Capstone Project Report

Name: Sumit Patel

Course: AI and ML (Batch – AUG 2020)

Duration: 10 months

Gaussian Mixture Models: Bag of Words Representation

Problem Statement:

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.

Prerequisites

What things you need to install the software and how to install them:

Python 3.6 This setup requires that your machine has latest version of python. The following url https://www.python.org/downloads/ can be referred to download python. 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-externalcommand/. Setting up PATH variable is optional as you can also run program without it and more instruction are given below on this topic. 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/ You will also need to download and install below 3 packages after you install either python or anaconda from the steps above Sklearn (scikit-learn) numpy scipy if you have chosen to install python 3.6 then run below commands in command prompt/terminal to install these packages:

pip install -U scikit-learn

pip install numpy

pip install pandas

pip install matplotlib

pip install scipy

If you have chosen to install anaconda then run below commands in anaconda prompt to install these packages:

conda install -c scikit-learn

conda install -c anaconda numpy

conda install -c anaconda pandas

conda install -c anaconda matplotlib

conda install -c anaconda scipy

Dataset used:

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

Method used for detection

Gaussian Mixture Model

Importing the libraries:

```
In [85]: import numpy as np
         import pandas as pd
         import matplotlib.pyplot as plt
         from scipy import stats
         import seaborn as sns
         from sklearn.mixture import GaussianMixture
         import warnings
         warnings.filterwarnings('ignore')
         sns.set()
```

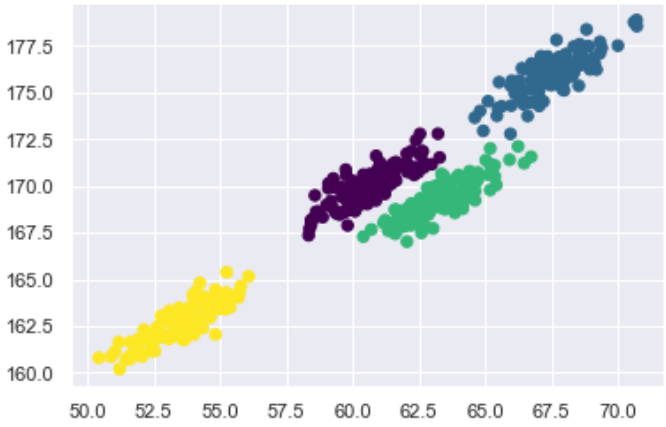
Training Data:

```
In [88]: gm = GaussianMixture(n_components=4, random_state=0).fit(x)
         gm.means_
```

```
Out[88]: array([[ 60.65944689, 170.03409967],
                [ 67.51014715, 175.97136573],
                [ 63.29374518, 169.26263398],
                [ 53.60032216, 162.76480188]])
```

```
In [89]: # Assign a label to each sample
         labels = gm.predict(data)
```

```
In [124]: plt.scatter(x[:,0], x[:,1], c = labels, cmap = 'viridis', s = 40);
```



Generating Dummy dataset and Using Gaussian Model to Cluster

In [92]:
```
# Generating a dummy dataset
from sklearn.datasets import make_blobs
from sklearn.metrics import classification_report
X, y_true = make_blobs(n_samples = 400, centers = 4, cluster_std = 0.60, random_state = 0)
X = X[:,::-1]
```

```
labels_dum = gm_dum.predict(X)
labels_dum[:10]
```

```
array([3, 0, 0, 3, 0, 1, 2, 3, 0, 2], dtype=int64)
```

```
fig, ax = plt.subplots(ncols = 2, figsize = (15,5))
ax[0].scatter(X[:,0], X[:,1], c = y_true, cmap = 'viridis', s = 40);
ax[1].scatter(X[:,0], X[:,1], c = labels_dum, cmap = 'viridis', s = 40);
```