**Fourth Industrial Revolution (4IR) Summer School**

**Machine Learning** – **D**ay **1** exercises

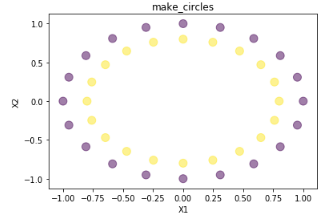
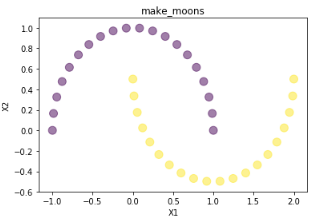
**Question 1 [Generate data using Scikit learn]**

Datasets package is a Scikit-learn feature that helps fetching commonly used benchmarked datasets used in ML (i.e. iris dataset). Moreover, Datasets-package can be used to evaluate ML models on controlled synthetic data. Such data is generated by controlling the scale of the dataset (n\_samples and n\_features) as well as the statistical properties of the data such the correlation of the features. In this exercise, we want to practice generating blobs, circles and moons synthetic datasets as shown below:

**A close up of a map

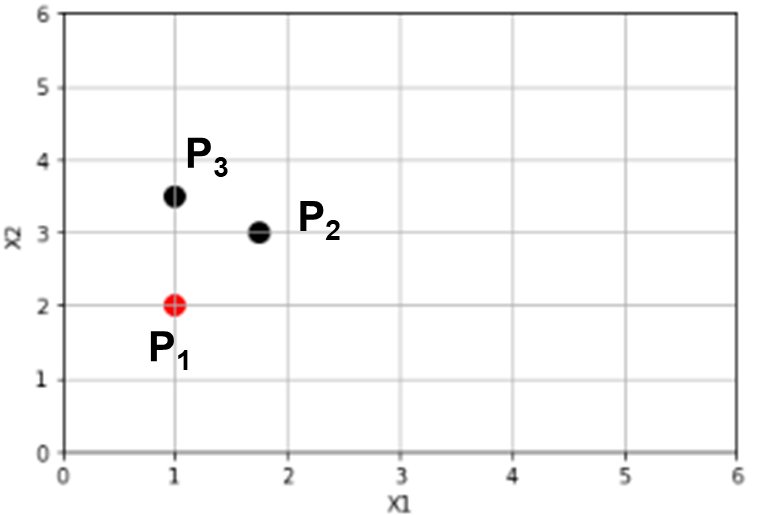
Description generated with very high confidenceA close up of a logo

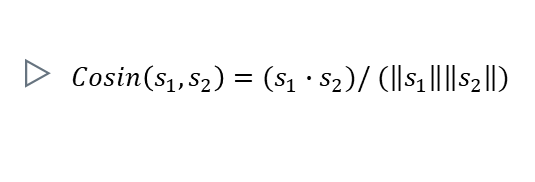
Description generated with high confidence**

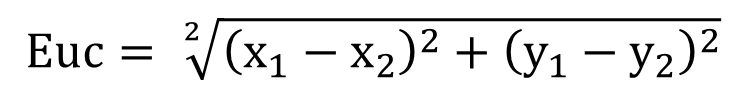
 

**Question 2 [Geometric Measurements]**

Suppose we have three points **P1=[1, 2], P2=[1.75, 3],**  and **P3=[1, 3.5],** and we want to know the nearest point to point P1**.**  To do that, one idea is to compute the distance between P1to the other two points and choose the one with the smallest distance. Another idea is to compute the similarity between P1and the other two points and select the most similar one. Write a python code to determine which of these two points P2 and P3 are the most similar to P1 and located near it. For similarity use Cosine similarity metric and for distance use Euclidean distance. Both formulas are given below







**Question 3 [Evaluation External Index]**

Generate two clusters using ***make\_blobs*** function that can be found in the following package

A screenshot of a cell phone

Description generated with high confidence

This function returns both the data samples in **X** and labels in **y.** To test Jaccard and Rand evaluation, let’s assume that we have already performed clustering on the data **X** by changing some values of y in . So, is not perfectly **y.** The following code snippet can be used to perform this task.

A screenshot of a cell phone

Description generated with very high confidence

To compute Jaccard Coefficient using the following equation:

Scikit learn uses a function called adjusted\_rand\_score() that can be used to compute the ARI

Rand =

For Rand index use a supplement “RandIndex\_Function.txt”. Open this file and copy/past the content to one cell and run it.

**Question 4 [Evaluation: Internal Index]**

Use the same created dataset X and the clustering results to compute the following metrics

* Within Sum of Squares (WSS) also called inertia
* Between Sum of Squares (BSS)
* Silhouette index

A close up of a device

Description generated with high confidenceA screenshot of a cell phone

Description generated with high confidence

**Question 5 [Perform Kmeans Clustering]**

Using KMeans from Scikit learn, create an instance of the model. Specify the following parameters:

* n\_clusters =2
* init="k-means++"

and perform clustering on the above created dataset. Finally, use scatter plot to show clustering results.

