3

3.1

Select the optimal number of clusters based on the SSE criterion,and calculate the NMI of the resulting clustering. Briey explain how you selected the optimal number of clusters.

Since overall, the sse curve goes down along the axis, so when trying k from 1-30 it's highly likely eventually the best k value (with lowest sse) near 30 will be given, which is way too complex and time-consuming.

So there is a trade off choice to consider an acceptable range (0.8 * best_sse to 1.0 * best_sse) instead of the optimal point. And within that range, we get the correspondingly min k value as the final choice.

By doing this, we can get relatively high performance on sse as well as achieve less complexities of computing.

I didn't take the elbow strategy because it will always give 2 as final answers. From 1-2 drops hugely.

3.2

1. For each dataset provide the plot of the SSE vs k (number of cluster).

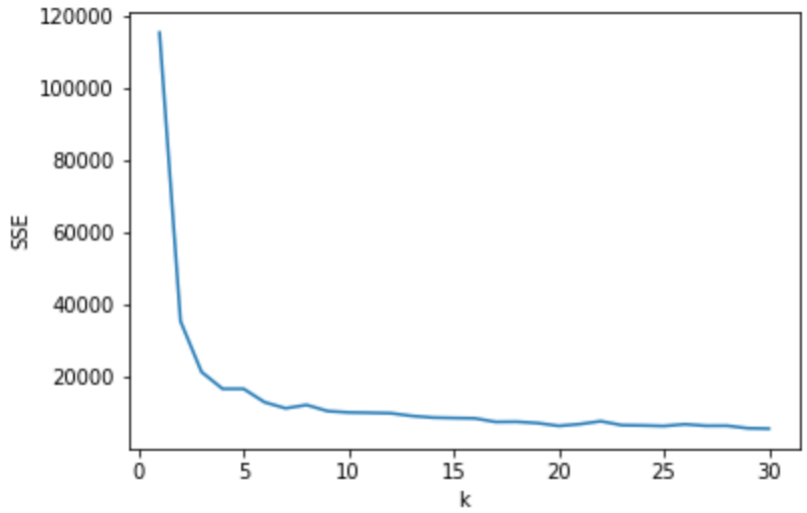

Figure 3-1 Dermatology

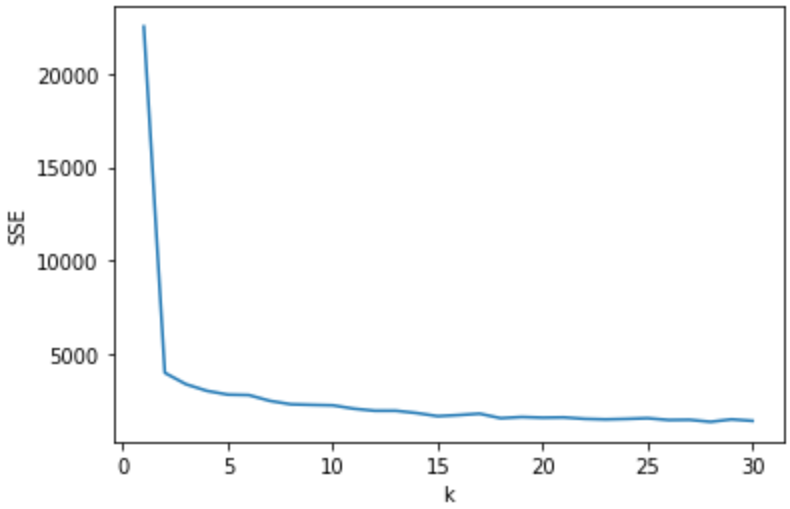

