Graphical user interface, text, application

Description automatically generatedGraphical user interface, application

Description automatically generated

Student Name

***Sanatkumar Rajmogali Ippalpalli***

Title of Project Report

***Guided Project 10 – Gaussian Mixture Models***

Table of Contents

[Introduction 4](#_Toc83044178)

[Eckovation theme & Question 5](#_Toc83044179)

[Prerequisites before starting coding 6](#_Toc83044180)

[program DEVELOPMENT steps 8](#_Toc83044181)

[Dataset/Image requirements 8](#_Toc83044182)

[Technique – GAUSSIAN MIXED MODEL CLUSTERING ANALYSIS 8](#_Toc83044183)

[PROGRAM / CODE DEVELOPMENT 9](#_Toc83044184)

[Analysis 17](#_Toc83044185)

[CONCLUSION 18](#_Toc83044186)

[Figure 1 Import libraries and datasets/modules 9](#_Toc83044122)

[Figure 2 Load Pandas Dataset 9](#_Toc83044123)

[Figure 3 Visualization of Dataset 10](#_Toc83044124)

[Figure 4 Elbow Method to find optimum clusters 11](#_Toc83044125)

[Figure 5 k-means clustering algorithm on Dataset 12](#_Toc83044126)

[Figure 6 Gaussian Mixture Model Algorithm 13](#_Toc83044127)

[Figure 7 Matplot lib plots 15](#_Toc83044128)

[Figure 8 Matplot lib plots using K-means 17](#_Toc83044129)

***EXECUTIVE SUMMARY***

Science and technology improved many technologies and has guided numerous innovative features which advanced clustering techniques in unsupervised learning domain.

As part of guided project, one of the important unsupervised learning domain is clustering. **Gaussian Mixture Models** used for the unsupervised clustering problems to produce unerring accuracy in the real-world datasets.

Linear Discriminant Analysis is a well-known scheme for

feature extraction and dimension reduction

Linear Discriminant Analysis is a well-known scheme for

feature extraction and dimension reduction

Linear Discriminant Analysis is a well-known scheme for

feature extraction and dimension reduction

# Introduction

Science and technology improved many technologies and has guided numerous innovative features which advanced the techniques in data classification & clustering.

Clustering refers to grouping of similar objects/variables data points together, based on their attributes or features. There are two main clustering algorithms which we shall study in this guided project: K-means and Gaussian Mixture Models.

k-means clustering is a distance-based algorithm. This means that it tries to group the closest points to form a cluster.  If the distribution of points is *not* in a circular form, k-means clustering algorithm fails to identify right clusters. **So instead of using a distance-based model, we will now use a distribution-based model i.e.** Gaussian Mixture Models!

Gaussian Mixture Models (GMMs) assume that there are a certain number of Gaussian distributions, and each of these distributions represent a cluster. Hence, a Gaussian Mixture Model tends to group the data points belonging to a single distribution together. **Gaussian Mixture Models are probabilistic models and use the soft clustering approach for distributing the points in different clusters.**

Hence Eckovation includes this guided project in the courseware for students to understand, implementation / execute the code themselves.

This report includes the 5W1H about the theme of development of code and running the code with database available over the internet. At the end of the report, the conclusions share the k-means & Gaussian Mixed Model algorithms accuracy in clustering by plots.

# Eckovation theme & Question

**Theme : Gaussian Mixture Model**

A **Gaussian mixture model is a probabilistic model** that assumes all the data points are generated from a mixture of a **finite number of Gaussian distributions** with unknown parameters. It attempts to find a mixture of **multi-dimensional Gaussian probability distributions** that best model any input dataset allowing the model to learn automatically, i.e. in an unsupervised manner. **The bag-of-words model** is a way of representing text data when **modelling text** with machine learning algorithms which can be combined with **GMM** to get a useful model representation.

**Question:**

Using a **gaussian mixture model**, perform a simple clustering on the given **2D Dataset**. Try to find the optimal number of clusters using python (you may use any module to implement this). Now implement the same from scratch using python and a dummy dataset generated using **scikit learn dataset** generating functions such as **make blob.**

**Dataset Link: Clustering\_GMM**

<https://cdn.analyticsvidhya.com/wp-content/uploads/2019/10/Clustering_gmm.csv>

# Prerequisites before starting coding

1. Who - Software needed?
2. What - Version / Release of software?
3. Any Prerequisites
4. How - to install the software
5. Which -libraries are needed to execute the problem statement
6. Where – dataset requirements, path location to include in the code
7. When – to use the above feature extraction
8. Who – Software neeed?

