Spatial Data Mining: Homework 2 (Due 10/07 at 11:59PM)

Use blue color to write your answers and submit on ELMS.

**Writing Problems**

1. Concepts:

1.1 (20 points)

(1) Which of the following is true?

A. Spatial autocorrelation refers to the correlation between random variables belonging to different types of observations (e.g., temperature and precipitation) at a single location

B. Spatial autocorrelation refers to the correlation between random variables belonging to a single type of observation (e.g., temperature) at multiple locations True

C. Spatial autocorrelation refers to the correlation between random variables belonging to different types of observations (e.g., temperature and precipitation) at multiple locations

(2) Which of the following statistical assumption is violated by spatial autocorrelation?

A. Independence assumption, i.e., data samples are independent to each other True

B. Identical distribution assumption, i.e., data follows a single distribution

C. Both of above

(3) Fill in the following table with choices A to F (datasets D, E and F all have a 4x4 layout. Consider each cell in D, E or F as a spatial object, and its neighbors are determined by the Rook neighborhood, i.e., cells sharing boundaries with the cell are neighbors with 1 in the W-matrix):

|  |  |  |
| --- | --- | --- |
|  | Moran’s I > 0 | Moran’s I < 0 |
| Corresponding Geary’s C value | A | C |
| Corresponding dataset | E, F | D |

Hint: For the three datasets, you do not need to calculate the actual Moran’s I values, and should be able to tell the right choice by looking at the dataset pattern. Select all that apply.

A. Geary’s C close to 0; B. Geary’s C < 0; C. Geary’s C much greater than 1

D. E. F.

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | 9 | 1 | 6 |
| 9 | 0 | 8 | 3 |
| 2 | 6 | 0 | 7 |
| 7 | 1 | 9 | 2 |

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | 1 | 0 | 0 |
| 1 | 1 | 0 | 0 |
| 1 | 1 | 0 | 0 |
| 1 | 1 | 0 | 0 |

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | 2 | 3 | 4 |
| 2 | 3 | 4 | 3 |
| 3 | 4 | 3 | 2 |
| 4 | 3 | 2 | 1 |

(4) Which of the following are true to k-means but not to DBSCAN?

A. Requiring the number of clusters True

B. Requiring clusters to have similar sizes (e.g., similar radii if cluster shapes are circular) True

C. Requiring clusters to have regular shapes (e.g., circles) True

D. Requiring clusters to have similar densities

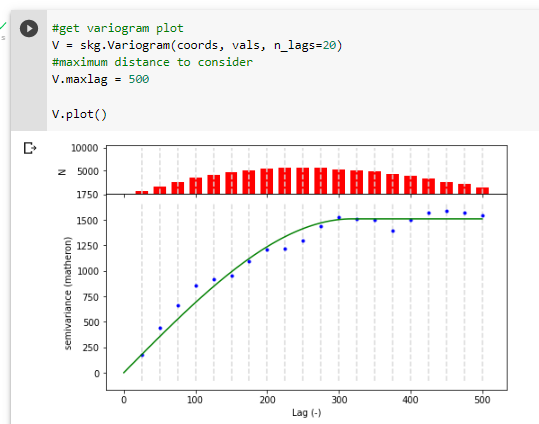
**Practice/Programming (80 points)**

Using Google CoLab, load the two iPython notebooks (.ipynb) provided in ELMS for HW2. Modify the code as demoed in class to answer the following questions. Copy related output here.

1. (20 points) Variogram. Data and functions are in the shared notebook file.

(1) After visualizing the data points, what do you feel is the maximum distance under which there is spatial autocorrelation? (hint: see at what distance you start to see the difference between values no longer increases, e.g., some pairs of points start getting similar values) more like 300. This is because, ideally, sample locations separated by distance closer than the range are spatially autocorrelated. And the range here is 300.

(2) Copy the output variogram plot here (set n\_lags = 20). Visually estimate the sill and range values and put the numbers here.



Sill = 1500,

Range = 300

(3) The n\_lags parameter refers to the number of bins we use to group distances. Why do we need bins (hint: explained in lecture)? Change the number of n\_lags to 200 and regenerate the plot. Do you start to see some noises in the plot? What do you think is the cause? (hint: again think about why we need bins)

We need bins to \*\*\*\*

Yes. I start to see some noise in the dataset. The noise is caused by the high number of bins. The number of bins is a necessary parameter for variograms and it depends on the size of the dataset -- if one has a small dataset then they need to be careful with the number of bins because they may risk having bins with no datapoints in them so estimation becomes difficult or having very low confidence because there are too few datapoints in the bins.

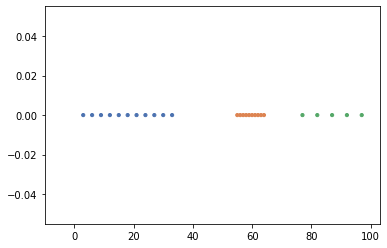
2. (60 points) Clustering. This tries to strengthen your understanding about clustering algorithms, especially how they compare to each other. In the following, you will be asked to construct a simple dataset for each question to realize its goal.

**Dataset creation instructions:** For each dataset, you will place 20 to 30 points in a one-dimensional range [0,100]. For simplicity, you can use an integer value for each point in [0,100] and the 20 points should all have different values. Put the numbers in the array that are marked in the notebook, and run the clustering algorithm on it to get result. **You should place the points in a way so that the true clusters are visually apparent.** Hint: Draw the data on a paper first to think about what results you might get before putting the numbers into the program, which can help save some efforts.