Figure 3-2 Vowels

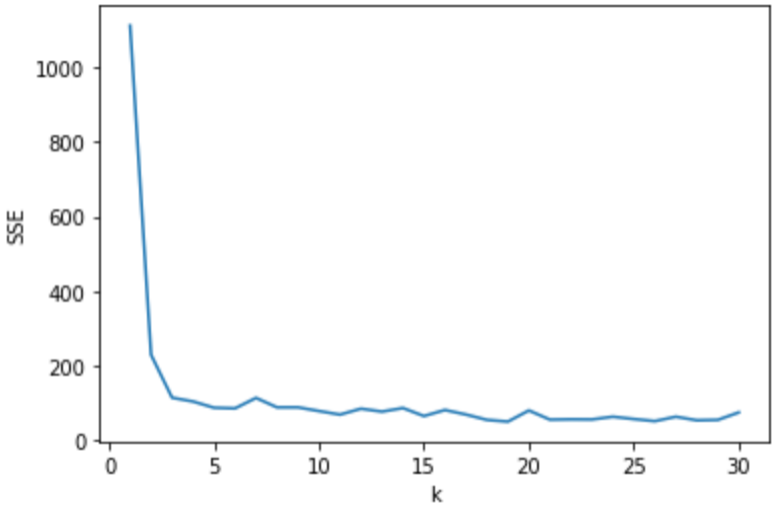

Figure 3-3 Glass

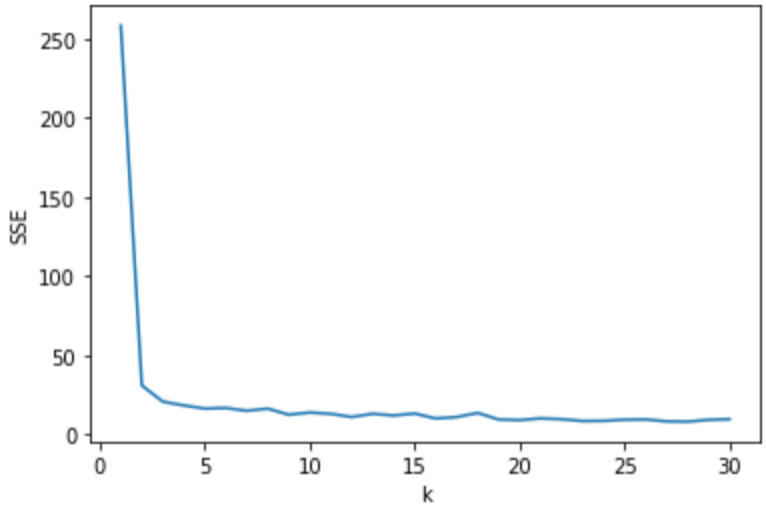

Figure 3-4 Ecoli

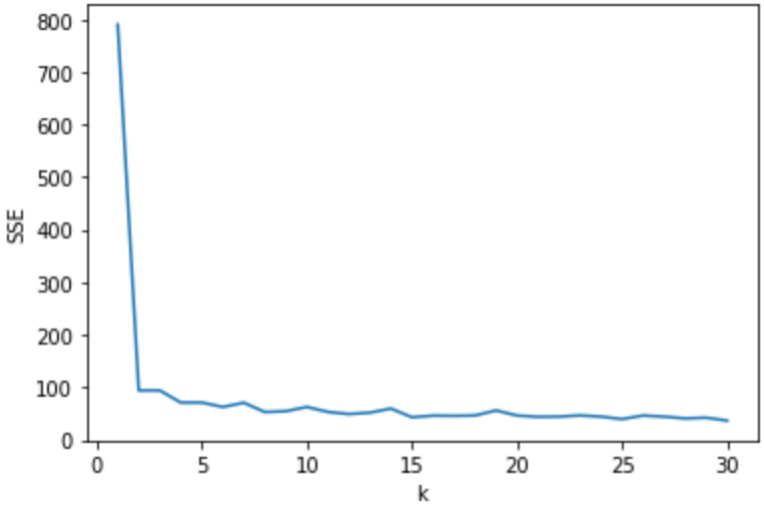

Figure 3-5 Yeast

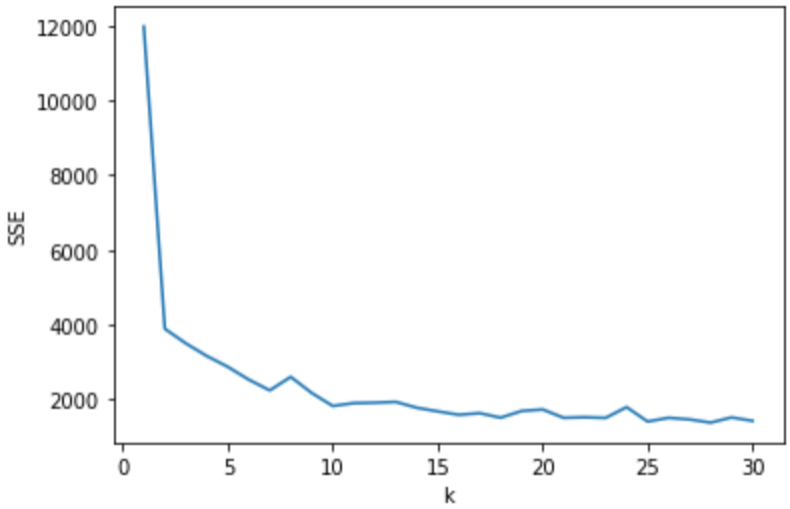

Figure 3-6 Soybean

2. In a table provide the optimal number of clusters for each dataset based on the SSE criterion, and the corresponding NMI.

	Dermatology	Vowels	Glass	Ecoli	Yeast	Soybean
K	20.000000	18.000000	18.000000	19.000000	15.000000	16.000000
SSE	6284.330749	1534.141627	54.906744	9.458943	43.151097	1582.683515
NMI	0.417946	0.494878	0.337523	0.532981	0.248423	0.663247

3. Set the number of clusters equal to the number of classes for each dataset and run the k-means algorithm. List the resulting NMI and SSE for each dataset in a table.

	Dermatology	Vowels	Glass	Ecoli	Yeast	Soybean
K	6.000000	11.000000	6.000000	5.000000	9.000000	15.000000
SSE	12860.877658	2044.392513	85.792215	16.335078	54.748263	1673.751007
NMI	0.164842	0.432501	0.412476	0.667179	0.226311	0.682566

4.

4.1

4. Briey explain which criterion i.e., SSE or NMI is better for GMM and why.

NMI is better for GMM. GMM is a probabilistic model and essentially it optimize NMI since NMI is also based on probabilities.

While K-means is the opposite, it mainly optimizes SSE and it already includes the process in its algorithm logic.

And it can be analyzed from graphs below that overall the SSE values given by GMM are much higher (bad and not indicative) than results calculated by k-means while the NMI values remain high (good).

4.2

Notice: the plot of the SSE vs k and the plot of the NMI vs k, both of which show huge fluctuations but they increase or decrease overall. The huge fluctuations are simply caused by different initialized parameters generated by "random generator".

Which means that when you try the same k value, it will give very different results for sse and nmi each run because different initial parameters (generated by random number generator) cause different local optimal solutions.

