# TW9: Clustering

Xiaomei Xie: [xiaomeiX/TW9 (github.com)](https://github.com/xiaomeiX/TW9/)

Lili Hao: [lhaoSeattleu/TW9 (github.com)](https://github.com/lhaoSeattleu/TW9)

Submit Assignment

* **Due** Sunday by 11:59pm
* **Points** 10
* **Submitting** a text entry box or a file upload
* **File Types** doc, docx, txt, jpeg, png, py, and zip

**Learning objectives:**

* Be able to understand clustering models: k-Means, DBSCAN and Gaussian Mixture models
* Be able to understand clustering problems and select an appropriate clustering algorithms.

**Part 0: Basic applications of clustering models.**

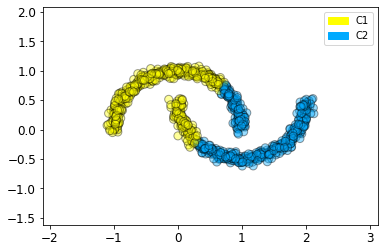
We will work together on basic applications of common clustering models.

* K-means and DBSCAN
* Download the starter notebook: [TW9-clustering.zip](https://seattleu.instructure.com/courses/1596306/files/66696255?wrap=1)
  + Use the following notebook (Part 0 is completed). Save it in TW9 folder.
    - [clustering\_basic\_part0\_completed.ipynb](https://seattleu.instructure.com/courses/1596306/files/66698860?wrap=1)

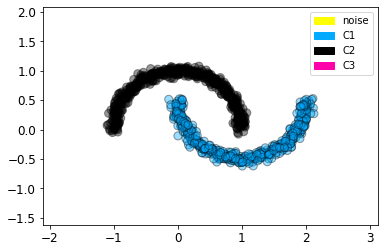
**Part 1: K-means vs. DBSCAN**

### (1) Apply the following clustering models on the generated data above. [(Links to an external site.)](http://localhost:8888/notebooks/A-CPSC4310/TW/week8/TW10/clustering_basic.ipynb#(1)-Apply-the-following-clustering-models-on-the-generated-data-above.)

* K-mean
  + Apply k-means model
  + use k = 2



* DBSCAN
  + Apply DBSCAN model: eps=0.2, min\_samples=5



(2) Apply k-means model on breast cancer dataset and check the model performance

* check also notebook, [clustering\_Kmeans.ipynb](https://seattleu.instructure.com/courses/1596306/files/66691530?wrap=1) for implementation details of k-means model

|  |  |
| --- | --- |
| KMeans(n\_clusters=2) | KMeans(n\_clusters=3) |
|  |  |

|  |
| --- |
| **DBSCAN**(eps=0.2, min\_samples=5) |
|  |

(3) Evaluate cluster models

* evaluation methods are described in the starter notebook.

|  |  |
| --- | --- |
| Accuracy of the model | Silhouette scores |
|  |  |

**Part 2: Optimal parameters of clustering models:**

The given notebooks of k-means and DBSCAN models includes examples of finding optimal parameters. Apply these techniques to find optimal parameters of K-means and DBSCAN models for the given dataset.

* K-means: [clustering\_Kmeans.ipynb](https://seattleu.instructure.com/courses/1596306/files/66691530?wrap=1)
* DBSCAN: [clustering\_DBSCAN.ipynb](https://seattleu.instructure.com/courses/1596306/files/66691529?wrap=1)

|  |  |
| --- | --- |
| **Optimal parameters of clustering models by Kmeans:** inertia\_ | **Silhouette score vs kmeans** |
|  |  |

|  |
| --- |
| **Optimal parameters of clustering models by clustering\_DBSCAN** |

*We learn* Basic applications of clustering models.

K-means and DBSCAN.

For the K-means: the parameter n\_clusters value can affect the result of the model.

* The dataset need be scaled by using MinMaxScaler() for normalized values of X.
* Evaluate Model performance by checking Accuracy of the model, which is y\_pred/ y\_real \* 100 %.
* Optimal parameters of clustering models by Kmeans: by increasing or decreasing the k value/n\_clusters; or can use **Silhouette score**

For the DBSCAN: the parameter eps and min\_samples value can affect the result of the model.

* The dataset need be scaled by using MinMaxScaler() for normalized values of X.
* Optimal parameters of clustering models by DBSCAN: by increasing or decreasing the eps value, or the min\_samples ; or can use **Silhouette score**

***Silhouette Score:*** The silhouette score is calculated utilizing the mean intra- cluster distance between points, AND the mean nearest-cluster distance. A silhouette score ranges from -1 to 1, with -1 being the worst score possible and 1 being the best score. Silhouette scores of 0 suggest overlapping clusters.

***Submission(s)***

*Each student should make individual submissions.*

* **Part 1**:
  + Push an updated notebook file to his/her/their Git repo.
    - **You do not need to submit any notebook files to Canvas**.
    - I will visit your Github to check the file.
* **Part 2**:
  + Submit a summary of your learning to Canvas. Your document should include:
    - Full names of your team members who work on the assignment.
    - URL links to the notebook of each student on GitHub repo.
    - A summary of what you learned from the teamwork assignment.