Report of EM Algorithm for Gaussian Mixture Distributions

Zhenyang Zhu
zhenyangzhu@buaa.edu.cn

Abstract

This is the second assignment report for the NLP course. In this report, the EM algorithm is used to estimate the parameters of the Gaussian mixture distribution model and the model is used to make predictions. The data set was randomly generated for 2000 mixed-gender heights. Finally, it converges well to the actual value through calculation iteration. In addition, this report also analyzes the performance of EM algorithm with different initial parameters.

Methodology

The Expectation-Maximization (EM) algorithm is a powerful iterative technique that is widely used in statistics and machine learning. It is an algorithm for finding maximum likelihood estimates of parameters in statistical models, where the data is incomplete or partially missing. In other words, EM algorithm is used when there are unobserved or hidden variables in the data.

The EM algorithm consists of two steps: the E-step and the M-step. In the E-step, the expected values of the missing data are computed, given the current estimates of the parameters. In the M-step, the parameters are updated to maximize the expected log-likelihood found in the E-step.

The EM algorithm is particularly useful for models that involve mixtures of distributions, such as Gaussian mixture models. In this report, the EM algorithm is used to estimate the mean, variance, and mixing coefficients of each Gaussian component.

The Gaussian mixture model (GMM) is a probabilistic model that assumes that a population is composed of several subpopulations, each of which follows a Gaussian distribution with its own mean and variance. The probability density function of a GMM with K components is given by:

$$p(x) = \sum_{k=1}^{K} \pi_k N(x \mid \mu_k, \Sigma_k)$$

where x is a data point, π_k is the mixing coefficient for the k-th component, and $N(x|\mu_k, \Sigma_k)$ is the Gaussian distribution with mean μ_k and covariance matrix Σ_k.

The EM algorithm for GMMs involves two steps: the E-step and the M-step.

In the E-step, we compute the posterior probabilities that each data point x_i belongs to each component k, given the current estimates of the parameters. These probabilities are known as the "responsibilities" of each component for each data point, and are denoted by r_{ik}:

$$r_{ik} = \frac{\pi_k N(x_i \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_i \mid \mu_k, \Sigma_k)}$$

In the M-step, we update the estimates of the parameters by maximizing the expected complete data log-likelihood with respect to each parameter. The expected complete data log-likelihood is given by:

$$Q(\theta \mid \theta^{(t)}) = \sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik} \log(\pi_k N(x_k \mid \mu_k, \Sigma_k)$$

where θ represents the set of model parameters, and $\theta^{(t)}$ represents the current estimate of the parameters at iteration t.

To update the mixing coefficients π_k, we set the derivative of Q with respect to π_k to zero and solve for π_k,. This yields:

$$\pi_k = \frac{1}{N} \sum_{i=1}^{N} r_{ik}$$

To update the mean vectors μ_k, we set the derivative of Q with respect to μ_k to zero and solve for μ_k. This yields:

$$\mu_k = \frac{1}{N_k} \sum_{i=1}^{N} r_{ik} x_i$$

where $N_k = \sum_{i=1}^{N} r_{ik}$ is the total responsibility of component k.

To update the covariance matrices Σ_k, we set the derivative of Q with respect to Σ_k to zero and solve for Σ_k. This yields:

$$\Sigma_k = \frac{1}{N_k} \sum_{i=1}^{N} r_{ik} (x_i - \mu_k)(x_i - \mu_k)^T$$

where N_k is as defined above.

The EM algorithm iterates between the E-step and the M-step until convergence is reached. Typically, convergence is achieved when the change in the log-likelihood or in the parameters falls below a certain threshold.

Experimental Studies

In this report, the EM algorithm is used to calculate the parameters of the Gaussian mixture distribution model with the height of boys and girls as data samples. The corresponding mathematical symbols are shown below.