1. For each dataset provide the plot of the SSE vs k (number of clusters).

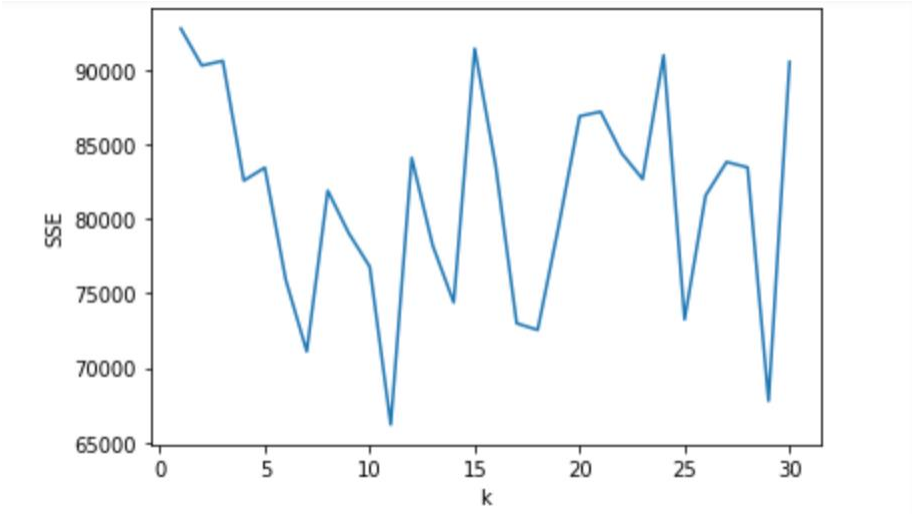

Figure 4-1 Dermatology

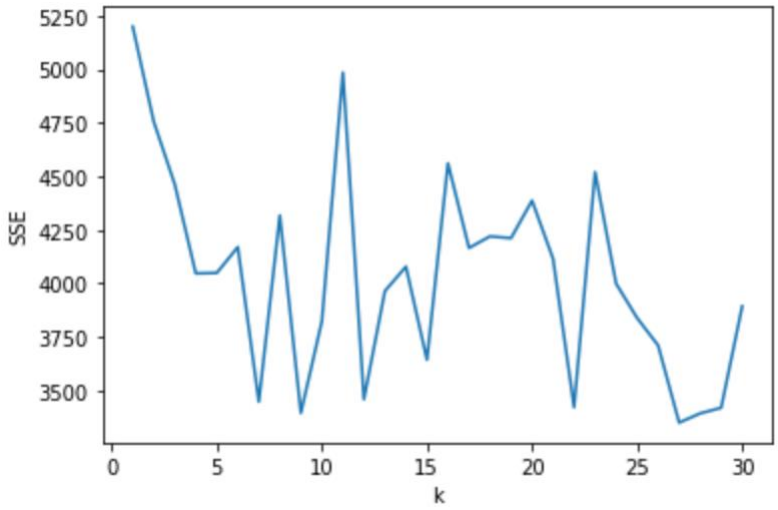

Figure 4-2 Vowels

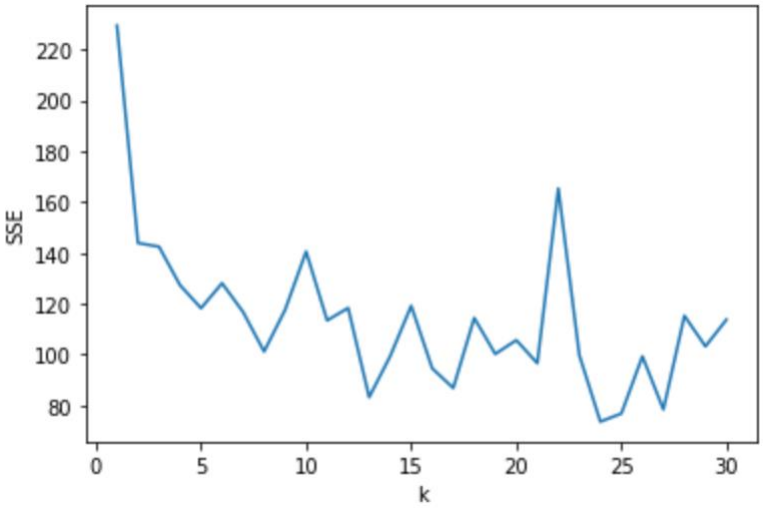

Figure 4-3 Glass

