# Assignment 9

# Clustering

*The purpose of this assignment is to use R/Python to learn how to perform K-means clustering in R/Python.*

This assignment provides you with an opportunity to demonstrate the achievement of the following course learning outcomes:

* Understand and apply the Python programming language
* Understand and be able to implement K-means clustering on a given dataset

Key Information

* **Type:** *Individual*
* **Weight:** 6.25%
* **Delivery:** Course website upload
* **Due Date:** End of lab session

## Expectations

You are expected to complete this assignment individually.

Respect for academic integrity is crucial to your success. Make sure you understand what constitutes acts of academic dishonesty in the page: [What is Academic Dishonesty?](http://mcmaster.ca/academicintegrity/students/whatis.html)

## Instructions

*Using R/Python, you are to complete the following questions.* ***Please submit your answers (CODE USED AND OUTPUT) as PDF* *files to the course website submission folder.***

*Please answer following questions:*

1. *Define K-means clustering?*
2. *Following the example in the K-means Jupyter Notebook file do the following (to be done 100 times in a loop structure):*
   * *Use a random number generator to generate the centre value between 2-10*
   * *Use a random number generator to generate cluster\_std between 0.01 and 0.99*
   * *Use the make\_blob function to generate a new data set*

*from sklearn.datasets.samples\_generator import make\_blobs*

*X, y = make\_blobs(n\_samples=300, centers=”Randomly generated number between 2-10”,*

*random\_state=0, cluster\_std==”Randomly generated number between 0.01 and 0.99”)*

*plt.scatter(X[:, 0], X[:, 1], s=50);*

* + *Use a loop structure (for or while) that applies K-means with different number of means (k takes the value between 2 and 15)*
  + *Choose the k-value that has the highest silhouette score*
  + *Print the following values:*
    - *Centres and the k value with the highest silhouette score*
  + *Repeat for 100 times*

***Option:***

*Instead of doing the previous steps, write a program that performs k-means*

**Appendix**

If the ground truth labels are not known, evaluation must be performed using the model itself. The Silhouette Coefficient (**[sklearn.metrics.silhouette\_score](http://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html" \l "sklearn.metrics.silhouette_score" \o "sklearn.metrics.silhouette_score)**) is an example of such an evaluation, where a higher Silhouette Coefficient score relates to a model with better defined clusters. The Silhouette Coefficient is defined for each sample and is composed of two scores:

* **a**: The mean distance between a sample and all other points in the same class.
* **b**: The mean distance between a sample and all other points in the *next nearest cluster*.

The Silhouette Coefficient *s* for a single sample is then given as:

s = \frac{b - a}{max(a, b)}

The Silhouette Coefficient for a set of samples is given as the mean of the Silhouette Coefficient for each sample.

#### 2.3.9.5.1. Advantages

* The score is bounded between -1 for incorrect clustering and +1 for highly dense clustering. Scores around zero indicate overlapping clusters.
* The score is higher when clusters are dense and well separated, which relates to a standard concept of a cluster.

#### 2.3.9.5.2. Drawbacks

* The Silhouette Coefficient is generally higher for convex clusters than other concepts of clusters, such as density based clusters like those obtained through DBSCAN.

Note: The silhouette score is one of a few evaluation metrics that are appropriate for use with k-means clustering.

## Rubric

To achieve full marks on this assignment, you must have answered all questions above correctly with code submitted that has no errors.