# **CS6140 Assignment5 Guanglei“Garrett” Wu**

## **3. K-Means Clustering**

3.1.3

To select the optional number of clusters based on SSE, we can apply Elbow Method, which is: choose a number of clusters so that adding another cluster doesn't give much better modeling of the data.

3.2.1 SSE vs k:

**K-Means for set: dermatologyData.csv**

1 clusters, SSE: 92806.72, NMI: 0.0000

2 clusters, SSE: 35335.63, NMI: 0.0414

3 clusters, SSE: 21262.41, NMI: 0.0639

4 clusters, SSE: 16940.16, NMI: 0.0856

5 clusters, SSE: 14840.54, NMI: 0.1018

6 clusters, SSE: 13101.24, NMI: 0.1531

7 clusters, SSE: 11639.53, NMI: 0.1932

8 clusters, SSE: 10861.30, NMI: 0.1446

9 clusters, SSE: 10369.73, NMI: 0.1814

10 clusters, SSE: 9520.18, NMI: 0.3074

11 clusters, SSE: 9234.13, NMI: 0.1996

12 clusters, SSE: 8720.89, NMI: 0.3153

13 clusters, SSE: 8698.65, NMI: 0.3059

14 clusters, SSE: 8137.86, NMI: 0.3673

15 clusters, SSE: 7770.46, NMI: 0.3797

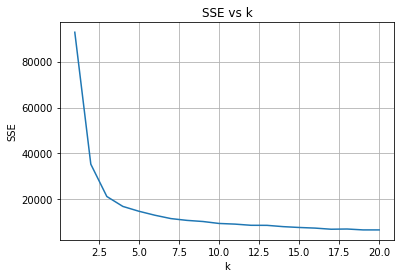
16 clusters, SSE: 7497.11, NMI: 0.3654

17 clusters, SSE: 7005.85, NMI: 0.3584

18 clusters, SSE: 7102.51, NMI: 0.3910

19 clusters, SSE: 6709.20, NMI: 0.4011

20 clusters, SSE: 6706.56, NMI: 0.4226



**K-Means for set: vowelsData.csv**

1 clusters, SSE: 5200.20, NMI: 0.0000

2 clusters, SSE: 3982.86, NMI: 0.2079

3 clusters, SSE: 3402.33, NMI: 0.2393

4 clusters, SSE: 3175.29, NMI: 0.2590

5 clusters, SSE: 2915.51, NMI: 0.3550

6 clusters, SSE: 2671.24, NMI: 0.3343

7 clusters, SSE: 2572.75, NMI: 0.3786

8 clusters, SSE: 2474.13, NMI: 0.4156

9 clusters, SSE: 2272.69, NMI: 0.3633

10 clusters, SSE: 2222.30, NMI: 0.3844

11 clusters, SSE: 2049.30, NMI: 0.4208

12 clusters, SSE: 2030.24, NMI: 0.4072

13 clusters, SSE: 1993.80, NMI: 0.4112

14 clusters, SSE: 1843.95, NMI: 0.4366

15 clusters, SSE: 1846.63, NMI: 0.4320

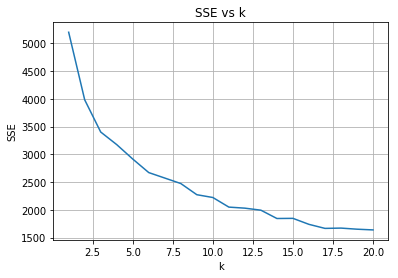
16 clusters, SSE: 1738.57, NMI: 0.4228

17 clusters, SSE: 1665.74, NMI: 0.4478

18 clusters, SSE: 1671.01, NMI: 0.4285

19 clusters, SSE: 1651.26, NMI: 0.4376

20 clusters, SSE: 1638.25, NMI: 0.4354



**K-Means for set: glassData.csv**

1 clusters, SSE: 229.46, NMI: 0.0000

