**EE 219 Project 2 - Report**

**Clustering**

Fangyao Liu 204945018

Xuan Hu 505031796

Zhechen Xu 805030074

**Content**

**Introduction …………………………………………………………………..... 1**

**Problem 1: Transform the Documents Into TF-IDF Vectors ……………….. 1**

**Problem 2: K-means Clustering (k = 2) ………………………………………. 1**

**Problem 3: K-means Clustering (Dimension Reduction) ……………………. 3**

**Problem 4: Performance of Clustering Visualization ………………………... 9**

**Problem 5: K-means Clustering (20 Original sub-classes) ………………….. 20**

**Introduction**

In this project, we work with “20 Newsgroups” dataset. It is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups, each corresponding to a different topic. We try to find proper representation of the data and evaluate the performance of the clustering. clustering algorithms are unsupervised methods of learning aimed at finding groups of data points that are similar to each other. Clustering differs from classification in the sense that in case of clustering no a priori labelling or grouping of data points is available.

We assume that the class labels are not available and aim is to find the grouping of the documents where documents in same group are more similar to each other than to the documents in other groups. The clustering algorithms we are using is K means clustering such that each data point belongs to one and only one cluster, and the sum of the squares of the distances between each data point and the center of the cluster it belongs to is minimized.

**Problem 1: Transform the Documents Into TF-IDF Vectors**

In this section, we have done similar work in Project 1. We create a TFxIDF vector representations, tokenize the documents and exclude the stop words, punctuations, and different stems of a word.

The dimension of TF-IDF matrix is (4732, 20269). Thus, the number of terms in the matrix is 20269.

**Problem 2: K-means Clustering (k = 2)**

In this question, we applied K-means clustering to the entire data set with k = 2 by using the k-means module of Sklearn package.

1. The confusion matrix obtained from running the K-means clustering is as follows:

|  |  |  |
| --- | --- | --- |
| Class Label | 0 | 1 |
| 0 | 2341 | 2 |
| 1 | 1355 | 1034 |

*Table 1 Confusion matrix (k = 2)*

Clearly, the confusion matrix here is not an ideal clustering confusion matrix due to non-diagonal property.

1. To make a concrete comparison of different clustering results, there are various measures of purity for a given partition of the data points with respect to the ground truth. The measures we examine in this project are the homogeneity score, the completeness score, the V-measure, the adjusted Rand score and the adjusted mutual info score.

|  |  |
| --- | --- |
| Measures of Purity | Value |
| Homogeneity Score | 0.255051569662 |
| Completeness Score | 0.336360010418 |
| V measure score | 0.290116566933 |
| Adjusted Rand Score | 0.181723830086 |
| Adjusted Mutual Info Score | 0.254937926183 |

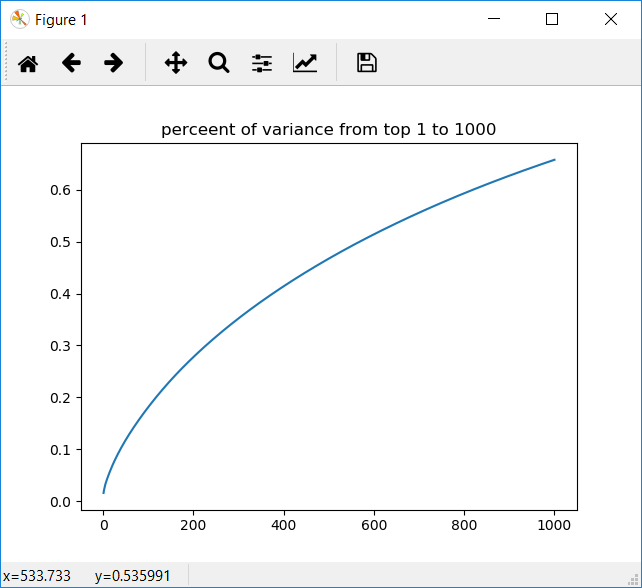
*Table 2 Measures of Purity*

The score span between 0 and 1. 1 stands for perfect clustering. Thus, from the measures of purity, we can find that the clustering result we obtained did not perform well.

**Problem 3: K-means Clustering (Dimension Reduction)**

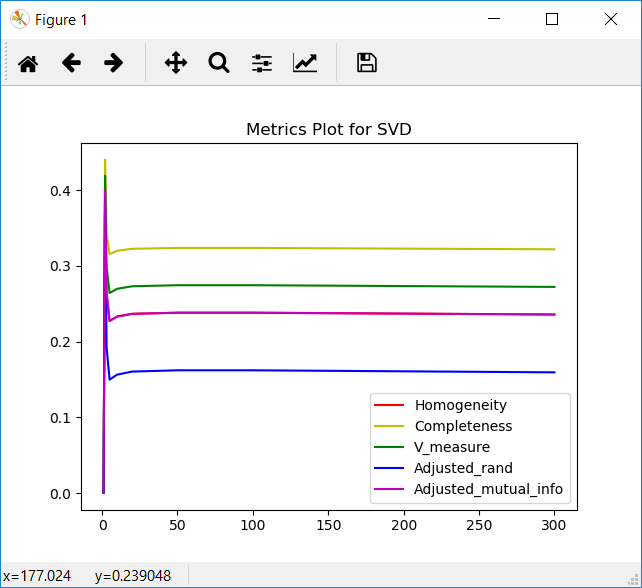
We have found that high dimensional sparse TF-IDF vectors do not yield a good clustering result. Thus, in this part we want to find a “better” representation tailored to the way that K-means clustering algorithm works by using Latent Semantic Indexing (LSI) and Non-Negative Matrix Factorization (NMF) to reduce TF-IDF matrix dimensionality.

1. Plot of the percent of variance the top r principle components can retain vs. r, for r = 1 to 1000.



*Figure 1 percent of variance from top 1 to 1000*

**LSI:**



*Figure 2 Metrics Plot for SVD*

In Figure 2, homogeneity overlaps with adjusted\_mutual\_info.

