

Introduction

This code performs the Expectation-Maximization (EM) algorithm on a dataset that is assumed to be generated from a mixture of two Gaussian distributions. The algorithm estimates the parameters of the two Gaussian distributions, including their means, standard deviations, and weights, as well as calculates the log likelihood of the estimated model on a separate test dataset.

Methodology

The code begins by importing the necessary libraries including NumPy, Pandas, and Matplotlib. It then defines the parameters of the two Gaussian distributions (mean and standard deviation) and generates 500 samples from the first distribution and 1500 samples from the second distribution using the `numpy.random.normal()` function. The two datasets are concatenated together to form the final dataset, which is then split into training and testing datasets using a 7:3 ratio.

The EM algorithm is then applied to the training dataset to estimate the parameters of the two Gaussian distributions. The algorithm consists of two steps: the E-step, which computes the probability of each sample belonging to each distribution, and the M-step, which updates the parameters of the distributions based on the probabilities computed in the E-step. The algorithm

iterates over these two steps until convergence, which is determined by checking if the change in the parameters between iterations is less than a predefined threshold.

Finally, the algorithm calculates the log likelihood of the estimated model on the testing dataset and prints out the estimated parameters of the two Gaussian distributions, including their means, standard deviations, and weights, as well as the log likelihood on the testing dataset.

Results

```
C:\Users\31217\AppData\Local\Programs\Python\Python37\python.exe E:\Github\projects\DL-NLP-Work\EM2.py
Log likelihood on test data: -1677.5786969993774
Gaussian distribution 1:
Mean: 163.60
Standard deviation: 2.79
Weight: 0.22
Gaussian distribution 2:
Mean: 175.85
Standard deviation: 5.13
Weight: 0.78

Process finished with exit code 0
```

Gaussian Distribution	Mean	Standard deviation	Weight
1	163.60	2.79	0.22
2	175.85	5.13	0.78
Log likelihood on the test data	value		
default	-1677.5786		

Conclusion

The capability of the modle is excellent due to the high log likelihood.