CSE506 - Data Mining

REPORT

Assignment 3

Submitted By
Vidhi Sharma, 2019286
Dolly Sidar, 2019304

QUESTION 1:

Assumption: Clustering technique used

- K Means Clustering
- K Median Clustering
- BIRCH Clustering (BIRCH is short for Balanced Iterative Reducing and Clustering using Hierarchies)
- Gaussian Mixture

Pre-processing

1. Label encoding
2. Splitting into features and labels
3. Scaling using MinMax Scaler
4. Feature transformation using PCA

PART 1: Centroid/representative object/prototype of each cluster for every model

Gaussian clustering: Representative object is means_

	0	1	2	3	4	5	6	7	8	9
0	0.374499	-0.352381	-0.045478	-0.023474	-0.049090	0.058427	0.361452	0.923244	-0.140462	-0.014274
1	-0.316636	0.013611	-0.123355	0.339407	-0.123825	0.024863	-0.006386	-0.002221	-0.000277	-0.000186
2	-0.012727	0.052118	0.598731	0.031016	0.066321	-0.008704	-0.000110	-0.003967	0.001856	-0.000582
3	-0.169690	0.128627	-0.153196	-0.175078	0.039003	-0.009133	-0.001756	-0.002073	0.001148	-0.000722
4	-0.098816	-0.497621	-0.154174	-0.104401	0.030126	0.001608	0.002978	-0.024669	0.001340	-0.000138
5	0.601018	0.032439	0.125063	0.048649	-0.010960	-0.004448	-0.010741	-0.021514	0.002782	-0.000865
6	0.186271	0.093838	-0.024507	-0.042538	0.007640	-0.033118	0.326274	0.164159	0.051610	0.993997

K Median clustering: Representative object is get_medians()

	0	1	r2	3	4	5	6	7	8	9
0	-0.438971	0.174450	0.057652	-0.170511	-0.046830	-0.132728	0.062663	0.007518	0.031105	-0.001433
1	0.055941	-0.516487	-0.307895	-0.073961	0.074360	-0.213633	0.139954	-0.039338	0.019898	-0.001578
2	0.586968	-0.455835	0.240852	-0.020292	-0.074274	0.209355	0.557943	0.901975	-0.074959	-0.016088
3	0.034140	0.155212	-0.402274	-0.063917	0.192925	0.051116	0.121338	-0.046124	-0.085022	-0.000849
4	-0.414286	-0.483794	0.089631	-0.138766	0.024831	0.098607	0.037774	0.019066	0.161079	-0.001634
5	-0.003736	0.250331	0.704179	0.156074	0.681896	0.209823	0.097760	0.001162	0.038620	-0.001675
6	0.598827	-0.425560	0.249172	-0.020240	-0.108704	0.113196	0.110815	-0.068076	-0.042980	-0.000689

K Mean clustering: Representative object is cluster_centers_

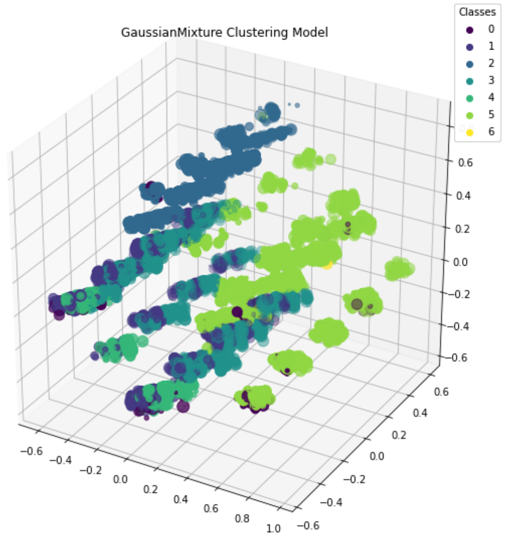
	0	1	2	3	4	5	6	7	8	9
0	0.641312	-0.296443	0.109913	0.040972	-0.048323	0.013356	-0.004054	0.016066	-0.001461	-0.000752
1	-0.421894	0.253227	0.019579	-0.050032	-0.066239	0.016941	-0.009528	0.008722	-0.000843	0.000228
2	0.043121	0.289019	-0.394947	0.010657	0.078138	-0.024306	0.012971	-0.001763	0.000543	-0.000562
3	0.591404	0.327890	0.097094	0.029945	-0.022620	-0.017554	0.000536	-0.004302	-0.000186	0.001402
4	0.006414	0.046192	0.595773	0.050057	0.097623	-0.008724	0.006127	-0.001242	0.000028	-0.000116
5	-0.408484	-0.279774	0.069915	-0.033688	-0.083470	0.019235	-0.018346	0.000710	0.003611	0.000110
6	0.048398	-0.371015	-0.314276	-0.004194	0.056422	-0.005160	0.014940	-0.015995	-0.002341	-0.000198