Contingency matrices for r = 1,2,3,5,10,20,50,100,300:

|  |  |
| --- | --- |
| 1110 | 1233 |
| 1085 | 1304 |

*Table 3 Contingency Matrix (r = 1)*

|  |  |
| --- | --- |
| 2327 | 17 |
| 870 | 1519 |

*Table 4 Contingency Matrix (r = 2)*

|  |  |
| --- | --- |
| 2341 | 2 |
| 1320 | 1069 |

*Table 5 Contingency Matrix (r = 3)*

|  |  |
| --- | --- |
| 2341 | 2 |
| 1462 | 927 |

*Table 6 Contingency Matrix (r = 5)*

|  |  |
| --- | --- |
| 2341 | 2 |
| 1415 | 974 |

*Table 7 Contingency Matrix (r = 10)*

|  |  |
| --- | --- |
| 2341 | 2 |
| 1412 | 977 |

*Table 8 Contingency Matrix (r = 20)*

|  |  |
| --- | --- |
| 2341 | 2 |
| 1425 | 964 |

*Table 9 Contingency Matrix (r = 50)*

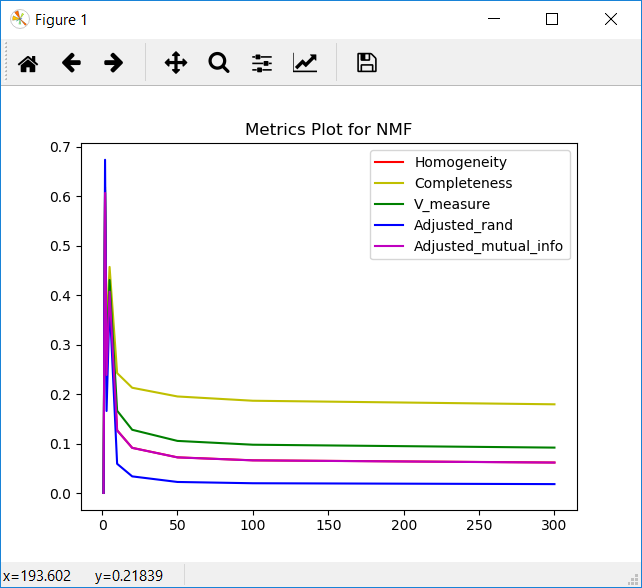
|  |  |
| --- | --- |
| 2341 | 2 |
| 1409 | 980 |

*Table 10 Contingency Matrix (r = 100)*

|  |  |
| --- | --- |
| 2 | 2341 |
| 965 | 1424 |

*Table 11 Contingency Matrix (r = 300)*

**NMF:**



*Figure 3 Metrics using NMF with reduced dimensions*

In Figure 3, homogeneity overlaps with adjusted\_mutual\_info.

Contingency matrices for r = 1,2,3,5,10,20,50,100,300:

|  |  |
| --- | --- |
| 1110 | 1233 |
| 1085 | 1304 |

*Table 12 Contingency Matrix (r = 1)*

|  |  |
| --- | --- |
| 1951 | 392 |
| 32 | 2357 |

*Table 13 Contingency Matrix (r = 2)*

|  |  |
| --- | --- |
| 2339 | 4 |
| 1398 | 991 |

*Table 14 Contingency Matrix (r = 3)*

|  |  |
| --- | --- |
| 1452 | 891 |
| 4 | 2385 |

*Table 15 Contingency Matrix (r = 5)*

|  |  |
| --- | --- |
| 1788 | 555 |
| 2387 | 2 |

*Table 16 Contingency Matrix (r = 10)*

|  |  |
| --- | --- |
| 1929 | 414 |
| 2387 | 2 |

*Table 17 Contingency Matrix (r = 20)*

|  |  |
| --- | --- |
| 2009 | 334 |
| 2387 | 2 |

*Table 18 Contingency Matrix (r = 50)*

|  |  |
| --- | --- |
| 2029 | 314 |
| 2386 | 3 |

*Table 19 Contingency Matrix (r = 100)*

|  |  |
| --- | --- |
| 300 | 2043 |
| 4 | 2385 |

*Table 20 Contingency Matrix (r = 300)*

**Report best r choice for SVD and NMF respectively.**

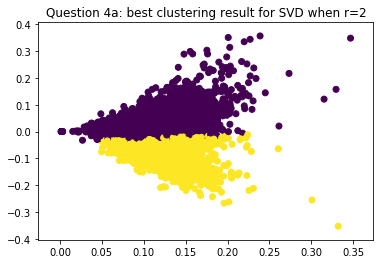
From the plots, it can be found out easily that all the metrics have same trend when r increases. We can reach the conclusion from the results that r=2 works best for SVD and NMF at the same time.

**Q: How do you explain the non-monotonic behavior of the measures as r increases?**

**A:** The result of k-means depends on the dimensions. Low dimensions are good for clustering but contains less information about the original data. High dimension contains enough data(information), but it presents terrible clustering result. Both factors will influence the result, which cause the non-monotonic behavior of the measures.

