**Project 2: Clustering**

Prof. Vwani Roychowdhury

ECE 219 Winter 2020

UCLA, Department of ECE

Siyuan Peng 805426447 Haonan Huang 605322629

Yinghan Cui 005430212 Yunzheng Zhu 505231998

**Introduction**

In this project, we are implementing the clustering on the “20 Newsgroups” dataset by the K-means clustering algorithm. The “20 Newsgroup” dataset has approximately 20,000 documents, partitioned evenly across 20 different newsgroups, each corresponding to a different category (topic). Then, we are using the K-means clustering to partition into the K clusters such that each point belongs to exactly 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. Our goal is to evaluate how purely the *a priori* known classes, defined as “Computer Technology” and “Recreational Activity”, can be reconstructed through the K-means clustering algorithm.

**Part 1 - Clustering of Text Data**

**Question 1 - Building the TF-IDF matrix**

Before we take all the documents belonging to the computer technology and recreational activity and perform unsupervised clustering into two clusters. We need to build the TF-IDF matrix first. So we use the CountVectorizer() and TfidfTransformer() to make the TF-IDF transformation. We use min-df=3 and its built-in English stop-words. Then we get the dimensions of the TF-IDF matrix which is (7882, 27768).

**Question 2 - The contingency table of your clustering result**

We use K-means clustering and set k=2 to cluster two classes. We change the max\_iter from 1000 to 1500,the result does not change. So we set n\_init=30, max\_iter=1000. Then we use contingency\_matrix() to get the contingency table of the clustering result, which is shown below.

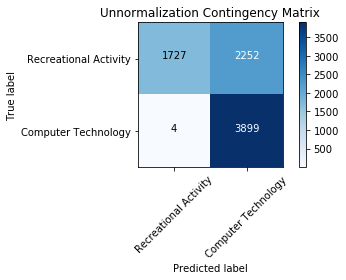
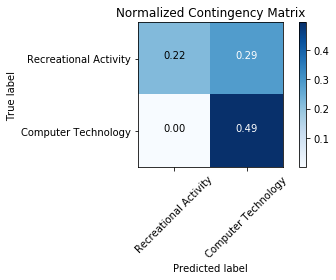


Figure 1. The contingency table of clustering result.

**Question 3 - The 5 measures for the K-means clustering**

In order to evaluate clustering results, there are various measures for a given partition of the data points with respect to the ground truth. We will use the measures homogeneity score, completeness score, V-measure, adjusted Rand score and adjusted mutual info score.

Homogeneity: 0.255

Completeness: 0.336

V-measure: 0.290

Adjusted Rand-Index: 0.183

Adjusted Mutual Information Score: 0.290

While the range of these evaluation index is [0, 1] or [-1, 1], and 1 stands for perfect clustering. The 5 measures evaluation shows that the K-means clustering does not have a good performance.

**Question 4 - Dimensionality reduction plot**

In this part, we are trying to provide a better clustering result instead of the TF-IDF since the high dimensional sparse TF-IDF vectors do not yield a good clustering result. The reason behind is that the Euclidean distance is not a good metric, which has the distance between data points tends to be almost the same. K-means clustering must be isotropically shaped. It will fail to identify the clusters properly if it is either not round-shaped or with unequal variances. As a result, we are applying Singular Value Decomposition (SVD) and Non-negative Matrix Factorization (NMF) for dimensionality reduction.

The percent of variance with the top r principal components v.s. r = 1 to 1000 plot is shown below:

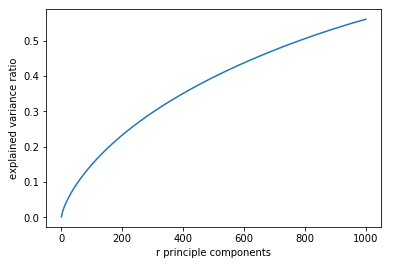


Figure 2. Percent of variance with the top r principal components v.s. r = 1 to 1000.

