Expectation Maximization Algorithm

EM Algorithm - Expectation Maximization Algorithm is a method to model the data by specifying a joint distribution P(X,Z) = P(X/Z) P(Z). The parameters of the model - mean, covariance and alpha are identified using max log likelihood function. Basically there are two steps in EM Algorithm. They are E - (Expected) Step and M - (Maximization) Step. The E-Step guesses the value of Z - multinomial by calculating the weight vector 'W' using gaussian distribution. The M- step will update the values of mean, variance and alpha and the process is iterative one and it breaks only when the max.log likelihood is same.

EM for general GMMs with unknown variance:

Run 1:

k = 3 (number of clusters)

```
the initial cluster mean is [[ 25.37275696  17.21684919  17.17307221]]
the initial cluster variance np is [[ 74.26643213  81.01792596  87.76941979]]
the inital cluster alpha np is [[ 0.33333333  0.33333333  0.33333333]]
inital max_log_likelihood is -21921.1192501
final_max_log_likelihood is -15098.2504919
the final variance is [[ 0.99809662  0.96711594  1.03025768]]
the final  mean is  [[ 25.48665443  15.44916079   5.5092794 ]]Run 2:
```

k = 3 (number of clusters)

```
the initial cluster mean is [[ 14.21493331  16.98919877  16.58099508]]
the initial cluster variance np is [[ 74.26643213  81.01792596  87.76941979]]
the inital cluster alpha np is [[ 0.33333333  0.33333333  0.33333333]]
inital max_log_likelihood is -21214.1430018
final_max_log_likelihood is -15098.2504919
the final variance is [[ 1.03025768  0.99809662  0.96711594]]
the final  mean is  [[  5.5092794   25.48665443  15.44916079]]Run 3:
```

k = 3 (number of clusters)

```
the initial cluster mean is [[ 15.82224953  27.16786136  15.97730712]]
the initial cluster variance np is [[ 74.26643213  81.01792596  87.76941979]]
the inital cluster alpha np is [[ 0.33333333  0.33333333  0.33333333]]
```

```
inital max_log_likelihood is -21888.1598867
final_max_log_likelihood is -15098.2504919
the final variance is [[ 1.03025768  0.99809662  0.96711594]]
the final  mean is  [[  5.5092794   25.48665443  15.44916079]]
```

Strategy used for Initialization:

The first heuristic method is considered to initialize the parameters. The Means of each cluster is identified randomly from the data points. The variance of the data points is calculated and the co-variance of the clusters initialized to the multiples of the calculated variance. The alpha values are calculated as 1/k. Since the program initializes only 3 clusters therefore the alpha values are 0.33,0.33 and 0.33

Parameter Settings

The EM algorithm performance depends on the initial parameters that are initialized for mean, covariance and alpha. Since the alpha values are same in the program , the cluster mean values calculated after classifying the data points does not change.

Maximum log-likelihood from different initializations:

The maximum log-likelihood observed after running the algorithm multiple times from a number of different initializations points is -15098.2504919 and the final means are well spread from each other (5.50927940017642, 15.449160792031073 and 25.486654429329228).

EM for general GMMs with known variance (variance = 1.0):

Run 1:

```
the initial cluster mean is [[  3.31284742  26.71589812  25.69380241]]
the initial cluster variance np is [[1 1 1]]
the inital cluster alpha np is [[ 0.33333333  0.33333333  0.33333333]]
inital max_log_likelihood is -119805.72653
```

```
final_max_log_likelihood is -15098.2504919
the final variance is [[ 1.03025768  0.99809662  0.96711594]]
the final  mean is  [[  5.5092794   25.48665443  15.44916079]]Run 2:
```

k = 3 (number of clusters)

```
the initial cluster mean is [[  5.25451902  25.86313166  25.76205156]]
the initial cluster variance np is [[1 1 1]]
the inital cluster alpha np is [[ 0.33333333  0.33333333  0.33333333]]
inital max_log_likelihood is -102574.476388
final_max_log_likelihood is -15098.2504919
the final variance is [[ 1.03025768  0.99809662  0.96711594]]
the final  mean is  [[  5.5092794   25.48665443  15.44916079]]Run 3:
```

k = 3 (number of clusters)

```
the initial cluster mean is [[ 13.17702536  26.01719181   6.00490361]]
the initial cluster variance np is [[1 1 1]]
the inital cluster alpha np is [[ 0.33333333  0.33333333  0.33333333]]
inital max_log_likelihood is -20788.964225
final_max_log_likelihood is -15098.2504919
the final variance is [[ 0.96711594  0.99809662  1.03025768]]
the final  mean is  [[ 15.44916079  25.48665443   5.5092794 ]]
```

EM for GMMs with unknown variance VS EM for GMMs with known variance:

It is inferred that the Expected Maximization for GMM with known variance is better in convergence than the expected maximization for GMM with unknown variance. The Expected Maximization for GMM with known variance converges in less than 5 iterations and the data points are spread equally among the 3 clusters.