**Problem 4: Performance of Clustering Visualization**

1. In this problem, we visualize the performance of the case with best clustering results in the previous part your clustering by projecting final data vectors onto 2-dimensional plane and color-coding the classes. The clustering figures are shown below:

**

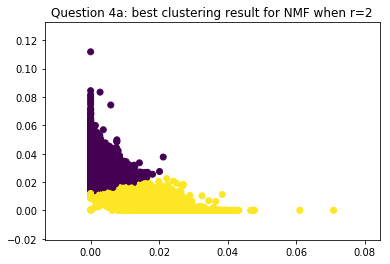
*Figure 4 clustering result for SVD (r = 2)*

|  |  |  |
| --- | --- | --- |
| Class Label | 0 | 1 |
| 0 | 2326 | 17 |
| 1 | 873 | 1516 |

*Table 21 Confusion matrix for SVD (r = 2)*

|  |  |
| --- | --- |
| Measures of Purity | Value |
| Homogeneity Score | 0.399822144328 |
| Completeness Score | 0.439992270447 |
| V measure score | 0.389053201367 |
| Adjusted Rand Score | 0.389053201367 |
| Adjusted Mutual Info Score | 0.399730598334 |

*Table 22 Measures of Purity for SVD (r = 2)*



*Figure 5 clustering result for NMF (r = 2)*

|  |  |  |
| --- | --- | --- |
| Class Label | 0 | 1 |
| 0 | 1951 | 392 |
| 1 | 32 | 2357 |

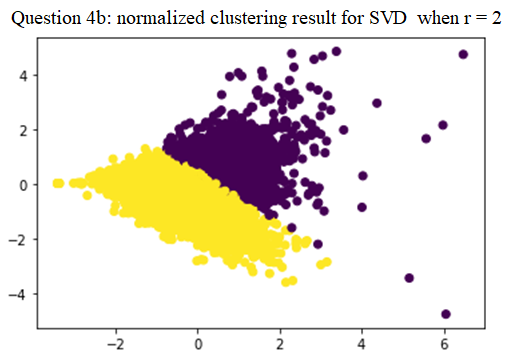
*Table 23 Confusion matrix for NMF (r = 2)*

|  |  |
| --- | --- |
| Measures of Purity | Value |
| Homogeneity Score | 0.606701062862 |
| Completeness Score | 0.618400434277 |
| V measure score | 0.612494885732 |
| Adjusted Rand Score | 0.673635780674 |
| Adjusted Mutual Info Score | 0.606641075318 |

*Table 24 Measures of Purity for NMF (r = 2)*

1. In part b, we applied three methods to see if they increase the clustering performance.

* Normalizing features such that each feature has unit variance
* Applying a non-linear transformation to the data vectors only after NMF.
* combining both transformations (in different orders) on NMF- reduced data.



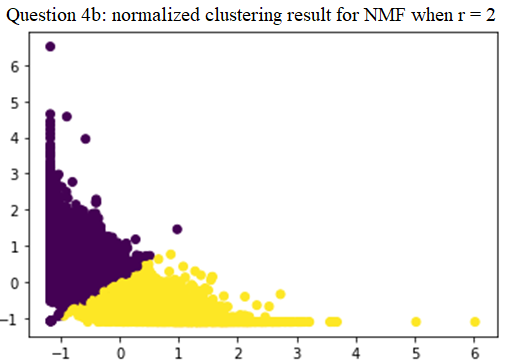
*Figure 6 normalized clustering result for SVD (r = 2)*

|  |  |  |
| --- | --- | --- |
| Class Label | 0 | 1 |
| 0 | 1289 | 1054 |
| 1 | 191 | 2198 |

*Table 25 Confusion matrix for SVD, normalized (r = 2)*

|  |  |
| --- | --- |
| Measures of Purity | Value |
| Homogeneity Score | 0.201856833288 |
| Completeness Score | 0.225186218949 |
| V measure score | 0.212884283301 |
| Adjusted Rand Score | 0.22433088572 |
| Adjusted Mutual Info Score | 0.201735090247 |

*Table 26 Measures of Purity for SVD, normalized (r = 2)*

**

*Figure 7 normalized clustering result for NMF (r= 2)*

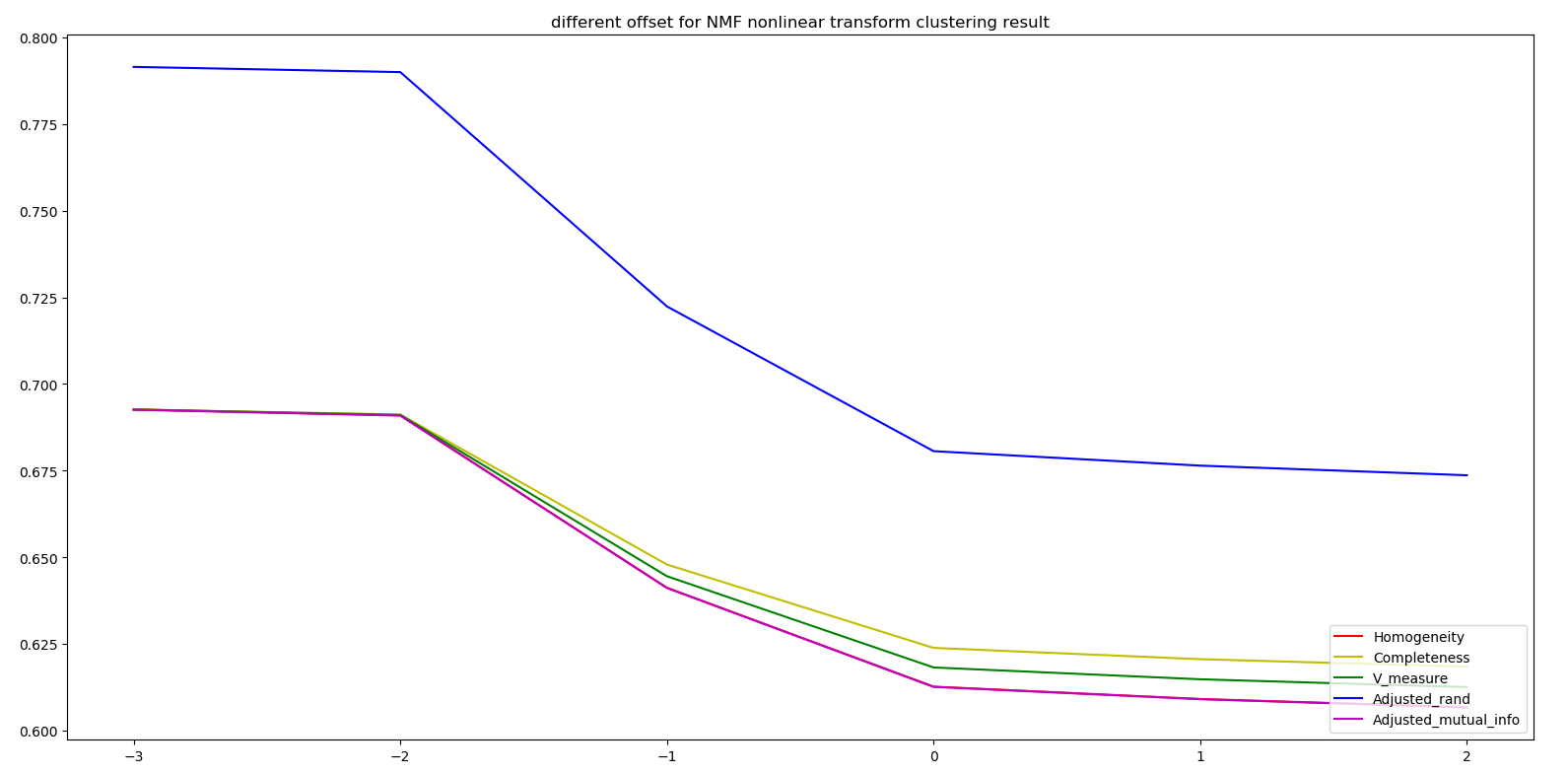
|  |  |  |
| --- | --- | --- |
| Class Label | 0 | 1 |
| 0 | 2188 | 155 |
| 1 | 100 | 2289 |

