EM 算法

王铭泽 ZY2203115

wmz20000729@buaa.edu.cn

1 Introduction

The experiment involves using the Expectation-Maximization (EM) algorithm to estimate parameters for a Gaussian Mixture Model (GMM) based on existing height data, as well as making predictions. The experiment generates two GMMs, composed of male and female height distributions, using the available code. By applying the EM algorithm, we estimate the parameters of the model, and the results obtained are found to be very close to the actual values. In the experiment, we discovered that the choice of different initial values impacts model performance.

2 Methods

The Expectation-Maximization (EM) algorithm is a commonly used method for solving Gaussian Mixture Models (GMMs). A GMM contains multiple Gaussian distributions, each with its own mean and variance. The EM algorithm iteratively calculates the mean and variance of each Gaussian distribution, as well as the weight of each distribution in the total population.

The basic steps for using the EM algorithm to solve GMMs are as follows:

Initialize the model parameters: First, the model parameters need to be initialized, including the mean, variance, and weights of each Gaussian distribution. When the data is not complex, random values can be used as the mean, a relatively fitting variance can be set based on experience, and the weights of each Gaussian distribution can be initialized as an equal distribution, i.e., the weight of each Gaussian distribution is 1/k.

E-step: For each data point x_i, calculate the probability of it belonging to each Gaussian distribution $p(z_i = j|x_i, \theta)$, where $j \in 1, \ldots, k$ and θ represents the current model parameters, i.e., the mean, variance, and weights of each Gaussian distribution.

M-step: Utilizing the probabilities of each data point belonging to each Gaussian distribution calculated in the E-step, re-estimate the mean, variance, and weights for each Gaussian distribution.

The mean and variance of each Gaussian distribution can be estimated using the weighted average method.

Termination condition: Repeat the E-step and M-step until the model converges, i.e., the model parameters do not change significantly or the maximum number of iterations is reached. A maximum number of iterations can be set, or the log-likelihood function can be used to check for model convergence. If the increment in the log-likelihood function is smaller than a pre-defined threshold, the model is considered to have converged.

Prediction: After the model has converged, for a new data point x, classification can be done by calculating its probabilities in all Gaussian distributions, and selecting the Gaussian distribution with the largest probability as its belonging category.

3 Experiment

3.1 Implement details

First, using the provided code, generate a total of 2000 data points with means of 164 and 176, variances of 3 and 5, and counts of 500 and 1500, respectively. The distribution of the generated height data is illustrated in the following plot:

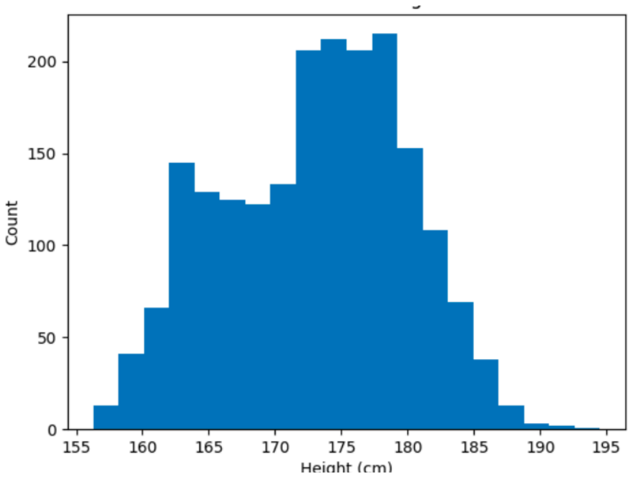

图 1: Distribution of Heights

In the EM algorithm, we first read the height data from the .csv file. Then, we set the initial weights to 0.5 and 0.5, the initial means are obtained by randomly extracting two values, and the initial variances are both set to 1. We run the algorithm for 300 iterations.

The results are as follows:

Final means: [163.01002858, 174.01056606]

Final standard deviations: [3.12794657, 4.99734404]

Final weights: [0.24359935, 0.75640065]

Comparing these results with the true values, they are found to be very close. The distribution plot is shown below:

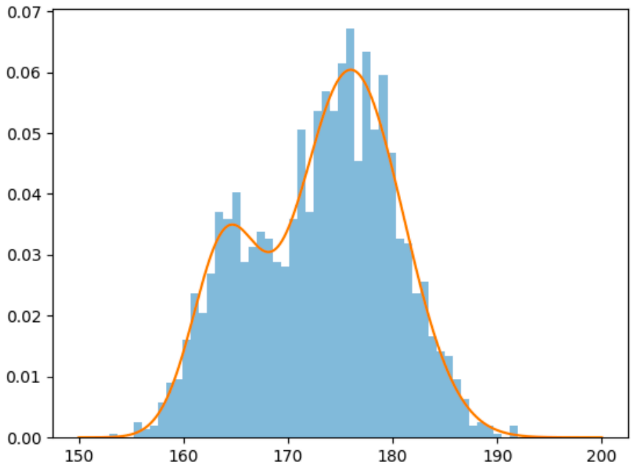

图 2: Distribution of Heights

4 Conclusion

After the calculations, we obtained the parameters for the Gaussian Mixture Model. When compared to the true values, the results obtained using the EM algorithm are very close.

During the experiment, we found that different parameters affect the number of iterations required for convergence. With the parameter generation method used in this experiment, most cases converge within 100 iterations, while some cases require up to 150 iterations for convergence. In this experiment, the maximum number of computation iterations was set based on practical considerations. Alternatively, a minimal difference value can be set, and when the results' differences are smaller than this value for several consecutive calculations, the model is considered to have converged.