Based on the plot, we find that the explained variance is increased as r increases. The plot is similar to a logarithm. The slope of increment is decreasing with respect to r.

**Question 5 - Choice of r for SVD and NMF**

We want to apply the two methods, truncated SVD/PCA and NMF, to reduce the dimension of the data. We are trying r, the dimension that we want to reduce the data, that equals to 1, 2, 3, 5, 10, 20, 50, 100, and 300. Then, we plot the 5 measure scores v.s. r for both SVD and NMF, and we don’t have to choose a specific here. The plots for the 5 measure scores v.s. r are shown below:

**SVD NMF**

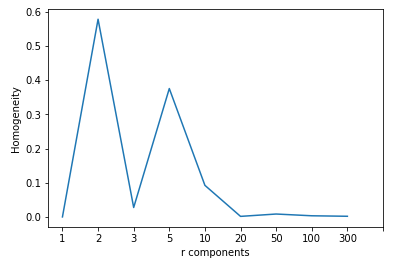
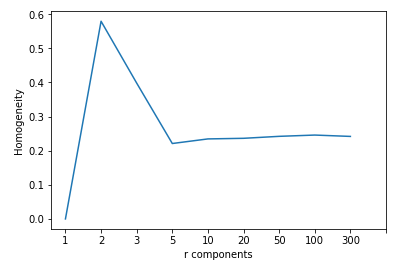


Figure 3. Homogeneity v.s. r components. Homogeneity is the measure of the purity for the clustering.

**SVD NMF**

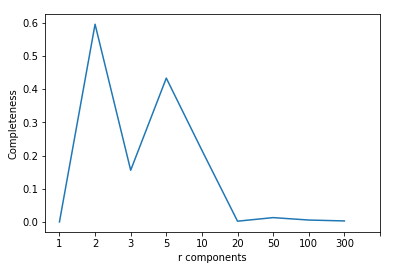
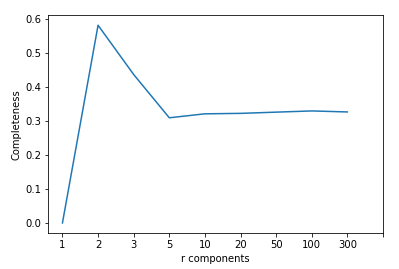


Figure 4. Completeness v.s. r components. Completeness is the measure of how complete each cluster covers the classes.

**SVD NMF**

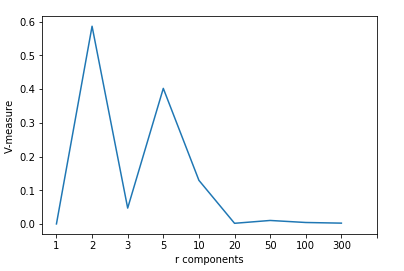
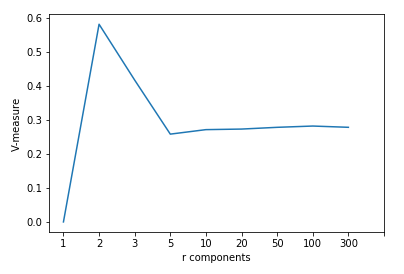


Figure 5. V-measure v.s. r components. V-measure is an harmonic mean of homogeneity and completeness which can be used as an overall measurement of the clustering.

**SVD NMF**

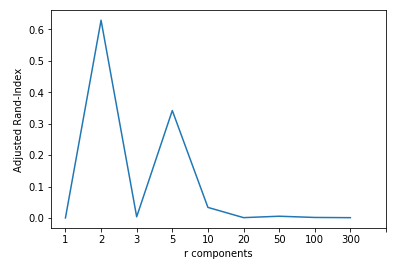
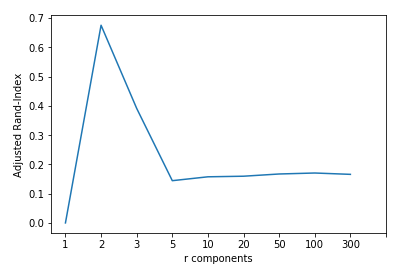
****

Figure 6. Adjusted Rand-Index v.s. r components.Adjusted rand index measures the similarity between two data clusterings.