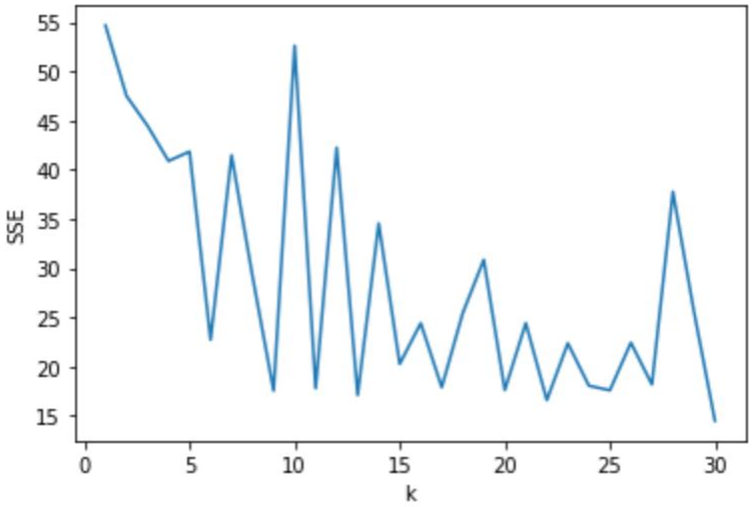

Figure 4-4 Ecoli

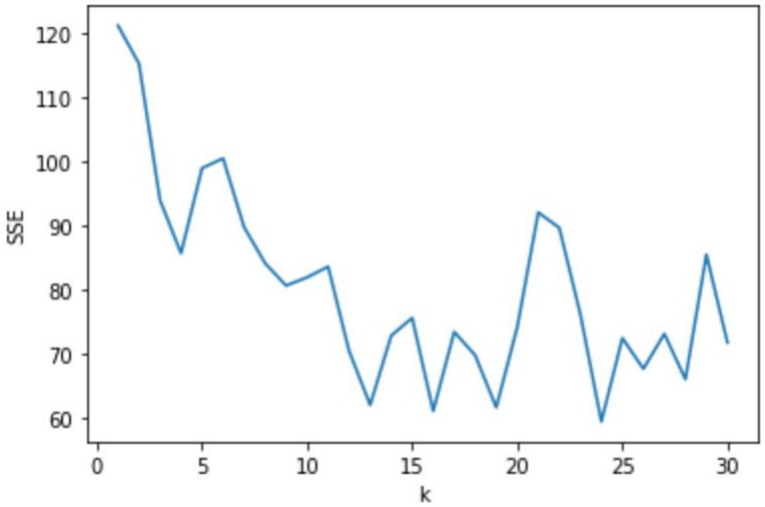

Figure 4-5 Yeast

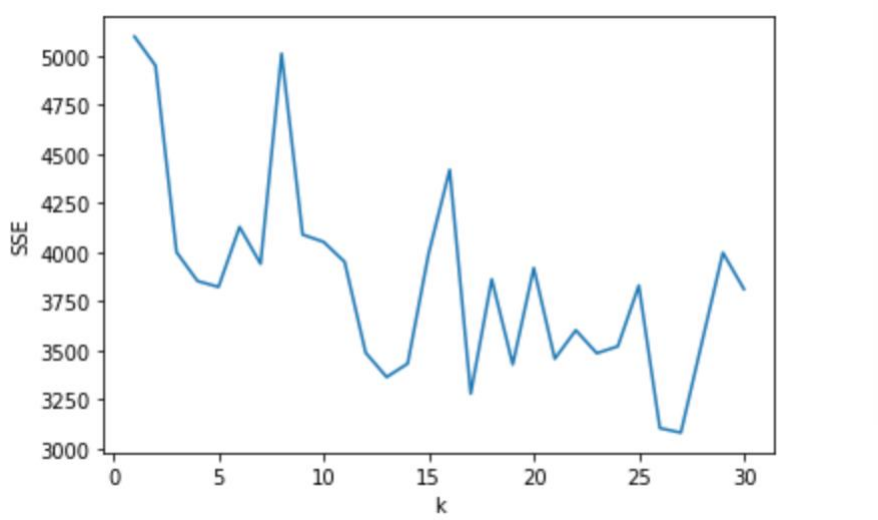

Figure 4-6 Soybean

2. For each dataset provide the plot of the NMI vs k (number of clusters).

