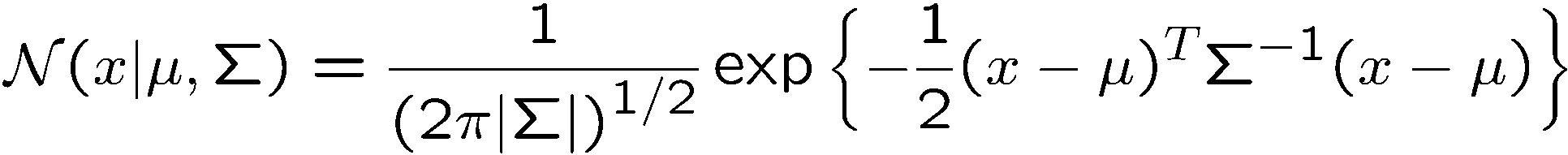
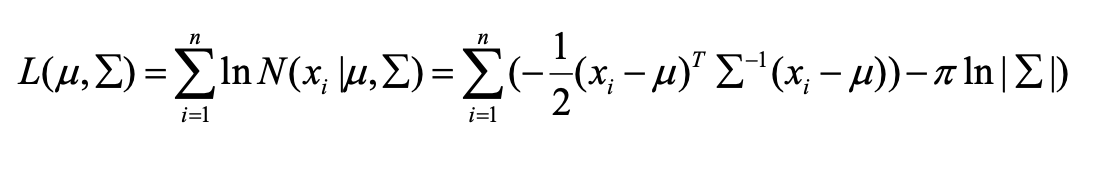
Gaussian mixture model

2.1 Abstract

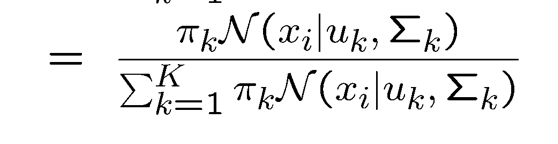
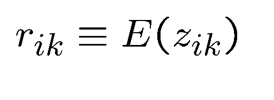
The algorithm of Gaussian mixture model is based on the probabilistic model. According to the datasets, it is assumed that the datasets are generated from a mixture of a provided number of Gaussian distributions with specific parameters.

The algorithm of gaussian mixture model implements the Expectation-Maximization algorithm.

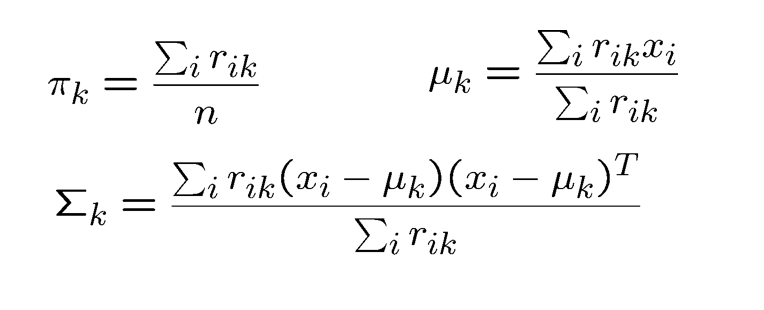




In the E-step, for given parameter values we can compute the expected values of the latent variables.



In the M-step, we maximize the expected complete log likelihood. We iterate E-step and M-step until the log likelihood of data does not increase any more.



In this way, based on the optimized parameters, we can compute posterior probability of membership and assign data points to each cluster.

2.2 Algorithm

1. Read the datasets

2. Initiate the parameters of GMM. For means in our case, we use K-means to get the initial value. For covariance in our case, we use diagonal matrix to get the initial value.

3.Repeat

4.Calculate the probability of corresponding gaussian models’ parameters.

5.Implement the E-step to get posterior probability.

6.Implement the M-step to optimize parameters.

7.Until the parameters converge to certain value.

2.3 Implementations

The python code takes two arguments. The first one is our dataset of txt format and the second argument is the number of clusters.

1. gmm\_init: it initiates the parameters of GMM.
2. gaussian: it calculates the probability of corresponding gaussian models’ parameters.
3. e\_step: it implements the E-step to get posterior probability.
4. m\_step: it implements the M-step to optimize parameters.
5. incidence\_mat\_gen: the function takes the result label as the only argument and output an incidence matrix based on that.
6. ja\_rand\_cal: it calculates rand and jaccard results. Ground truths and my results are the first and second argument respectively.

2.4 Results

Pros for GMM clustering:

1. it can handle clusters with varying sizes, variance.
2. It can give probabilistic cluster assignments.
3. It has probabilistic interpretation.
4. It has higher Rand index and Jaccard coefficient compared to other clustering.

Cons for GMM clustering:

1. The GMM clustering is the most time consuming compared to other clustering algorithms.

2. The initial parameters of GMM influence the performance of clustering a lot.

3. It can have overfitting issues.