**SVD NMF**

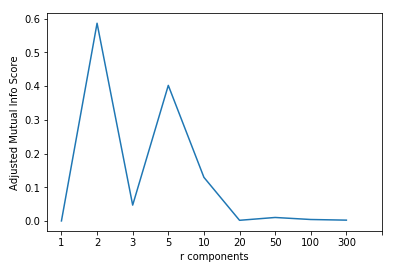
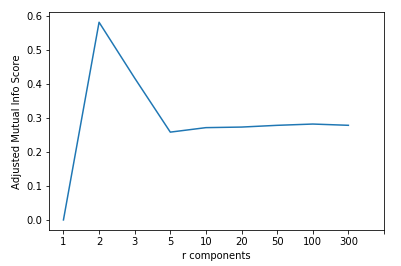
****

Figure 7. Adjusted mutual information score v.s. r components.Adjusted mutual information score measures the similarity between two data clusterings as well.

Based on the plots, we find that the tendencies of all measurements in both SVD and NMF are exactly the same.

**Question 6 - Explain the non-monotonic behavior of the measures as r increases**

The measures for the clustering results are non-monotonic. For low dimensional features, the measures are close to zero. As r goes up, the measures increase initially and drop afterwards. The reason is that low dimensional features are with limited information contained thus clustering cannot be properly processed. For high dimensional features, the Euclidean distance used in the K-means will not be a good metric since the distance for each pair of data points are too small and the data points are not performed uniformly well.

**Question 7 - Visualizing the clustering results for SVD and NMF with the choice of r**

We think r = 2 is a proper number here due to the reasons explained in the previous part. The plots of the SVD and NMF with r = 2 are shown below:

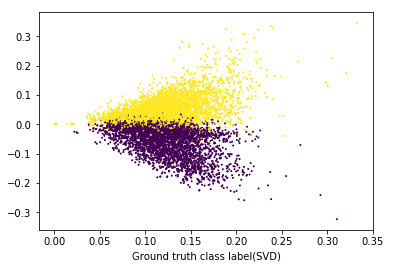
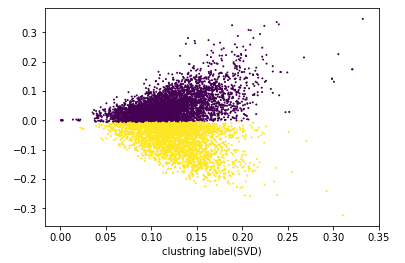
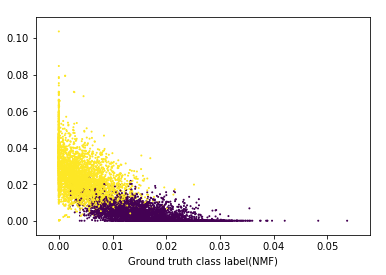
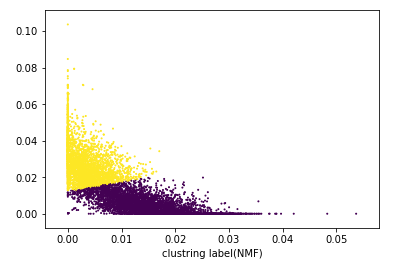
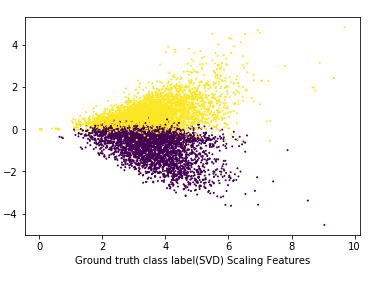
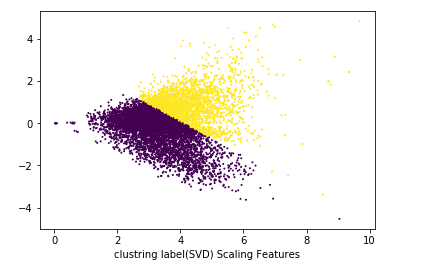


Figure 8. The Visualizing clustering results with r = 2 for SVD in 2D. Different clusters are having different colors (purple and yellow).