2 clusters, SSE: 136.55, NMI: 0.2736

3 clusters, SSE: 116.78, NMI: 0.2310

4 clusters, SSE: 97.41, NMI: 0.3217

5 clusters, SSE: 86.19, NMI: 0.3348

6 clusters, SSE: 82.64, NMI: 0.3370

7 clusters, SSE: 79.59, NMI: 0.3013

8 clusters, SSE: 74.93, NMI: 0.2817

9 clusters, SSE: 63.87, NMI: 0.3469

10 clusters, SSE: 64.60, NMI: 0.3370

11 clusters, SSE: 61.35, NMI: 0.3281

12 clusters, SSE: 54.69, NMI: 0.3870

13 clusters, SSE: 54.29, NMI: 0.3461

14 clusters, SSE: 46.35, NMI: 0.4086

15 clusters, SSE: 53.37, NMI: 0.3392

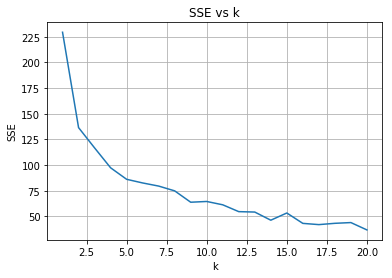
16 clusters, SSE: 43.26, NMI: 0.3501

17 clusters, SSE: 42.05, NMI: 0.3745

18 clusters, SSE: 43.32, NMI: 0.3736

19 clusters, SSE: 44.04, NMI: 0.3723

20 clusters, SSE: 36.89, NMI: 0.3469



**K-Means for set: ecoliData.csv**

1 clusters, SSE: 54.67, NMI: 0.0000

2 clusters, SSE: 30.99, NMI: 0.4754

3 clusters, SSE: 20.76, NMI: 0.6829

4 clusters, SSE: 18.58, NMI: 0.6031

5 clusters, SSE: 16.44, NMI: 0.5959

6 clusters, SSE: 14.83, NMI: 0.6211

7 clusters, SSE: 14.69, NMI: 0.5841

8 clusters, SSE: 13.23, NMI: 0.5789

9 clusters, SSE: 12.62, NMI: 0.5729

10 clusters, SSE: 11.99, NMI: 0.5489

11 clusters, SSE: 11.60, NMI: 0.5476

12 clusters, SSE: 11.55, NMI: 0.5581

13 clusters, SSE: 11.33, NMI: 0.5236

14 clusters, SSE: 10.71, NMI: 0.5052

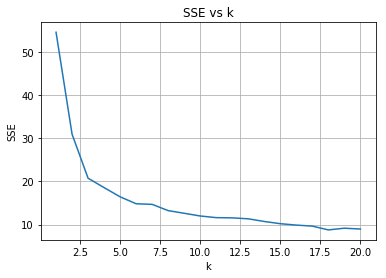
15 clusters, SSE: 10.20, NMI: 0.5240

16 clusters, SSE: 9.87, NMI: 0.4756

17 clusters, SSE: 9.64, NMI: 0.5359

18 clusters, SSE: 8.75, NMI: 0.5055

19 clusters, SSE: 9.15, NMI: 0.5019

20 clusters, SSE: 8.96, NMI: 0.4977

**K-Means for set: yeastData.csv**

1 clusters, SSE: 121.25, NMI: 0.0000

2 clusters, SSE: 94.60, NMI: 0.1376

3 clusters, SSE: 80.34, NMI: 0.1444

4 clusters, SSE: 75.16, NMI: 0.1683

5 clusters, SSE: 68.18, NMI: 0.1909

6 clusters, SSE: 64.24, NMI: 0.1597

7 clusters, SSE: 60.80, NMI: 0.1975

8 clusters, SSE: 51.48, NMI: 0.2214

9 clusters, SSE: 48.46, NMI: 0.2631

10 clusters, SSE: 53.25, NMI: 0.2067

11 clusters, SSE: 52.18, NMI: 0.2207

12 clusters, SSE: 44.48, NMI: 0.2446

13 clusters, SSE: 45.33, NMI: 0.2461