π — the ratio of girls

μ_F — mean height of girls

Σ_F — standard deviation of girls

μ_M — mean height of boys

Σ_M — standard deviation of boys

There are 2000 pieces of data in the original data set, among which boys accounted for 0.75, the mean height is 176, and the standard deviation is 5. The ratio of girls is 0.25, the mean height is 165, and the standard deviation is 3. When the initial parameters are selected, after running the EM algorithm, we obtained the final values of the parameters and the log-likelihood of the data, as is shown in table 1. I also plotted the data points and the estimated Gaussian mixture distributions in Figure 1, which provided a visual representation of the clustering.

Table 1 The results of EM algorithm for Gaussian mixture model

	Initial Parameters	Final Parameters	Actual Parameters	Log-likelihood
π	0.5	0.228	0.25	
μ_F	160	163.53	165	
Σ_F	1	2.81	3	-1660.9
μ_M	170	175.83	176	
Σ_M	1	5.21	5	

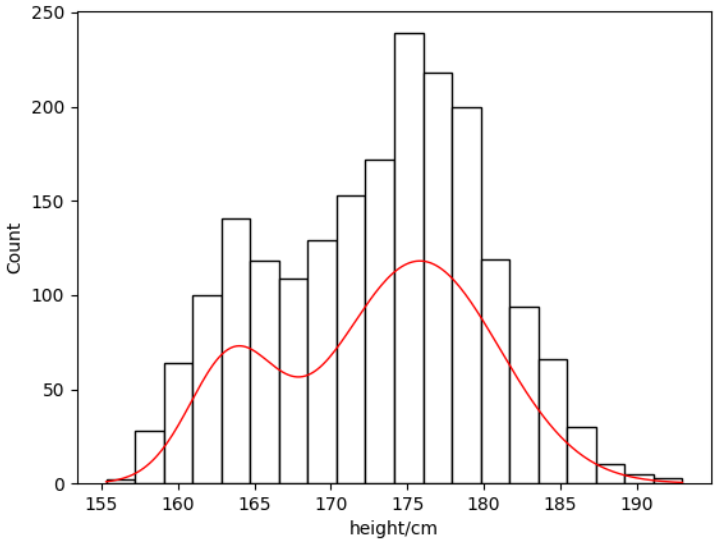

Fig 1 The histogram of height data and the estimated Gaussian mixture distributions

It can be seen that EM algorithm can converge well to the parameters of the model itself.

Considering that EM algorithm is sensitive to initial parameters and cannot guarantee global convergence, different initial values are selected for iterative calculation in th is paper, and the results are shown in Table 2, the maximum number of iterations is 200.

Table 2 Results of EM algorithm under different initial parameters

Initial Parameters- $\{\pi\ \mu_F\ \Sigma_F\ \mu_M\ \Sigma_M\}$	Final Parameters- $\{\pi\ \mu_F\ \Sigma_F\ \mu_M\ \Sigma_M\}$	Log-likelihood	Number of iterations
{0.5 160 1 170 1}	{0.23 163.5 2.8 175.8 5.21}	-1660.9	129
{0.5 175 1 175 1}	{0.5 173 7 173 7}	-3368.9	1
{0.5 160 100 180 100}	{0.25 163.8 2.97 176.1 5.1}	-1791.0	200
{0.99 160 10 180 10}	{0.23 163.6 2.83 175.9 5.2}	-1677.5	200

Table 2 shows that different initial parameters affect the computational performance of the EM algorithm. When the initial parameters deviate too much from the actual values, as in the third and

fourth experiments, the convergence speed of the algorithm becomes slow and more iterations are needed to converge to the final result. However, the initial value chosen in the second experiment makes the algorithm fall into a local optimum at the beginning, which seriously affects the accuracy of the algorithm. Therefore, for the EM algorithm, the choice of initial parameters is extremely important. Moreover, when the computation converges to a certain point, one can consider adding a small perturbation to take it away from the current extreme point.

Conclusion

In this report, the EM algorithm is used to estimate the parameters of the bivariate Gaussian mixture distribution model, and the results are compared with the distribution of the original data set. The calculation results show that the EM algorithm can effectively estimate the parameters of the model with hidden parameters, and can converge to the extreme value. But at the same time, because the EM algorithm is very dependent on the selection of initial parameters, when the initial parameters are not appropriate, the convergence speed of the algorithm will be reduced, even into a local optimum.