Figure 9. The Visualizing clustering results with r = 2 for NMF in 2D. Different clusters are having different colors (purple and yellow).

**Question 8 - Visualizing the transformed data**

In this part, we are performing the transformation on SVD-reduced data and NMF-reduced data with the same r we have chosen in the last part (r = 2). We apply scaling features and logarithmic non-linear transformation on both types of data. The plot results are shown below:

Figure 10. SVD scaling features results with r = 2. Different clusters are having different colors (purple and yellow).

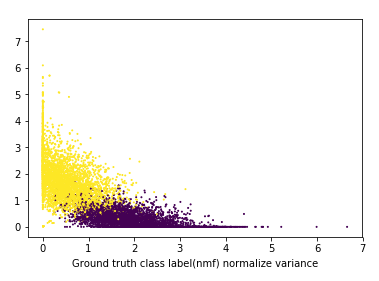
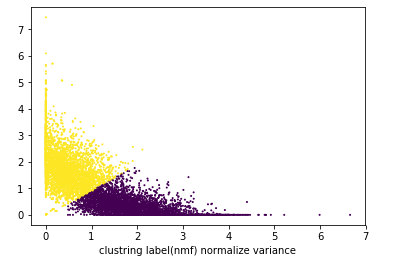
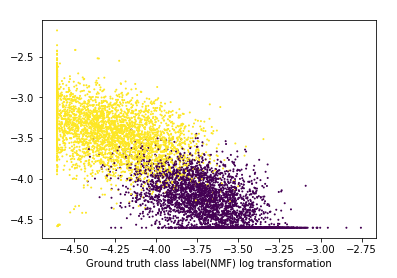
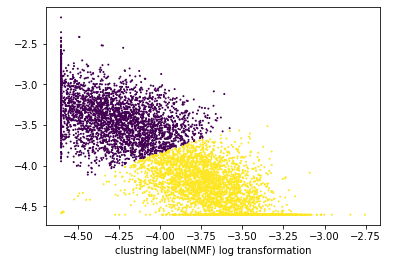
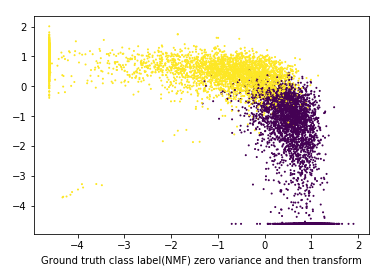
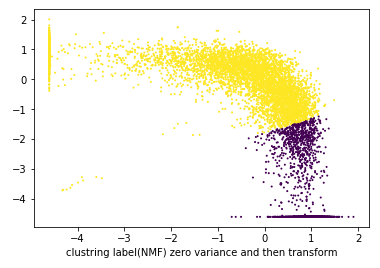


Figure 11. NMF scaling features results with r = 2. Different clusters are having different colors (purple and yellow).

Figure 12. NMF logarithmic transformation results with r = 2. Different clusters are having different colors (purple and yellow).

Figure 13. NMF scaling featuring first then logarithmic transformation results with r = 2. Different clusters are having different colors (purple and yellow).

**Question 9 - Why logarithmic transformation may improve the clustering results?**

In K-means algorithm, it uses the Euclidean distance that implicitly assumes the clusters are isotropically shaped. The results will be bad if the data is skewed severely. However, logarithm transformation will distribute the data more uniformly, which will avoid the skewed data results. Thus, logarithm transformation method can improve the clustering results.

