ECE219 Project2 Clustering

Zhilai Shen, Yufei Hu, Zheang Huai, and Tianyi Liu

UCLA

1 Building the TF-IDF Matrix

Question 1:
The dimensions of the TF-IDF matrix is: (7882, 27768).

2 Apply K-means Clustering

2.1 Contingency Table

Question 2:
The contingency table of the clustering result is:

[[4 3899]
[1718 2261]]

As we can see, the result of K-means clustering using the TF-IDF data without dimensionality reduction is not that bad. Most points of one class can be recognized as in one cluster. But many points of the other class can't be clustered correctly.

2.2 Measures

Question 3: The 5 measures for the K-means clustering results is shown in Table 1.

Table 1. 5 measures for the K-means clustering results

Homogeneity	0.2536
Completeness	0.3348
V-measure	0.2886
Adjusted Rand Index	0.1808
Adjusted mutual info score	0.2535

As can be seen from the table, the scores of 5 measures are not very high, which means that the result of clustering is not very good. And we can see some reasons from question2: many points of one class are recognized as in another cluster.

3 Dimensionality Reduction

3.1 Principle Components

Question 4:

A plot of what ratio of the variance of the original data is retained after the dimensionality reduction is shown in Figure 1. It's pretty obvious that this curve increases monotonically.

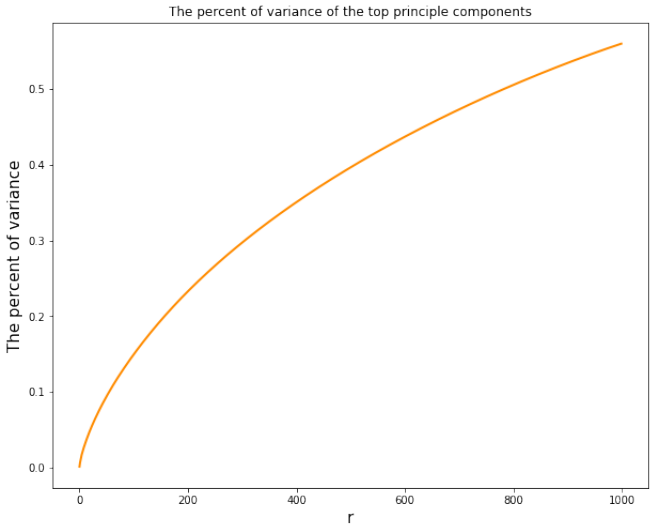

Fig. 1. The percent of variance of the top principle components

3.2 PCA and NMF

Question 5:

This time, we try several r values for SVD and NMF respectively and their measure scores are shown in Figure 2and 3. When it comes to the problem of choosing the best r, we need to decide which measure scores we care the most. Overall speaking, homogeneity and completeness didn't measure the result very well since they only show part of the properties of the result. Therefore we choose the other three methods to decide which parameter is the best. We can see from the plot result that the best r for both SVD and NMF is 2.

Question 6:

What we can also see from the plot is that the curve is non-monotonic. This can be explained as when the dimension of the data is pretty high the Euclidean distance metric becomes useless since the distances between data points tends