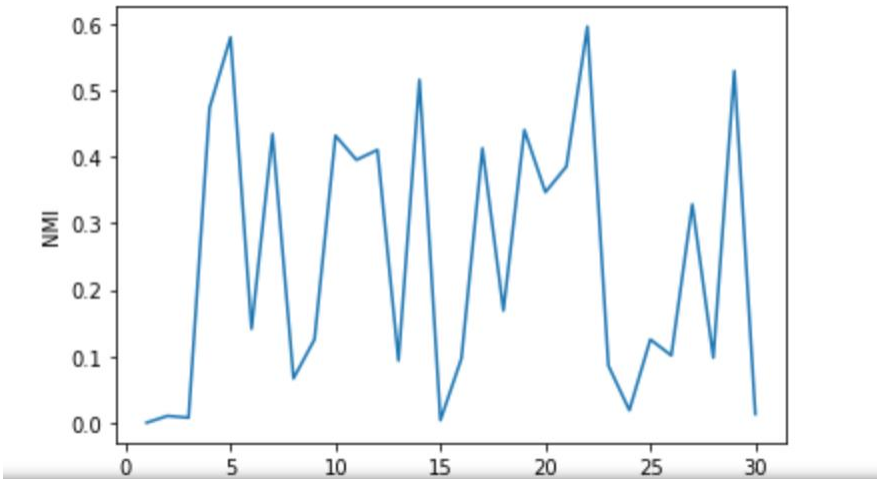

Figure 4-7 Dermatology

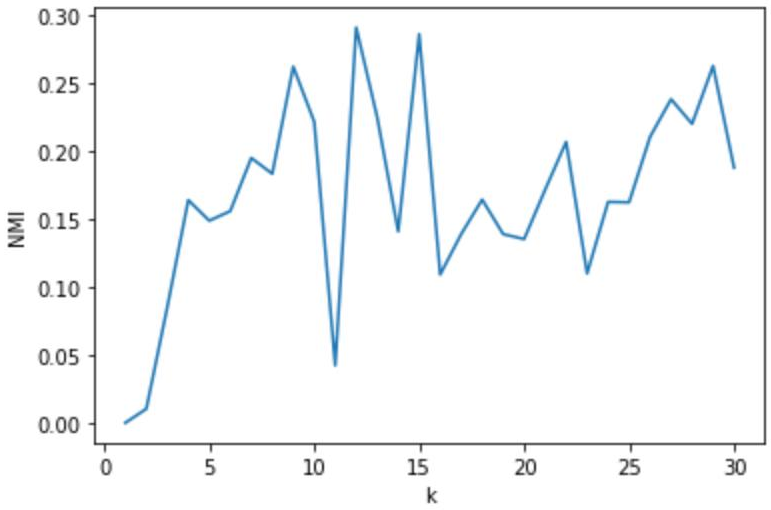

Figure 4-8 Vowels

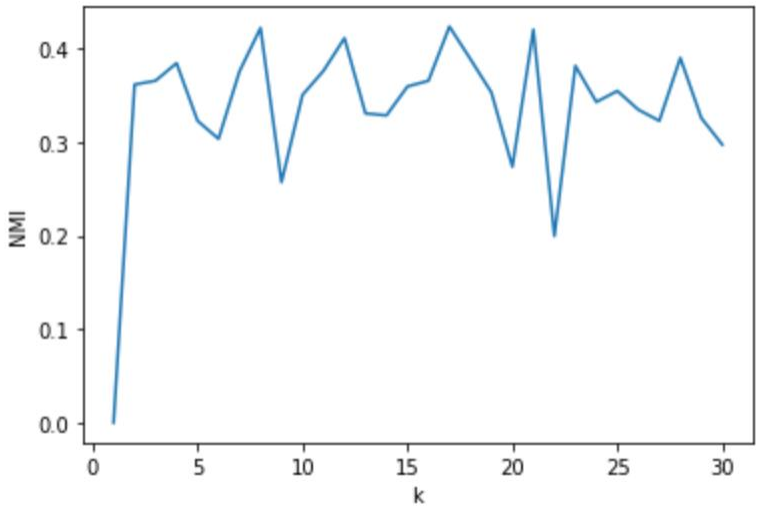

Figure 4-9 Glass

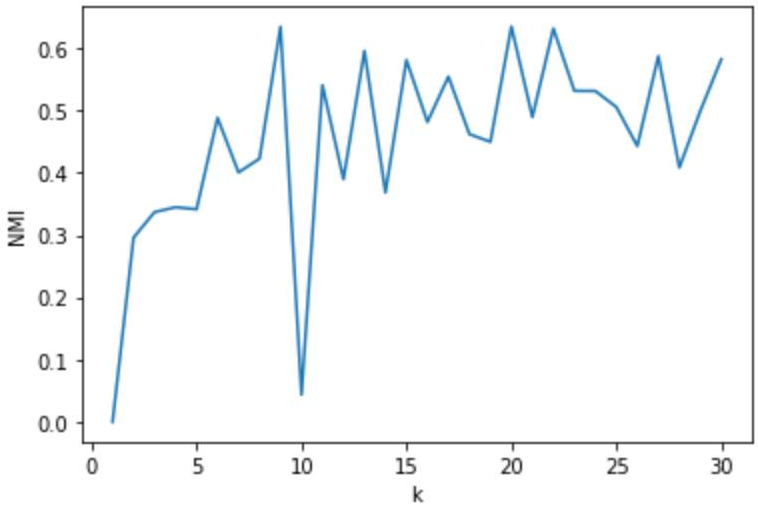

Figure 4-10 Ecoli