**Question 10 - Clustering measures**

The results of the clustering measures for the SVD and NMF with r = 2 in the six different cases are shown in the tables below (since the results can be negative after using the logarithmic transformation, we are only testing with or without the scaling features method):

|  |  |  |
| --- | --- | --- |
| Measure | Scaling Features | None |
| Homogeneity | 0.235 | 0.579 |
| Completeness | 0.264 | 0.582 |
| V-measure | 0.249 | 0.581 |
| Adjusted Rand-Index | 0.254 | 0.675 |
| Adjusted Mutual Info Score | 0.248 | 0.581 |

Table 1. Clustering Measures for SVD with r = 2 in two different cases (with scaling features only and without any of the method).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Measure | Log Transformation | Scaling Features | Scaling Features then Log Transformation | None |
| Homogeneity | 0.701 | 0.684 | 0.313 | 0.578 |
| Completeness | 0.702 | 0.686 | 0.383 | 0.595 |
| V-measure | 0.701 | 0.685 | 0.344 | 0.586 |
| Adjusted Rand-Index | 0.796 | 0.774 | 0.249 | 0.629 |
| Adjusted Mutual Info Score | 0.701 | 0.685 | 0.344 | 0.586 |

Table 2. Clustering Measures for NMF with r = 2 in four different cases (with logarithmic transformation only, with scaling features only, with both scaling features and logarithmic transformation, and without any of the method).

Based on the results, in SVD, it is better to cluster the data without applying the scaling features method. In NMF, it is better to cluster the data with logarithmic transformation since it could smooth the dataset with uniformly distributed results.

**Part 2 - Own Dataset**

We choose a very typical dataset, CIFAR-10 dataset to perform a clustering analysis based on work mentioned above. CIFAR-10 dataset consists of 60000 32x32 color images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images. Inspired by Part 1, which we achieve data importing, K-means clustering, contingency table of two classes clustering, 5 measures, namely homogeneity score, completeness score, V-measure, adjusted Rand score and adjusted mutual info score, 2 ways of dimension reduction, namely principle component analysis (Truncated SVD specifically), and non-negative matrix factorization, visualization, and different methods for performance improvement.

With regard to this pipeline, we achieve a realization for the whole clustering process for CIFAR-10. Since we do 10 classes clustering, it is more convenient to use 5 measures, homogeneity score, completeness score, V-measure, adjusted Rand score and adjusted mutual info score, to measure the performance than contingency tables.

Because we are doing 10 classes clustering, the 5 measures we achieve are much lower than those in Part 1. However, the improvement still holds, which is that scaling features can enhance the performance. For example, the homogeneity of 10 classes clustering after PCA (Truncated SVD) is 0.078, and the homogeneity after scaling is 0.105, which gains an increase of about 35%. The resulting plots and tables for the clustering are shown below. For more details, please refer to our jupyter notebook document.

1. The 5 measures for the clustering:

Homogeneity: 0.078

Completeness: 0.080

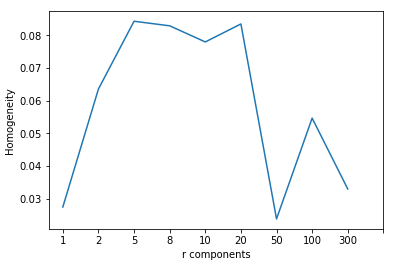
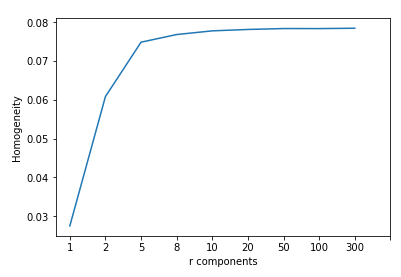
V-measure: 0.079

Adjusted Rand-Index: 0.042

Adjusted Mutual Information Score: 0.079

1. The plots for the 5 measure scores v.s. r:

**SVD NMF**

Figure 14. Homogeneity v.s. r components. Homogeneity is the measure of the purity for the clustering.