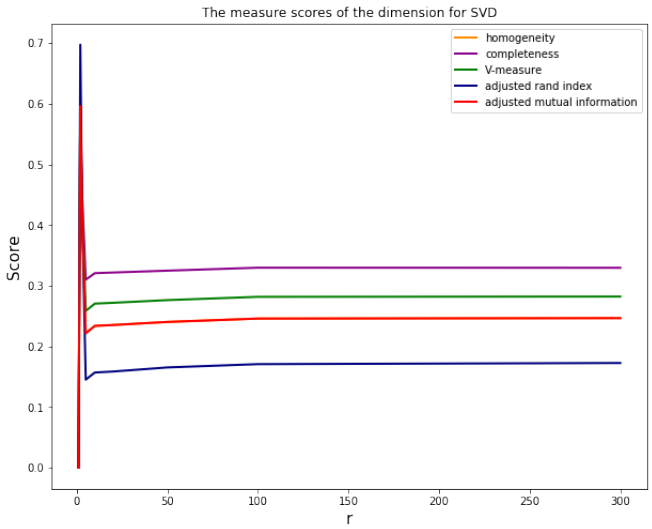

Fig. 2. 5 measure scores for different r value of SVD

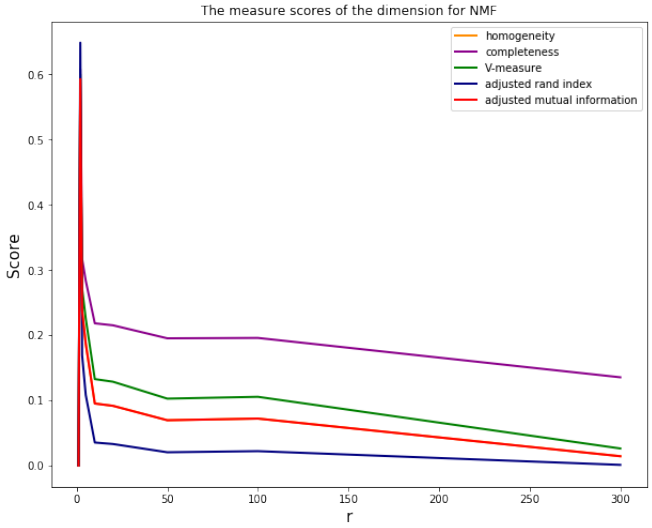

Fig. 3. 5 measure scores for different r value of NMF

to be almost the same in the pretty sparse high dimensional space. Also as the number of dimensions grows, the relative euclidean distance between a point in a set and its closest neighbour, and between that point and its furthest neighbour, changes in some non-obvious ways. Therefore we need a best parameter r which is not too small and not too big.

4 Optimization

4.1 Visualization

Question 7:

In this problem, we visualize the performance of clustering data with best clustering parameters derived from the previous questions. Dimension reduction are conducted using SVD and NMF respectively. The visualization can be seen below:

Fig. 4. Clustering result for SVD (r=2)

Table 2. Confusion matrix for SVD (r=2)

	label = 0	label = 1
label = 0	3656	247
label = 1	363	3616

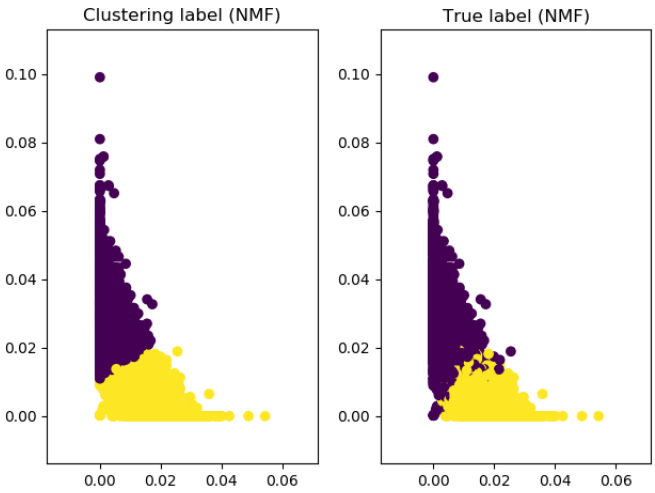

Fig. 5. Clustering result for NMF (r=2)

Table 3. Performance for SVD (r=2)

Homogeneity	0.609
Completeness	0.609
V measure	0.609
Adjusted rand	0.714
Adjusted mutual info	0.609

From the distribution of the data points after SVD, we can observe that it is relatively harder to cluster them into 2 classes comparing to NMF, because most of the points are very close to each other and there is no clear boundary. In contrast, NMF is slightly better, showing a triangular distribution. These can be seen from 6 and 10.

4.2 Transformation Methods

Question 8:

In this question, we applied different normalization methods to see if they increase the clustering performance: 1) normalizing features such that each feature has unit variance; 2) applying a non-linear transformation to the data vectors; 3) combining both transformations (in different orders) on the data.