***Note: You may change n\_clusters to 3, and 4 and test the results.***

**Question 6 [Determine value for K (elbow method)]**

Elbow method is used to determine the right number of clusters for a given dataset. The turning point ( that forms an elbow shape) should be used as K values. The following Figure shows a curve with an elbow shape.

A screenshot of a cell phone

Description generated with very high confidence

Given the data used in Question 3, write a python code that performs kmeans clustering of the data for several iterations with different **K** each time. Compute the WSS (inertia) in each iteration. Then, plot the curve with number of clusters at the horizontal access, while the inertia in the vertical axis.

**Question 7 [Determine value of K using Silhouette score]**

Use make\_blobs to generate a dataset with 5 clusters. Set the parameters as follows:

* n\_samples = 1000
* n\_features =2
* centers = 5
* random\_state = 40

1. Write a python code that can loop through a list [2, 3, 4, 5, 6, 7, 8, 9]. Use each value from this list to initialize a new instance of **KMeans** estimator. Then, run the clustering and compute the silhouette score of each of them. Use a list to save both the models and silhouette scores. Print the silhouette score along with the number of clusters as shown below:

A picture containing indoor

Description generated with high confidence

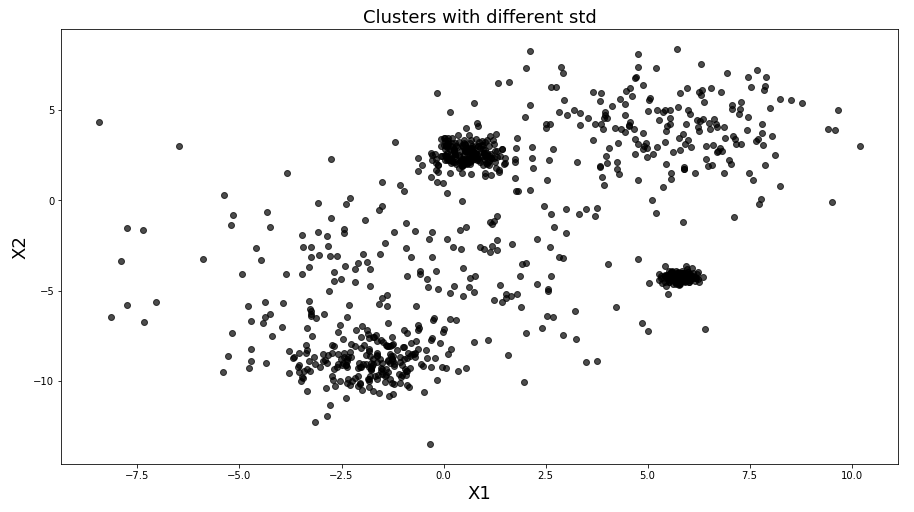
1. Determine the best clustering model and use it to predict the data labels.
2. Use scatter plot to show the clusters in both original data and our prediction

A screenshot of a cell phone

Description generated with high confidence

**Question 8[ Kmeans weakness]**

Consider the following dataset of two variables X1, and X2:



1. Create a dataset with 5 clusters, and each cluster has different standard deviation; std = {1, 0.2 ,3 ,0.5, 2}. Then, uses scatter plot to show the data
2. Perform the procedure developed in exercise 5 and 6 to determine the number of clusters in the data
3. Change number of samples per cluster. Make them differ from one cluster to another. Then, repeat clustering using Kmeans. Was it successful?
4. Generate another dataset with 2 features and 2 clusters. Use the anisotropic transformation (shown below) to make the two clusters skewed. Then, repeat clustering using Kmeans. Observe the results.

A close up of a map

Description generated with high confidence

A screenshot of a cell phone

Description generated with high confidence

1. Create circles and moons datasets. Then, use silhouette method to determine a value for K. After that, perform clustering using found K in Kmeans. Observe the results.

**Question 9 [Real dataset (Credit Card Dataset for Clustering)] (Optional)**

Examine the supplemented csv file named ([CC GENERAL](http://localhost:8888/edit/Jupyter_notebooks/data/CC%20GENERAL.csv) ), the dataset is almost clean but require some imputation for some missing values. The target of this dataset is to segment (cluster) customers into a number of groups to define marketing strategies. More details about the dataset can be found in <https://www.kaggle.com/arjunbhasin2013/ccdata>.

The data has 30 features as shown below:



* Load this data set to your colab space
* Clean the data, by filling missing values
* Remove the customer ID feature from the data before clustering
* Perform clustering and find the best possible value for K.
* Remove some features and keep others and repeat the previous step to find K

A screenshot of a cell phone

Description generated with high confidence

The above clustering results is computed using ‘Balance’ and ‘Payments’ features.