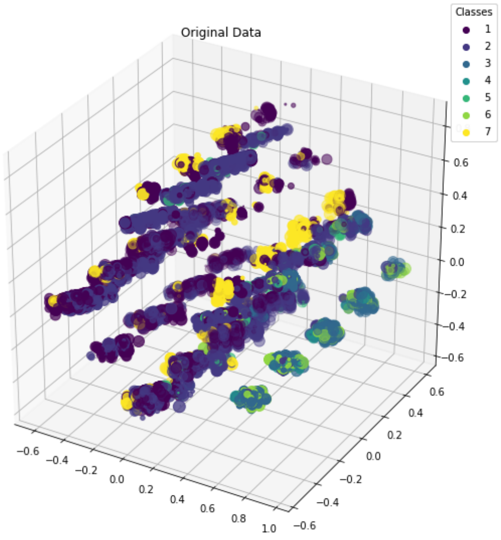
BIRCH clustering: Representative object is subcluster_centers_

0	1	2	3	4	5	6	7	8	9	
0	0.187618	0.522233	-0.422880	-0.123937	0.006443	0.613125	0.026188	0.028415	0.175191	-0.002526
1	-0.149167	0.510755	-0.204652	-0.139939	0.025822	0.212644	0.049513	0.024016	0.103017	-0.002200
2	-0.052724	0.199297	-0.158329	-0.171806	-0.130139	0.470034	0.016055	-0.003932	0.051143	-0.000918
3	-0.153232	0.514746	-0.223849	-0.126234	0.086621	0.355733	0.041428	0.009596	0.050633	-0.001505
4	-0.115804	0.508567	-0.212816	-0.132344	0.030396	0.183016	-0.277844	0.061005	0.069998	0.000349
5	-0.064019	0.530373	-0.188109	-0.177906	-0.121262	0.454387	0.016099	0.007214	0.051047	-0.001261
6	-0.066114	0.526465	-0.160529	-0.197451	-0.210678	0.254347	0.040900	-0.026625	-0.078735	-0.000374

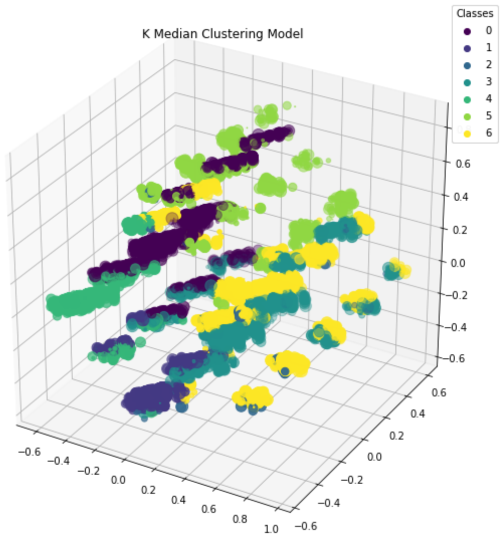
PART 2: Visualization of the clusters

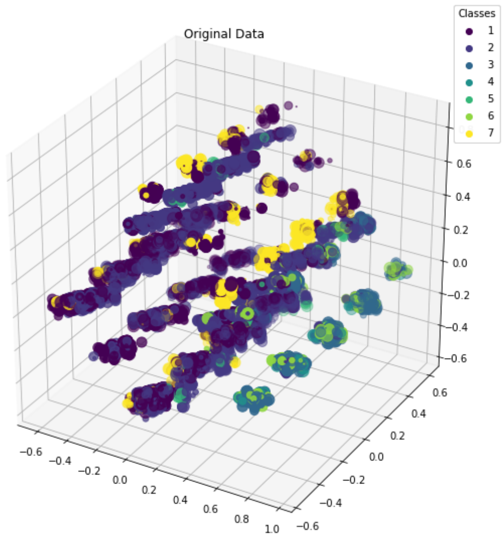
Gaussian clustering:



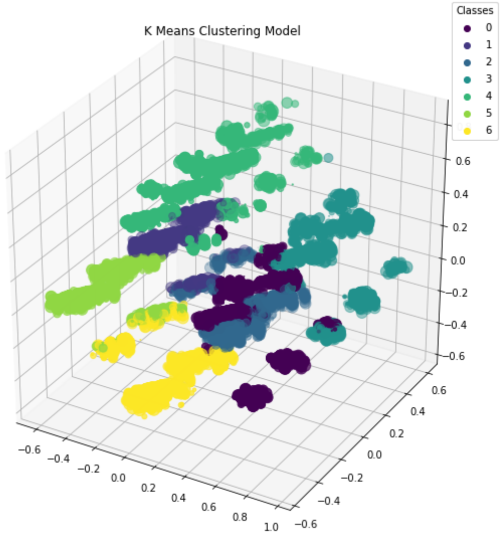


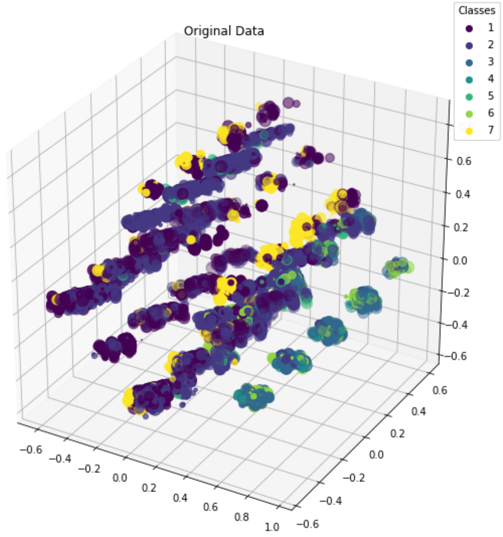
K Median clustering:



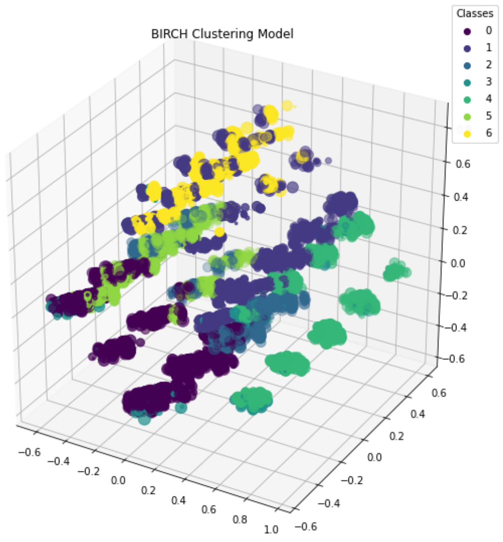


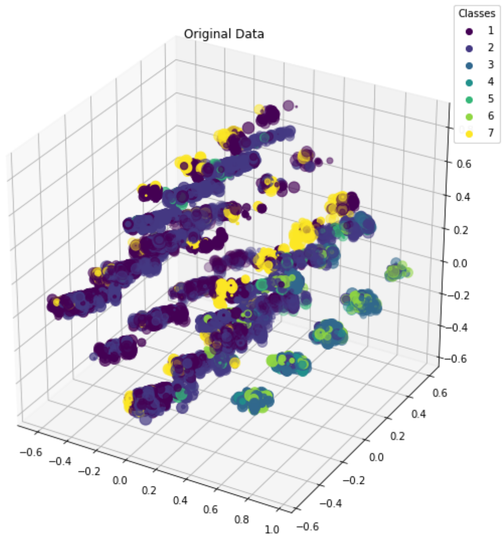
K Mean clustering:





BIRCH clustering:





PART 3: Comparing cluster distribution with the true label count

Assumption - for comparing the cluster distribution with the true label count, we are finding the percentage of true labels 1-7 in each of the 7 clusters(0-6).
For example, in cluster 0 we have shown percentage of true labels present in descending order.