14 clusters, SSE: 43.81, NMI: 0.2231

15 clusters, SSE: 38.97, NMI: 0.2568

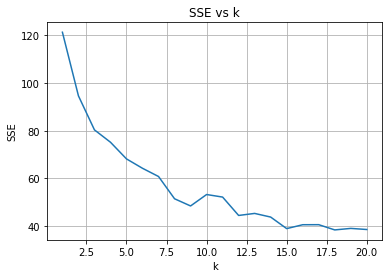
16 clusters, SSE: 40.60, NMI: 0.2439

17 clusters, SSE: 40.62, NMI: 0.2461

18 clusters, SSE: 38.42, NMI: 0.2258

19 clusters, SSE: 39.06, NMI: 0.2450

20 clusters, SSE: 38.59, NMI: 0.2508



**K-Means for set: soybeanData.csv**

1 clusters, SSE: 5097.58, NMI: 0.0000

2 clusters, SSE: 3899.31, NMI: 0.2987

3 clusters, SSE: 3391.27, NMI: 0.3197

4 clusters, SSE: 3124.66, NMI: 0.4261

5 clusters, SSE: 2903.21, NMI: 0.4909

6 clusters, SSE: 2603.64, NMI: 0.4944

7 clusters, SSE: 2380.72, NMI: 0.5962

8 clusters, SSE: 2253.80, NMI: 0.5719

9 clusters, SSE: 2157.38, NMI: 0.6484

10 clusters, SSE: 1990.89, NMI: 0.6008

11 clusters, SSE: 2000.90, NMI: 0.6122

12 clusters, SSE: 1941.25, NMI: 0.5979

13 clusters, SSE: 1957.92, NMI: 0.6317

14 clusters, SSE: 1716.26, NMI: 0.6233

15 clusters, SSE: 1744.10, NMI: 0.6582

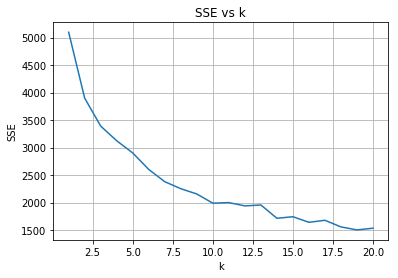
16 clusters, SSE: 1642.75, NMI: 0.6472

17 clusters, SSE: 1679.50, NMI: 0.6472

18 clusters, SSE: 1558.87, NMI: 0.6630

19 clusters, SSE: 1504.54, NMI: 0.6659

20 clusters, SSE: 1535.80, NMI: 0.6820



3.2.2 KMeans Optional number of clusters:

|  |  |  |  |
| --- | --- | --- | --- |
| Data | K for SSE | SSE | NMI |
| Dermatology | 5 | 14840.54 | 0.1018 |
| Vowels | 11 | 2049.30 | 0.4208 |
| Glass | 9 | 63.87 | 0.3469 |
| Ecoli | 6 | 14.83 | 0.6211 |
| Yeast | 9 | 48.46 | 0.2631 |
| Soybean | 14 | 1716.26 | 0.6233 |

3.2.3 KMeans n\_clusters == n\_classes

|  |  |  |  |
| --- | --- | --- | --- |
| Data | K | SSE | NMI |
| Dermatology | 6 | 13101.24 | 0.1531 |
| Vowels | 11 | 2049.30 | 0.4208 |
| Glass | 6 | 82.64 | 0.3370 |
| Ecoli | 5 | 16.44 | 0.5959 |
| Yeast | 9 | 48.46 | 0.2631 |
| Soybean | 15 | 1744.10 | 0.6582 |

## 4.Gaussian Mixture Models

4.1.4

For GMM, SSE may not be a good measure. Since GMM is optimized by log-likelihood, instead of SSE, and uses soft cluster instead of hard cluster. So data other than current class also impose influence on GMM. In this case, NMI can be a better measure for GMM.

