Name – Akshay Sawant
NUID – 001948537
Mail ID – sawant.ak@husky.neu.edu

Question 1 :

Actual / Predicted	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
Cluster 1	0	1	4	0
Cluster 2	5	0	0	0
Cluster 3	0	5	0	0
Cluster 4	1	0	1	3

$$purity(C, \Omega) = \frac{1}{N} \sum_k \max_j |c_k \cap \omega_j|$$

a) Purity is calculated using above formula.
Thus, calculating purity for our output we get,
Purity = (1/20) * (4 + 5 + 5 + 3)
Purity = 0.85

b) Precision is calculated as
Precision = TruePositive / (TruePositive + FalsePositive)
Thus, precision for our output will be,
Precision = 3 / 15 = 0.2

c) Recall is calculated as
Recall = TruePositive / (TruePositive + FalseNegative)
Hence, recall for our output will be,
Recall = 3 / 8 = 0.375

d) F-Measure is calculated as

*F-Measure = (2 * Precision * Recall) / (Precision + Recall)*
Hence, we calculate the purity of our given output as

F-Measure = (2 * 0.2 * 0.375) / (0.2 + 0.375)
F-Measure = 0.2609

e) Using formula for NMI given in Clustering slide 2 –

Value for our output we get
ΣI = 0.4303
H(**Ω**) = 0.6021
H(C) = 0.5878
Hence, NMI = **0.7234**

Question 2)

1. For K-means, DBSCAN and EM algorithm please refer to Python (".py") files in submitted file list.
2. For scatter plot associated with K-means, DBSCAN and EM algorithm please refer to image (.png) files in submitted file list.

Following is the calculation for the purity and NMI for each algorithm and each dataset –

Algorithm	**K-Means**		**DBSCAN**		**EM**	
Measure	**Purity**	**NMI**	**Purity**	**NMI**	**Purity**	**NMI**
Dataset1	1	1	1	1	1	1
Dataset2	1	1	0.974	0.97	0.973	0.961
Dataset3	0.719	0.227	0.982	0.991	0.951	0.962

(3) Can you give the reasoning why some algorithm works better (if any) than others for each of these datasets?

Dataset1
Looking at dataset1 we see that the data is densely distributed into different clusters. Hence all of the algorithms worked correctly and predicted the required output without any errors. Thus all of the 3 algorithms can be used to cluster any data with near perfect results.

Dataset2
Looking at dataset2 we see that the data has both densely and sparsely distributed data. Running K-means algorithm on it didn't give good results. Thus we can say that K-means algorithm is not suitable for data where the distribution has irregular density among different clusters. As compare to K-means, DBSCAN and EM algorithms produced better results. DBSCAN also removed many outliers.

Dataset3
Looking at dataset3 we see that the data has a unique structure that contains two U-shaped clusters. K-means has a limitation over such data making it work less efficiently while DBSCAN giving the best results. Thus we can say that DBSCAN works well where the cluster density is reasonable and the data is in some form where other algorithms get weak. EM algorithm also produced comparable results.

Question 3)

1) **Major Steps**

- Algorithm used: Matrix data clustering using mixture model
- We wrote a program that compares the abbreviations with venues in AP_Train.txt file to give the most relevant full form.
- Created 4 cluster labels and associated a list of relevant words with each cluster label.
- Processed AP_Train.txt and extracted the publication entries of data associated with each matching conferences.
- Processed every publications associated with one conference with the list of words in all the four clusters.
- Assigned every conference a cluster label after matching the publications with the relevant words in each cluster and selecting the one that gave the maximum match.

2) We applied the above-mentioned algorithm on the training data set AP_Train.txt to cluster the conferences into 4 different clusters. Here the four cluster labels and their related class label are as given below -

Cluster 1 – Database
Cluster 2 – Data analysis and management
Cluster 3 – Machine Learning / AI / Data Mining
Cluster 4 – Information processing / Web technologies

The output after running the algorithm is obtained as below,

Conference	Ground Truth	Algorithm Output
IJCAI	3	3
AAAI	3	3
ICDE	1	2
VLDB	1	1
SIGMOD	1	1
SIGIR	4	4
ICML	3	3
NIPS	3	4
CIKM	4	4
KDD	2	3
WWW	4	4
PAKDD	2	1
PODS	1	1
ICDM	2	2
ECML	3	3

PKDD	2	1
EDBT	1	3
SDM	2	2
ECIR	4	3
WSDM	4	4

Purity = 0.65
NMI = 0.4347

3)

Why is the solution reasonable?

As we can see that the data that is given to us has all the information related to the paper, information like abstract, venue, title etc. Hence clustering the conferences into different categories can be considered as problem that resembles document clustering.

After analyzing we can see that we can extract some relevant set of words that would be helpful in classifying the conferences into different categories. Some of these categories are closely related to each other (example artificial intelligence, machine learning, data mining etc.). Thus grouping them into 4 categories we can form the class labels for our 4 clusters.

To classify the documents we extract the relevant words from the entry in AP_Train.txt and then match the set against the set of words of every cluster. The cluster that gives the maximum matching words and the maximum count is assigned as the cluster for the given document set and the whole conference. And as this approach takes the whole context into consideration the solution is reasonable.