Gaussian clustering:

```
Percentage of true labels in Cluster  0
Total Instances :  3957
3    0.341168
2    0.304018
1    0.257266
6    0.086176
7    0.011120
5    0.000253

Percentage of true labels in Cluster  1
Total Instances :  73953
1    0.503266
2    0.462253
7    0.034062
5    0.000419

Percentage of true labels in Cluster  2
Total Instances :  47290
2    0.771897
1    0.158575
```

```
Percentage of true labels in Cluster  4
Total Instances :  47805
1    0.561803
2    0.417613
7    0.012990
5    0.006987
6    0.000586
3    0.000021

Percentage of true labels in Cluster  5
Total Instances :  86694
2    0.358675
3    0.271899
6    0.135869
1    0.105636
7    0.074065
5    0.031675
4    0.022181

Percentage of true labels in Cluster  6
```

```
5    0.046860
7    0.022669

Percentage of true labels in Cluster  3
Total Instances :  146722
2    0.513120
1    0.453006
7    0.024829
5    0.008833
3    0.000211
```

```
Total Instances :  287
3    0.257840
2    0.257840
1    0.250871
7    0.128920
5    0.073171
6    0.031359

Total True Label Count :
1    148288
2    198310
3     25028
4      1923
5      6645
6     12157
7     14357
```

K Median clustering:

```
Percentage of true labels in Cluster  0
Total Instances :  127409
2    0.534413
1    0.437143
7    0.016796
5    0.011648

Percentage of true labels in Cluster  1
Total Instances :  48217
1    0.526599
2    0.428977
7    0.033992
5    0.009851
6    0.000581

Percentage of true labels in Cluster  2
Total Instances :  3532
3    0.385334
2    0.311721
1    0.194224
6    0.097112
7    0.011325
5    0.000283

Percentage of true labels in Cluster  3
Total Instances :  90989
2    0.519689
1    0.373913
3    0.037158
7    0.028421
6    0.021222
5    0.014738
4    0.004858
```

```
Percentage of true labels in Cluster  4
Total Instances :  43702
2    0.521555
1    0.463823
7    0.010022
5    0.004599

Percentage of true labels in Cluster  5
Total Instances :  21509
2    0.386024
1    0.321958
7    0.249709
5    0.042308

Percentage of true labels in Cluster  6
Total Instances :  71350
2    0.421219
3    0.284317
6    0.138122
1    0.074254
5    0.031296
7    0.030035
4    0.020757

Total True Label Count :
1    148288
2    198310
3     25028
4      1923
5      6645
6     12157
7     14357
```

K Mean clustering:

```
Percentage of true labels in Cluster  0
Total Instances :  41821
3    0.346668
2    0.297339
6    0.150522
1    0.073289
7    0.070228
5    0.032639
4    0.029315

Percentage of true labels in Cluster  1
Total Instances :  74863
1    0.506485
2    0.474373
7    0.016764
5    0.002378

Percentage of true labels in Cluster  2
Total Instances :  64918
2    0.499030
1    0.458024
7    0.035137
5    0.007640
3    0.000169

Percentage of true labels in Cluster  3
Total Instances :  44083
2    0.434998
3    0.238142
6    0.132341
1    0.099993
7    0.047365
5    0.031350
4    0.015811
```

```
Percentage of true labels in Cluster  4
Total Instances :  50929
2    0.722437
1    0.184119
7    0.049540
5    0.043904

Percentage of true labels in Cluster  5
Total Instances :  67253
2    0.493405
1    0.482715
7    0.020371
5    0.003509

Percentage of true labels in Cluster  6
Total Instances :  62841
1    0.498448
2    0.458522
7    0.030283
5    0.011967
6    0.000446
3    0.000334

Total True Label Count :
1    148288
2    198310
3     25028
4      1923
5      6645
6     12157
7     14357
```

BIRCH clustering:

```
Percentage of true labels in Cluster  0
Total Instances :  127535
1    0.518289
2    0.446285
7    0.026714
5    0.007331
3    0.000800
6    0.000580

Percentage of true labels in Cluster  1
Total Instances :  43428
2    0.549346
1    0.222184
7    0.149351
```

```
Percentage of true labels in Cluster  4
Total Instances :  55998
3    0.408318
2    0.327565
6    0.182685
4    0.034341
5    0.028858
1    0.018233

Percentage of true labels in Cluster  5
Total Instances :  65001
2    0.508100
1    0.475454
7    0.012846
```

```
6    0.034609
5    0.030096
3    0.014415

Percentage of true labels in Cluster  2
Total Instances :  76077
2    0.519171
1    0.440817
7    0.033400
5    0.006467
3    0.000145

Percentage of true labels in Cluster  3
Total Instances :  4244
3    0.335533
2    0.300895
1    0.256833
6    0.082469
7    0.019086
5    0.005184
```