**SVD NMF**

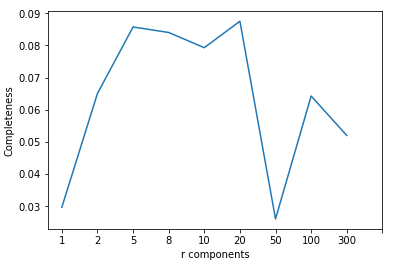
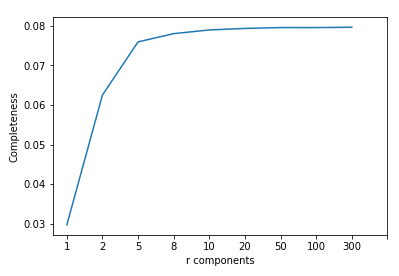
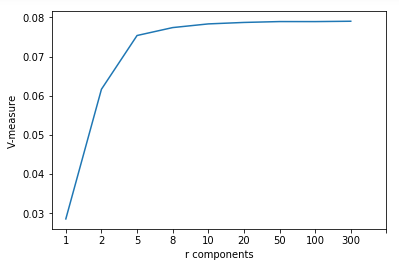
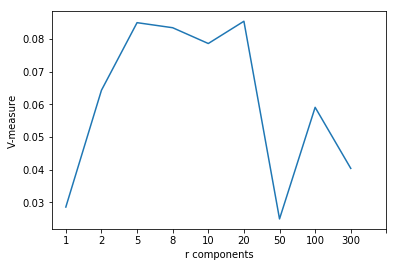


Figure 15. Completeness v.s. r components. Completeness is the measure of how complete each cluster covers the classes.

**SVD NMF**

****Figure 16. V-measure v.s. r components. V-measure is an harmonic mean of homogeneity and completeness which can be used as an overall measurement of the clustering.

**SVD NMF**

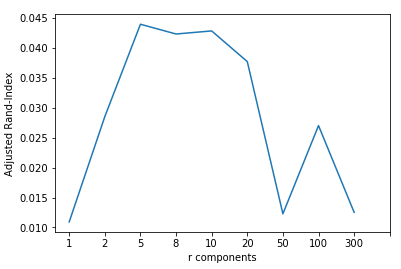
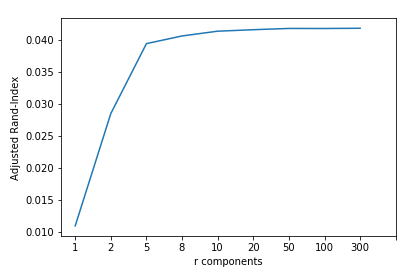
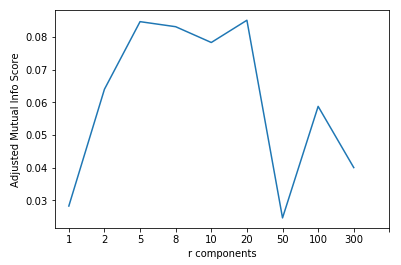
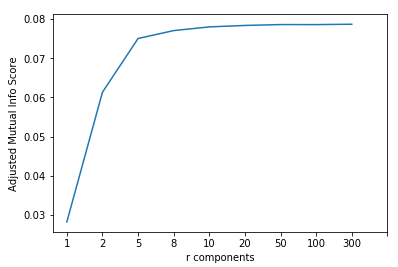
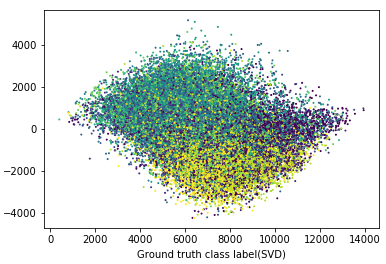
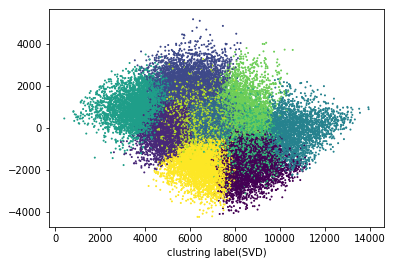
****

Figure 17. Adjusted Rand-Index v.s. r components.Adjusted rand index measures the similarity between two data clusterings.