4.2.1 & 4.2.2 SSE vs k and NMI vs k

**GaussianMixture for set: dermatologyData.csv**

1 components, SSE: 92806.72, NMI: 0.0000

2 components, SSE: 90124.93, NMI: 0.3885

3 components, SSE: 42039.21, NMI: 0.1637

4 components, SSE: 25667.26, NMI: 0.1349

5 components, SSE: 21587.92, NMI: 0.1706

6 components, SSE: 17320.75, NMI: 0.1630

7 components, SSE: 12948.67, NMI: 0.1640

8 components, SSE: 16191.39, NMI: 0.2487

9 components, SSE: 41728.47, NMI: 0.4323

10 components, SSE: 20674.45, NMI: 0.3241

11 components, SSE: 14529.45, NMI: 0.2959

12 components, SSE: 8960.51, NMI: 0.2945

13 components, SSE: 8532.93, NMI: 0.3217

14 components, SSE: 19959.30, NMI: 0.2724

15 components, SSE: 42607.13, NMI: 0.3020

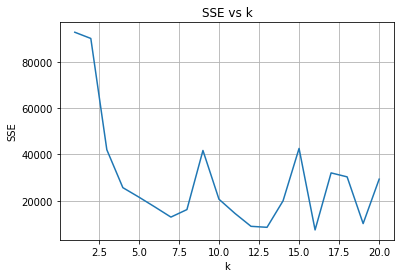
16 components, SSE: 7407.79, NMI: 0.3564

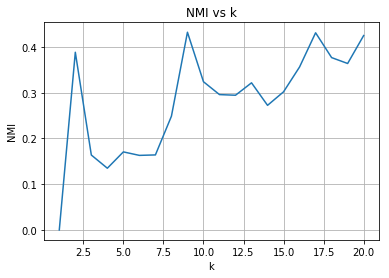
17 components, SSE: 32042.13, NMI: 0.4311

18 components, SSE: 30329.51, NMI: 0.3768

19 components, SSE: 10145.02, NMI: 0.3640

20 components, SSE: 29332.16, NMI: 0.4250





**GaussianMixture for set: vowelsData.csv**

1 components, SSE: 5200.20, NMI: 0.0000

2 components, SSE: 4332.92, NMI: 0.1587

3 components, SSE: 3607.95, NMI: 0.1840

4 components, SSE: 3451.13, NMI: 0.2222

5 components, SSE: 3078.27, NMI: 0.3350

6 components, SSE: 3151.51, NMI: 0.3217

7 components, SSE: 2820.74, NMI: 0.3146

8 components, SSE: 2694.95, NMI: 0.3083

9 components, SSE: 2529.25, NMI: 0.3997

10 components, SSE: 2346.35, NMI: 0.3638

11 components, SSE: 2332.62, NMI: 0.3773

12 components, SSE: 2176.10, NMI: 0.3849

13 components, SSE: 2134.65, NMI: 0.3914

14 components, SSE: 1910.93, NMI: 0.3932

15 components, SSE: 1958.90, NMI: 0.4528

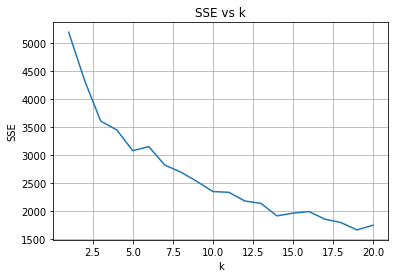
16 components, SSE: 1987.73, NMI: 0.4001