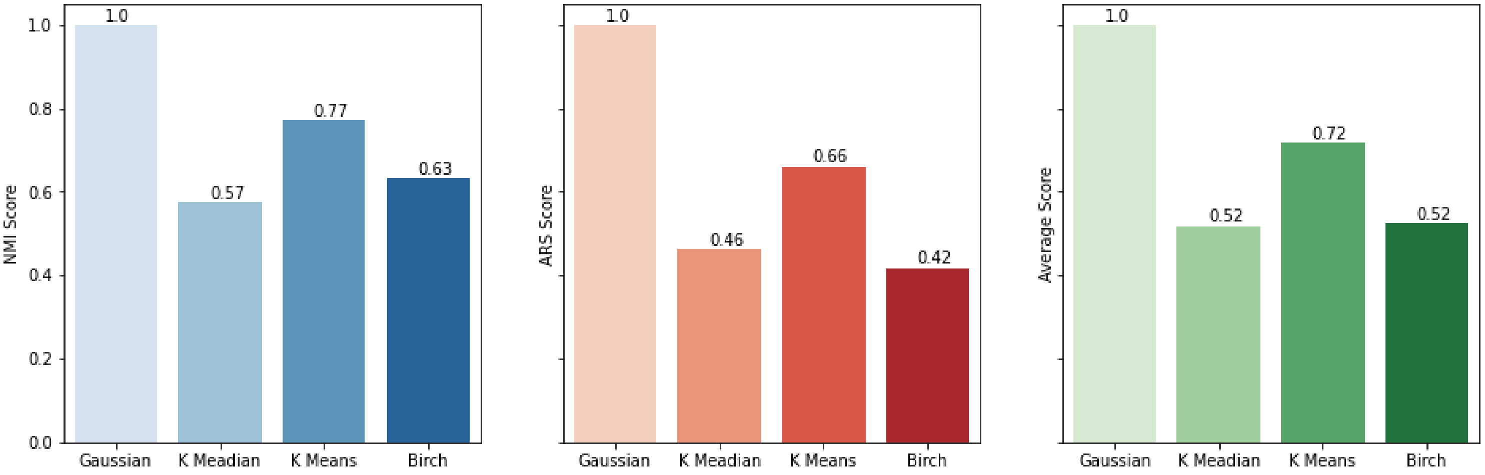
```
5    0.003600

Percentage of true labels in Cluster  6
Total Instances :  34425
2    0.737603
1    0.173914
5    0.059230
7    0.029252

Total True Label Count :
1    148288
2    198310
3     25028
4      1923
5      6645
6     12157
7     14357
```

PART 4: Comparing the cluster formation of the gaussian based method with the other three clustering

Assumption - We are using Adjusted Rand Score (ARS) and the Normalized Mutual Information (NMI) metrics for comparing gaussian predicted labels with other clustering predicted labels.

Adjusted Rand Score (ARS) - It computes a measure of similarity between two clusters. In the predicted and true clusters, ARS considers all pairings of samples and counts pairs that are assigned to the same or different clusters.

Normalized Mutual Information (NMI) - It's a measure of how dependent the two variables are on each other. NMI is a normalization of the Mutual Information (MI) score to scale the results between 0 (no mutual information) and 1 (perfect correlation). To put it another way, 0 denotes dissimilarity, and 1 denotes a perfect match.



These scores are calculated as compared to cluster formation by the Gaussian Mixture model. So, the score for the Gaussian model is 1.

Observations:

- From the ARS scores it can be observed that K means the highest cluster similarity with the Gaussian model. Birch and K median have comparatively less similarity.
- K means cluster formation is 72%(average score) similar to Gaussian mixture. K-Median and Birch cluster formation are 52% similar to the Gaussian model.
- Visualization of clusters for K Means clustering is better as compared to other models. So, we have chosen K Means as the best clustering model for this dataset.

QUESTION 2:

Usage: Run predict() function in inference.py
Command: predict('test.csv')
Return: list of predictions

Create your own train and validation set and measure your performance against it:

Assumption -

1. Split the dataset into train and validation set in ratio 70:30 using stratified test train split which ensures equal class distribution.
2. Clustering technique used is K Means
3. Train the model on X_train and measure the performance of X_val

Performance Measure-

Balanced F1 Score for Validation data: 0.6078778490816552