Project 4 - Clustering

Brendon Faleiro - 704759004
Anurag Pande - 604749647
Sachin Bhat - 304759727
March 9, 2017

The main aim of this project is to understand clustering algorithms, which are unsupervised methods for grouping data points with similiar representations in a proper space with no a priori labeling available.

1 Transform Documents to TF-IDF Vectors

Since there are lot of common words in each document we need to preprocess the data so that we can find significant terms in the dataset. For this, we first remove punctuations, common stop words and finding which words share the same stem so that they can be counted together while finding their TFXIDF. In order to do the latter we used a SnowBall stemmer (**Python NLTK**) to achieve this. In addition to removing the above mentioned terms, we also removed non ASCII characters. This helped in increasing the over-all accuracy when we tired training the classifier. Once the data has been pre-processed, the next step is to find the TFXIDF of each term. For this we convert the document into a set of numerical features. This is done using CountVectorizer. Next we get the TFxIDF Representations using a Transformer. A matrix is obtained with the number of documents (records) as the row and the number of terms obtained as the number of columns.

2 Basic K-means Clustering

In this part, we applied k-means with k=2 to the data. The resulting confusion matrix is as follows:

Table 2.1: Standard Confusion Matrix for K-means

Actual \ Predicted	Computer Tech	Recreation
Computer Tech	1	3902
Recreation	1751	2228

The measures we examine in this project are homogeneity score, completeness score, adjusted rand score and the adjusted mutual info score. Homogeneity is a measure of how purely clusters contain only data points that belong to a single class. On the other hand, a clustering result satisfies completeness if all of its clusters contain only data points that belong to a single class. Both of these scores span between 0 and 1; where 1 stands for perfect clustering. The Rand Index is similar to accuracy measure, which computes similarity between the clustering labels and ground truth labels. This method counts all pairs of points that both fall either in the same cluster and the same class or in different clusters and different classes. Finally, adjusted mutual information score measures mutual information between the cluster label distribution and the ground truth label distributions.

Table 2.2: Measures of Purity

Metric	**Value**
Accuracy	0.280
Precision	0.361
Recall	0.555
Homogeneity	0.267
Completeness	0.347
Adjusted Random Index	0.193
Adjusted Mutual Info Index	0.267

3 K-means Clustering with Dimensionality Reduction

We have to find a better way of representing input vectors to clustering algorithm based on its working in order to boost its performance. This is inspired by our knowledge that, performance is hindered when high-dimensional sparse TF-IDF vectors are used in clustering algorithm. In this problem, we use Latent Semantic Indexing(LSI) and Non-negative Matrix Factorization(NMF) for dimensionality reduction. However, we need to initially guess an appropriate dimension for the vectors to be used as input to the clustering algorithm. To accomplish task, we have calculated the singular values of the TF-IDF matrix. Then we have tried to find out the number of these singular values are actually significant in reconstructing the original high-dimensional matrix by using truncated SVD representation. Also, we have normalized the vectors prior to feeding them to clustering algorithm.

As mentioned in the problem, first we explore LSI method beginning with lower values as 2 to 3 and explore till an effective dimension.In order to find the range for the reduced dimension we perform SVD and find the number of significant singularities. The below plot shows the range of singularities for the top 50 dimensions.

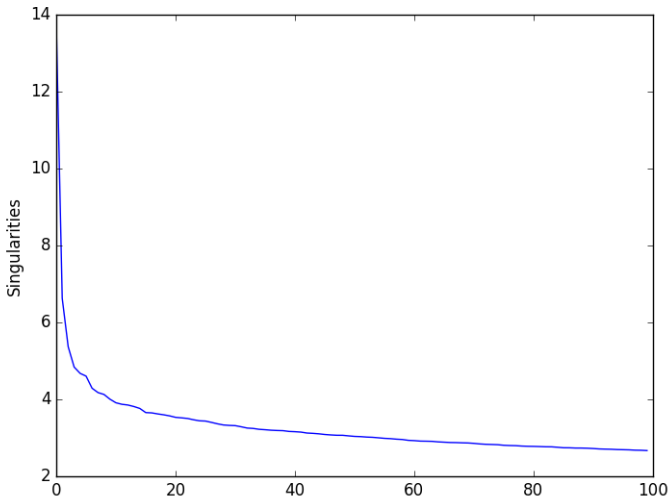

Figure 3.1: Top 50 singularities

Since the singularities flatten before 40 dimensions, we can look for a possible optimum dimension below 40. Table 3.1 indicates the Purity Measures for some of the dimensions. Similar Purity Measures were also calculated for NMF as shown in Table 3.2.

