Report

Analysing clustering results

K-means: Yes, the points belonging to the same letter form 3 different clusters depending on the initial choice of points, generally if the points are well distributed (distance wise) then the letters form distinct clusters but if the initial choice of centroids is bad (like if they are all very close to each other) then the letters sometimes don't form distinct clusters

DB-scan: Yes, the points belonging to the same letter form 3 different clusters if the choice of ε and min_samples is good. When I chose ε to be (> half the diagonal of the rectangle of the letter ~ 9.01 (height 15, width 10)) and (< 10 distance between letters) and min_samples to be 6,7 I am getting good results. if the ε >= 10 then point from different letters are intermixing so you will have to keep min_samples high, if it less than 9 then some point are not reachable from other points so you will have to keep min_samples low to get good result.

In K-means you have to choose a good set of initial centroids, in DB-scan it is a density-based clustering, since our letters are well – separated and there is no noise both methods can do it well for good parameters but K-means is simpler (less complex) to implement, if there had been some overlap or noise then DB-scans density-based clustering would distinguish better (K-means can't handle noise and irregular shape well) .

Conclusion

Both methods are able to classify the letters into different clusters for good parameters, since our letters are well-separated (10 units) dense in the rectangles (height 15, width 10) both methods are working well.

For our set of letters both methods work well but K-means is simpler (less complex) to implement. If the no of points had been lesser than 7 like 4 etc, then DB-means would have struggled (it doesn’t handle sparse graphs well) K-means would be better choice, if the distance between the letters had been closer or the letter were more weirdly shaped or had some noise then K-means would have struggled to handle it, DB-means would be a better choice.

Using Silhouette score or Davies-Bouldin index to check clustering quality while tuning parameters (ε,min-samples,k). Running K-means with multiple sets of initial centroids and selecting the best result, selecting initial centroids that are furthest from each other. These methods might help improve cluster accuracy.