Performance of K-Means Algorithm on Woman's Fertility Dataset

Shrikant patro,

School of Computer Science and Engineering, VIT Chennai, Tamil Nadu, India 600127

Email:shrikantjagannath.2018@vitstudent.ac.in

Abstract

K-Means algorithm is very popular clustering algorithm. It belong to the class of unsupervised machine learning model. The clustering algorithm seeks to learn from the properties of the data point, an optimal division or discrete labeling of group of point. This algorithm is implemented in scikit-learn, it is simplest to understand and implement. This paper will discuss the working principle of the K-Means algorithm and performance on the woman fertility dataset. This will determine based on the property feature or the data point whether the data point is expected belong to donor class or not.

Introduction

The K-Means algorithm searches for the predetermined number of the cluster within an unlabeled multi-dimensional dataset. It accomplishes this using the concept of optimal clustering look like:

(a) The "Cluster Center" is the arithmetic mean of all the point belong to the cluster.
(b) Each point belonging to cluster center is closer to its own cluster center rather than other cluster center.

These two assumption for the foundation of the K-Means algorithm. The K-Means algorithm cluster the data point and identify the cluster center, similar to assumption of human eye. The number of possible combination of data point in cluster is exponential in nature. The K-Means algorithm involve an intuitive and iterative approach known as expectation maximization.

Methodology

Expectation-maximization (E-M) is a powerful algorithm that comes with data science in verity of context. K-means is a particularly simple and easy to understand application of the algorithm. Expectation Maximization approach here is consisting of the following steps:

(a) Guess some cluster centers.
(b) Repeat until converged
    a. E-Step: assign points to nearest cluster center.
    b. M-Step: set the cluster center to the mean.

Here, "Expectation Maximization" is named so because, here "E-Step" it involve updating our expectation of which our each cluster belongs to. The "M-Steps" or maximization involve some fitness function that define the location of the cluster centers. In this maximization is achieved by taking the mean the point that form the cluster center. Each repetition of E-Step and M-Step will produce better estimate of the cluster and cluster center.

Limitation of Expectation-maximization

1. The globally optimal result may not be achieved:

Although, The EM procedure is guaranteed to produce the optimize result. But, it may not be globally best optimization result. If we use

different seed in our simple procedure, the particular starting guess will lead to the poor solution. Here, the EM has converged but not converged to globally optimal solution.

2. The number of cluster must be selected beforehand:

Another common challenge with k-means is that we must tell beforehand about the number of the cluster expected from the data point. For example, if we ask the algorithm to identify six cluster, it will proceed and find six cluster. Whether, the result is meaningful or not, this cannot be answered.

3. K-Means is limited to the linear cluster boundaries:

The fundamental model assumption of k-means (the point will be closer to their cluster center rather than their own cluster. The algorithm is often not effective if cluster have complicated boundaries.

4. K-Means can be slow for large number of sample:

Each iteration of k-means must access every point dataset. The algorithm will go slow as number of the sample grow. You might wonder, the usage of the subset of data point instead of all the data point to form the cluster. It is called as MiniBatchKMeans

Experiment

The implementation of the K-Means in scikit-learn package require us to import KMeans from sklearn.cluster. The expected parameter are n_cluster and random state. The n_cluster specify the number of cluster to be formed. The random state is set to 0. The model with fit_predict method require the dataset and it will predict the number of cluster with data point. The data point associated with the particular cluster can be identified.

Conclusion

The number of cluster formed after the data point is good enough to classify them in number of group. All the data point belong to some or the other group.