Table 3.1: **Metrics Using SVD with dimensions reduced**

Dimensions	Homogeneity	Completeness	Adjusted Rand	Adjusted Mutual Info
3	0.673	0.673	0.775	0.673
5	0.732	0.732	0.824	0.732
10	0.719	0.719	0.814	0.719
20	0.738	0.738	0.832	0.738
30	0.746	0.746	0.827	0.746

Table 3.2: **Metrics Using NMF with dimensions reduced**

Dimensions	Homogeneity	Completeness	Adjusted Rand	Adjusted Mutual Info
3	0.245	0.330	0.168	0.245
5	0.211	0.261	0.184	0.211
10	0.233	0.321	0.153	0.233
20	0.117	0.235	0.043	0.117
30	0.083	0.207	0.022	0.083

The optimal value of the number of dimensions was found to be 31, since it showed the best purity metrics.

Table 3.3: Measures of Purity for 31 dimensions using SVD

Metric	**Value**
Accuracy	0.957
Precision	0.959
Recall	0.956
Homogeneity	0.747
Completeness	0.747
Adjusted Random Index	0.836
Adjusted Mutual Info Index	0.747

Question → Can you justify why logarithm is a good candidate for your TFxIDF data?
Answer → The aspect that the frequency of the term does not increase proportionally with the significance or relevance of the term is emphasized here. We can reduce this effect by using a sub-linear function. Also, the influence of rare or large or small words is also amortized. Thus, the scoring function which is mostly perceived as additive using the log by many, will make the probability of various independent terms from $P(A,B) = P(A)P(B)$ to appear more like $log(P(A,B)) = log(P(A)) + log(P(B))$.

Table 3.4: Purity Metrics for Polynomial in degree 2 for d=31

Purity Measure	Value
Homogeneity Score	0.256
Completeness Score	0.338
Adjusted Rand Score	0.180
Adjusted Mutual Info Score	0.255
Accuracy	0.711
Precision	0.999
Recall	0.429

Table 3.5: Confusion Matrix for Polynomial in degree 2

Actual \ Predicted	Computer Tech	Recreation
Computer Tech	3902	1
Recreation	2273	1706

Table 3.6: Purity Metrics for Natural Log for n=37

Purity Measure	Value
Homogeneity Score	0.735
Completeness Score	0.735
Adjusted Rand Score	0.828
Adjusted Mutual Info Score	0.735
Accuracy	0.955
Precision	0.952
Recall	0.959

Table 3.7: Confusion Matrix for Natural Log

Actual\Predicted	Computer Tech	Recreation
Computer Tech	3712	191
Recreation	164	3815

4 Visualization of Clusters Formed

In this problem, to help understand the data more thoroughly, we visualize the performance of the clustering by projecting final data vectors onto 2 dimensions.

