**Machine Learning**

**Impl – 1**

1. Randomly generate 120 values of x in the range [0,1]. Let them be x1, x2, · · · , x120
2. **Clustering:**

i. Cluster the entire set of 120 points using the Leader algorithm. Choose different values for threshold θ and cluster.

ii. Let the clustering obtained using some threshold θi be Ci = {cluster1, cluster2, · · · , clusterli }.

iii. Compute the purity value for each clustering. The purity value is given by

Purity (Ci) = X li j=1 maximum (|clusterj ∩ Class1|, |clusterj ∩ Class2|).

1. **SOLUTION**

**CODE:**

|  |
| --- |
| Please find the code for Clustering Submitted as *KNNC\_Clustering\_impl\_1.py*   * Leader Clustering algorithm is implemented * Leader clustering algorithm is evaluated for different threshold values * **RESULT** captures the threshold value, number of clusters created, value of leader of each cluster and purity value as per the above equation. |

**RESULT:**

|  |
| --- |
| Threshold = 0.1  Number of Clusters = 8  Leaders of Clusters = [0.8535611253071369, 0.4761119072487371, 0.19392614571337774, 0.6454102440090569, 0.02636938907744546, 0.30204063996597785, 0.7455415408820913, 0.9741555710823887]  Purity = 112  Threshold = 0.2  Number of Clusters = 4  Leaders of Clusters = [0.2611734472855062, 0.8089884736499725, 0.5907621763440891, 0.03248354449480306]  Purity = 110  Threshold = 0.25  Number of Clusters = 3  Leaders of Clusters = [0.48524435983076986, 0.1972478763196739, 0.7485130427819631]  Purity = 111  Threshold = 0.4  Number of Clusters = 2  Leaders of Clusters = [0.23693419742187538, 0.8450886739163418]  Purity = 113  Threshold = 0.5  Number of Clusters = 2  Leaders of Clusters = [0.6841896716699505, 0.09295104149782729]  Purity = 101 |

**PLOT:**

Threshold

Threshold

**INFERENCE/ANALYSIS:**

|  |
| --- |
| * The program determines the **purity** value which is an indication of how many data points are **correctly classified**. * When the threshold is as high as 0.5 the purity is the least. Since 0.5 is the basis on which the samples are classified in 2 classes, such high threshold can result in wrong classification and hence lower purity. * When the **threshold is high**, **number of clusters are low and vice versa** * **Purity of Leader algorithm depends largely on the initial value** in the sample set |

**RESOURCES USED FOR THE ASSIGNMENT:**

|  |
| --- |
| * **Environment:**   Anaconda, Jupyter notebook |
| * **Software :**   Python  **Python libraries/modules:** Pandas, Numpy, SkLearn etc |