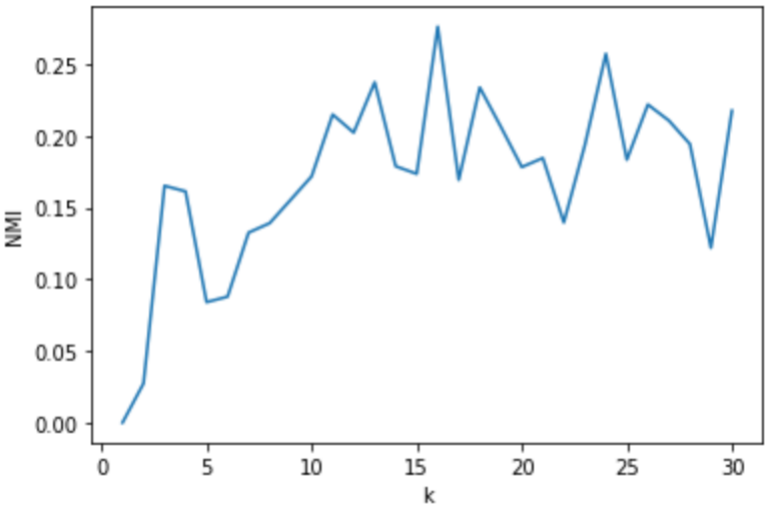

Figure 4-11 Yeast

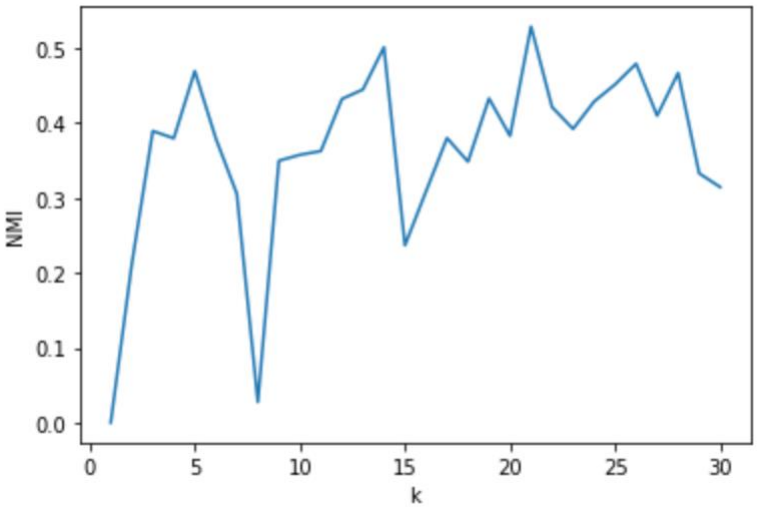

Figure 4-12 Soybean

3. In a table provide the optimal number of clusters for each dataset based on the SSE criterion, and the corresponding NMI.