For clustering algorithms, feel free to change the model parameters yourself. Copy the visualized clustering results to the corresponding questions (function provided). **Also, briefly explain your design (e.g., why it is suitable for one approach but not the other) of the dataset for each question.**

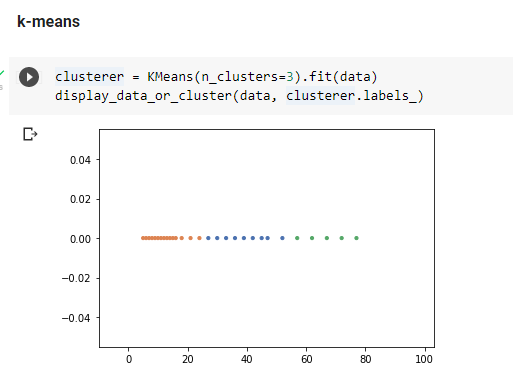
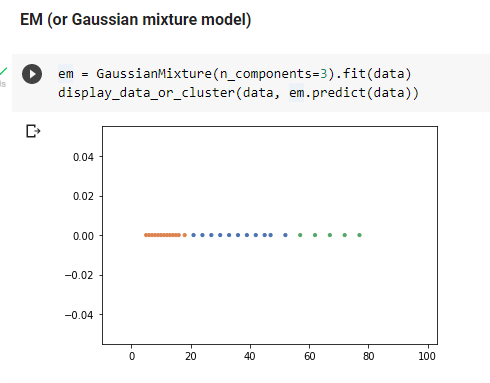
1. Create one dataset where k-means works well.

Data1 = [3,6,9,12,15,18,21,24,27,30,33, 55,56,57,58,59,60,61,62,63,64, 77,82,87,92,97]



KMeans clustering algorithm works well for this dataset scenario. As we can see that it correctly groups the three different clusters and assigns them different colors

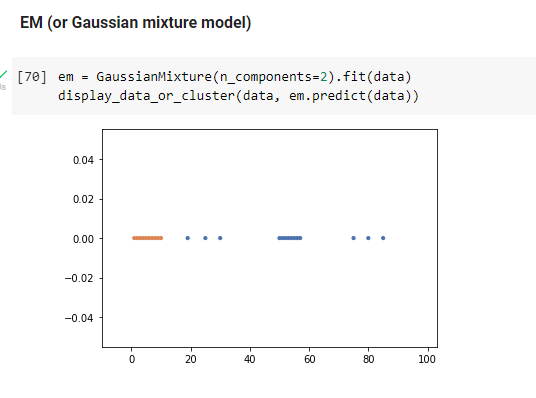
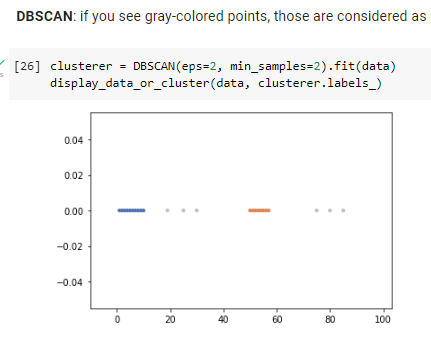
(2) Create one dataset where EM works well but k-means does not work (i.e., cannot get the right clusters).

data2 = [5,6,7,8,9,10,11,12,13,14,15,16,18, 21,24,27,30,33,36,39,42,45, 47,52,57,62,67,72,77]

EM is suitable for different radii in the clusters. As you can see in the dataset and from the figure generated, EM (figure to the right) does a better job of grouping the clusters with different radii.

Create one dataset where DBSCAN works but EM does not work well.

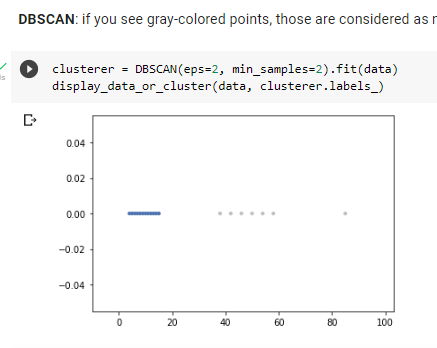
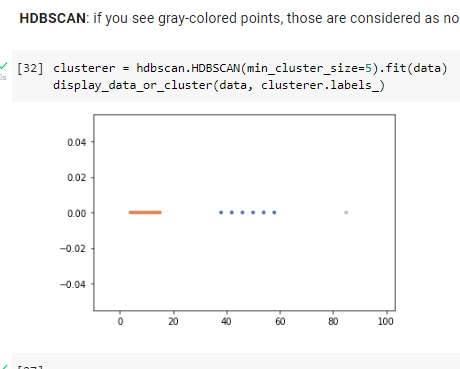
data3 = [1,2,3,4,5,6,7,8,9,10, 19,25,30, 50,51,52,53,54,55,56,57,75,80,85]



From the figures above we have just two clusters and it is obvious that EM is very susceptible to outliers and cannot detect them. On the other hand, DBSCAN correctly identifies the clusters and the outliers (grey colored points)

1. Create one dataset where HDBSCAN works but DBSCAN does not work well.

data4 = [3,4,5,6,7,8,9,10,11,12,13,14,15,38,42,46,50,54,58,85]

`

HDBSCAN shown on the right is robust to noise and different radii, but not DBSCAN, as we can see it treats everything in the second cluster as noise (to the left).