Question 9:

We found log transformation greatly improves the result. The reason is because the original data points are skewed with wide distribution, and taking the log of the features may restore symmetry to it. As we can see from the distribu-

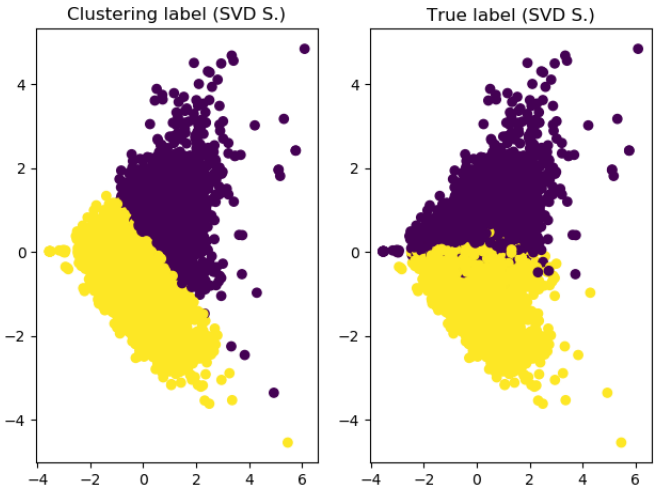

Fig. 6. Normalized (scaling) clustering result for SVD (r=2)

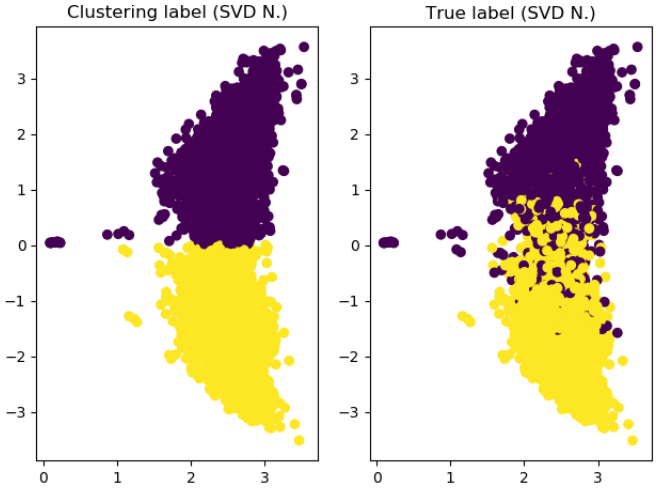

Fig. 7. Normalized (non-linear transformation) clustering result for SVD (r=2)

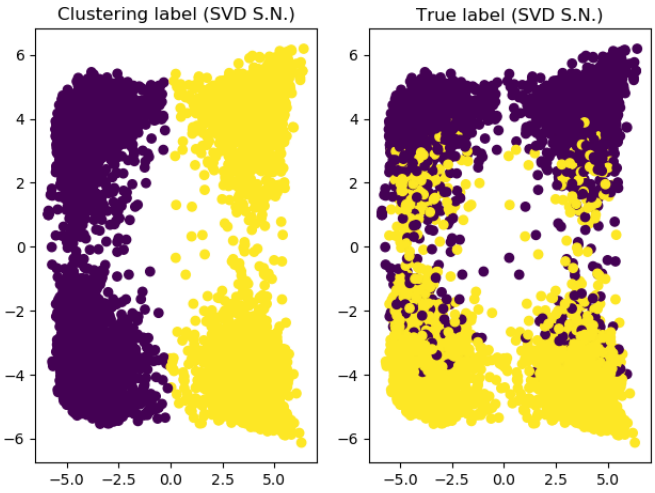

Fig. 8. Normalized (scaling + non-linear transformation) clustering result for SVD (r=2)