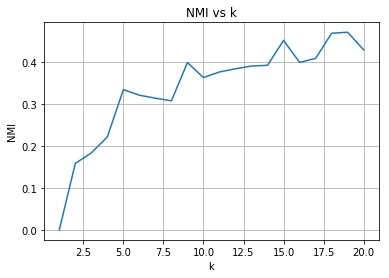
17 components, SSE: 1851.24, NMI: 0.4096

18 components, SSE: 1790.13, NMI: 0.4697

19 components, SSE: 1658.93, NMI: 0.4721

20 components, SSE: 1742.55, NMI: 0.4300





**GaussianMixture for set: glassData.csv**

1 components, SSE: 229.46, NMI: 0.0000

2 components, SSE: 153.36, NMI: 0.3338

3 components, SSE: 130.50, NMI: 0.2543

4 components, SSE: 121.92, NMI: 0.3858

5 components, SSE: 117.09, NMI: 0.2977

6 components, SSE: 83.08, NMI: 0.3567

7 components, SSE: 83.82, NMI: 0.3583

8 components, SSE: 82.52, NMI: 0.2984

9 components, SSE: 81.84, NMI: 0.3227

10 components, SSE: 68.62, NMI: 0.3507

11 components, SSE: 68.78, NMI: 0.3203

12 components, SSE: 79.48, NMI: 0.3317

13 components, SSE: 70.44, NMI: 0.3078

14 components, SSE: 63.25, NMI: 0.3290

15 components, SSE: 69.94, NMI: 0.3542

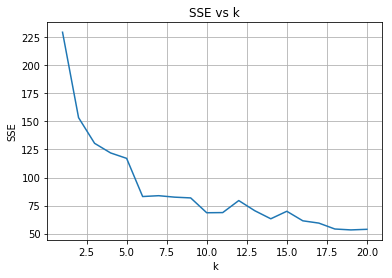
16 components, SSE: 61.44, NMI: 0.3699

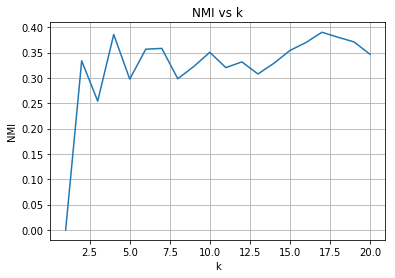
17 components, SSE: 59.40, NMI: 0.3902

18 components, SSE: 54.15, NMI: 0.3803

19 components, SSE: 53.33, NMI: 0.3711

20 components, SSE: 53.87, NMI: 0.3467





**GaussianMixture for set: ecoliData.csv**

1 components, SSE: 54.67, NMI: 0.0000

2 components, SSE: 32.62, NMI: 0.4205

3 components, SSE: 21.75, NMI: 0.6391

4 components, SSE: 23.27, NMI: 0.4876

5 components, SSE: 21.51, NMI: 0.5529

6 components, SSE: 18.80, NMI: 0.6213

7 components, SSE: 21.19, NMI: 0.6546

8 components, SSE: 17.77, NMI: 0.6085

9 components, SSE: 19.90, NMI: 0.5244

10 components, SSE: 20.48, NMI: 0.5499

11 components, SSE: 17.13, NMI: 0.5306

12 components, SSE: 16.72, NMI: 0.5005

13 components, SSE: 20.18, NMI: 0.5397

14 components, SSE: 19.11, NMI: 0.4456

15 components, SSE: 14.08, NMI: 0.5930

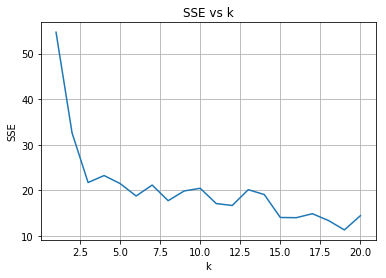
16 components, SSE: 14.04, NMI: 0.5379

