CSC3022H: Machine Learning

Lab 1: K-Means Clustering

Department of Computer Science
University of Cape Town, South Africa

July 28, 2018

Due: Friday, 3 August, 2018, 10.00 AM

Problem Description

Implement (in C++) the *K-means* clustering algorithm [MacQueen, 1967] with a Euclidean distance metric. See online tutorials at:

- http://www.saedsayad.com/clustering_kmeans.htm

Use the implemented K-means algorithm to cluster the following 8 examples (table 1) into 3 clusters.

When running K-means, set the initial seeds (initial centroid of each cluster) as examples 1, 4 and 7.

Table 1: Data (examples have two attributes: X, Y, both in range: [1, 10]).

Example Number	X	Y
1	2	10
2	2	5
3	8	4
4	5	8
5	7	5
6	6	4
7	1	2
8	4	9

Question 1: How many iterations are needed for k-means to converge?

In a text file output the results of each iteration (for each cluster, list the examples that fall into each cluster), and the centroids of each cluster, **e.g.:**

Iteration 1

Cluster 1: 1, 2, 3
Centroid: (3.0, 9.5)

Cluster 2: 4, 5, 6
Centroid: (6.5, 5.25)

Cluster 3: 7, 8
Centroid: (1.5, 3.5)

...

Iteration N

Cluster 1: 8, 7, 6
Centroid: (1.5, 3.5)

Cluster 2: 5, 4, 3
Centroid: (6.5, 5.25)

Cluster 3: 2, 1
Centroid: (3.0, 9.5)

In a ZIP file, place the source code, makefile, and the output text file (answer to question 1). Upload the ZIP file to *Vula* before 10.00 AM, Friday, 3rd of August, 2018.

References

[MacQueen, 1967] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Berkeley Symposium on Mathematics, Statistics and Probability*, pages 281–297, Berkeley, USA. University of California Press.