CLUSTERING

Maria Ingold
12693772
Unit 5
Machine Learning
University of Essex Online
25 November 2024

CONTENTS

Clustering Overview

Grus (2016) shows *k*-means clustering consists of:

Initialise:

1) Choose *k* – this will be used to create *k* clusters
2) Randomly select *k* centroids (points to represent cluster centres).

Loop until centroid positions stop changing:

1) Calculate nearest centroid to each point and assign point to that centroid (expectation)
2) Compute new centroid (mean) based on the cluster of points assigned to old centroid (maximization)

Animation 1

Shabalin (N.D.) effectively illustrates the importance of selecting the initial k centroids. Figures 1 and 2 show starting from left-most points. Figures 2, 3, 4 and 5 show the result of each starting point. Each cluster is different although 4 (top) and 5 (random) are almost identical, but 3 shows one centroid with no cluster. These will all provide a different meaning to the data analysis.

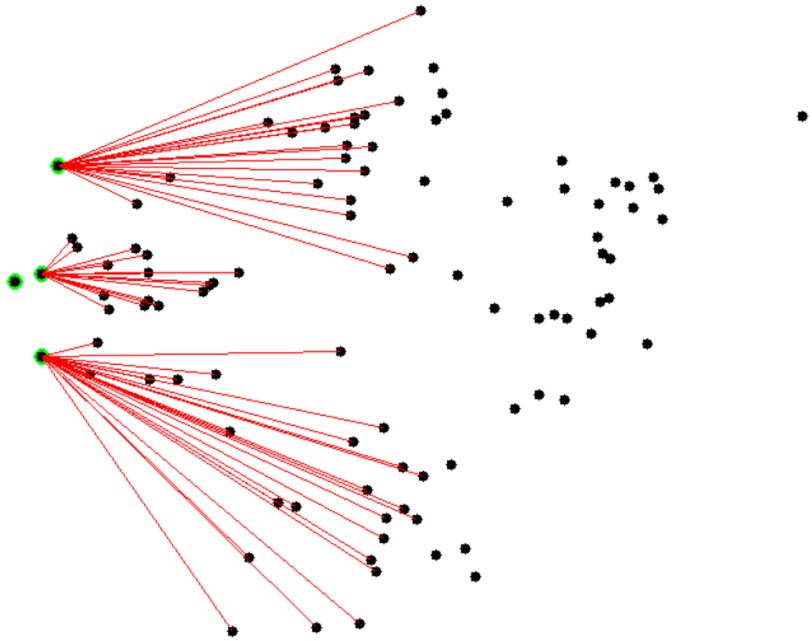

FIGURE 1 | Initialisation with centroids starting from 4 left-most points

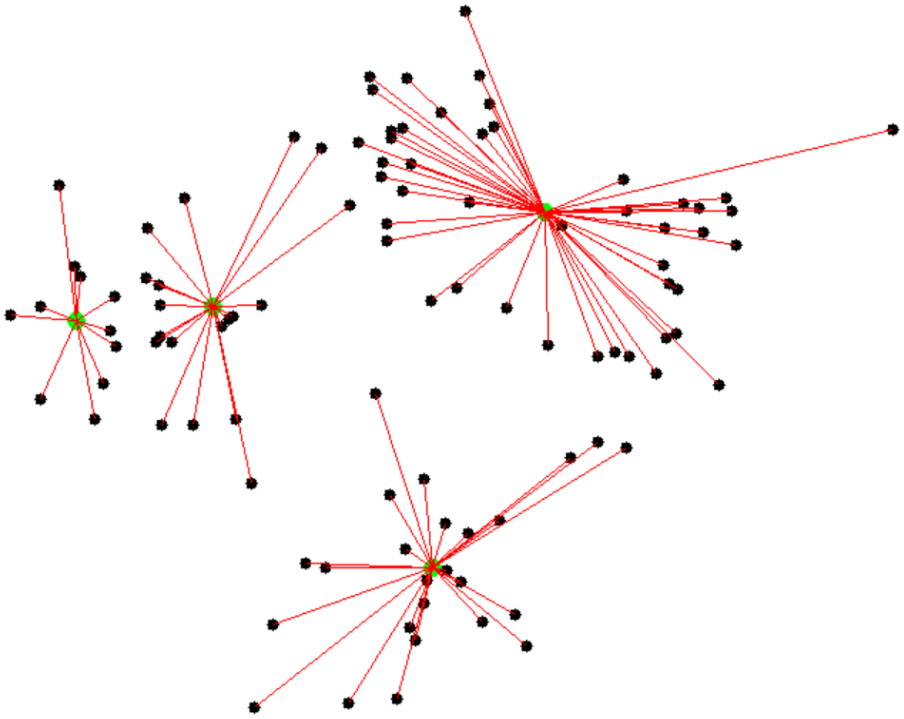

FIGURE 2 | Clusters with centroid initialisation starting from 4 left-most points

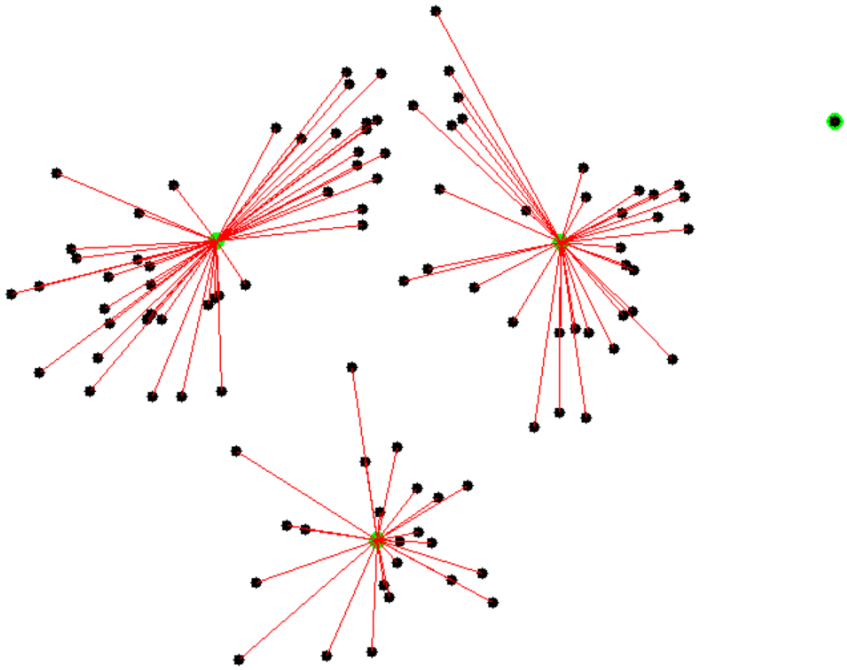

FIGURE 3 | Clusters with centroid initialisation starting from 4 right-most points.