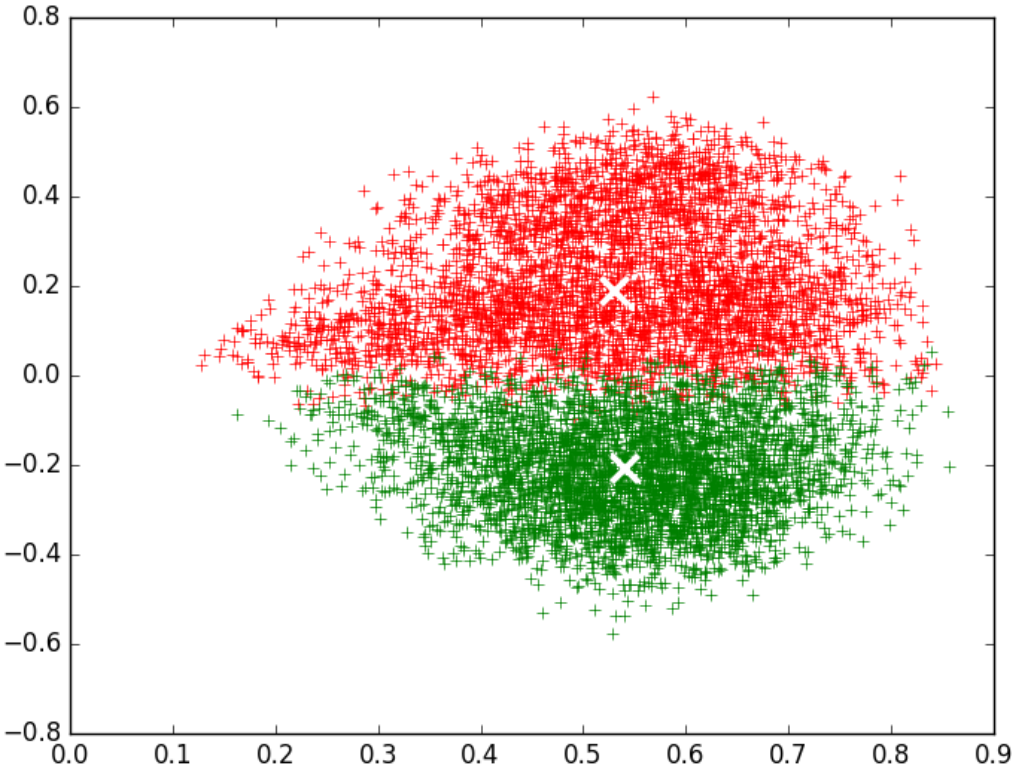

Figure 4.1: TFxIDF in 2 Dimensions

The clusters we obtained with KMeans, plotted in 2d space are as shown in Fig 4.1. When the logarithmic normalization from part 3 was applied, we got a clustering as shown in Fig 4.2.

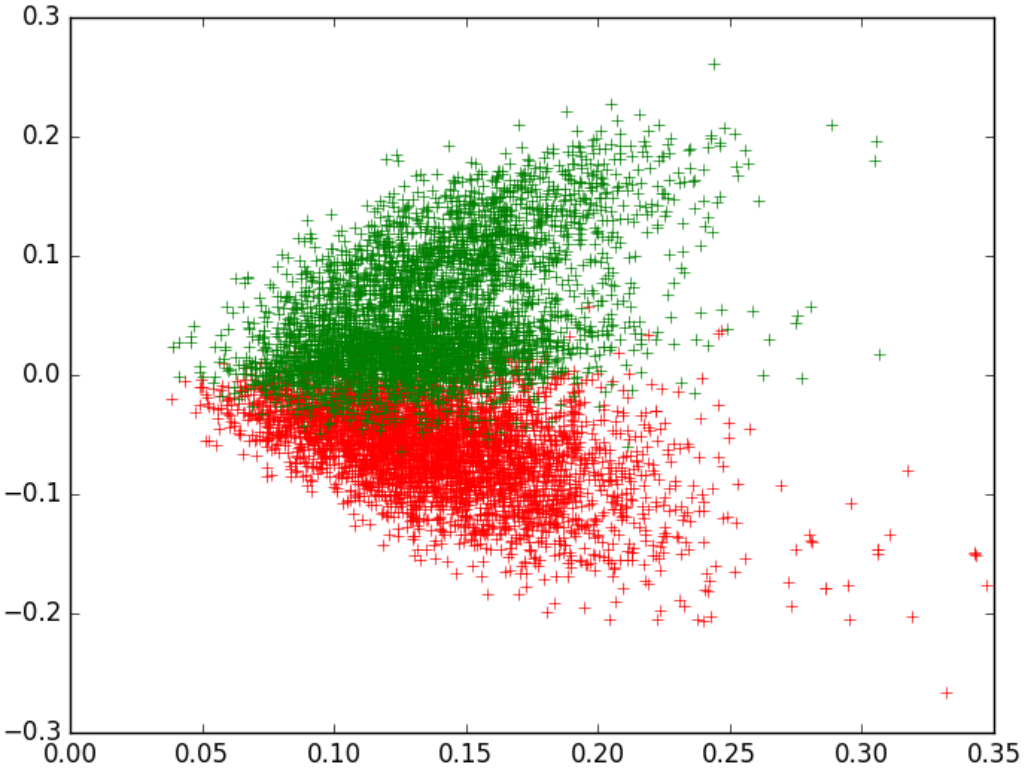

Figure 4.2: TFxIDF clustering 2 Dimensions with logarithmic normalization

Question → Can you justify why non-linear transform is a good candidate for your TFxIDF data?
Answer → For the 20-newsgroups dataset, the features are independent, implying that we are actually dealing with non-linear transformation. Thus, using non-linear transformation, we can solve a non-linear problem as a linear problem.

