Information Retrieval CSF469

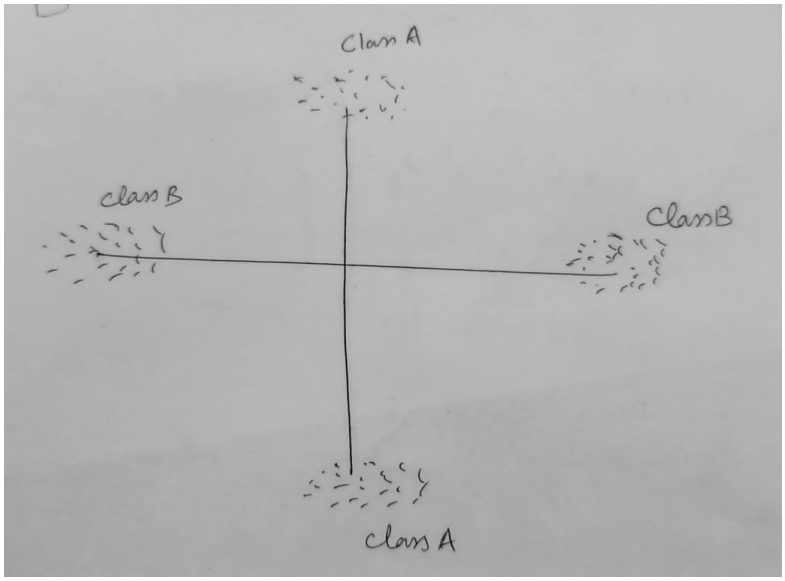
Lab Session - 6
Date - 27/03/2024
Marks: 10

Objective:
To perform classification using Rocchio and KNN classification models.

Dataset: Generate a customised dataset with 500 data points belonging to two classes.

Tasks:

1. Generate and plot the dataset.



2. Rocchio classification model:
    a. Divide the dataset into train (400 data points) and test (100 data points).
    b. Compute the centroid for each individual class using the formula.

$$\vec{\mu}(c) = \frac{1}{|D_c|} \sum_{d \in D_c} \vec{v}(d)$$

where D_c is the set of all documents that belong to class c and

$\vec{v}(d)$ is the vector space representation of d.

    c. Now, predict the class for test data by computing the distance between the target data point and centroids of each class.
    d. Report accuracy.
3. KNN classification model:
    a. Compute the distance of each datapoint in test with each datapoint in train.
    b. Sort the distances in increasing order.
    c. Choose the top-k (k=10) closest data points.

d. Assign the label of the target datapoint to the class with the maximum datapoints in the top-k closest datapoints.
e. Report accuracy.