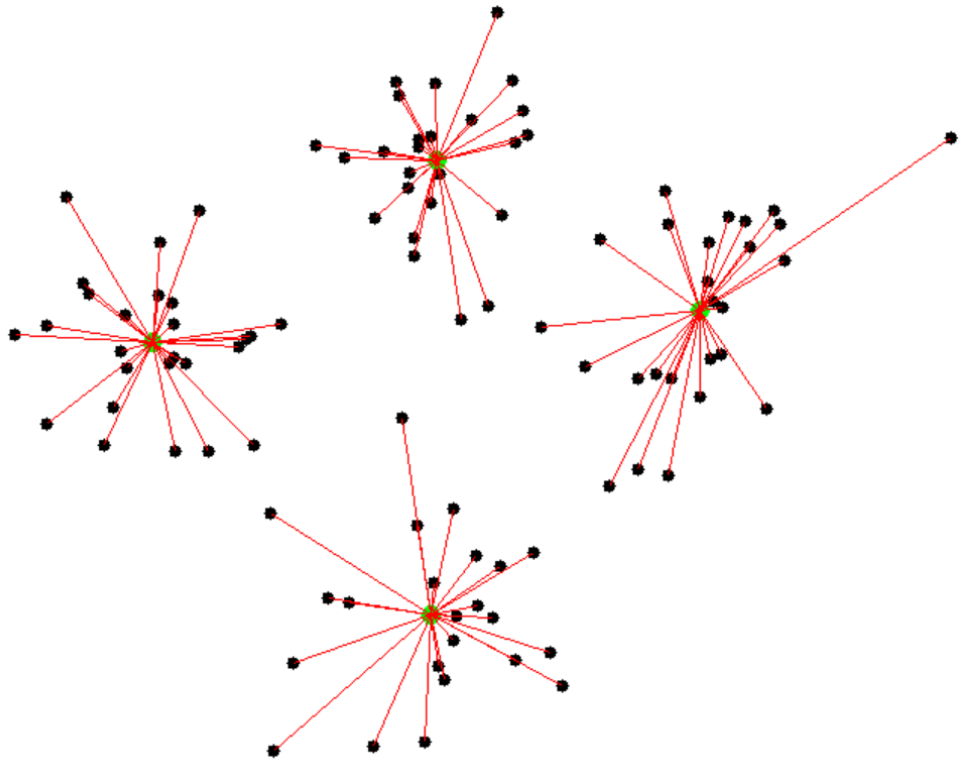

FIGURE 4 | Clusters with centroid initialisation starting from 4 top-most points.

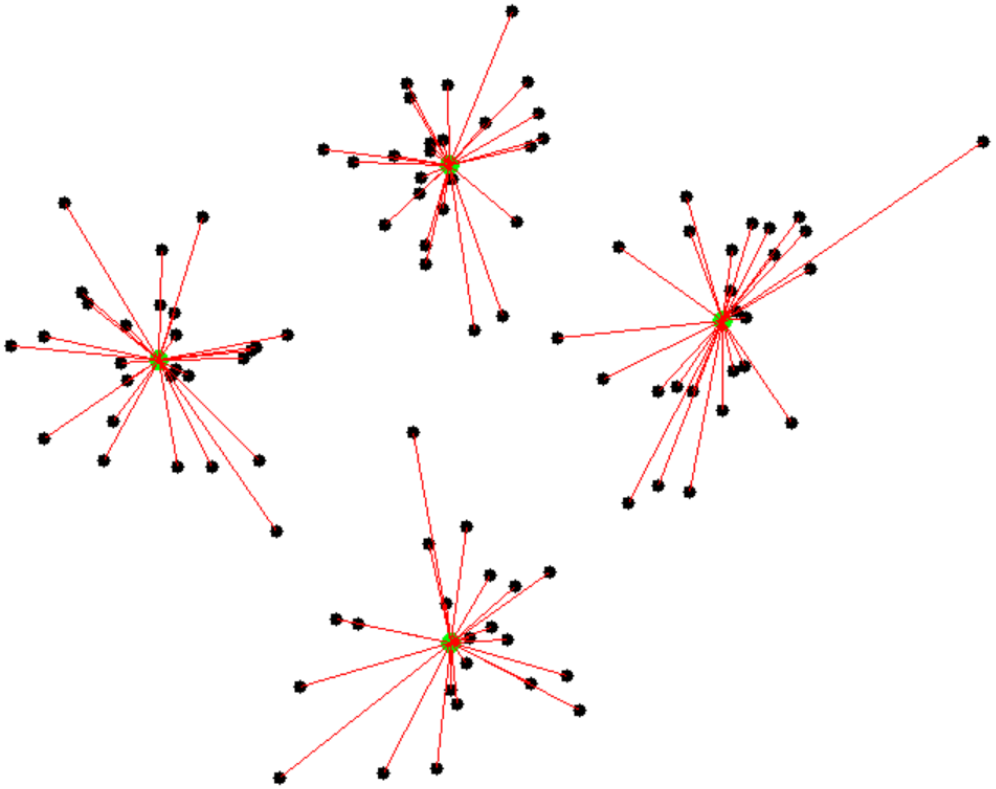

FIGURE 5 | Clusters with centroid initialisation starting from 4 random points in one cluster