5 K-means Clustering on Original Subclasses

We examine how purely we can retrieve all the 20 original sub-class labels with clustering. Therefore, we include all the documents and the corresponding terms in the data matrix and find proper representation through reducing the dimension of TF-IDF matrix representation. We first retrieve all the 20 original sub-class documents and generate TF-IDF as before. And then we cluster them without any reduction in dimensionality. Below table shows corresponding purity measures.

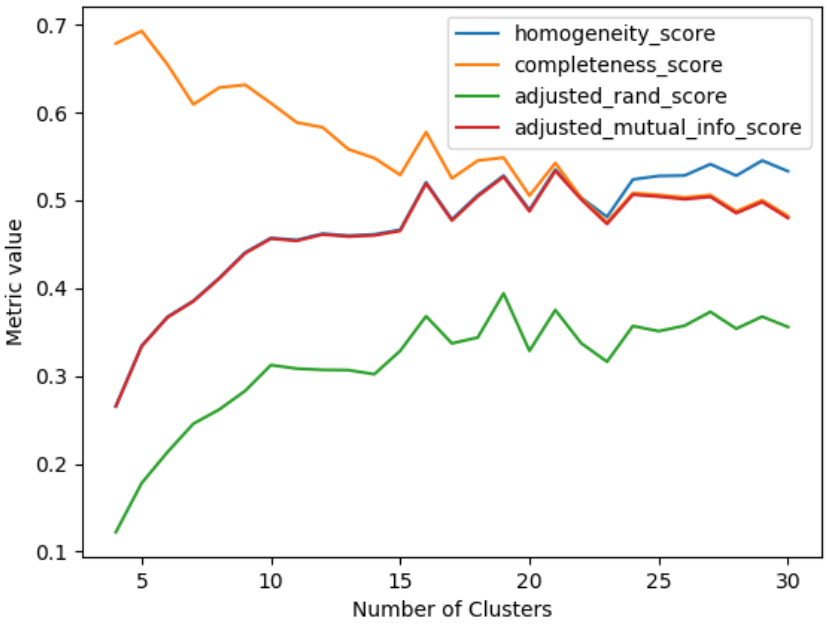

Figure 5.1: Performance Metrics of SVD with dimensions fixed at 50

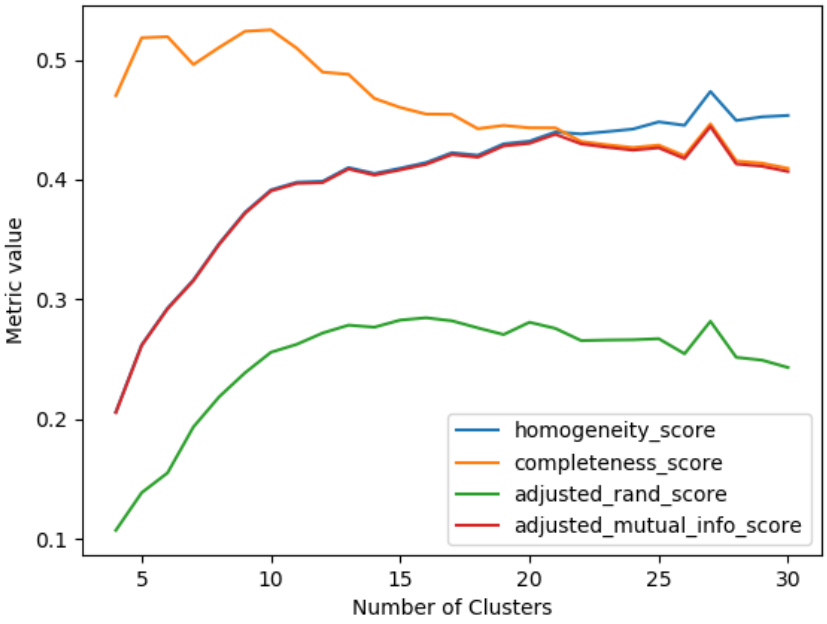

Figure 5.2: Performance Metrics of NMF with dimensions fixed at 50

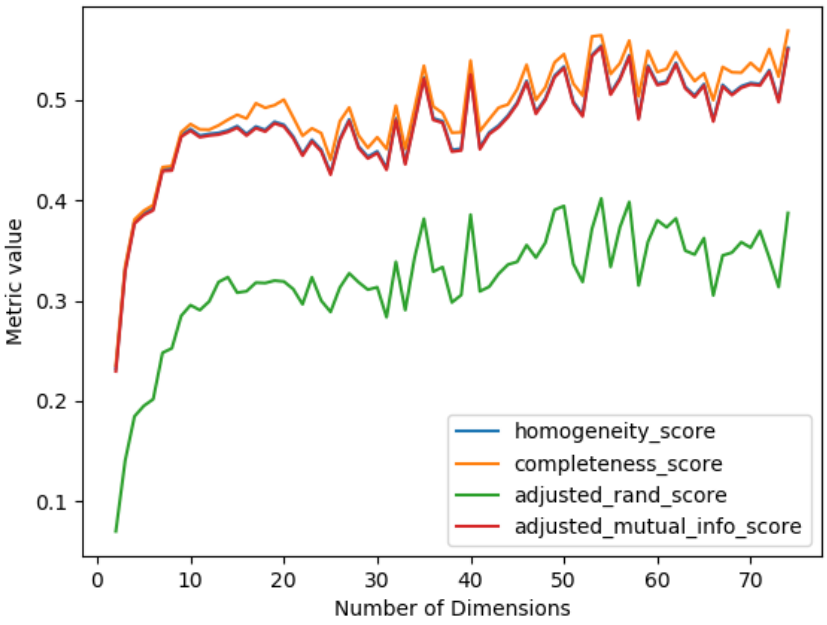