**SVD NMF**

****Figure 18. Adjusted mutual information score v.s. r components.Adjusted mutual information score measures the similarity between two data clusterings as well.

1. The visualizing results of clustering for SVD and NMF with r = 20:

Figure 19. The Visualizing clustering results with r = 20 for SVD in 2D. 10 different clusters are having different colors.

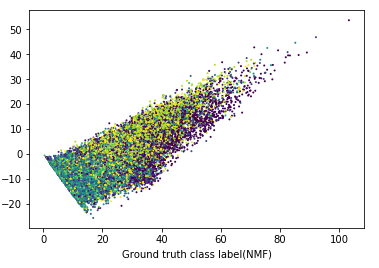
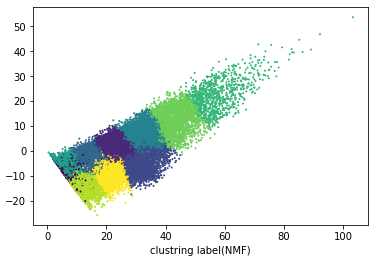


Figure 20.The Visualizing clustering results with r = 20 for NMF in 2D. 10 different clusters are having different colors.

1. The visualizing results of clustering for SVD and NMF with scaling features with r = 20:

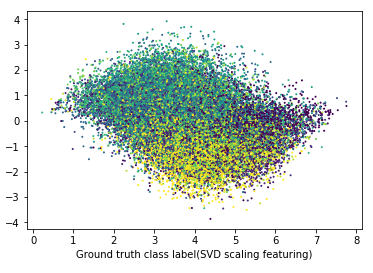
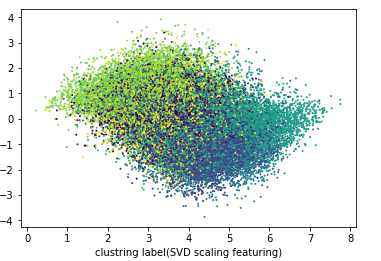


Figure 21. SVD scaling features results with r = 20. 10 different clusters are having different colors.

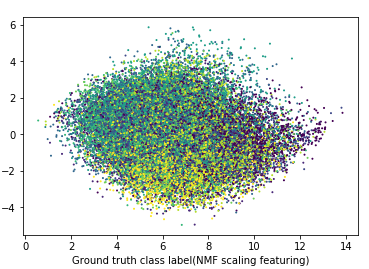
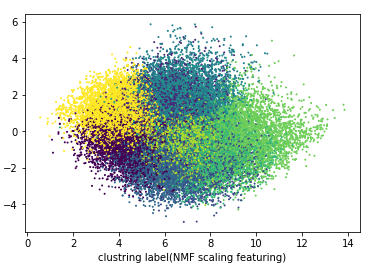


Figure 22. NMF scaling features results with r = 20. 10 different clusters are having different colors.

1. The visualizing results of clustering for SVD and NMF with logarithmic transformation with r = 20:

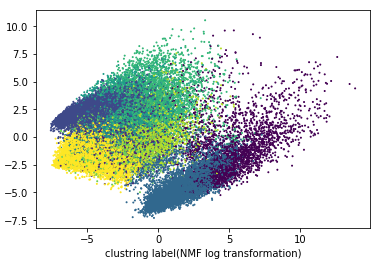


Figure 23. NMF logarithmic transformation results with r = 20. 10 different clusters are having different colors.

