May 16th, 2018

Lab 3 Report

Devon Martin and Shiv Sulkar

Study Design

All algorithms were implemented in Python 3.5.2 using built-in modules, as well as two external packages, numpy and matplotlib. The external packages were used for data visualization purposes only.

The approach to the K-Means algorithm was to create a Cluster object that would hold all of the data points, as well as the centroid location. For our initial centroid selection, we chose to take the initial points directly from the dataset, and have two options on the index: either completely random, or evenly distributed (when deterministic centroids are needed).

Our approach to DBSCAN was to have a Cluster object that could expand given a dataset. For example, once a Cluster has been created with an initial point, that cluster will recursively expand until it runs out of points. This way, all points in a chain will be covered, and outliers will be determined at the end of the algorithm.

Results - Description

See end of document for raw program output.

The following lists each dataset with the type of algorithm that created the most meaningful clusters and how many clusters it found.

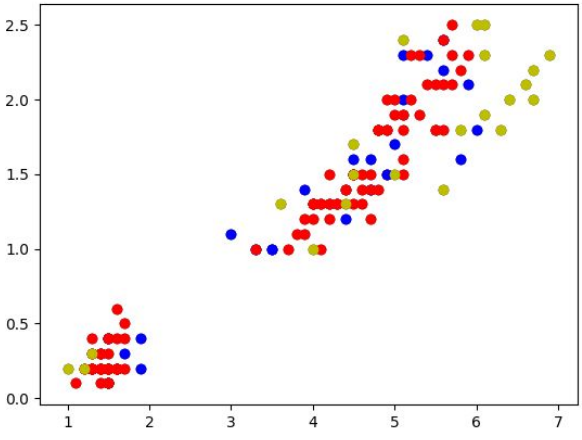
Dataset	**Algorithm**	**Clusters**
4 Clusters	K-Means	4
Accidents 1	K-Means	4
Accidents 2	Hierarchical	3
Accidents 3	DBSCAN	2
Economy	Hierarchical	3
Iris	K-Means	3
Mammal Milk	K-Means	3

Many Clusters	K-Means	5
Planets	DBSCAN	1

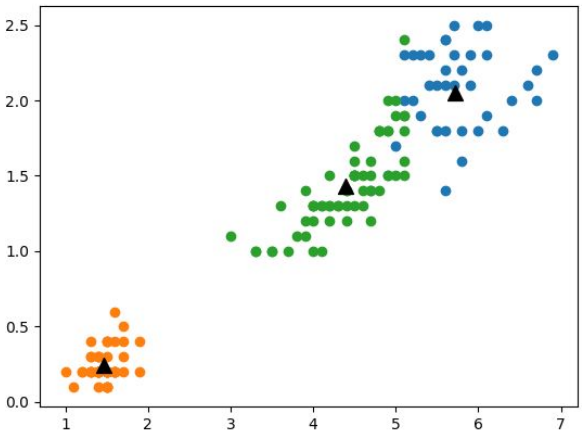
Visualizations

Iris Dataset: data[2] vs data[3]