*Table 27 Confusion matrix for NMF, normalized (r = 2)*

|  |  |
| --- | --- |
| Measures of Purity | Value |
| Homogeneity Score | 0.69867486191 |
| Completeness Score | 0.699175472984 |
| V measure score | 0.698925077805 |
| Adjusted Rand Score | 0.796019084001 |
| Adjusted Mutual Info Score | 0.698628903084 |

*Table 28 Measures of Purity for NMF, normalized (r = 2)*

When we tried to perform nonlinear transformation, an unexpected situation coming to us was that log function couldn’t be used on zeros or very small number because of our limited memory. However, NMF matrix contains lots of zero elements. In order to avoid this limitation over here, our group decided to add an extra constant offset into NMF matrix. As for what constant to be added, we conducted some experiments by choosing constant from [0.001, 0.01, 0.1, 1, 10, 100] and compared the performances. Results are presented below:



*Figure 8 Different offset for NMF nonlinear transform clustering result*

As we could see here, when offset constant was smaller, clustering result became better. When we printed out the NMF matrix, the reason of this trend above could be explained well.

NMF matrix = [[ 0.02047681 0. ]

[ 0.02620815 0. ]

[ 0.01769416 0.01108869]

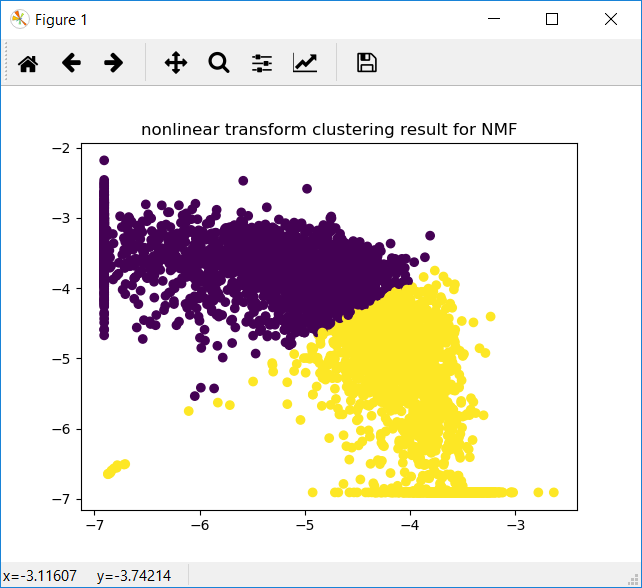
...,

[ 0.02856409 0.00130083]

[ 0.03385415 0. ]

[ 0.00556458 0.01682442]]

The order of NMF matrix is about 0.01. When we add a large offset into the NMF matrix, like 10, the relative distance among each sample is becoming smaller, which decreases the differences of each sample. Therefore, in log transformation, we adopt constant offset to be 0.001.



*Figure 9 nonlinear transform clustering result for NMF (r = 2)*

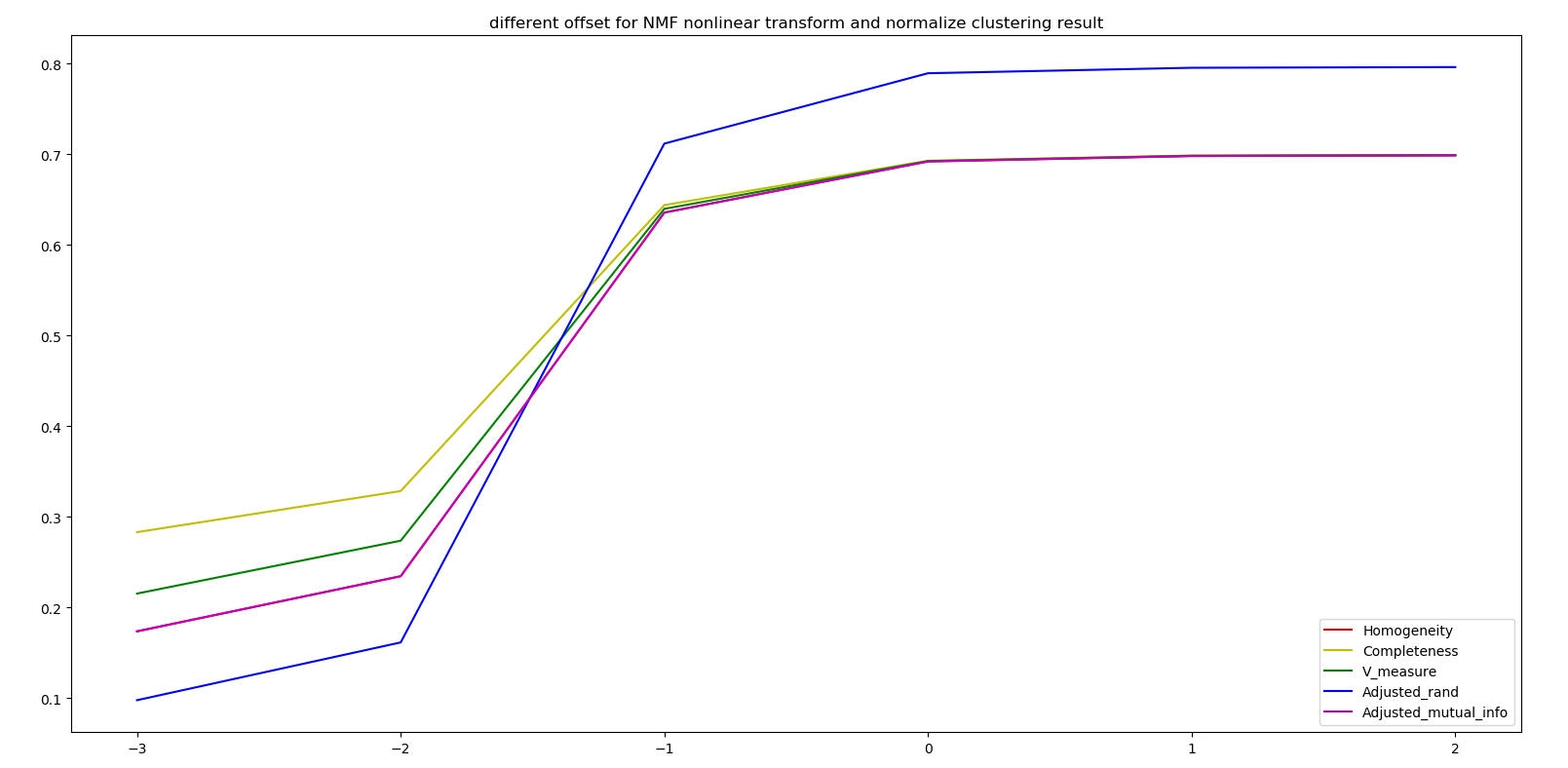
|  |  |  |
| --- | --- | --- |
| Class Label | 0 | 1 |
| 0 | 2228 | 115 |
| 1 | 146 | 2243 |

*Table 29 Confusion matrix for NMF, nonlinear transform (r = 2)*

|  |  |
| --- | --- |
| Measures of Purity | Value |
| Homogeneity Score | 0.692631456308 |
| Completeness Score | 0.692589953118 |
| V measure score | 0.692610704091 |
| Adjusted Rand Score | 0.791499310895 |
| Adjusted Mutual Info Score | 0.692543069039 |

*Table 30 Measures of Purity for NMF, nonlinear transform (r = 2)*

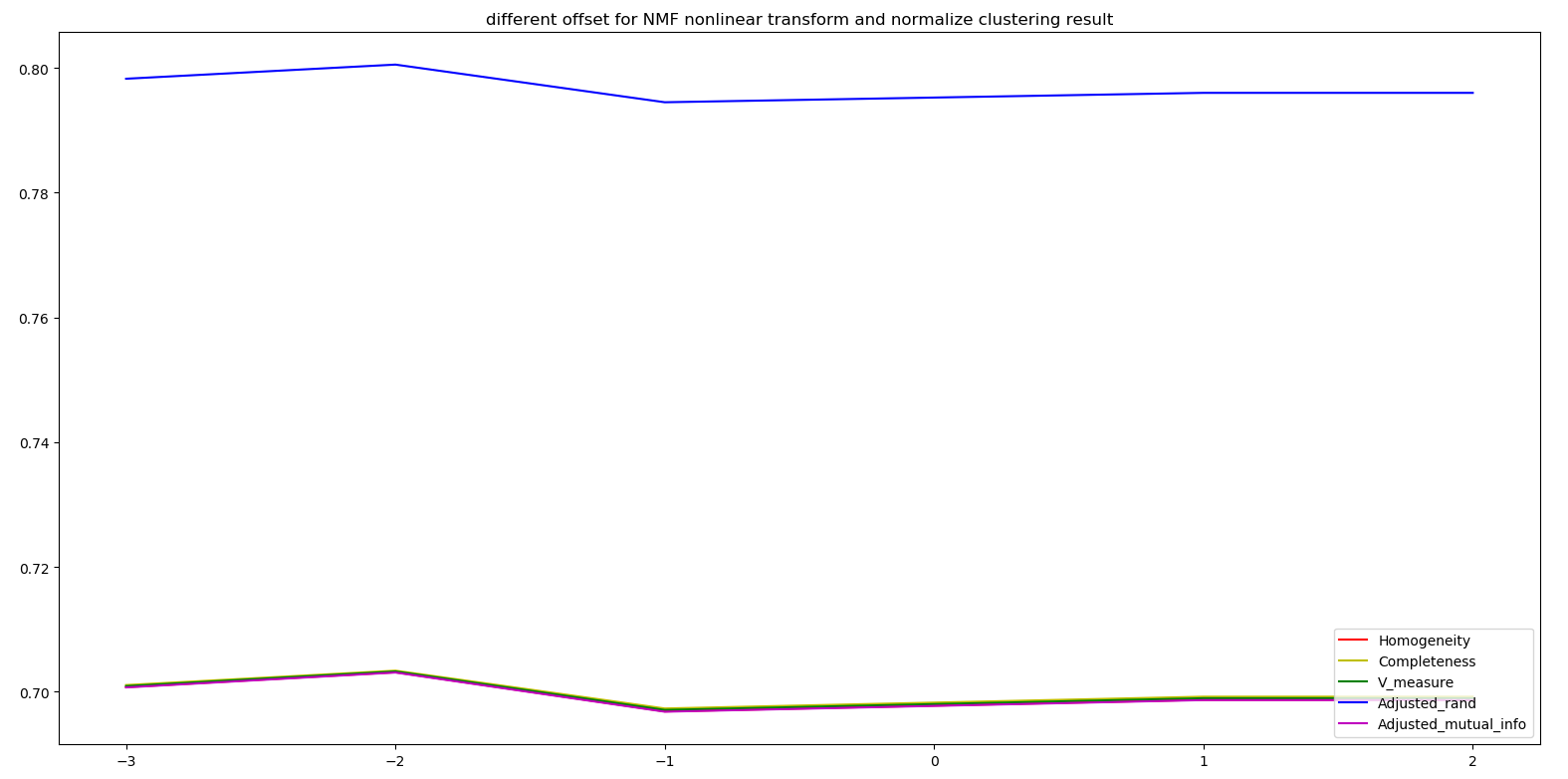
Because the following two questions also are related to nonlinear transformation, so we take the same steps above. For “first scale, then log transformation”, we minus matrix by the minimum value in the matrix after scaling, to assure that each element in scaled matrix is non-negative. Then we add the offset constant to the matrix and get the plot below:



*Figure 10 Different offset for NMF nonlinear transform and normalize clustering result*

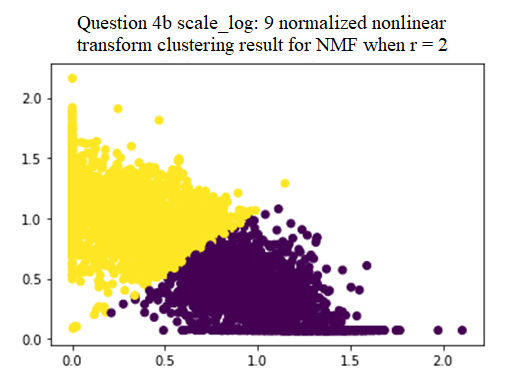
This plot presents a completely contrary trend to figure 8. After our group’s discussion, we think the result can be explained as this:

For “first log transformation, then scale”, we get the plot below:



*Figure 11 Different offset for NMF nonlinear transform and normalize clustering result*

It can be observed that there are not large variations to the results when we change the offset constant. We believe that the scale process has remove the offset effect in samples. Therefore, we adopt 1 as our offset constant for “first scale, then log transformation”.



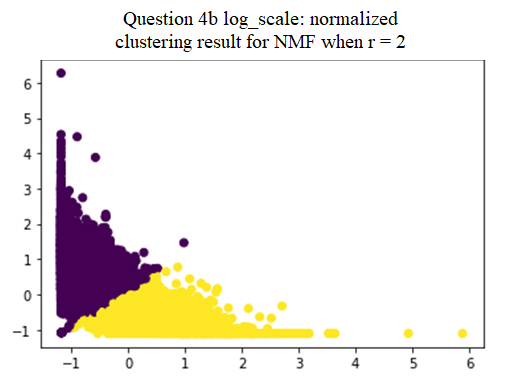
*Figure 12 normalized nonlinear transform clustering result for NMF (r = 2)*

|  |  |  |
| --- | --- | --- |
| Class Label | 0 | 1 |
| 0 | 169 | 2174 |
| 1 | 2294 | 95 |

*Table 31 Confusion matrix for NMF, normalized nonlinear transform (r = 2)*

|  |  |
| --- | --- |
| Measures of Purity | Value |
| Homogeneity Score | 0.691959893949 |
| Completeness Score | 0.692752878161 |
| V measure score | 0.692356158996 |
| Adjusted Rand Score | 0.789244302491 |
| Adjusted Mutual Info Score | 0.691912910929 |

*Table 32 Measures of Purity for NMF, normalized nonlinear transform (r = 2)*



*Figure 13 normalized clustering result for NMF*

|  |  |  |
| --- | --- | --- |
| Class Label | 0 | 1 |
| 0 | 2188 | 155 |
| 1 | 101 | 2288 |

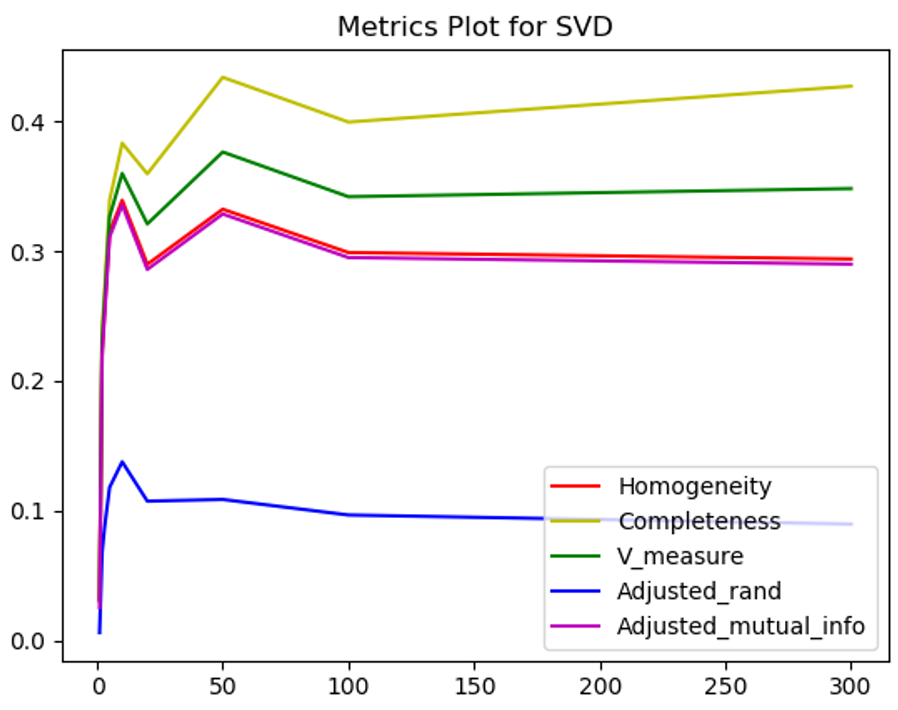
*Table 33 Confusion matrix for NMF, normalized (r = 2)*

|  |  |
| --- | --- |
| Measures of Purity | Value |
| Homogeneity Score | 0.697741874808 |
| Completeness Score | 0.698227855629 |
| V measure score | 0.697984780626 |
| Adjusted Rand Score | 0.795264898816 |
| Adjusted Mutual Info Score | 0.6976957736818 |