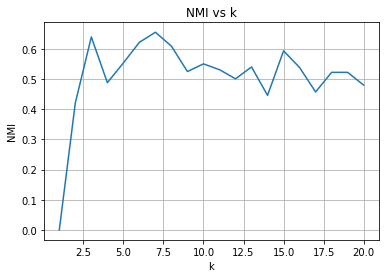
17 components, SSE: 14.91, NMI: 0.4568

18 components, SSE: 13.43, NMI: 0.5218

19 components, SSE: 11.36, NMI: 0.5218

20 components, SSE: 14.48, NMI: 0.4796





**GaussianMixture for set: yeastData.csv**

1 components, SSE: 121.25, NMI: 0.0000

2 components, SSE: 114.43, NMI: 0.1082

3 components, SSE: 104.11, NMI: 0.0970

4 components, SSE: 79.82, NMI: 0.2028

5 components, SSE: 78.36, NMI: 0.2049

6 components, SSE: 163.51, NMI: 0.0992

7 components, SSE: 77.34, NMI: 0.2086

8 components, SSE: 143.08, NMI: 0.1791

9 components, SSE: 76.84, NMI: 0.2226

10 components, SSE: 85.12, NMI: 0.2004

11 components, SSE: 91.54, NMI: 0.2017

12 components, SSE: 79.03, NMI: 0.2460

13 components, SSE: 86.43, NMI: 0.2136

14 components, SSE: 86.63, NMI: 0.2171

15 components, SSE: 86.11, NMI: 0.2421

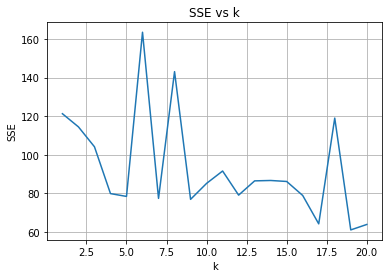
16 components, SSE: 78.85, NMI: 0.2231

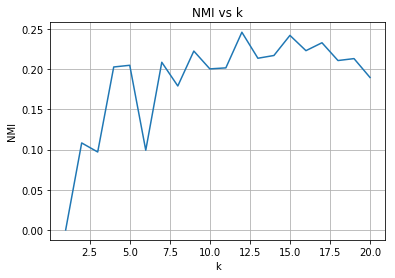
17 components, SSE: 64.13, NMI: 0.2329

18 components, SSE: 118.99, NMI: 0.2108

19 components, SSE: 60.99, NMI: 0.2132

20 components, SSE: 63.82, NMI: 0.1897





**GaussianMixture for set: soybeanData.csv**

1 components, SSE: 5097.58, NMI: 0.0000

2 components, SSE: 3889.00, NMI: 0.3234

3 components, SSE: 3440.81, NMI: 0.3003

4 components, SSE: 3049.48, NMI: 0.4362

5 components, SSE: 3173.42, NMI: 0.4397

6 components, SSE: 2628.96, NMI: 0.5238

7 components, SSE: 2600.07, NMI: 0.5596

8 components, SSE: 2450.59, NMI: 0.5464

9 components, SSE: 2420.76, NMI: 0.5874

10 components, SSE: 2801.96, NMI: 0.5782

11 components, SSE: 2855.15, NMI: 0.5501

12 components, SSE: 2365.95, NMI: 0.6088

13 components, SSE: 2149.30, NMI: 0.5977

14 components, SSE: 2072.35, NMI: 0.6281

15 components, SSE: 2189.87, NMI: 0.5720

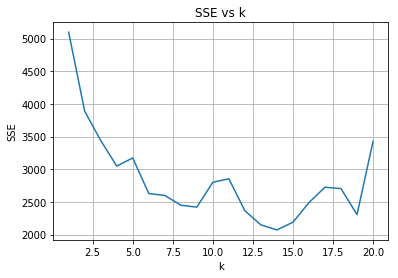
16 components, SSE: 2490.25, NMI: 0.6019