Figure 5.3: Performance Metrics of SVD with fixed k=20

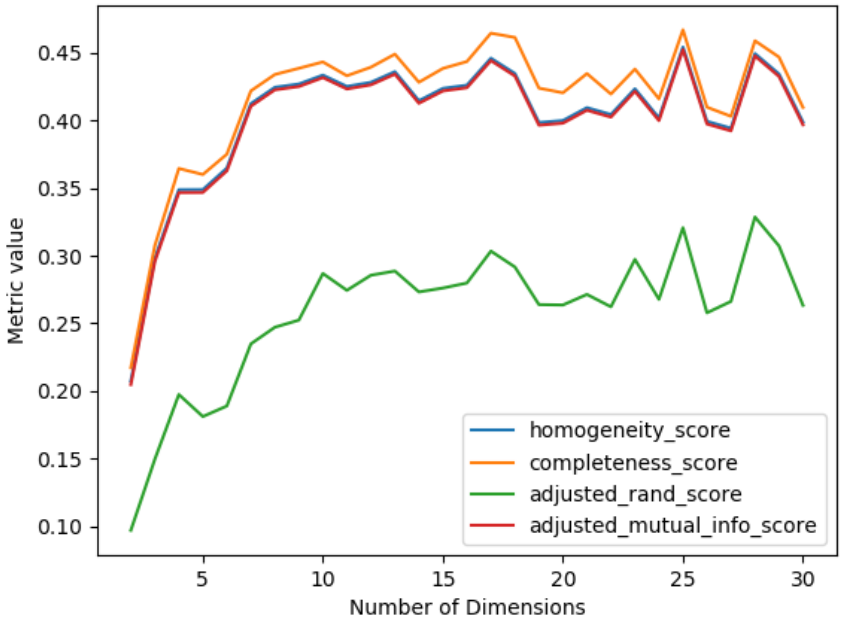

Figure 5.4: Performance Metrics of NMF with fixed k=20

Then we apply truncated SVD (LSI) in order to achieve dimensionality reduction for TF-IDF matrix. We plugged in different values of dimensions for a fixed k = 20. We vary the value of dimensions in a range between 2 and 75. Below table shows corresponding purity measures for the best value of dimension:

Table 5.1: Measures of Purity after LSI Dimensionality Reduction

Components	Homogeneity Score	Completeness Score	Adjusted Rand Score
37	0.217	0.092	0.216

6 K-means Clustering on Topic-wise Classes

Here, we examine how effectively we can get the topic-wise classes with the original classes grouped in the following way:

1. Computer Technology
    a) comp.graphics
    b) comp.os.ms-windows.misc
    c) comp.sys.ibm.pc.hardware
    d) comp.sys.mac.hardware
    e) comp.windows.x
2. Recreational Activity
    a) rec.autos
    b) rec.motorcycles
    c) rec.sport.baseball
    d) rec.sport.hockey
3. Science
    a) sci.crypt
    b) sci.electronics
    c) sci.med
    d) sci.space
4. Miscellaneous
    a) misc.forsale
5. Politics
    a) talk.politics.misc
    b) talk.politics.guns
    c) talk.politics.mideast
6. Religion
    a) talk.religion.misc
    b) alt.atheism
    c) soc.religion.christian

We used these super classes to relabel the dataset, and followed the following procedure:

- We first computed an SVD of the TF-IDF obtained, and plotted the singular values. We observed that the knee in the graph occurred at around d = 50.
- Next keeping k fixed at 6, we ran SVD for varying number of components(2 to 75), and plotted the resulting metrics.

At 2 dimensions, the performance metrics were as follows:

- Homogeneity: 0.184
- Completeness: 0.175
- Normalized Mutual Info: 0.111
- Adjusted Rand-Index: 0.175
- Confusion Matrix: Refer to Table 6.1

Class Labels	0	1	2	3	4	5
0	257	662	1569	210	1543	461
1	226	988	991	831	905	8
0	731	1075	564	1337	206	55
0	363	173	75	336	12	31
0	802	334	134	597	25	1002
0	782	92	40	296	3	1130

Table 6.1: Confusion Matrix

We observed that the best performance metrics were obtained at k = 6 and number of dimensions = 58. The accuracy at such a dimension was 0.320.

- Homogeneity: 0.331
- Completeness: 0.339
- Normalized Mutual Info: 0.198
- Adjusted Rand-Index: 0.331
- Confusion Matrix: Refer to Table 6.2

The plot for the variation of Purity measures with dimensions, is shown below.

Class Labels	0	1	2	3	4	5
0	2176	1558	59	2	334	573
1	1160	875	1802	7	0	105
2	774	154	253	1743	1	1043
3	186	18	54	0	12	720
4	196	33	28	5	865	1767
5	100	8	8	3	359	1865

Table 6.2: Confusion Matrix

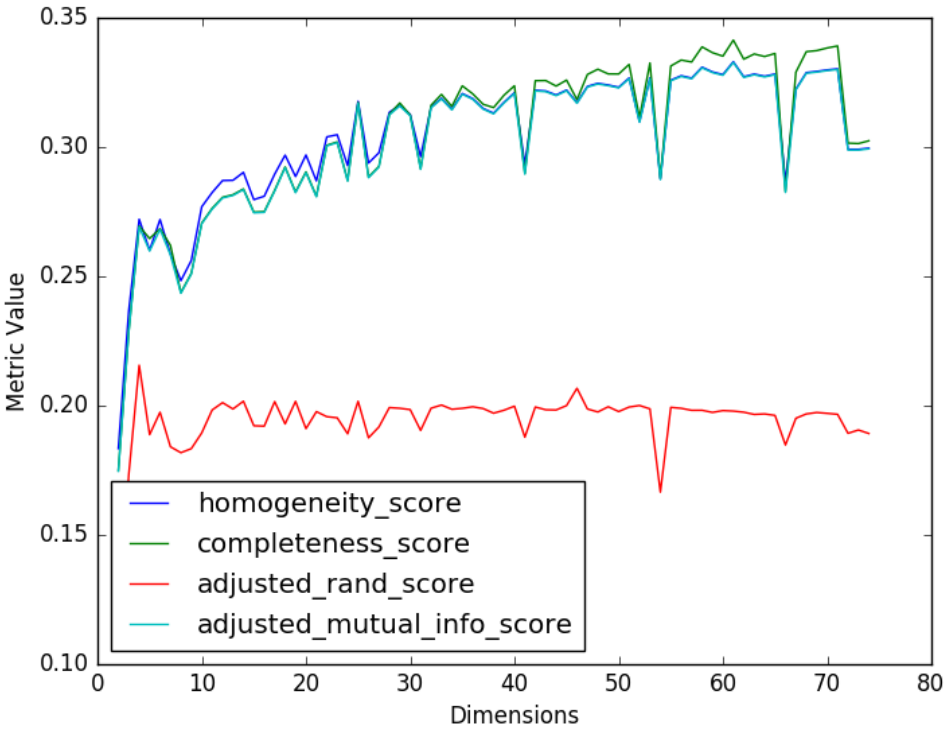

Figure 6.1: Variation in Purity Measures with Dimension