CS145 Howework 4

Important Note: HW4 is due on **11:59 PM PT, Nov 20 (Friday, Week 7)**. Please submit through GradeScope.

Print Out Your Name and UID

Name: Ali Mirabzadeh, UID: 305179067

Before You Start

You need to first create HW4 conda environment by the given `cs145hw4.yml` file, which provides the name and necessary packages for this tasks. If you have `conda` properly installed, you may create, activate or deactivate by the following commands:

```
conda env create -f cs145hw4.yml
conda activate hw4
conda deactivate
```

OR

```
conda env create --name NAMEOFYOURCHOICE -f cs145hw4.yml
conda activate NAMEOFYOURCHOICE
conda deactivate
```

To view the list of your environments, use the following command:

```
conda env list
```

More useful information about managing environments can be found [here (https://docs.conda.io/projects/conda/en/latest/user-guide/tasks/manage-environments.html)](https://docs.conda.io/projects/conda/en/latest/user-guide/tasks/manage-environments.html).

You may also quickly review the usage of basic Python and Numpy package, if needed in coding for matrix operations.

In this notebook, you must not delete any code cells in this notebook. If you change any code outside the blocks (such as some important hyperparameters) that you are allowed to edit (between STRART/END YOUR CODE HERE), you need to highlight these changes. You may add some additional cells to help explain your results and observations.

In [1]:
```
import numpy as np
import pandas as pd
import sys
import random
import math
import matplotlib.pyplot as plt
from scipy.stats import multivariate_normal
%load_ext autoreload
%autoreload 2
```

If you can successfully run the code above, there will be no problem for environment setting.

1. Clustering Evaluation

This workbook will walk you through an example for calculating different clustering metrics.

Note: This is a "question-answer" style problem. You do not need to code anything and you are required to calculate by hand (with a scientific calculator).

Questions

Suppose we want to cluster the following 20 conferences into four areas, with ground truth label and algorithm output label shown in third and fourth column. Please evaluate the quality of the clustering algorithm according to four different metrics respectively.

ID	Conference Name	Ground Truth Label	Algorithm output Label
1	IJCAI	3	2
2	AAAI	3	2
3	ICDE	1	3
4	VLDB	1	3
5	SIGMOD	1	3
6	SIGIR	4	4
7	ICML	3	2
8	NIPS	3	2
9	CIKM	4	3
10	KDD	2	1
11	WWW	4	4
12	PAKDD	2	1
13	PODS	1	3
14	ICDM	2	1
15	ECML	3	2
16	PKDD	2	1
17	EDBT	1	2
18	SDM	2	1
19	ECIR	4	4
20	WSDM	4	4

Questions (please include intermediate steps)

1. Calculate purity.
2. Calculate precision.
3. Calculate recall.

4. Calculate F1-score.
5. Calculate normalized mutual information.

Your answer here:

Note: you can use several code cells to help you compute the results and answer the questions. Again you don't need to do any coding.

Please type your answer here!

answer 1

Cluster	Truth	Name	Cluster	Output label	Name
w_1	3	IJCAI	C_1	2	IJCAI
	3	AAAI		2	AAAI
	3	ICML		2	ICML
	3	NIPS		2	NIPS
	3	ECML		2	EDBT
w_2	1	ICDE		2	ECML
	1	SIGMOD	C_2	3	ICDE
	1	PODS		3	VLDB
	1	EDBT		3	SIGMOD
	1	VLDB		3	CIKM
w_3	2	KDD		3	PODS
	2	PAKDD	C_3	4	SIGIR
	2	ICDM		4	WWW
	2	PKDD		4	ECIR
	2	SDM		4	WSDM
w_4	4	SIGIR	C_4	1	KDD
	4	CIKM		1	PAKDD
	4	WWW		1	ICDM
	4	ECIR		1	PKDD
	4	WSDM		1	SDM

① $C_k = \{C_1, C_2, C_3, C_4\}$, $w_j = \{w_1, w_2, w_3, w_4\}$

$C_1 \to w_3$: majority 5, $C_2 \to w_1$: majority : 5

$C_3 \to w_4$: majority 4, $C_4 \to w_2$: majority : 5

$$\text{Purity} = \frac{5+5+5+4}{20} = \frac{19}{20} = \underline{\underline{0.95}}$$

answer 2, 3, 4

- Random Index (RI) = (TP+TN)/(TP+FP+FN+TN)
- F-measure: 2Precision*Recall/(Precision+Recall)
  - Precision = TP/(TP+FP)
  - Recall = TP/(TP+FN)