*Table 34 Measures of Purity for NMF, normalized (r = 2)*

**Problem 5: K-means Clustering (20 Original sub-classes)**

1. In question5, we expand what we do in question 3 and 4 to 20 datasets group. It needs to include all the documents and we try different dimension of truncated SVD and NMF to find proper parameters. For data processed by truncated SVD, we obtain better performance in both 20 and 50 dimensional space. For NMF, we get best performance when the dimension of matrix equals to 20.



*Figure 14 clustering result for SVD*

|  |  |
| --- | --- |
| Measures of Purity | Value |
| Homogeneity Score | 0.0303805740802 |
| Completeness Score | 0.0330925029545 |
| V measure score | 0.0316786040468 |
| Adjusted Rand Score | 0.00582183114211 |
| Adjusted Mutual Info Score | 0.025251706307 |

*Table 35 Measures of Purity for SVD (r = 1)*

|  |  |
| --- | --- |
| Measures of Purity | Value |
| Homogeneity Score | 0.22374133554 |
| Completeness Score | 0.239333003527 |
| V measure score | 0.231274684561 |
| Adjusted Rand Score | 0.0678107298008 |
| Adjusted Mutual Info Score | 0.219490450245 |

*Table 36 Measures of Purity for SVD (r =2 )*

|  |  |
| --- | --- |
| Measures of Purity | Value |
| Homogeneity Score | 0.255153733285 |
| Completeness Score | 0.269368550168 |
| V measure score | 0.262068527394 |
| Adjusted Rand Score | 0.0873335592807 |
| Adjusted Mutual Info Score | 0.251103018169 |

*Table 37 Measures of Purity for SVD (r = 3 )*

|  |  |
| --- | --- |
| Measures of Purity | Value |
| Homogeneity Score | 0.314897954026 |
| Completeness Score | 0.339421069478 |
| V measure score | 0.32669996284 |
| Adjusted Rand Score | 0.11791826017 |
| Adjusted Mutual Info Score | 0.311152601424 |

*Table 38 Measures of Purity for SVD (r = 5)*

|  |  |
| --- | --- |
| Measures of Purity | Value |
| Homogeneity Score | 0.339417815535 |
| Completeness Score | 0.383393857183 |
| V measure score | 0.360068079713 |
| Adjusted Rand Score | 0.137618887724 |
| Adjusted Mutual Info Score | 0.335796411643 |

*Table 39 Measures of Purity for SVD (r = 10 )*

|  |  |
| --- | --- |
| Measures of Purity | Value |
| Homogeneity Score | 0.289937811751 |
| Completeness Score | 0.359754974327 |
| V measure score | 0.321095053718 |
| Adjusted Rand Score | 0.107248016063 |
| Adjusted Mutual Info Score | 0.286017761724 |

*Table 40 Measures of Purity for SVD (r =20 )*

|  |  |
| --- | --- |
| Measures of Purity | Value |
| Homogeneity Score | 0.332461328151 |
| Completeness Score | 0.43422743531 |
| V measure score | 0.37659044124 |
| Adjusted Rand Score | 0.108556699632 |
| Adjusted Mutual Info Score | 0.328753175561 |

*Table 41 Measures of Purity for SVD (r = 50)*

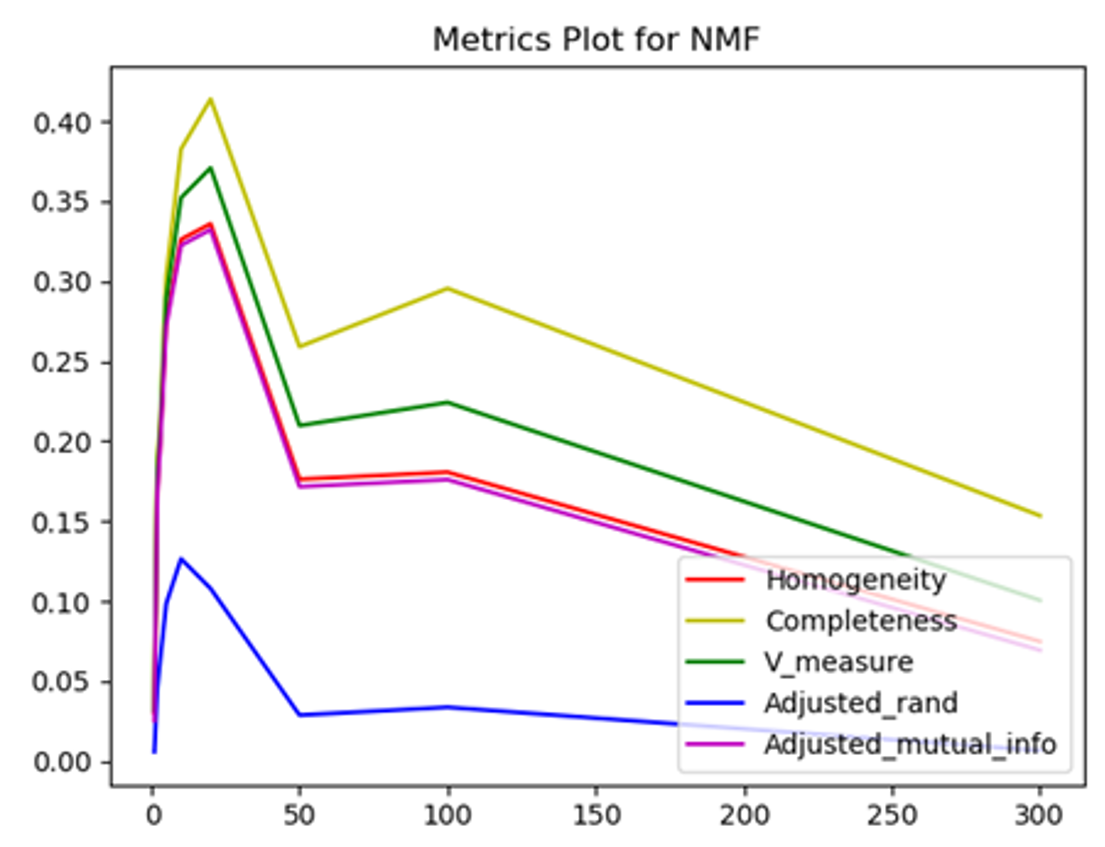
|  |  |
| --- | --- |
| Measures of Purity | Value |
| Homogeneity Score | 0.299039056575 |
| Completeness Score | 0.399699107114 |
| V measure score | 0.342118550601 |
| Adjusted Rand Score | 0.0965749963836 |
| Adjusted Mutual Info Score | 0.295165741685 |