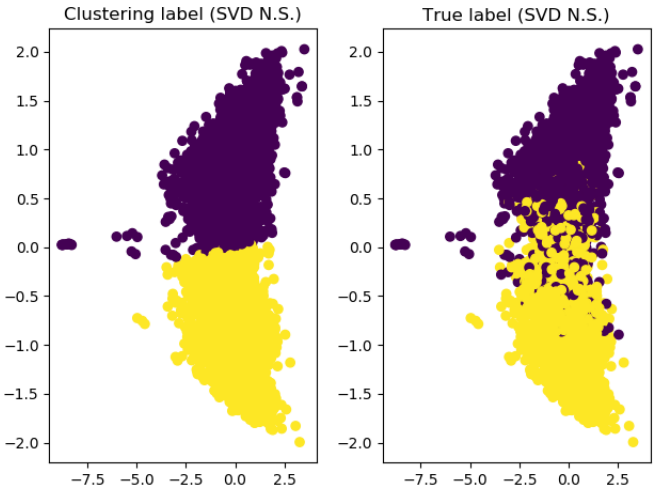

Fig. 9. Normalized (non-linear transformation + scaling) clustering result for SVD (r=2)

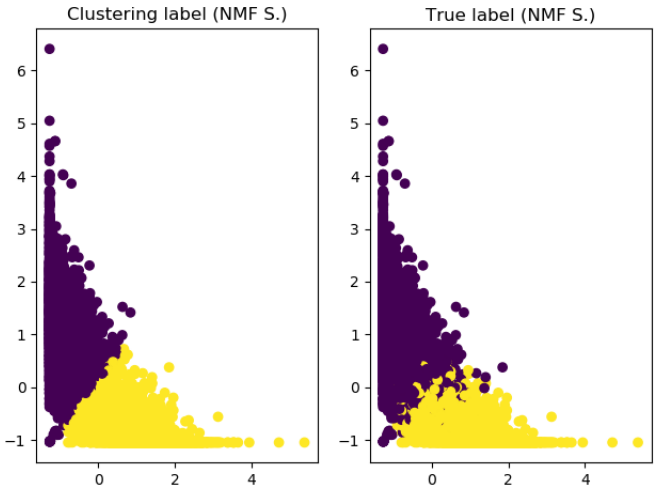

Fig. 10. Normalized (scaling) clustering result for NMF (r=2)

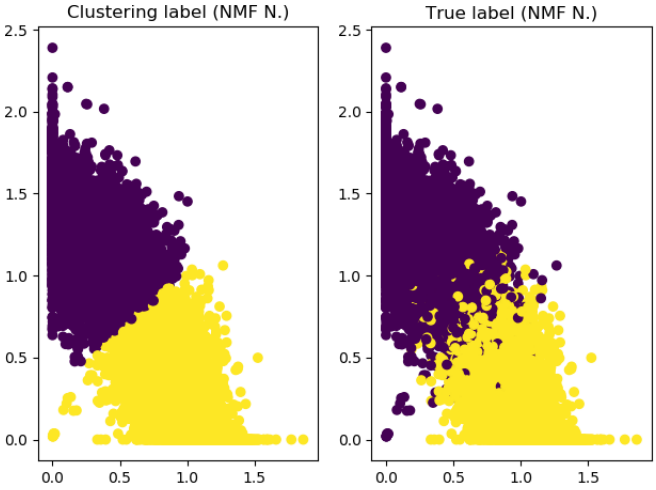

Fig. 11. Normalized (non-linear transformation) clustering result for NMF (r=2)

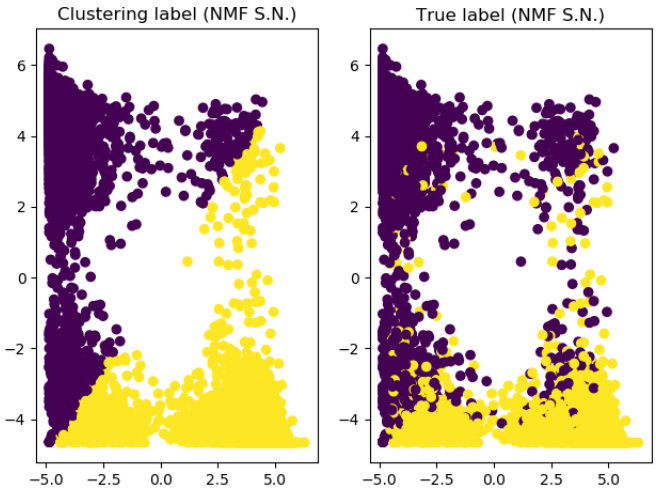

Fig. 12. Normalized (scaling + non-linear transformation) clustering result for NMF (r=2)