1. The visualizing results of clustering for SVD and NMF with scaling features then logarithmic transformation with r = 20:

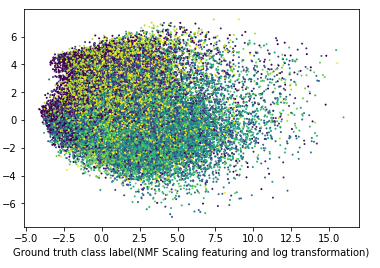
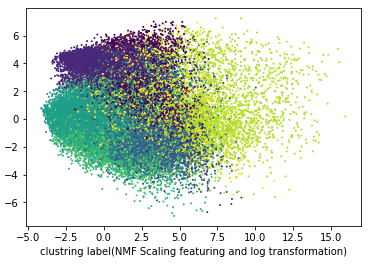


Figure 24. NMF scaling features then logarithmic transformation results with r = 20. 10 different clusters are having different colors.

1. The results of the clustering measures for the SVD and NMF with r = 20 in the six different cases are shown in the tables below (since the results can be negative after using the logarithmic transformation, we are only testing with or without the scaling features method):

|  |  |  |
| --- | --- | --- |
| Measure | Scaling Features | None |
| Homogeneity | 0.105 | 0.078 |
| Completeness | 0.106 | 0.079 |
| V-measure | 0.106 | 0.079 |
| Adjusted Rand-Index | 0.057 | 0.042 |
| Adjusted Mutual Info Score | 0.105 | 0.078 |

Table 3. Clustering Measures for SVD with r = 20 in two different cases (with scaling features only and without any of the method).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Measure | Log Transformation | Scaling Features | Scaling Features then Log Transformation | None |
| Homogeneity | 0.068 | 0.095 | 0.061 | 0.083 |
| Completeness | 0.074 | 0.097 | 0.064 | 0.088 |
| V-measure | 0.071 | 0.096 | 0.062 | 0.085 |
| Adjusted Rand-Index | 0.031 | 0.050 | 0.027 | 0.038 |
| Adjusted Mutual Info Score | 0.071 | 0.096 | 0.062 | 0.085 |

Table 4. Clustering Measures for NMF with r = 20 in four different cases (with logarithmic transformation only, with scaling features only, with both scaling features and logarithmic transformation, and without any of the method).

Based on the results, we find that the scaling features is improving the score for both SVD and NMF methods in the 10 classes clustering for the CIFAR-10 dataset. However, logarithmic transformation does not improve the score in the NMF method.

**Part 3- Color Clustering**

In this part, we use an image of G.E.M., a famous singer in Hong Kong, China, as shown in Fig. 25, which is found in<http://j.17qq.com/article/hcsutuax.html>. We do a 3 classes clustering using K-means to every pixel in this image and obtain a new image, as shown in Fig. There are exactly 3 colors in this new image, corresponding to (255,0,0), (0,255,0), and (0,0,255) in RGB vector space.

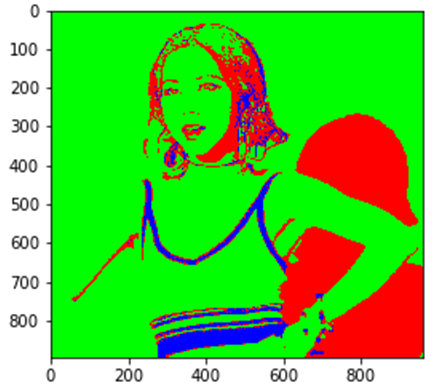


Figure 25. Original Image(Left) and the result of K-means clustering(Right).

**Question 11 (Bonus) - Methodology to make appropriate choice of k and initial seeds of cluster centers**

There are three main factors we need to consider when constructing the methodology. One is the 5 measures mentioned, namely homogeneity score, completeness score, V-measure, adjusted Rand score and adjusted mutual info score. The second one is time t that each epoch consumes. And the last one is the number of classes we want to get, namely k. We denote the sum of 5 measures as s. And we can construct a grade as s/(k\*log(t+1)). The bigger the grade is, the more appropriate k and initial seeds are.