	Dermatology	Vowels	Glass	Ecoli	Yeast	Soybean
K	6.000000	7.000000	13.000000	13.000000	12.000000	12.000000
SSE	75910.291542	3447.347441	83.200351	17.082069	70.447222	3488.397540
NMI	0.141384	0.195272	0.330663	0.595754	0.202216	0.431863

4. In another table provide the optimal number of clusters for each dataset based on the NMI criterion, and the corresponding SSE.

	Dermatology	Vowels	Glass	Ecoli	Yeast	Soybean
K	5.000000	9.000000	2.000000	9.000000	13.000000	5.000000
SSE	83458.138302	3394.754339	143.865729	17.529189	62.047045	3823.270058
NMI	0.580023	0.262585	0.361518	0.634509	0.237327	0.469223

5. Set the number of clusters equal to the number of classes for each dataset and cluster the

data using GMM. List the resulting NMI for each dataset in a table.

	Dermatology	Vowels	Glass	Ecoli	Yeast	Soybean
K	6.000000	11.000000	6.000000	5.000000	9.000000	15.000000
SSE	75910.291542	4985.590587	128.061345	41.850974	80.671901	3994.775399
NMI	0.141384	0.042337	0.303468	0.341405	0.155339	0.236625

5 Comparing k-Means and GMM

1. (25 points)For each dataset which algorithm would you use to cluster? why?

   Notice: Overall I choose based on NMI since GMM needs NMI instead of SSE to determine if a k value is good or not. To show consistencies, we compare K-means and GMM based on NMI values.
   And for the same performance, the k lower, the better (less complex).

   Dermatology: GMM is better. Because K-means gives best k 20 and best NMI 0.417946, while GMM gives best k 5 and best NMI 0.580023. So for GMM, it has lower k and higher NMI, definitely it's better.

   Vowels: K-means is better. Because K-means gives best k 18 and best NMI 0.494878, while GMM gives best k 9 and best NMI 0.262585. Even though GMM gives lower k but the NMI is too low compared to K-means, so overall I choose K-means.

   Glass: GMM is better. Because K-means gives best k 18 and best NMI 0.337523, while GMM gives best k 2 and best NMI 0.361518. So for GMM, it has lower k and similar NMI, it's better.

   Ecoli: GMM is better. Because K-means gives best k 19 and best NMI 0.532981, while GMM gives best k 9 and best NMI 0.634509. So for GMM, it has lower k and higher NMI, it's better.

   Yeast: GMM is better. Because K-means gives best k 15 and best NMI 0.248423, while GMM gives best k 13 and best NMI 0.237327. So for GMM, it has lower k and similar NMI, it's better.

   Soybean: K-means is better. Because K-means gives best k 16 and best NMI 0.663247, while GMM gives best k 5 and best NMI 0.469223. Even though GMM gives lower k but the NMI is too low compared to K-means, so overall I choose K-means.

2. (25 points)Does the clustering for each dataset gives you any insight about the separability of the classes?

   Dermatology: Best k for GMM is 5. The actual number of classes is 6. So overall, the classes for this dataset are not separable.

Vowels: Best k for K-means is 18. The actual number of classes is 11. That means classes of this dataset can be further separated into 18 clusters.

Glass: Best k for GMM is 2. The actual number of classes is 6. Not only that classes cannot be further separated but also different classes need to join together to create only 2 clusters eventually.

Ecoli: Best k for GMM is 9. The actual number of classes is 5. So classes can be further separated into 9 clusters.

Yeast: Best k for GMM is 13. The actual number of classes is 9. So classes can be further separated into 13 clusters.

Soybean: Best k for K-means is 16. The actual number of classes is 15. So overall classes are not separable.

To sum up, different datasets have different separability of classes. It solely depends on the data itself.