DBSCAN e=0.4, n=3
Red: Core, Blue: Border, Yellow: Outlier

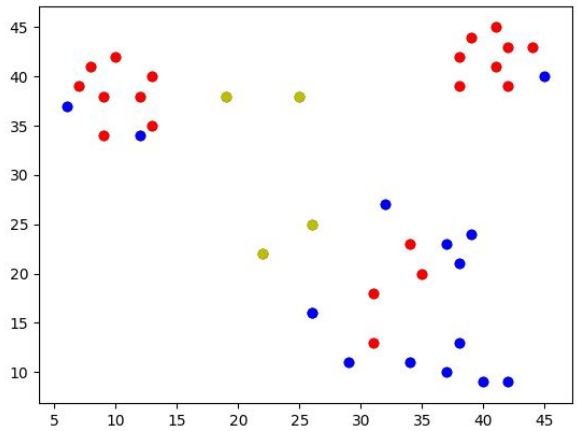


K-Means k=3
Colors: Clusters, Black: Centroids

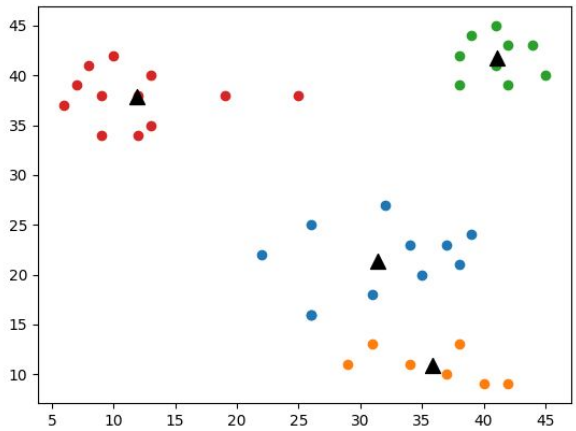


4 Clusters Dataset: data[0] vs data[1]

DBSCAN e=6.0, n=5
Red: Core, Blue: Border, Yellow: Outlier



K-Means k=4
Colors: Clusters, Black: Centroids

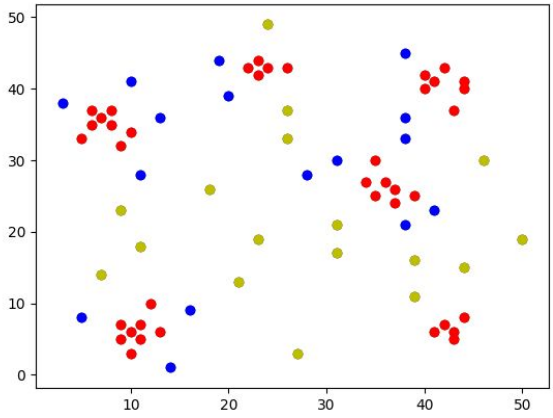


Many Clusters Dataset: data[0] vs data[1]

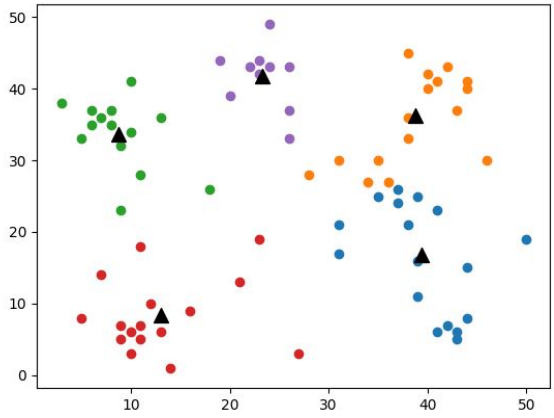
DBSCAN e=5.0, n=4

K-Means k=5

Red: Core, Blue: Border, Yellow: Outlier



Colors: Clusters, Black: Centroids



Discussion

For the Iris dataset, I thought it was interesting that K-Means so accurately fit the data for k = 3, but was disastrous with any other value of k. This is a prime example of the disadvantage to using K-Means, that the number of expected clusters must be known ahead of time. If one did not know to expect 3 species in the Iris dataset, it would be difficult to find meaningful clusters for any other value of k.

In addition, it also shows an example of DBSCAN failing. The Iris dataset has a large area of tightly grouped that the algorithm groups into one. It was difficult to fine tune the epsilon value to get it to work with this dataset.

Analysis

In datasets where the data points were there was minimal clustering to start with, K-Means seemed to provide the most insight into how the data might be grouped. However, the clusterings can also be very wrong since our implementation starts with randomly selected centroids.

In datasets with a large number of outliers, as expected, DBSCAN out-performed the other algorithms. DBSCAN is able to find outliers because it develops clusters through chains of connected points, and isn't forced to group every point into a cluster.

An unexpected discovery of working with DBSCAN is that it doesn't seem to work well on data that has a large number of features. Aside from the issues with graphing data like this, it is difficult to reason about how DBSCAN is working through the data and creating these clusters of N-spheres, and it seems the distance metric begins to break down.