	Same cluster	Different clusters
Same class	TP	FN
Different classes	FP	TN

class

ID	Conference Name	Ground Truth Label	Algorithm output Label
1	IJCAI	3	2
2	AAAI	3	2
3	ICDE	1	3
4	VLDB	1	3
5	SIGMOD	1	3
6	SIGIR	4	4
7	ICML	3	2
8	NIPS	3	2
9	CIKM	4	3
10	KDD	2	1
11	WWW	4	4
12	PAKDD	2	1
13	PODS	1	3
14	ICDM	2	1
15	ECML	3	2
16	PKDD	2	1
17	EDBT	1	2
18	SDM	2	1
19	ECIR	4	4
20	WSDM	4	4

we group by ID's and follow table above to get the below results

grouping by ID's	TP	TN	FP	FN
	34	143	7	5

(2) Precision $= \frac{34}{34+7} = \frac{34}{41} = \underline{\underline{0.83}}$

(3) Recall $= \frac{34}{34+5} = \frac{34}{39} = \underline{\underline{0.87}}$

(4) F1-score $= \frac{34+143}{34+143+7+5} = \frac{177}{189} = \underline{\underline{0.94}}$

answer 5

⑤

$\lvert C_j \cap w_k \rvert$	C1	C2	C3	C4	Sum
1	0	0	0	5	5
2	6	0	0	0	6
3	0	5	0	0	5
4	0	0	4	0	4
Sum	6	5	4	5	20

$\lvert w_j \rvert$

$\lvert C_j \rvert$

$$I(C,\Omega) = \frac{5}{20}\log_2\left(\frac{100}{25}\right) + \frac{6}{20}\log_2\left(\frac{120}{36}\right) + \frac{5}{20}\log_2\left(\frac{100}{25}\right) + \frac{4}{20}\log_2\left(\frac{80}{16}\right) = 0.679$$

$$H(C) = -\frac{6}{20}\log_2\frac{6}{20} - \frac{5}{20}\log_2\frac{5}{20} - \frac{5}{20}\log_2\frac{5}{20} - \frac{4}{20}\log_2\frac{4}{20} = 1.98$$

$$H(\Omega) = 4\left(-\frac{1}{4}\log_2\frac{1}{4}\right) = 2$$

$$\sqrt{H(C)H(\Omega)} = 1.98$$

$$NMI = \frac{0.679}{1.98} = \underline{0.342}$$

2. K-means

In this section, we are going to apply K-means algorithm against two datasets (dataset1.txt, dataset2.txt) with different distributions, respectively.

For each dataset, it contains 3 columns, with the format: x1 \t x2 \t cluster_label. You need to use the first two columns for clustering, and the last column for evaluation.

In [2]:
```
from hw4code.KMeans import KMeans
k = KMeans()
# As a sanity check, we print out a sample of each dataset
dataname1 = "data/dataset1.txt"
dataname2 = "data/dataset2.txt"
k.check_dataloader(dataname1)
k.check_dataloader(dataname2)
```

```
For dataset1: number of datapoints is 150
          x         y  ground_truth_cluster
0 -0.163880 -0.219869                     1
1 -0.886274 -0.356186                     1
2 -0.978910 -0.893314                     1
3 -0.658867 -0.371122                     1
4 -0.072518  0.399157                     1

For dataset2: number of datapoints is 200
          x         y  ground_truth_cluster
0  1.068587  0.136921                     1
1  0.705440  0.393068                     1
2  0.840811 -0.054906                     1
3 -0.923447  0.598501                     1
4  0.784353  0.724743                     1
```

2.1 Coding K-means

Complete the `reassignClusters` and `getCentroid` function in `KMeans.py` .

Print out each output cluster's size and centroid (x,y) for dataset1 and dataset2 respectively.

In [3]:
```
k = KMeans()
#========================#
# STRART YOUR CODE HERE  #
#========================#
k.main(dataname1)
k.kmeans()
k.main(dataname2)
k.kmeans()
#========================#
#   END YOUR CODE HERE   #
#========================#
```

```
For dataset1
Iteration :3
Cluster 0 size :50
Centroid [x=2.5737264423871222, y=-0.027462568841232965]
Cluster 1 size :50
Centroid [x=-0.46333686463472107, y=-0.46611409698195816]
Cluster 2 size :50
Centroid [x=0.988876620573686, y=2.0104789651972013]
Iteration :3
Cluster 0 size :50
Centroid [x=2.5737264423871222, y=-0.027462568841232965]
Cluster 1 size :50
Centroid [x=-0.46333686463472107, y=-0.46611409698195816]
Cluster 2 size :50
Centroid [x=0.988876620573686, y=2.0104789651972013]

For dataset2
Iteration :4
Cluster 0 size :102
Centroid [x=1.2708406269481842, y=-0.08583389704900131]
Cluster 1 size :98
Centroid [x=-0.20185935062367868, y=0.5726963240559536]
Iteration :4
Cluster 0 size :102
Centroid [x=1.2708406269481842, y=-0.08583389704900131]
Cluster 1 size :98
Centroid [x=-0.20185935062367868, y=0.5726963240559536]
```

2.2 Purity and NMI Evaluation

Complete the `compute_purity` function in `KMeans.py` .

In order to compute NMI, you need to firstly compute NMI matrix and then do the calculation. That is to complete the `getNMIMatrix` and `calcNMI` functions in `KMeans.py` .

Print out the purity and NMI for each dataset respectively.

In [3]:
```
k = KMeans()
#========================#
# STRART YOUR CODE HERE  #
#========================#
k.main(dataname1)
k.kmeans(True)
k.main(dataname2)
k.kmeans(True)
#========================#
#   END YOUR CODE HERE   #
#========================#
```

```
For dataset1
Iteration :3
Cluster 0 size :50
Centroid [x=2.5737264423871222, y=-0.027462568841232965]
Cluster 1 size :50
Centroid [x=-0.46333686463472107, y=-0.46611409698195816]
Cluster 2 size :50
Centroid [x=0.988876620573686, y=2.0104789651972013]
Iteration :3
Purity is 1.000000
NMI is 1.000000
Cluster 0 size :50
Centroid [x=2.5737264423871222, y=-0.027462568841232965]
Cluster 1 size :50
Centroid [x=-0.46333686463472107, y=-0.46611409698195816]
Cluster 2 size :50
Centroid [x=0.988876620573686, y=2.0104789651972013]

For dataset2
Iteration :4
Cluster 0 size :102
Centroid [x=1.2708406269481842, y=-0.08583389704900131]
Cluster 1 size :98
Centroid [x=-0.20185935062367868, y=0.5726963240559536]
Iteration :4
Purity is 0.760000
NMI is 0.145025
Cluster 0 size :102
Centroid [x=1.2708406269481842, y=-0.08583389704900131]
Cluster 1 size :98
Centroid [x=-0.20185935062367868, y=0.5726963240559536]
```