*Table 42 Measures of Purity for SVD (r = 100 )*

|  |  |
| --- | --- |
| Measures of Purity | Value |
| Homogeneity Score | 0.294056770078 |
| Completeness Score | 0.427333441448 |
| V measure score | 0.348383688968 |
| Adjusted Rand Score | 0.0895781600162 |
| Adjusted Mutual Info Score | 0.290085673813 |

*Table 43 Measures of Purity for SVD (r = 300)*

1. Same as what we do in problem 4, we try different methods to process reduced-dimensional data matrix and see whether we can get better performance. For truncated SVD, we only normalize them. As for NMF, we also apply non-linear transformation( logarithm transformation) and combine it with normalization. Here, as we discussed in problem4, we try different offset(1 and 0.001) in logarithm transformation and the result shows that smaller offset can obtain better performance, exactly as what we predict.



|  |  |
| --- | --- |
| Measures of Purity | Value |
| Homogeneity Score | 0.0303640306657 |
| Completeness Score | 0.0330754403832 |
| V measure score | 0.0316617925551 |
| Adjusted Rand Score | 0.00581392384775 |
| Adjusted Mutual Info Score | 0.0252350628477 |

*Figure 15 clustering result for NMF*

*Table 44 Measures of Purity for NMF (r = 1)*

|  |  |
| --- | --- |
| Measures of Purity | Value |
| Homogeneity Score | 0.171986777301 |
| Completeness Score | 0.188534865339 |
| V measure score | 0.17988103938 |
| Adjusted Rand Score | 0.0456158982168 |
| Adjusted Mutual Info Score | 0.167453110588 |

*Table 45 Measures of Purity for NMF (r =2 )*

|  |  |
| --- | --- |
| Measures of Purity | Value |
| Homogeneity Score | 0.202601920081 |
| Completeness Score | 0.211551704952 |
| V measure score | 0.206980110901 |
| Adjusted Rand Score | 0.0637095411581 |
| Adjusted Mutual Info Score | 0.198256237603 |

*Table 46 Measures of Purity for NMF (r = 3 )*

|  |  |
| --- | --- |
| Measures of Purity | Value |
| Homogeneity Score | 0.275955771557 |
| Completeness Score | 0.303529863465 |
| V measure score | 0.289086778346 |
| Adjusted Rand Score | 0.098427823631 |
| Adjusted Mutual Info Score | 0.271979039973 |

*Table 47 Measures of Purity for NMF (r = 5)*

|  |  |
| --- | --- |
| Measures of Purity | Value |
| Homogeneity Score | 0.326231817807 |
| Completeness Score | 0.382683123057 |
| V measure score | 0.352209845449 |
| Adjusted Rand Score | 0.12647932908 |
| Adjusted Mutual Info Score | 0.322500771734 |

*Table 48 Measures of Purity for NMF (r = 10 )*

|  |  |
| --- | --- |
| Measures of Purity | Value |
| Homogeneity Score | 0.335967382618 |
| Completeness Score | 0.413960308428 |
| V measure score | 0.370908190192 |
| Adjusted Rand Score | 0.107877552157 |
| Adjusted Mutual Info Score | 0.332298789827 |

*Table 49 Measures of Purity for NMF (r =20 )*

|  |  |
| --- | --- |
| Measures of Purity | Value |
| Homogeneity Score | 0.176234588925 |
| Completeness Score | 0.259102019613 |
| V measure score | 0.209781291169 |
| Adjusted Rand Score | 0.0287453710237 |
| Adjusted Mutual Info Score | 0.171625592622 |

*Table 50 Measures of Purity for NMF (r = 50)*

|  |  |
| --- | --- |
| Measures of Purity | Value |
| Homogeneity Score | 0.180639833078 |
| Completeness Score | 0.295545740007 |
| V measure score | 0.224229107975 |
| Adjusted Rand Score | 0.0336674718109 |
| Adjusted Mutual Info Score | 0.175990380928 |

*Table 51 Measures of Purity for NMF (r = 100 )*

|  |  |
| --- | --- |
| Measures of Purity | Value |
| Homogeneity Score | 0.074792304002 |
| Completeness Score | 0.153443954209 |
| V measure score | 0.100566202412 |
| Adjusted Rand Score | 0.00666903597814 |
| Adjusted Mutual Info Score | 0.0693866451228 |

*Table 52 Measures of Purity for NMF (r = 300)*