Animation 2

Harris (2014) shows the full loop. Generate initial data points (Figure 6), manually select three initial centroids (Figure 7) as three device types, and then iterate to update centroids and reassign points until the centroids stop moving (Figure 8).

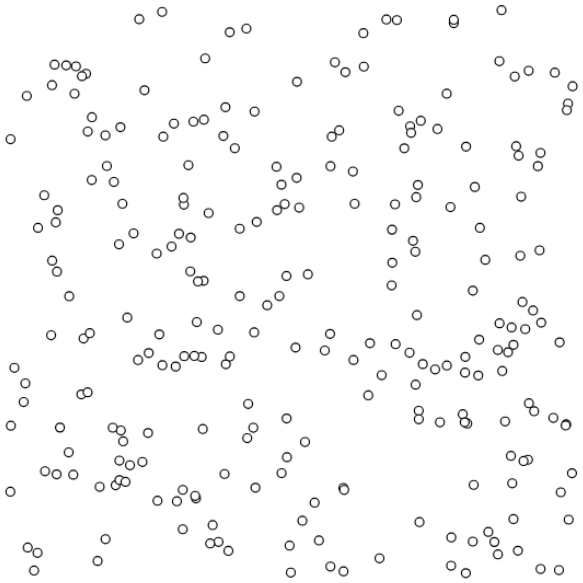

FIGURE 6 | Initial data points

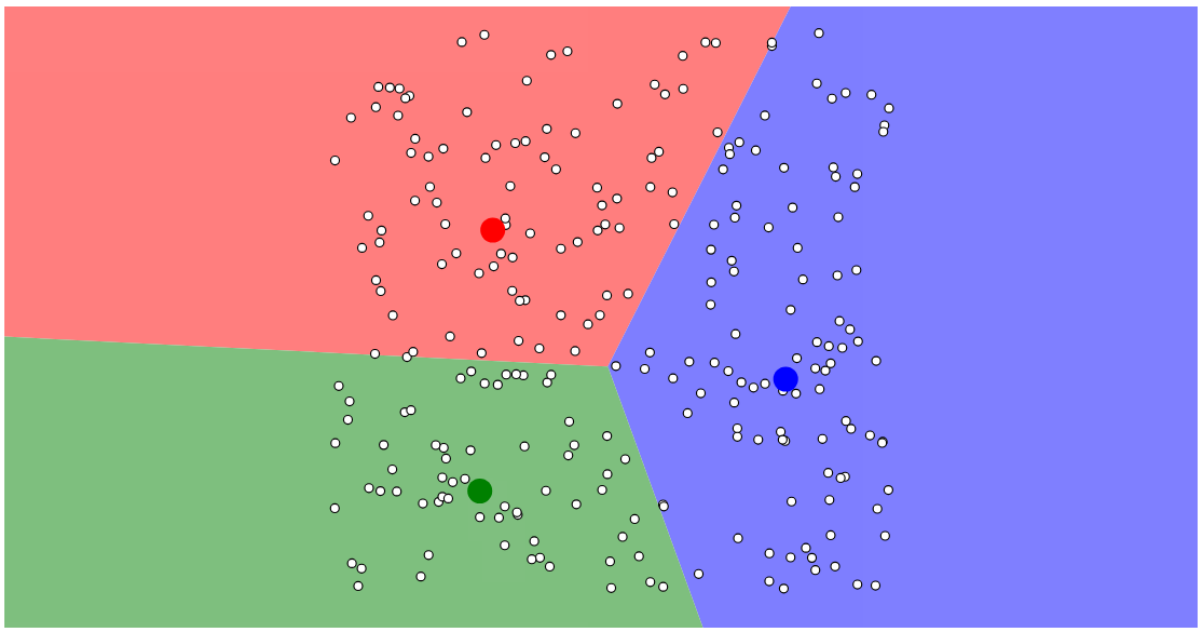

FIGURE 7 | Initial centroids

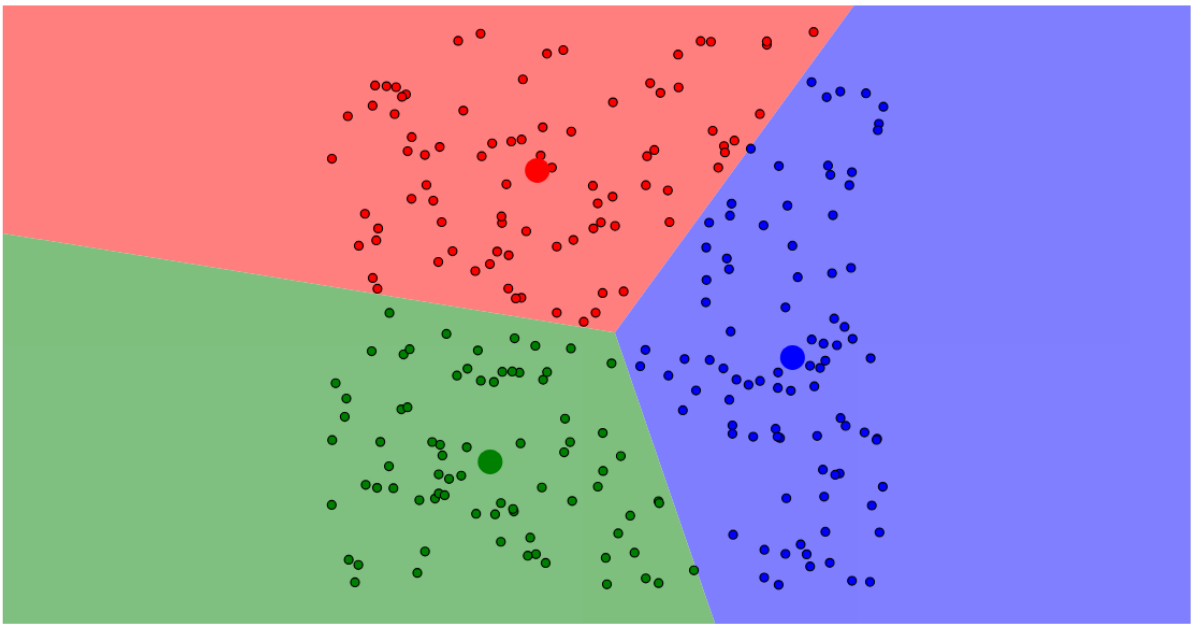

FIGURE 8 | Final centroids and clusters

Conclusion

k-means is an iterative process. In both examples, selection of *k* and initial centroids influences the outcome. Increasing *k*, created four clusters in Harris (2014), but the outset had said that only three would typically be expected, so understanding the data is essential to a more useful result.

References

Grus, J. (2016) *K-means and hierarchical clustering with Python | K-means and hierarchical clustering with Python*, *O'Reilly Media, Inc*. Available from: https://learning.oreilly.com/library/view/k-means-and-hierarchical/9781491965306/ch01.html#idm139773482897232 [Accessed 24 November 2024].

Harris, N. (2014) *Visualizing K-Means Clustering*. Available from: https://www.naftaliharris.com/blog/visualizing-k-means-clustering/ [Accessed 24 November 2024].

Shabalin, A.A. (N.D.) *Visuals and Animations: K-means clustering*. Available from: https://shabal.in/visuals.html [Accessed 24 November 2024].