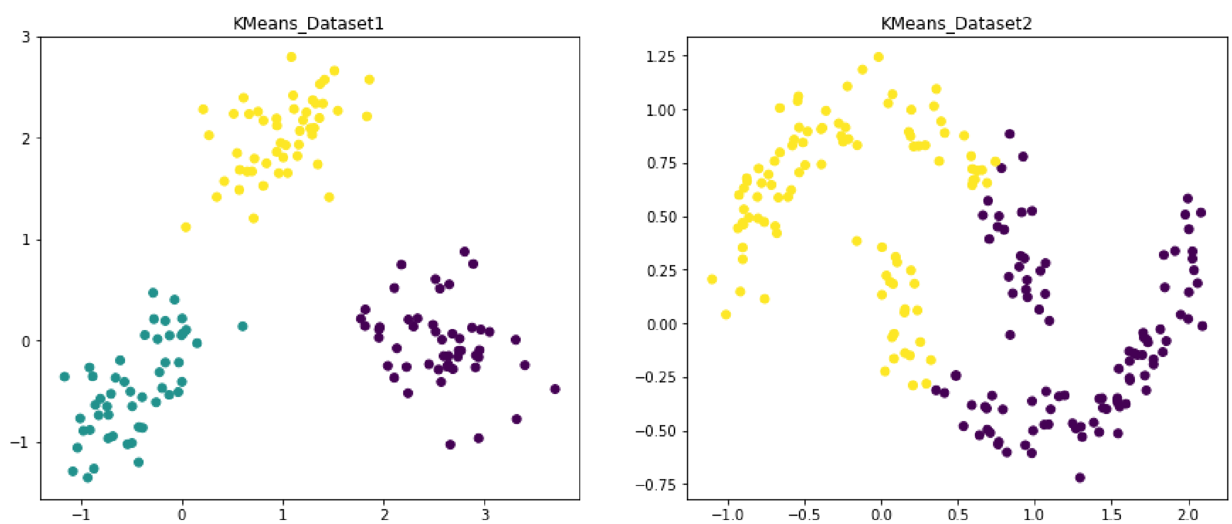
2.3 Visualization

The clustering results for KMeans are saved as `KMeans_dataset1.csv` and `KMeans_dataset2.csv` respectively under your root folder. Plot the clustering results for the two datasets, with different colors representing different clusters.

In [22]:
```
CSV_FILE_PATH1 = 'Kmeans_dataset1.csv'
CSV_FILE_PATH2 = 'Kmeans_dataset2.csv'

df1 = pd.read_csv(CSV_FILE_PATH1,header=None,names=['x','y','pred'])
df2 = pd.read_csv(CSV_FILE_PATH2,header=None,names=['x','y','pred'])
fig, [ax0,ax1] = plt.subplots(1, 2, figsize=(15, 6))
ax0.title.set_text("KMeans_Dataset1")
ax1.title.set_text("KMeans_Dataset2")

#=========================#
# STRART YOUR CODE HERE  #
#=========================#
ax0.scatter(df1.iloc[:, 0], df1.iloc[:, 1], c=df1.iloc[:, 2])
ax1.scatter(df2.iloc[:, 0], df2.iloc[:, 1], c=df2.iloc[:, 2])
#=========================#
#   END YOUR CODE HERE    #
#=========================#
plt.show()
```



Question

Give the pros and cons of K-means algorithm. (At least one for pro and two for cons to get full marks)

Your answer here

Please type your answer here!

Pros: 1. It's efficient as it has a linear run time

2. It's easy to interpret

Cons: Not suitable to discover clusters with non-convex shapes. 2 It's sensitive to noisy data

3 DBSCAN

In this section, we are going to use DBSCAN for clustering the same two datasets.

3.1 Coding DBSCAN

Complete the `dbscan` function in `DBSCAN.py`. Print out the purity, NMI and cluter size for each dataset respectively.

In [10]:
```
from hw4code.DBSCAN import DBSCAN
d = DBSCAN()
#========================#
# STRART YOUR CODE HERE  #
#========================#
d.main(dataname1)
d.main(dataname2)
#========================#
#   END YOUR CODE HERE   #
#========================#
```

```
For dataset1
Esp :0.3560832705047313
Number of clusters formed :4
Noise points :9
Purity is 0.940000
NMI is 0.959065
Cluster 0 size :49
Cluster 1 size :41
Cluster 2 size :47
Cluster 3 size :4

For dataset2
Esp :0.18652096476712493
Number of clusters formed :3
Noise points :3
Purity is 0.985000
NMI is 0.817349
Cluster 0 size :99
Cluster 1 size :51
Cluster 2 size :47
```

