**TPR2251 Pattern Recognition**

**Tutorial 9**

Part 1 - Theory

1. Briefly explain the clustering technique.

Lecture 8, slide 3

1. What are the 3 fundamental steps involved in the *k*-means clustering algorithm?
2. Determine the centroid coordinate
3. Determine the distance of each object to the centroid.
4. Group the object based on minimum distance.
5. *K*-means algorithm generally goes into a number of iterations for finding clusters for the objects. What are the termination criteria for the algorithm to stop the clustering process?

Lecture 8, slide 24

1. Consider the application of the *k*-means clustering algorithm to the two dimensional data set *D* = {} for *k* = 3 clusters.
2. Start with the three centroids: m1(0) = , m2(0) =  and m3(0) = . What are the values of the centroids at the next iteration?

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | 1st | 2nd | 3rd | 4th | 5th | 6th |  |  |
| D0= | | | 2.24 | 5 | 5.39 | 12.53 | 13 | 13.9 | | | M1=(-7,4) |
|  | | | 12 | 10.44 | 9.2 | 13.6 | 7.9 | 5.1 | | | M2=(7,4) |
|  | | | 10.6 | 7.81 | 11.7 | 3.6 | 2.8 | 5.7 | | | M1=(2,-5) |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | 1st | 2nd | 3rd | 4th | 5th | 6th |  |  |
| G0= | | | 1 | 1 | 1 | 0 | 0 | 0 | | | M1=(-7,4) |
|  | | | 0 | 0 | 0 | 0 | 0 | 1 | | | M2=(7,4) |
|  | | | 0 | 0 | 0 | 1 | 1 | 1 | | | M1=(2,-5) |

Cluster 1: (-5,3), (-3,1), (-2,6)

Cluster 2: (6,-1)

Cluster 3: (-1,-7), (4,-3)

The new centroids:

M1 = (-3.3, 3.3)

M2 = (6,-1)

M3 = (1.5, -5)

1. Given the final clusters’ centroids after convergence are m1 = , m2 =  and m3 = . Cluster the points, *x*1 = and *x*1 =.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | 1st |  |  |
| D= | | | 6.3 | | | M1=(-3.3, 3) |
|  | | | 5.4 | | | M2=(5, -2) |
|  | | | 10.77 | | | M3=(-1, -7) |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | 1st |  |  |
| G= | | | 0 | | | M1=(-3.3, 3) |
|  | | | 1 | | | M2=(5, -2) |
|  | | | 0 | | | M3=(-1, -7) |

So, this point belongs to cluster 3.

Part 2 – Practical

**Objective**: You are going to implement the k-means algorithm on a synthetic dataset.

1. Generate a two-dimensional dataset containing four distinct blobs. You can use the make\_blobs module in sklearn.datasets.samples\_generator to generate 300 samples having 4 centers. The standard deviation of the clusters is 0.60. Set the random\_state to 0. Use a scatter plot to visualize the data.
2. Instantiate the KMeans class, and set the number of clusters to 4. After that, call the fit method with the data.
3. During the algorithm, each training data point in X is assigned a cluster label. You can find these labels in the kmeans.labels\_ attribute.
4. You can assign cluster labels to new points using the predict method. Each new point is assigned to the closest cluster center when predicting. For this, run predict on the training set X. What is the result returned?
5. Visualize the results by plotting the data colored by the labels. Also plot the cluster centers as determined by the k-means estimator.

Part 3 – Self-challenge

**Objective**: Now, implement the k-means algorithm on a digits recognition problem.

1. Load the digits dataset from sklearn.datasets using the load\_digits function. This dataset consist of 1,797 samples with 64 features, where each of the 64 features is the brightness of one pixel in an image.
2. Perform k-means clustering on the dataset. Note the number of clusters to use. Print the size of the cluster centers.
3. Visualize the cluster centers.
4. Because k-means knows nothing about the identity of the cluster, the 0–9 labels may be permuted. We can fix this by matching each learned cluster label with the true labels found in them.
5. Check the accuracy of the clustering algorithm by using the accuracy\_score function in sklearn.metrics.
6. Check the confusion matrix for the method by calling the confusion\_matrix function in sklearn.metrics. What are the meanings of the figures contained in the confusion matrix?