Weekly Report 2 - K-Means

Ganji Varshitha
AI20BTECH11009

Introduction

K-Means is an unsupervised learning algorithm which performs partitioned clustering. Clustering helps us to understand the structure of the data by grouping it into distinct sub-groups.

Algorithm

It is a parametric method where we need to specify the number of clusters K, which we want to divide the data into.
It is a simple and iterative algorithm described as follows:

> Initialize K random points from the dataset as centroids
>
> **repeat**
>
> > Form K clusters by assigning all points to the closest centroid
> >
> > Recompute the centroid of each cluster
>
> **until** The centroids don't change

It can be seen as Expectation-Maximisation problem where the E-step is assigning the data points to the closest cluster and the M-step is computing the centroid of each cluster. It minimises the intra-cluster sum of squared distance from its centroid and keeps the clusters distant from each other.

Evaluation metric

Since it is an unsupervised learning model, we don't have ground truth values to evaluate its performance. Hence, we select the model with a K value which minimises sum of squared distances between data points and its centroids.
When SSE is plotted against K, select the K where the graph flatten out and forms an elbow. This method is referred to as elbow method.

Key points

- The algorithm clusters the data into distinct sub groups which will not work for overlapping clusters.
- Normalization is required as it deals with distances.
- It assumes spherical shapes of clusters with center as centroid and fails in case of complex designs or even elliptical shape.