Python

1. What- Version / Release of software?

Python version 3.6 (latest version of python)

1. Any Prerequisites

RAM space availability & hard disk space availability

Admin rights to install the software

1. How - to install the software
2. The following url <https://www.python.org/downloads/>can be referred to download python.
3. Second and easier option is to download anaconda and use its anaconda prompt to run the commands. To install anaconda check this url <https://www.anaconda.com/download/>
4. Which -libraries are needed to execute the problem statement
5. Sklearn dataset
6. Sklearn mixture
7. Sklearn cluster
8. Numpy (pip install numpy)
9. Matplotlib (pip install matplotlib)
10. Pandas
11. Seaborn if needed
12. Where – dataset requirements, path location to include in the code
13. Once you have python downloaded and installed, you will need to setup PATH variables (if you want to run python program directly, detail instructions are below in how to run software section). To do that check this: [https://www.pythoncentral.io/add-python-to-path-python-is-not- recognized-as-an-internal-or-external-](https://www.pythoncentral.io/add-python-to-path-python-is-not-recognized-as-an-internal-or-external-command/) [command/](https://www.pythoncentral.io/add-python-to-path-python-is-not-recognized-as-an-internal-or-external-command/).
14. Setting up PATH variable is optional as you can also run program without it and more instruction are given below on this topic.
15. When – to use the above feature extraction
16. When – to use the above technique

It is used for grouping of similar objects/variables data points together, based on their attributes or features.

It’s time to dive into the code!

# program DEVELOPMENT steps

* Dataset file
* Technique selections
* Program / code development
* Analysis

### Dataset/Image requirements

The source file used for this project are downloaded / collected from analytics Vidhya website as instructed in the question.

**Dataset Link: Clustering\_GMM**

<https://cdn.analyticsvidhya.com/wp-content/uploads/2019/10/Clustering_gmm.csv>

### Technique – GAUSSIAN MIXED MODEL CLUSTERING ANALYSIS

For performing a Gaussian Mixed Model Clustering, following simple steps to be implemented.

1. Load the dataset in to the pandas dataframe
2. Plot the dataset using Matplot lib to understand the data distribution
3. Use Clustering technique k-means algorithm
4. Use Clustering technique Gaussian Mixed Model algorithm
5. Compare the results & state analysis results

Let us hop to the inscribing carving!

### PROGRAM / CODE DEVELOPMENT

As explained step by step during the lecture by mentor, we would approach steps and understand the basics with brief explanation as needed.

#### Step 1: Import the relevant libraries and applicable datasets/modules

Graphical user interface, text, application, email

Description automatically generated

Figure 1 Import libraries and datasets/modules

#### Step 2: Load dataset/Image

As Eckovation advised to use analytics Vidhya clustering.csv file (https://archive.ics.uci.edu/ml/datasets/Iris).

Graphical user interface, text, application, email

Description automatically generated

Figure 2 Load Pandas Dataset

Visulization of Dataset

Graphical user interface, text

Description automatically generated

Chart, scatter chart

Description automatically generated

Figure 3 Visualization of Dataset

#### Step 3: Elbow Method

Text

Description automatically generated

Chart, line chart

Description automatically generated

Figure 4 Elbow Method to find optimum clusters

#### Step 4: k-means clustering algorithm

Apply the k-means clustering algorithm technique

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Figure 5 k-means clustering algorithm on Dataset

#### Step 5: Gaussian Mixture Model Clustering Algorithm

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

Figure 6 Gaussian Mixture Model Algorithm

Graphical user interface, text, application, email

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Figure 7 Matplot lib plots

#### Step 6: Make Blob & Analyse from Scratch

Chart, scatter chart

Description automatically generated

Chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Figure Matplot lib plots using K-means

### Analysis

Clustering algorithm is used on clustering dataset advised by Eckovation and make blobs techniques.

This entire program runs within few seconds.

# CONCLUSION

In this guided project, we built clustering algorithms(clusters).

This is done first using k-means followed with Gaussian Mixture Model algorithm. However, the improvements in the code with time with multiple attempts may be checked for various comparisons with other datasets and justified for the usefulness of GMM technique.

This entire program runs within few seconds.

references:

1. <https://www.analyticsvidhya.com/blog/2019/10/gaussian-mixture-models-clustering/>
2. <https://towardsdatascience.com/gaussian-mixture-model-clusterization-how-to-select-the-number-of-components-clusters-553bef45f6e4>
3. <https://www.analyticsvidhya.com/blog/2020/02/quick-introduction-bag-of-words-bow-tf-idf/>
4. <https://www.geeksforgeeks.org/gaussian-mixture-model/>