# K-Means Clustering for Breast Cancer Wisconsin dataset

Shashank Pathak(15BCE1287)*,
**School of Computing Sciences and Engineering, VIT Chennai, Tamilnadu, India 600127*
*Email: shashank.pathak2015@vit.ac.in**

***Abstract*—In machine learning, K-Means Clustering is a unsupervised learning model which gives a simple way to cluster the data into K clusters. The purpose of this experiment is that we will be able to categorize a set of data into 2 clusters without having the knowledge about the class for each datapoint**

## 1. Introduction

K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity Rather than defining groups before looking at the data, clustering allows you to find and analyze the groups that have formed organically. The ”Choosing K” section below describes how the number of groups can be determined.

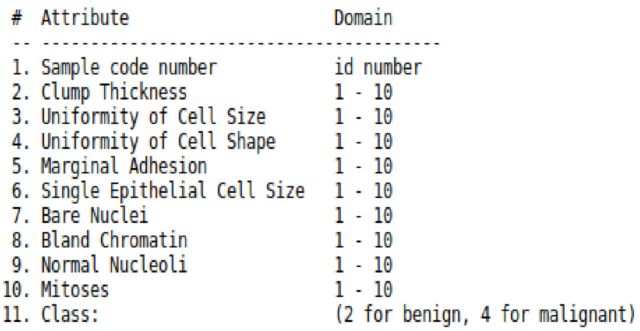
Each centroid of a cluster is a collection of feature values which define the resulting groups. Examining the centroid feature weights can be used to qualitatively interpret what kind of group each cluster represents.

## 2. Methodology

k-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other.

## 3. Database - Breast Cancer Wisconsin Dataset(Original)

The Wisconsin Breast Cancer datasets from the UCI Machine Learning Repository is used,to distinguish malignant (cancerous) from benign (non-cancerous) samples.This dataset consist of 699 instances and 11 attributes that would help the classiy the data into the two classes.The attributes are described in the fig.1



```
 #  Attribute                     Domain
 -- -----------------------------------------
 1. Sample code number            id number
 2. Clump Thickness               1 - 10
 3. Uniformity of Cell Size       1 - 10
 4. Uniformity of Cell Shape      1 - 10
 5. Marginal Adhesion             1 - 10
 6. Single Epithelial Cell Size   1 - 10
 7. Bare Nuclei                   1 - 10
 8. Bland Chromatin               1 - 10
 9. Normal Nucleoli               1 - 10
10. Mitoses                       1 - 10
11. Class:                        (2 for benign, 4 for malignant)
```

Figure 1. Breast Cancer Wisconsin Dataset Attributes

## 4. Algorithm

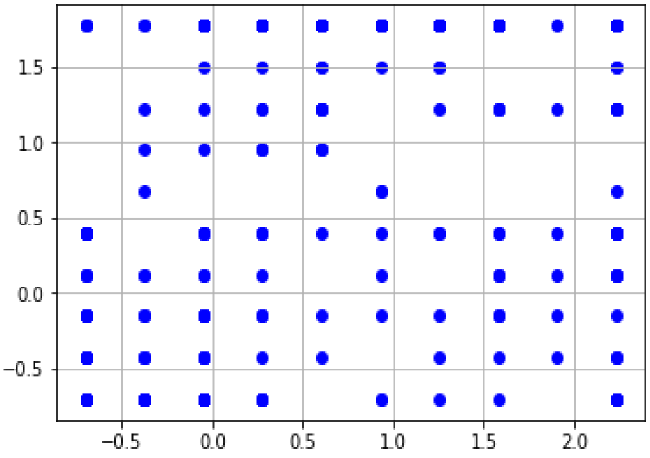
The main idea is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more. Finally, this algorithm aims at minimizing an objective function know as squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left( \left\| x_i - v_j \right\| \right)^2$$

## 5. Experiments

First we used pairplot from seaborn library to get a idea of correlation between the a reduced set of attributes of the dataset .The reduced set of attributes is namely:

’ClumpThickness’, ’UCellSize’, ’UCellShape’,’MargAdhes’, ’EpiCellSize’, ’BareNuc’, ’Chromatin’, ’Nucleoli’ and ’Mitoses’. After this I did some preprocessing i.e. scaling the data and taking the standardized values.Then I used K-mens clustering to cluster the datapoint into 2 clusters



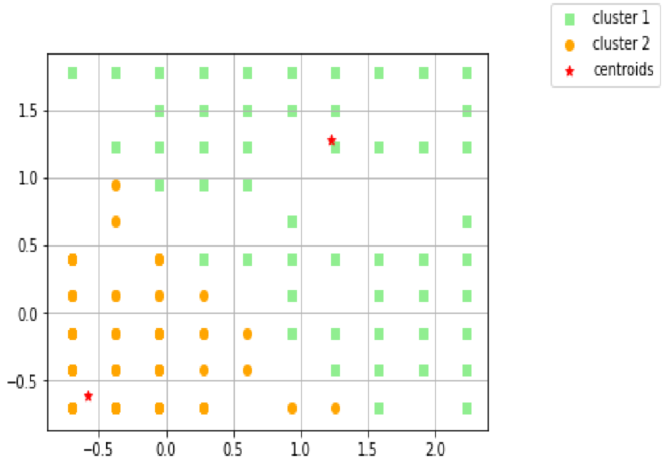


Figure 2.

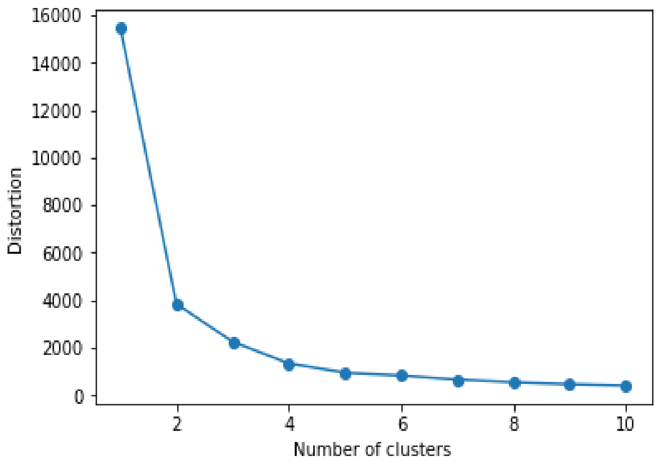


Figure 3.

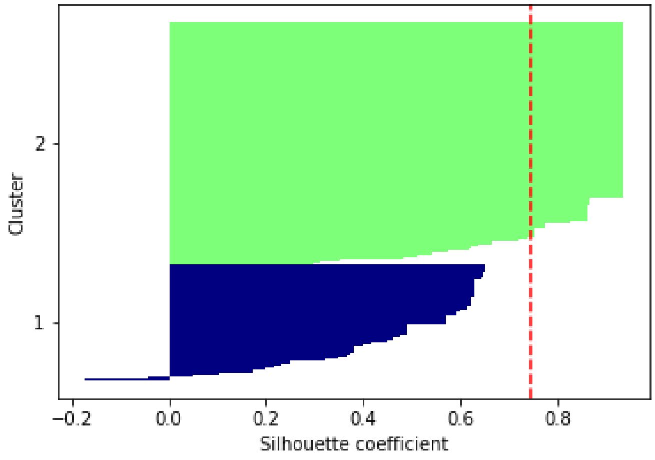


Figure 4.

## 6. Conclusion

After running K-means clustering we see that we are able to cluster the Breast Cancer Wisconsin Dataset into two categories rather than knowing what each category or cluster is we are able to form clusters.This indeed is a simple model to do unsupervised learing although we get a very high value of distortions for the Breast Cancer Wisconsin Dataset.We can also see from fig.3 that as we increase the number of clusters the distortions decreases.