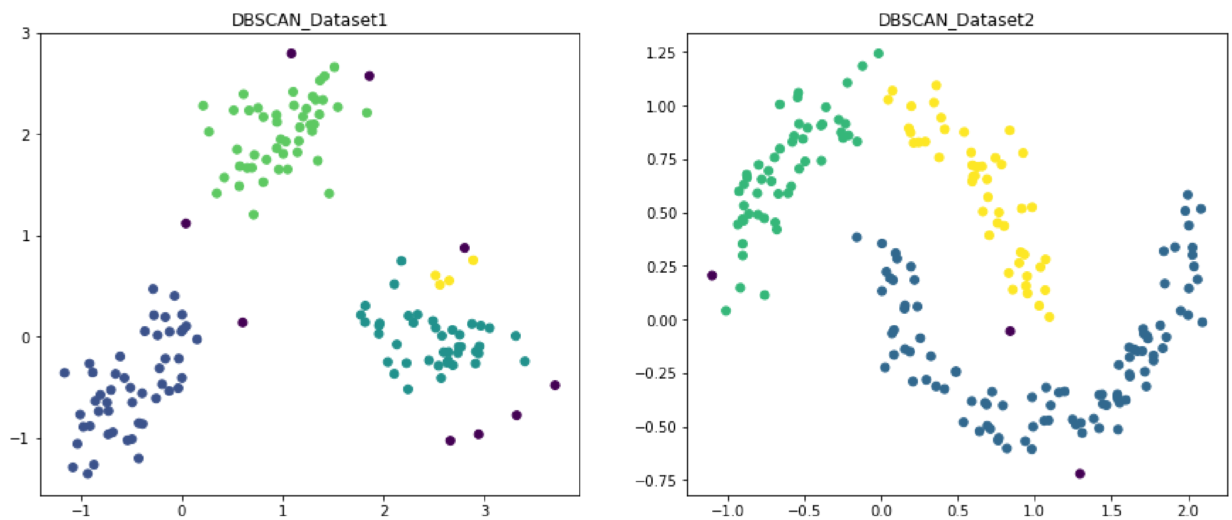
3.2 Visualization

The clustering results for DBSCAN are saved as `DBSCAN_dataset1.csv` and `DBSCAN_dataset2.csv` respectively under your root folder. Plot the clustering results for the two datasets, with different colors representing different clusters.

In [11]:
```
CSV_FILE_PATH1 = 'DBSCAN_dataset1.csv'
CSV_FILE_PATH2 = 'DBSCAN_dataset2.csv'

df1 = pd.read_csv(CSV_FILE_PATH1,header=None,names=['x','y','pred'])
df2 = pd.read_csv(CSV_FILE_PATH2,header=None,names=['x','y','pred'])
fig, [ax0,ax1] = plt.subplots(1, 2, figsize=(15, 6))
ax0.title.set_text("DBSCAN_Dataset1")
ax1.title.set_text("DBSCAN_Dataset2")

#========================#
# STRART YOUR CODE HERE  #
#========================#
ax0.scatter(df1.iloc[:, 0], df1.iloc[:, 1], c=df1.iloc[:, 2])
ax1.scatter(df2.iloc[:, 0], df2.iloc[:, 1], c=df2.iloc[:, 2])
#========================#
#   END YOUR CODE HERE   #
#========================#
plt.show()
```



Question

Give the pros and cons of DBSCAN algorithm. (At least two for pro and one for cons to get full marks)

Your answer here

Please type your answer here!

Pros: It can find clust with aribitarely shapes. 2 It's robust to outliers

Cons: Its outputs relies on its parameters and it's difficult to find the optimal parameters

4 GMM

In this section, we are going to use GMM for clustering the same two datasets.

4.1 Coding GMM

Complete the `Estep` and 'Mstep' function in `GMM.py` . Print out the purity, NMI, final mean, covariance and cluter size for each dataset respectively.

In [12]:
```
from hw4code.GMM import GMM
g = GMM()
#========================#
# STRART YOUR CODE HERE  #
#========================#
g.main(dataname1)
g.main(dataname2)
#========================#
#   END YOUR CODE HERE   #
#========================#
```

```
For Cluster : 1
0.7692790765358335
-0.28782809642382123

-0.28782809642382123
0.1901249384356512


For Cluster : 2
0.6828574757628689
-0.30058915994390517

-0.30058915994390517
0.17583559485120062


Purity is 0.690000
NMI is 0.075948
Cluster 0 size :106
Cluster 1 size :94
```

