Convergence with the number of clusters equal to 3

Heuristic 1

Enter the number of clusters: 3

Enter the maximum number of iteration below which the convergence could be reached: 1000

1. Randomized K-Mean Selection Heuristic

2. Randomized Class Distribution Heuristic

3. Randomized K-Mean Selection with Variance equal to 1

4. Randomized Class Distribution Heuristic with Variance equal to 1

5. Randomized Weight Distribution without E-Step

6. Exit

Please enter any of the five above heuristics: 1

The values for Mean and Variance are as under:

14.7572305909 27.67262932254875

6.6344195013 8.962670135925778

15.2551156919 22.939944481757774

Values for Alpha:

0.40268848492225595

0.18103697173130862

0.4162745433464353

The log likelihood values are as under:

-25386.104

-24334.953

-23895.664

-23518.402

-22380.025

-21770.736

-21606.303

-21162.123

-21150.676

-21150.664

-21150.664

The value of convergence is: -21150.664

Number of iterations took for convergence: 10

Here we specified the total number of clusters and also the number of iterations that you would want if the convergence did not occur for a considerable amount of time. In simple terms you keep a bound on the number of iterations so as to avoid running the program for a prolonged period of time. After this, we provide which heuristic to select.

Option 1 provides the heuristic with random selection of few data items and then finding out the mean and variance from it. It also assigns various alpha values. The initialization values are similarly shown when the heuristic is run. First value in mean and variance section is that of mean and the other one is that for variance.

Heuristic 2

Please enter any of the five above heuristics: 2

The values for Mean and Variance are as under:

15.866940992000806 69.21319415378159

15.231379982091427 69.52155616741724

15.48994635988101 67.49962257204764

The Values for Alphas are as under:

0.34057804464495006

0.3269359616425218

0.3324859937125282

The log likelihood values are as under:

-26683.547

-26566.428

-25689.145

-23289.04

-22076.309

-21514.764

-21154.756

-21150.672

-21150.664

-21150.664

The value of convergence is: -21150.664

Number of iterations took for convergence: 9

Initial bold values indicate the initialized values while running heuristic 2. Heuristic 2 is randomly allocating the class to each data record of the data set and then computing each class' mean and variance. It also computes the values for alpha.

This heuristic takes less time than heuristic 1.

Heuristic 3

No initialization of parameters

The log likelihood values are as under:

-26695.252

-26691.998

-26658.398

-26336.37

-24550.01

-22289.137

-21867.602

-21170.535

-21150.682

-21150.664

-21150.664

The value of convergence is: -21150.664

Number of iterations took for convergence: 10

Heuristic 3 which is denoted by 5 in the output is the one where we actually skip the E-step of the EM algorithm. We randomly allot the w_{ik} values by selecting random values from the dataset. After each M-step, alpha, mean, variance and GMM values are computed. It has the same convergence iteration as heuristic 1.

Convergence with the number of clusters equal to 5

Heuristic 1

Enter the number of clusters: 5

Enter the maximum number of iteration below which the convergence could be reached: 1000

1. Randomized K-Mean Selection Heuristic

2. Randomized Class Distribution Heuristic

3. Randomized K-Mean Selection with Variance equal to 1

4. Randomized Class Distribution Heuristic with Variance equal to 1

5. Randomized Weight Distribution without E-Step

6. Exit

Please enter any of the five above heuristics: 1

The values for Mean and Variance are as under:

6.33379297976 16.136752234543525

15.1453923929 50.719135601253704

25.3668345462 8.68783550372636

13.6924593346 28.611885789272897

6.53094238101 22.357430976690495

Values for Alpha:

0.09443637391536383

0.2258166542040986

0.3782175830358263

0.20415353227919936

0.09737585656551204

The log likelihood values are as under:

-27079.906

-22898.578

-22105.047

-21185.48

-21150.686

-21150.664

-21150.664

The value of convergence is: -21150.664

Number of iterations took for convergence: 6

Heuristic 2

Please enter any of the five above heuristics: 2

The values for Mean and Variance are as under:

15.622060118805786 67.81661629281787

15.426367630674681 68.28750795373047

15.467094179473477 67.07146868755488

15.522510849695065 68.83817520764067

15.730668350899062 63.6490894542124

The Values for Alphas are as under:

0.200878501143826

0.19836216121158107

0.19888584938185092

0.19959843258584803

0.20227505567689394

The log likelihood values are as under:

-31872.438

-31734.982

-29793.988

-25116.219

-22320.559

-21769.316

-21211.164

-21150.68

-21150.664

-21150.664

The value of convergence is: -21150.664

Number of iterations took for convergence: 9

Heuristic 3

Please enter any of the five above heuristics: 5

No initialization of parameters

The log likelihood values are as under:

-31877.992

-31864.967

-31605.896

-29401.295

-26455.365

-24224.957

-22491.271

-21764.984

-21552.729

-21156.973

-21150.67

-21150.664

-21150.664

The value of convergence is: -21150.664

Number of iterations took for convergence: 12

Heuristic 1 shows the best possible result while 3 shows the least efficient one with the highest number of iterations.

Convergence for cluster size equal to 10

Heuristic 1

Enter the number of clusters: 10

Enter the maximum number of iteration below which the convergence could be reached: 1000

1. Randomized K-Mean Selection Heuristic

2. Randomized Class Distribution Heuristic

3. Randomized K-Mean Selection with Variance equal to 1

4. Randomized Class Distribution Heuristic with Variance equal to 1

5. Randomized Weight Distribution without E-Step

6. Exit

Please enter any of the five above heuristics: 1

The values for Mean and Variance are as under:

13.5322447643 34.463338868431464

14.5125836907 51.49489310639736

5.50767493898 19.505925643369498

15.8934101938 38.71794684786793

15.3451098337 23.164202148754622

4.82367363965 13.949216592348089

4.54081464085 48.5659660544301

6.72985958463 56.57845951251815

6.68932913439 58.57612775693711

5.42000799731 21.50129620125332

Values for Alpha:

0.145516287909943

0.15605816650791898

0.05922568103773502

0.17090660817288716

0.16501056989903584

0.051870409851193774

0.048828742173419676

0.07236819921363329

0.07193236312221092

0.05828297211202221

The log likelihood values are as under:

-30825.094

-25490.357

-22820.984

-21787.955

-21397.627

-21151.18

-21150.668

-21150.664

-21150.664

The value of convergence is: -21150.664

Number of iterations took for convergence: 8

Heuristic 2

The values for Mean and Variance are as under:

15.776042964445187 68.212567825923

15.573206684325331 66.9528459952885

15.25061577306493 69.90615860781031

15.441322168009945 68.31619796697566

15.484004474759029 67.02321335214745

15.438137052961936 66.69301374672207

15.567559514264847 67.46112861428033

15.442275671541992 67.11056722245274

15.52810872802501 68.60247260459752

15.860825886965232 63.75409568184658

The Values for Alphas are as under:

0.10154370386521919

0.10023813267680187

0.09816175167071163

0.0993892479279888

0.09966397585099088

0.09936874669203626

0.1002017842359672

0.09939538522620173

0.09994785624120854

0.1020894156128739

The log likelihood values are as under:

-45230.336

-40976.32

-27200.21

-22959.18

-21732.924

-21510.45

-21154.41

-21150.67

-21150.664

-21150.664

The value of convergence is: -21150.664

Number of iterations took for convergence: 9

<u>Heuristic 3</u>

Please enter any of the five above heuristics: 5

The values for Mean and Variance are as under:

5.78546304021 1.0

15.5217229881 1.0

26.6379816143 1.0

25.8174007241 1.0

6.27896041458 1.0

26.1275444488 1.0

26.2033698462 1.0

6.71019859577 1.0

7.24379193246 1.0

13.9943101647 1.0

Values for Alpha:

0.03608680264444202

0.09681668524719018

0.1661543041033111

0.16103593407266045

0.03916499054931093

0.1629704543188143

0.16344341493275175

0.041854836984970945

0.04518312329493282

0.08728945385161549

The log likelihood values are as under:

-35838.92

-28561.49

-24106.695

-21998.97

-21780.826

-21203.934

-21150.684

-21150.664

-21150.664

The value of convergence is: -21150.664

Number of iterations took for convergence: 8

Heuristic 1 gives the best result out of thee three heuristics. For larger clusters heuristic 1 converges at a less number of iterations than both 2 and 3. Heuristic 2 is better for lower clusters like between 1 and 4. As the cluster size gets larger and larger the time of convergence and number of iterations gets increased too. Heuristic 2 has an overhead of assigning each data point a class which becomes a demerit for it when it compared with both 1 and 3.

Heuristic 3 is slow for lower clusters while taking the highest number of iterations. It takes less time for computation since there are no initial parameters assigned to it.

As more the number of clusters is given more is the size of the number of parameters. Thus heuristic 2 is much more sensitive to the number of parameters.

The initial randomization of parameters result in delivering higher log likelihood value in case of heuristic 2 while heuristic 1 provides the least. Hence the gradual decrease of log likelihood value is much higher in 2 as well.

In terms of stability, heuristic 1 gives better result over heuristic 2. The ranking can be given as under:

Heuristic 1 > Heuristic 2 > Heuristic 3

Convergence with variance equal to 1 only at initialization

Heuristic 3

Enter the number of clusters: 3

Enter the maximum number of iteration below which the convergence could be reached: 100

1. Randomized K-Mean Selection Heuristic
2. Randomized Class Distribution Heuristic
3. Randomized K-Mean Selection with Variance equal to 1
4. Randomized Class Distribution Heuristic with Variance equal to 1
5. Randomized Weight Distribution without E-Step
6. Exit

Please enter any of the five above heuristics: 3

The values for Mean and Variance are as under:

14.0042258263 1.0

4.2771348021 1.0

25.4126566431 1.0

Values for Alpha:

0.32050671237854894

0.09788833962149361

0.5816049479999574

The log likelihood values are as under:

-25835.312

-23062.43

-21934.217

-21480.121

-21152.723

-21150.668

-21150.664

-21150.664

The value of convergence is: -21150.664

Number of iterations took for convergence: 7

Heuristic 4

Please enter any of the five above heuristics: 4

The values for Mean and Variance are as under:

15.51916672706098 1.0 15.440368115624844 1.0 15.516161301262603 1.0

The Values for Alphas are as under:

0.33392004885722887

0.3322245688972929

0.3338553822454783

The log likelihood values are as under:

-26545.996

-25579.102

-24675.447

-23925.924

-23138.053

-21889.69

-21888.322

-21302.74

-21150.703

-21150.664

-21150.664

The value of convergence is: -21150.664

Number of iterations took for convergence: 10

The overall performance increases when heuristic 4 and 5 are run. These heuristics only assign variance equal to 1 at the time of initialization. During the computation of Gaussian mixture models (GMM) the probability values have tend to remain lower in most of the cases with the reason of exponent numerator remaining low as always. While performing the E-step for the computation of weighted values, some values may tend to become low and much moderate due to lower GMM values. Also there is no major decrease in the values across all the iterations as the sensitivity towards random initial values (for both alpha and mean) is also the least. Hence with heuristic 3 we can get much better results than with heuristic 1 (As it is concluded to be better than 2 and 3).

Although there is an increase of performance with variance equal to 1, if variance is kept 1 throughout and never changed as the number of iterations across EM algorithm kept going, the values appear to never converge. Hence the performance is compromised when variance is kept the constant. GMM value appears to descend further and further since the negative exponent decreases time after time with more and more number of iterations. This causes every GMM value to decrease and hence the huge fluctuation in weight calculation resulting in convergence at a very long time.

[Note: Result from variance equal to 1 throughout has not been posted since the values did not appear to converge at a particular point. However the code for this has been kept in a separate folder.]