Table of Contents

[CHAPTER 1: INTRODUCTION 2](#_Toc39440481)

[Problem Statements 2](#_Toc39440482)

[Research Question 2](#_Toc39440483)

[Research Objective 3](#_Toc39440484)

# CHAPTER 1: INTRODUCTION

# Problem Statements

Determining the optimal internal grouping based on unlabeled points is a huge challenge due to the ambiguity of the absolute ‘best’ criterion independent of the final goal. These types of problems are commonly known as an NP-Hard problem which involves a tradeoff of computation time versus accuracy to approximate. The goal here is to find an optimum point to achieve a sufficient accuracy in a reasonable amount of time. A classic example of an NP-Hard problem would be k-means clustering. Given a finite number of points, the k-means algorithm is tasked to find k number of centers to minimize the distance of each point to each of those centers. Logically, we would like the distance between the different centers to be as far away from each other as possible. This would uncover underlying patterns that form these clusters based on their distance. Being an unsupervised algorithm, there are no labels associated to the and therefore learn only from its features.

There are various assumptions that are required to be made when operationalizing the k-means algorithm namely, the clusters are assumed to be spherical in nature. A spherical dataset is a dataset that is roughly round in a sample space. Besides that, another assumption is that the data points are almost equally distributed among the clusters. The k-means algorithm also requires that the k which is the number of clusters to be approximated apriori.

The k-means algorithm is evaluated based on a metric called the Dunn Index. The Dunn index measures two things which are the intra-cluster distance which is the sum of distances between all the points to the centroid as well as the inter-cluster distance of the centroids. Ideally, intra-cluster distance should be low as possible creating compact clusters whereas the inter-cluster distance should be higher to show better separation between clusters. Therefore the Dunn index is defined as the ratio of the minimum of inter-cluster distances and maximum of intra-cluster distances. So we need to maximize the Dunn index and this may be contributed by centroid which maximizes the inter-cluster distance or minimizes the intra-cluster distance

The process k-means undergoes is relatively simple that is to assign a centroid in the sample space of the data and compute the distance between every single data point to the centroid. It then assigns the data to the centroid with the lower distance. Since all the data points belong to a cluster, the mean of each cluster is computed and the centroid is shifted to the new mean. The entire process is repeated until the mew mean of the centroid doesn’t change. For data with a clear separation boundary, the k-means will converge very quickly.

At this point we have identified the best cluster we have identified a good cluster but we’ll not know if there is a better clustering possible, so kmeans continues

Talk about distance metrics and when to use what

Next :methods

Assumptions

Pros and cons

Hook, what, why, how

Questions : intra and inter cluster distance. Better clustering can be obtained by trading off only?

# Research Question

# Research Objective