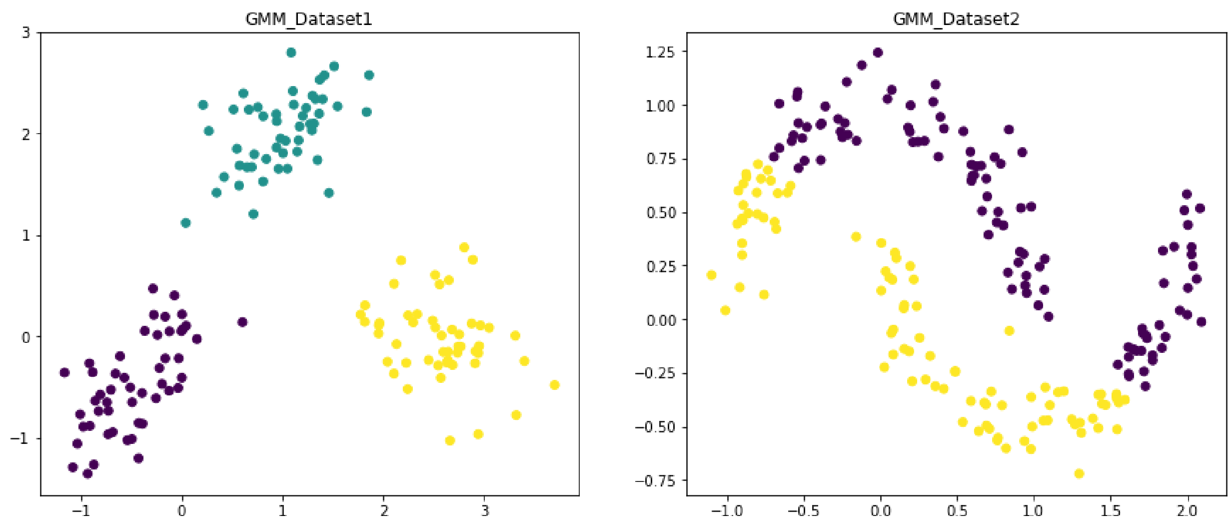
4.2 Visualization

The clustering results for GMM are saved as `GMM_dataset1.csv` and `GMM_dataset2.csv` respectively under your root folder. Plot the clustering results for the two datasets, with different colors representing different clusters.

In [13]:
```
CSV_FILE_PATH1 = 'GMM_dataset1.csv'
CSV_FILE_PATH2 = 'GMM_dataset2.csv'

df1 = pd.read_csv(CSV_FILE_PATH1,header=None,names=['x','y','pred'])
df2 = pd.read_csv(CSV_FILE_PATH2,header=None,names=['x','y','pred'])
fig, [ax0,ax1] = plt.subplots(1, 2, figsize=(15, 6))
ax0.title.set_text("GMM_Dataset1")
ax1.title.set_text("GMM_Dataset2")

#========================#
# STRART YOUR CODE HERE  #
#========================#
ax0.scatter(df1.iloc[:, 0], df1.iloc[:, 1], c=df1.iloc[:, 2])
ax1.scatter(df2.iloc[:, 0], df2.iloc[:, 1], c=df2.iloc[:, 2])
#========================#
#   END YOUR CODE HERE   #
#========================#
plt.show()
```



Questions

1. Give the pros and cons of GMM algorithm. (At least two for pro and two for cons to get full marks)
2. Compare the visualization results from three algorithms, analyze for each dataset why these algorithms would produce such result.

Your answer here:

Please type your answer here!

Pros of GMM: 1. GMM models are more general than partitioning: different densities and sizes of clusters. 2. Clusters can be characterized by a small number of parameters

Cons of GMM: 1. Converge to local optimal. 2. Hard to estimate the number of clusters

Reasoning over dataset1: Kmeans and GMM resulted in 1,0 purity whereras DBSCAN in 0.94. And we can perfectly see in the visualization as there are three perfect clusters. However, for DBSCAN we can see there are some points identified as Noise and there is a fourth cluster as well. Maybe by tunning DBSCAN parameters we can get a better purity and cluster creation for DBSCAN

Reasoning over dataset2: DBSCAN gets the best purity, 0.98, then Kmeans with 0.88 and lastly GMM with 0.69 as the lowest one. I think DBSCAN performed the best because clusters are dense and seperabalem Keamns is still perform relatively good and the reason it's not as good. Lastly we can see why GMM is not performing well due to its constraint on performing on non-convex shapes

5 Bonus Question

Prove that KMeans algorithm would guarantee covergence. (**Hint: prove for each step the loss would descrease.**)

Please type your answer here!

End of Homework 4 :)

After you've finished the homework, please print out the entire `ipynb` notebook and four `py` files into one PDF file. Make sure you include the output of code cells and answers for questions. Prepare submit it to GradeScope. Also this time remember assign the pages to the questions on GradeScope