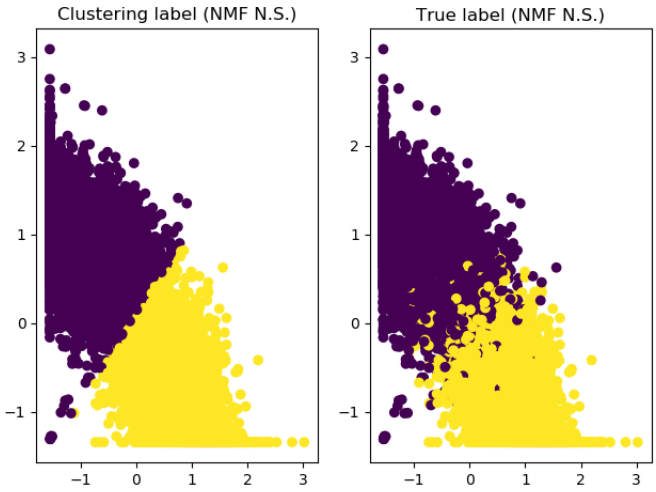

Fig. 13. Normalized (non-linear transformation + scaling) clustering result for NMF (r=2)

Table 4. Confusion matrix for NMF (r=2)

	label = 0	label = 1
label = 0	731	3172
label = 1	3943	36

Table 5. Performance for NMF (r=2)

Homogeneity	0.593
Completeness	0.608
V measure	0.600
Adjusted rand	0.649
Adjusted mutual info	0.593

tion of the observations after taking log, the points are mapped to a sector that is easier to cluster.

Question 10:

Table 6. Performance for SVD after normalization (scaling) (r=2)

Homogeneity	0.227
Completeness	0.257
V measure	0.241
Adjusted rand	0.244
Adjusted mutual info	0.227

All clustering measures are listed in 6, 7, 8, 9, 10, 11, 12, 13.

5 Expand Dataset into 20 Categories

Question 11:

Question 12:

Table 7. Performance for SVD after normalization (non-linear transformation) (r=2)

Homogeneity	0.609
Completeness	0.609
V measure	0.609
Adjusted rand	0.716
Adjusted mutual info	0.609

Table 8. Performance for SVD after normalization (scaling + non-linear transformation) (r=2)

Homogeneity	7.41e-05
Completeness	7.43e-05
V measure	7.42e-05
Adjusted rand	-1.32e-05
Adjusted mutual info	-1.74e-05

Table 9. Performance for SVD after normalization (non-linear transformation + scaling) (r=2)

Homogeneity	0.610
Completeness	0.610
V measure	0.610
Adjusted rand	0.717
Adjusted mutual info	0.609

Table 10. Performance for NMF after normalization (scaling) (r=2)

Homogeneity	0.682
Completeness	0.685
V measure	0.683
Adjusted rand	0.773
Adjusted mutual info	0.682

Table 11. Performance for NMF after normalization (non-linear transformation) (r=2)

Homogeneity	0.701
Completeness	0.702
V measure	0.702
Adjusted rand	0.795
Adjusted mutual info	0.701

Table 12. Performance for NMF after normalization (scaling + non-linear transformation) (r=2)

Homogeneity	0.695
Completeness	0.695
V measure	0.695
Adjusted rand	0.792
Adjusted mutual info	0.695

Table 13. Performance for NMF after normalization (non-linear transformation + scaling) (r=2)

Homogeneity	0.702
Completeness	0.704
V measure	0.703
Adjusted rand	0.797
Adjusted mutual info	0.702