Results - Data

4clusters.csv

K-Means

Command:
python3 kmeans.py datasets/4clusters.csv 4 -d

Output (points not shown:
Cluster 0:
Center: [11.916666666666666, 37.833333333333336],
Max Dist. to Center: 13.08439486139458
Min Dist. to Center: 0.18633899812498061
Avg Dist. to Center: 4.836295539238219
12 Points:

Cluster 1:
Center: [33.666666666666664, 12.0],
Max Dist. to Center: 8.856886837057617
Min Dist. to Center: 1.0540925533894605
Avg Dist. to Center: 5.57431903554341
9 Points:

Cluster 2:
Center: [41.111111111111114, 41.77777777777778],
Max Dist. to Center: 4.275973645531965
Min Dist. to Center: 0.7856742013183874
Avg Dist. to Center: 2.9117006199139848
9 Points:

Cluster 3:
Center: [32.666666666666664, 22.555555555555557],
Max Dist. to Center: 10.681124461080227
Min Dist. to Center: 1.4054567378526148
Avg Dist. to Center: 5.37782404030705
9 Points:

Hierarchical

Command:
python3 kmeans.py datasets/4clusters.csv 4

Output (points not shown:

DBSCAN

Command:
python3 dbscan.py datasets/4clusters.csv 6 5 -i 0 1

Output (points not shown):
Cluster 0:
Max Dist. to Center: 3.1622776601683795
Min Dist. to Center: 3.1622776601683795
Avg Dist. to Center: 3.1622776601683795
8 Core Points:
1 Reachable Points:

Cluster 1:
Max Dist. to Center: 2.23606797749979
Min Dist. to Center: 1.4142135623730951
Avg Dist. to Center: 1.8251407699364424
8 Core Points:
2 Reachable Points:

Cluster 2:
Max Dist. to Center: 11.704699910719626
Min Dist. to Center: 2.8284271247461903
Avg Dist. to Center: 5.710394361920049
4 Core Points:
11 Reachable Points:

4 Outliers:

AccidentsSet01.csv

K-Means

Command:
python3 kmeans.py datasets/AccidentsSet01.csv 4 -d

Output (points not shown):
Cluster 0:
Center: [5.0, 13.666666666666666, 1.0],
Max Dist. to Center: 1.666666666666666
Min Dist. to Center: 0.3333333333333339
Avg Dist. to Center: 1.1111111111111114
3 Points:

Cluster 1:
Center: [5.0, 8.857142857142858, 1.0],
Max Dist. to Center: 2.1428571428571423
Min Dist. to Center: 0.14285714285714235

Avg Dist. to Center: 1.0204081632653061
7 Points:

Cluster 2:
Center: [2.0, 19.0, 2.0],
Max Dist. to Center: 0.0
Min Dist. to Center: 0.0
Avg Dist. to Center: 0.0
1 Points:

Cluster 3:
Center: [2.0, 3.125, 2.0],
Max Dist. to Center: 2.7414640249326636
Min Dist. to Center: 0.875
Avg Dist. to Center: 1.6467460440475654
8 Points:

Hierarchical

DBSCAN

Command:
python3 dbscan.py datasets/AccidentsSet01.csv 3 6 -i 1 2

Output (points not shown):
Cluster 0:
Max Dist. to Center: 5.0
Min Dist. to Center: 1.0
Avg Dist. to Center: 2.5
2 Core Points:
6 Reachable Points:

Cluster 1:
Max Dist. to Center: 3.1622776601683795
Min Dist. to Center: 1.0
Avg Dist. to Center: 2.008104529590176
1 Core Points:
6 Reachable Points:

1 Outliers:

AccidentsSet02.csv

K-Means

Command:

python3 kmeans.py datasets/AccidentsSet02.csv 4 -d

Output (points not shown):
Cluster 0:
Center: [1.8333333333333333, 4.916666666666667, 0.5, 3.5833333333333335, 35.0, 1.3333333333333333, 0.4166666666666667],
Max Dist. to Center: 4.803789013777446
Min Dist. to Center: 1.2555432644432805
Avg Dist. to Center: 2.1969137983357245
11 Points:

Cluster 1:
Center: [1.4166666666666667, 3.0833333333333335, 0.9166666666666666, 4.0, 45.0, 1.1666666666666667, 0.4166666666666667],
Max Dist. to Center: 4.801620096962644
Min Dist. to Center: 0.6236095644623237
Avg Dist. to Center: 1.7548938881153708
12 Points:

Cluster 2:
Center: [1.3, 1.9, 0.5, 7.6, 35.0, 1.0, 0.0],
Max Dist. to Center: 2.628687885618983
Min Dist. to Center: 0.8426149773176357
Avg Dist. to Center: 1.7785811670095348
11 Points:

Cluster 3:
Center: [1.7333333333333334, 4.533333333333333, 0.0, 2.0, 70.0, 1.2666666666666666, 0.3333333333333333],
Max Dist. to Center: 20.607549857057514
Min Dist. to Center: 1.41578403877302
Avg Dist. to Center: 4.447912126242111
15 Points:

Hierarchical

DBSCAN

Command:
python3 dbscan.py datasets/AccidentsSet02.csv 3 6

Output (points not shown):
Cluster 0:
Max Dist. to Center: 2.6457513110645907
Min Dist. to Center: 1.0
Avg Dist. to Center: 1.970772949497902
11 Core Points:

7 Reachable Points:

Cluster 1:
No distance statistics, all core points
7 Core Points:
0 Reachable Points:

Cluster 2:
Max Dist. to Center: 2.8284271247461903
Min Dist. to Center: 2.8284271247461903
Avg Dist. to Center: 2.8284271247461903
6 Core Points:
1 Reachable Points:

8 Outliers:

AccidentsSet03.csv

K-Means

Command:
python3 kmeans.py datasets/AccidentsSet03.csv 4 -d

Output (points not shown):
Cluster 0:
Center: [4.4, 0.2, 2.8, 1.0, 0.4],
Max Dist. to Center: 5.761944116355173
Min Dist. to Center: 1.0
Avg Dist. to Center: 1.914768451084391
10 Points:

Cluster 1:
Center: [2.625, 0.0, 4.0, 1.375, 0.875],
Max Dist. to Center: 3.1299960063872287
Min Dist. to Center: 0.5448623679425842
Avg Dist. to Center: 1.2220870780613895
8 Points:

Cluster 2:
Center: [1.1904761904761905, 0.3333333333333333, 2.0, 1.0952380952380953, 0.8095238095238095],
Max Dist. to Center: 1.5893637398360605
Min Dist. to Center: 0.4390259265377565
Avg Dist. to Center: 0.8147566715779944
21 Points:

Cluster 3:
Center: [1.0, 0.9565217391304348, 4.0, 1.0434782608695652, 0.6956521739130435],

Max Dist. to Center: 1.670945987006577
Min Dist. to Center: 0.3104968881975152
Avg Dist. to Center: 0.6907879021782558
23 Points:

Hierarchical

DBSCAN

Command:
python3 dbscan.py datasets/AccidentsSet03.csv 2 6

Output (points not shown):
Cluster 0:
Max Dist. to Center: 1.0
Min Dist. to Center: 1.0
Avg Dist. to Center: 1.0
13 Core Points:
1 Reachable Points:

Cluster 1:
Max Dist. to Center: 1.4142135623730951
Min Dist. to Center: 1.0
Avg Dist. to Center: 1.1380711874576983
10 Core Points:
3 Reachable Points:

2 Outliers:

economy.csv

K-Means

Command:
python3 kmeans.py datasets/economy.csv 6 -d

Output (points not shown):
Cluster 0:
Center: [10.5, 9.625, 9.125, 9.875, 9.75, 11.5, 12.75, 13.25, 12.0, 12.0],
Max Dist. to Center: 4.044672421840859
Min Dist. to Center: 1.0532687216470449
Avg Dist. to Center: 2.77596466960322
9 Points:

Cluster 1:
Center: [9.0, 7.0, 6.0, 8.2, 8.2, 11.2, 13.0, 14.0, 12.4, 12.4],
Max Dist. to Center: 3.2310988842807022

Min Dist. to Center: 1.6852299546352711
Avg Dist. to Center: 2.340330817745085
4 Points:

Cluster 2:
Center: [11.5, 9.5, 8.5, 8.5, 8.5, 9.5, 9.5, 10.5, 8.5, 9.5],
Max Dist. to Center: 2.1213203435596424
Min Dist. to Center: 2.1213203435596424
Avg Dist. to Center: 2.1213203435596424
2 Points:

Cluster 3:
Center: [8.5, 8.0, 8.5, 8.5, 8.5, 10.0, 11.5, 15.0, 12.0, 11.5],
Max Dist. to Center: 2.345207879911715
Min Dist. to Center: 2.345207879911715
Avg Dist. to Center: 2.345207879911715
2 Points:

Cluster 4:
Center: [8.333333333333334, 5.666666666666667, 5.0, 5.666666666666667, 7.0, 9.333333333333334, 10.333333333333334, 10.0, 8.333333333333334, 10.666666666666666],
Max Dist. to Center: 5.011098792790969
Min Dist. to Center: 2.185812841434
Avg Dist. to Center: 3.493288078274557
3 Points:

Cluster 5:
Center: [13.5, 12.75, 12.0, 13.25, 13.5, 14.5, 16.75, 16.5, 14.25, 14.75],
Max Dist. to Center: 6.841966091702004
Min Dist. to Center: 3.848701079585163
Avg Dist. to Center: 5.500356854734989
4 Points:

Hierarchical

DBSCAN

Command:
python3 dbscan.py datasets/economy.csv 5 5

Output (points not shown):
Cluster 0:
Max Dist. to Center: 5.291502622129181
Min Dist. to Center: 2.8284271247461903
Avg Dist. to Center: 3.866744617729778
12 Core Points:
5 Reachable Points:

7 Outliers:

iris.csv

K-Means

Command:
python3 kmeans.py datasets/iris.csv 3 -d

Output (points not shown):
Cluster 0:
Center: [5.005999999999999, 3.4180000000000006, 1.464, 0.2439999999999999],
Max Dist. to Center: 1.2393514432960495
Min Dist. to Center: 0.05993329625508687
Avg Dist. to Center: 0.4841322496689401
50 Points:

Cluster 1:
Center: [5.901612903225806, 2.748387096774194, 4.393548387096776, 1.4338709677419357],
Max Dist. to Center: 1.6606403363591353
Min Dist. to Center: 0.21993519052961508
Avg Dist. to Center: 0.738152369268767
62 Points:

Cluster 2:
Center: [6.8500000000000005, 3.073684210526315, 5.742105263157894, 2.0710526315789473],
Max Dist. to Center: 1.5297103812210706
Min Dist. to Center: 0.259580953592793
Avg Dist. to Center: 0.7198385488470929
38 Points:

Hierarchical

DBSCAN

Command:
python3 dbscan.py datasets/iris.csv .5 4

Output (points not shown):
Cluster 0:
Max Dist. to Center: 0.45825756949558394
Min Dist. to Center: 0.33166247903553986
Avg Dist. to Center: 0.39069643465982223
43 Core Points:

4 Reachable Points:

Cluster 1:
Max Dist. to Center: 0.4898979485566356
Min Dist. to Center: 0.2449489742783171
Avg Dist. to Center: 0.38721191404487154
71 Core Points:
12 Reachable Points:

17 Outliers:

mammal_milk.csv

K-Means

Command:
python3 kmeans.py datasets/mammal_milk.csv 3 -d

Output (points not shown):
Cluster 0:
Center: [88.50000000000001, 2.57, 2.8, 5.68, 0.485],
Max Dist. to Center: 3.4881979588320338
Min Dist. to Center: 0.876541499302801
Avg Dist. to Center: 2.3077424120996444
10 Points:

Cluster 1:
Center: [81.18571428571428, 7.428571428571429, 6.9, 4.014285714285714, 0.9314285714285715],
Max Dist. to Center: 12.893461719105572
Min Dist. to Center: 1.528070011217969
Avg Dist. to Center: 3.9481151691395446
8 Points:

Cluster 2:
Center: [62.662499999999994, 9.7, 22.675, 2.3, 1.17],
Max Dist. to Center: 25.363693170553844
Min Dist. to Center: 3.049086953499361
Avg Dist. to Center: 11.969425961596945
7 Points:

Hierarchical

DBSCAN

Command:
python3 dbscan.py datasets/mammal_milk.csv 4 5

Output (points not shown):
Cluster 0:
Max Dist. to Center: 2.7981422408448107
Min Dist. to Center: 1.2257650672131268
Avg Dist. to Center: 1.9413024360193132
7 Core Points:
3 Reachable Points:

Cluster 1:
Max Dist. to Center: 3.3301651610693432
Min Dist. to Center: 1.2043255373859696
Avg Dist. to Center: 2.2419837995121417
3 Core Points:
3 Reachable Points:

9 Outliers:

many_clusters.csv

K-Means

Command:
python3 kmeans.py datasets/many_clusters.csv 5

Output (points not shown):
Cluster 0:
Center: [39.44444444444444, 16.77777777777778],
Max Dist. to Center: 12.302764920592365
Min Dist. to Center: 0.8958064164776165
Avg Dist. to Center: 8.25974666309695
17 Points:

Cluster 1:
Center: [38.8, 36.2],
Max Dist. to Center: 13.560235986147143
Min Dist. to Center: 0.82462112512353
Avg Dist. to Center: 7.1091604068973915
16 Points:

Cluster 2:
Center: [8.785714285714286, 33.642857142857146],
Max Dist. to Center: 11.971479713494581
Min Dist. to Center: 1.2657175104763807
Avg Dist. to Center: 5.028802899674039
14 Points:

Cluster 3:

Center: [13.0625, 8.375],
Max Dist. to Center: 14.9380230033964
Min Dist. to Center: 1.9415280708761333
Avg Dist. to Center: 6.604920976608036
16 Points:

Cluster 4:
Center: [23.3, 41.7],
Max Dist. to Center: 9.109335870413389
Min Dist. to Center: 0.424264068711927
Avg Dist. to Center: 4.005879958716357
10 Points:

Hierarchical

DBSCAN

Command:
python3 dbscan.py datasets/many_clusters.csv 5 4

Output (points not shown):
Cluster 0:
Max Dist. to Center: 4.242640687119285
Min Dist. to Center: 3.1622776601683795
Avg Dist. to Center: 3.702459173643832
5 Core Points:
2 Reachable Points:

Cluster 1:
Max Dist. to Center: 6.082762530298219
Min Dist. to Center: 2.8284271247461903
Avg Dist. to Center: 4.056256461827949
14 Core Points:
7 Reachable Points:

Cluster 2:
Max Dist. to Center: 4.47213595499958
Min Dist. to Center: 3.1622776601683795
Avg Dist. to Center: 3.928025211407882
8 Core Points:
4 Reachable Points:

Cluster 3:
Max Dist. to Center: 4.47213595499958
Min Dist. to Center: 4.123105625617661
Avg Dist. to Center: 4.239449068744967
8 Core Points:

3 Reachable Points:

Cluster 4:
No distance statistics, all core points
5 Core Points:
0 Reachable Points:

17 Outliers:

planets.csv

K-Means

Command:
python3 kmeans.py datasets/planets.csv 2

Output (points not shown):
Cluster 0:
Center: [68.55650000000001, 5.67325, 2.7604875],
Max Dist. to Center: 46.614386549649616
Min Dist. to Center: 1.4277812508070726
Avg Dist. to Center: 17.893269456769133
9 Points:

Cluster 1:
Center: [177.1534545454545, 6.510999999999999, 2.580972727272727],
Max Dist. to Center: 162.75239931551658
Min Dist. to Center: 13.527363455673838
Avg Dist. to Center: 62.242115104110056
10 Points:

Hierarchical

DBSCAN

Command:
python3 dbscan.py datasets/planets.csv 30 4

Output (points not shown):
Cluster 0:
Max Dist. to Center: 22.405955112201752
Min Dist. to Center: 11.555102768906904
Avg Dist. to Center: 16.98052894055433
13 Core Points:
2 Reachable Points:

3 Outliers: