**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**COMPUTER SCIENCE**

**CLASS**

**MACHINE LEARNING IN COMPUTER VISION**

**PRACTICAL EXCERCISE**

**CLUSTERING**

**Lecturer: DR. LÊ ĐÌNH DUY**

**Student: Trần Quốc Long**

**Ho Chi Minh, October 19th, 2017**

1. **What is clustering?**

**Clustering** if the process of grouping data into classes or clusters.   
The grouping is done in such a manner that the objects within the same cluster are very similar to each other but they are very dissimilar to the objects in some other cluster.

Clustering is a form of “learning by observation”. It is an unsupervised learning method and does not require a training data set to generate a model. Clustering can lead to the discovery of previously unknown groups within the data.

For example, in business intelligence, clustering can be used to organize a large number of customers into groups, where customers within a group share strong similar characteristics.

1. K-means;
2. *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).
3. *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the [Voronoi diagram](https://en.wikipedia.org/wiki/Voronoi_diagram" \o "Voronoi diagram) generated by the means.
4. The [centroid](https://en.wikipedia.org/wiki/Centroid) of each of the *k* clusters becomes the new mean.

4. Steps 2 and 3 are repeated until convergence has been reached.

1. **Clustering measurement methods:**
2. **Environments:**

* Python: ver. 3.6.2
* Jupyter notebook.

1. **Functions from scikit-learn library for clustering problem**

\* Taken examples were applied with dataset of handwriting digits from sklearn

1. K-means:

#Kmeans

nClusters = 10

kmeans\_model = KMeans(nClusters)

labels\_kmeans = kmeans\_model.fit\_predict(digits.data)

1. Spectral clustering

#Spectral\_clustering

graph = cosine\_similarity(digits.data)

labels\_spectral = spectral\_clustering(graph, n\_clusters=10)

1. DBSCAN

#DBSCAN

data = digits.data

labels\_dbscan = DBSCAN(eps=0.06, min\_samples=10,metric = 'cosine').fit\_predict(data)

1. Agglomerative

#Agglomerative Clustering

Agglomerative\_model = AgglomerativeClustering(n\_clusters = nClusters)

labels\_AgglomerativeClustering = Agglomerative\_model.fit\_predict(data)

1. Clustering measurement:

metrics.homogeneity\_score(digits.target, labels),

metrics.completeness\_score(digits.target, labels),

metrics.v\_measure\_score(digits.target, labels),

metrics.adjusted\_rand\_score(digits.target, labels),

metrics.adjusted\_mutual\_info\_score(digits.target, labels),

1. Applying clustering function on datasets
2. Brief comparison of clustering methods
3. Folder structure submitted on Github:

References:

* 1. <https://machinelearningcoban.com/2017/01/01/kmeans/>
  2. <https://www.quora.com/What-is-clustering>
  3. <https://en.wikipedia.org/wiki/K-means_clustering>