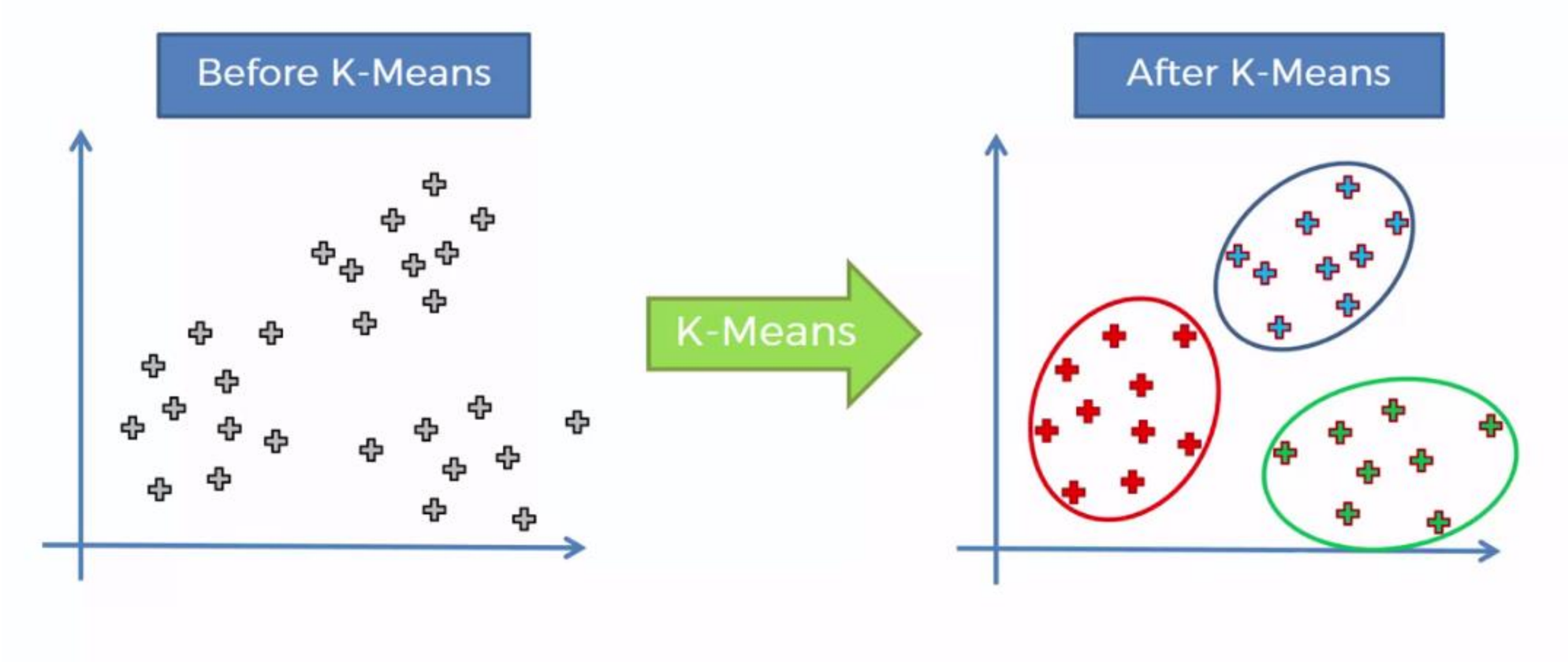
Practice 6

K-Means

Problem

- **Use K-means in mllib**
  - Use predefined function in pyspark.**mllib.clustering**



https://spark.apache.org/docs/latest/api/python/pyspark.mllib.html#pyspark.mllib.clustering.KMeans

Dataset for K-means

➢ Recognition of handwritten digits

- There are 10 handwritten digits(0~9) in bitmap format.

➢ 64 Features (pixel values)

1. Pixel 1
2. Pixel 2
…
…
…
64. Pixel 64
65. digit

❖ The last column of the data matrix indicates the class labels.

*** UCI Machine Learning Repository :**

https://archive.ics.uci.edu/ml/datasets/optical+recognition+of+handwritten+digits

➢ You can download the pre-processed dataset on iCampus

Practice 6

1. Use predefined classes in *pyspark.mllib.clustering : Kmeans()*

   Parameters for the method

   - *k=10, maxIterations=100, seed = given index*

2. Perform k-means 30 times and find the k-means model with the smallest objective

   function value. For example, like following.

```
kmeans_list = []
for i in range(30):
    kmeans_list.append(KMeans.train(trData,k=10,maxIterations=100, seed=i))
```

3. Then, with that model, calculate ***NMI score*** of the result to the test data points.

https://spark.apache.org/docs/latest/api/python/pyspark.mllib.html#pyspark.mllib.clustering.KMeans

https://spark.apache.org/docs/latest/mllib-clustering.html

Submission

1. You have to submit "**result.txt**" file on iCampus.
2. In your **result.txt** file, there must be ***NMI score*** of K-Means clustering result for digit dataset.
3. ***NMI*** means ***normalized mutual information*** which is a metric to measure some clustering results.
4. Deadline: May 21st 23:59 P.M.
5. Your result.txt file must be like following

```
NMI of K-Means clustering
0.7499
```

Windows

```
NMI of K-Means clustering
0.7499
```

Linux