1 \leq i \leq N.$$

This follows from a direct application of Bayes rule.

The EM Algorithm

We define the EM (Expectation-Maximization) algorithm for Gaussian mixtures as follows:

E-Step: Denote the current parameter values as Θ . Compute w_{ik} (using the equation above) for all data points $\underline{x}(i)$, $1 \leq i \leq N$ and all mixture components $1 \leq k \leq K$. Note that for each data point $\underline{x}(i)$ the membership weights are defined such that $\sum_{k=1}^K w_{ik} = 1$. This yields an $N \times K$ matrix of membership weights, where each of the rows sum to 1.

M-Step: Now use the membership weights and the data to calculate new parameter values. Specifically,

$$\alpha_k^{new} = \frac{1}{N} \sum_{i=1}^N w_{ik}, \quad 1 \leq k \leq K.$$

These are the new mixture weights.

$$\underline{\mu}_k^{new} = \left(\frac{1}{\sum_{i=1}^N w_{ik}} \right) \sum_{i=1}^N w_{ik} \cdot \underline{x}(i) \quad 1 \leq k \leq K.$$

The updated mean is calculated in a manner similar to how we could compute a standard empirical average, except that the i th measurement $\underline{x}(i)$ has a fractional weight w_{ik} . Note that this is a vector equation since $\underline{\mu}_k^{new}$ and $\underline{x}(i)$ are both d -dimensional vectors.

$$\Sigma_k^{new} = \left(\frac{1}{\sum_{i=1}^N w_{ik}} \right) \sum_{i=1}^N w_{ik} \cdot (\underline{x}(i) - \underline{\mu}_k^{new})(\underline{x}(i) - \underline{\mu}_k^{new})^t \quad 1 \leq k \leq K.$$

Again we get an equation that is similar in form to how we would normally compute an empirical covariance matrix, except that the contribution of each data point is weighted by w_{ik} . Note that this is a matrix equation with $d \times d$ terms on each side.

The equations in the M-step need to be computed in this order, i.e., first compute the K new α 's, then the K new $\underline{\mu}_k$'s, and finally the K new Σ_k 's.

After we have computed all of the new parameters, the M-step is complete and we can now go back and recompute the membership weights in the E-step, then recompute the parameters again in the E-step, and continue updating the parameters in this manner. Each pair of E and M steps is considered to be an iteration.

Initialization and Convergence Issues

The EM algorithm can be started by either initializing the algorithm with a set of initial parameters and then conducting an E-step, or by starting with a set of initial weights and then starting with

an M-step. The initial parameters or weights can be chosen randomly (e.g. select K random data points as initial means and select the covariance matrix of the whole data set for each of the initial K covariance matrices) or could be chosen via some heuristic method (such as by using k-means to cluster the data first and then defining weights based on k-means memberships).

Convergence is generally detected by computing the value of the log-likelihood after each iteration and halting when it appears not to be changing in a significant manner from one iteration to the next.

Note that the log-likelihood (under the IID assumption) is defined as follows:

$$\log l(\Theta) = \sum_{i=1}^N \log p(\underline{x}(i)|\Theta) = \sum_{i=1}^N$$

where $p(\underline{x}(i)|\Theta)$ is the equation for the Gaussian mixture model with K components defined earlier.