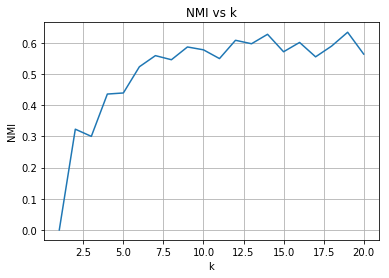
17 components, SSE: 2725.99, NMI: 0.5557

18 components, SSE: 2704.52, NMI: 0.5902

19 components, SSE: 2306.85, NMI: 0.6347

20 components, SSE: 3426.28, NMI: 0.5646





4.2.3 GMM Optimal by SSE

|  |  |  |  |
| --- | --- | --- | --- |
| Data | K for SSE | SSE | NMI |
| Dermatology | 7 | 12948.67 | 0.1640 |
| Vowels | 14 | 1910.93 | 0.3932 |
| Glass | 6 | 83.08 | 0.3567 |
| Ecoli | 3 | 21.75 | 0.6391 |
| Yeast | 4 | 79.82 | 0.2028 |
| Soybean | 14 | 2072.35 | 0.6281 |

4.2.4 GMM Optimal by NMI

For NMI measure, we simply choose the ones with the highest value.

|  |  |  |  |
| --- | --- | --- | --- |
| Data | K for SSE | SSE | NMI |
| Dermatology | 9 | 41728.47 | 0.4323 |
| Vowels | 19 | 1658.93 | 0.4721 |
| Glass | 17 | 59.40 | 0.3902 |
| Ecoli | 7 | 21.19 | 0.6546 |
| Yeast | 12 | 79.03 | 0.2460 |
| Soybean | 19 | 2306.85 | 0.6347 |

4.2.5 GMM n\_components == n\_classes

|  |  |  |  |
| --- | --- | --- | --- |
| Data | K | SSE | NMI |
| Dermatology | 6 | 17320.75 | 0.1630 |
| Vowels | 11 | 2332.62 | 0.3773 |
| Glass | 6 | 83.08 | 0.3567 |
| Ecoli | 5 | 21.51 | 0.5529 |
| Yeast | 9 | 76.84 | 0.2226 |
| Soybean | 15 | 2189.87 | 0.5720 |

## 5. Comparing K-Means and GMM

5.1

If no significant discrepancies, we should prefer K-Means than GMM, since it’s faster and easier to explain. Based on the NMI measure, it seems we should only apply GMM on Vowels data, which gets a relatively significant improve (best NMI from 0.4478 to 0.4721). For other datasets, the improvements are trivial, we would select K-Means.

5.2

If the classes are fully separable, in the best case, clustering with n\_clusters == n\_classes can give perfect results, NMI can be as god as close to 1. So we can assess the separability of the datasets by analyzing the performance of NMI. According to tables in 3.2.3 and 4.2.5, for our datasets, for n\_clusters == n\_classes, it seems:

|  |  |
| --- | --- |
| Dermatology | Less separable |
| Vowels | Neutral |
| Glass | Neutral |
| Ecoli | More separable |
| Yeast | Less separable |
| Soybean | More separable |

5.3

Yes, both of the algorithms are sensitive to initialization, since they can only achieve to local optimal, instead of global optimal. In order to solve the problem, we can run the algorithm multiple times with different (random) initializations, and select the best outcome. For K-Means, we may run the algorithm several time with different random seeds. For GMM, we may run with different K-Means outcomes as inputs.

5.4

If the datasets are less separable, by showing low NMIs, it may also because we implied inappropriate distance metric for the datasets. So we can get the same results with 5.2, as the confidence of the distance metric. While further investigation could be taken to analyze the appropriateness of the metric, like trying to compare difference metrics, analyzing the meaning of the features of each dataset.

|  |  |
| --- | --- |
| Dermatology | Less confident |
| Vowels | Neutral |
| Glass | Neutral |
| Ecoli | More confident |
| Yeast | Less confident |
| Soybean | More confident |