Assignment - 5

TITLE:
Gaussian mixture models clustering.

PROBLEM STATEMENT:
Problem clustering of the iris dataset based on all variables using Gaussian mixture models. Use PCA to visualise clusters.

OBJECTIVES:
To learn data processing techniques required to get applied on machine learning algorithm.

OUTCOMES:
Formulate suitable statistical method required as pre-processing technique for finding the solution of machine learning algorithm.

PREREQUISITES:
→ Concept of clustering

THEORY:
The goal is clustering is to find groups that share similar properties. The data in each group should be similar, but each cluster should be sufficiently different.

1) Gaussian Mixture Model (GMM)
The GMM is a simple but powerful model that performs clustering via density estimation. The features histogram is modelled as the sum of multiple multi-variate Gaussian distributions.

In one dimensions the probability density function of a Gaussian Distribution is given by:

$$G(X|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

Where μ and σ^2 are respectively mean and variance of distribution.
For multivariate Gaussian Distribution, the probability density function is given by:

$$G(X|\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (X-\mu)\right)$$

Here μ is a dimensional vector and Σ is the $d \times d$ covariance matrix.

2) Expectation - Maximisation (EM) Algorithm
It is an iterative way to find maximum-likelihood estimates for model parameters when the data is incomplete or has some missing data points.
These new values are then respectively used to estimate a better data, by filling up missing points, until the values get fixed.

→ Estimation step.
→ Maximisation step.

Date: / /20

ALGORITHM:-

1) Initialize the mean U_k, the covariance matrix Σ_k ~~used~~ and the mixing coeff. π_k by some random values.

2) Compute the γ_k values for all k.

3) Again estimate all the parameters using current γ_k values.

4) Complete log-likelihood function.

5) Put some convergence criterion.

6) If the log-likelihood value converges to some value then stop, else return to step 2.

CONCLUSION:

Gaussian Mixture Model (GMM) Clustering handles ellipsoidal distributions and makes '~~self~~ soft' assignments to clusters.

AIML Assignment5.ipynb - Colab | Combine your Documents and In

colab.research.google.com/drive/1aUYjO74QwyIFk6UmZzJlR1X-IfVFwNx4

Apps | HackerRank | W3Schools | Elearn | Codeforces | Django | ERDPlus | LeetCode | Tinkercad | Coursera | My Captain | Google Tech Dev G... | Raading list

AIML Assignment5.ipynb

File Edit View Insert Runtime Tools Help Last edited on January 10

Comment Share

Files

Connecting to a runtime to enable file browsing.

+ Code + Text

Connecting Editing

Perform clustering of the iris dataset based on all variables using Gaussian mixture models. Use PCA to visualize clusters.

```
[ ] import numpy as np
    import pandas as pd
    import seaborn as sns
    import matplotlib.pyplot as plt
    from sklearn.preprocessing import StandardScaler
    from sklearn.mixture import GaussianMixture
    from sklearn.metrics.cluster import adjusted_rand_score
```

```
[ ] irisdata = pd.read_csv("/content/drive/MyDrive/data/iris.csv")
```

```
[ ] x= irisdata.iloc[:,:4]
    y= irisdata.iloc[:,-1]
```

Standardizing Dataset useing sklearn

```
[ ] sc =StandardScaler()
    sc.fit(x)
    std_array =sc.transform(x)
    X = pd.DataFrame(std_array,columns = x.columns)
```

TCOB41 AIML Assi....pdf | TCOB41 AIML Assi....pdf | TCOB41 AIML Assi....pdf | Show all

Type here to search

24°C Light rain ENG 13:47 13-01-2022

AIML Assignment5.ipynb

File Edit View Insert Runtime Tools Help Last edited on January 10

Comment Share

Files

..
sample_data

+ Code + Text

Initializing Editing

```
[ ] sc =StandardScaler()
    sc.fit(x)
    std_array =sc.transform(x)
    X = pd.DataFrame(std_array,columns = x.columns)
```

Gaussian Mixture model

```
[ ] cluster =GaussianMixture(n_components=3)
    cluster.fit(X)
    y_pred =cluster.predict(X)
    score = adjusted_rand_score(y,y_pred)
    score
```

```
0.9038742317748124
```

Using PCA to visualize data

```
[ ] from sklearn.decomposition import PCA

    pca =PCA(n_components=2)
    pca_array =pca.fit_transform(irisdata.drop(['species'],axis=1))
    pca_df =pd.DataFrame(pca_array,columns=["PC1","PC2"])
    pca_df.head()
```

AIML Assignment5.ipynb - Colab | Combine your Documents and In

colab.research.google.com/drive/1aUYjO74QwyIFk6UmZzJlR1X-IfVFwNx4

Apps | HackerRank | W3Schools | Elearn | Codeforces | Django | ERDPlus | LeetCode | Tinkercad | Coursera | My Captain | Google Tech Dev G... | Raading list

AIML Assignment5.ipynb

File Edit View Insert Runtime Tools Help Last edited on January 10

Comment Share

Files

..

sample_data

+ Code + Text

Initializing Editing

Using PCA to visualize data

```
[ ] from sklearn.decomposition import PCA

    pca =PCA(n_components=2)
    pca_array =pca.fit_transform(irisdata.drop(['species'],axis=1))
    pca_df =pd.DataFrame(pca_array,columns=["PC1","PC2"])
    pca_df.head()
```

	PC1	PC2
0	-2.684126	0.319397
1	-2.714142	-0.177001
2	-2.888991	-0.144949
3	-2.745343	-0.318299
4	-2.728717	0.326755

```
[ ] col_code = {0:"yellow",1:"darkblue",2:"green"}
    label = {0:"setosa",1:"versicolor",2:"virginica"}

    pca_df["labels"]= pd.DataFrame(y_pred)
    groups = pca_df.groupby('labels')
```

Mounting Google Drive...

TCOB41 AIML Assi....pdf | TCOB41 AIML Assi....pdf | TCOB41 AIML Assi....pdf | Show all

Type here to search

24°C Light rain ENG 13:48 13-01-2022

AIML Assignment5.ipynb

File Edit View Insert Runtime Tools Help Last edited on January 10

Comment Share

Files

sample_data

Disk 65.98 GB available

+ Code + Text

RAM Disk Editing

3	-2.745343	-0.318299
4	-2.728717	0.326755

```
col_code = {0:"yellow",1:"darkblue",2:"green"}
label = {0:"setosa",1:"versicolor",2:"virginica"}

pca_df["labels"]= pd.DataFrame(y_pred)
groups = pca_df.groupby('labels')

# Grouping instances based on species
groups.mean()
```

labels	PC1	PC2
0	0.452518	-0.248189
1	-2.642415	0.190885
2	2.031954	0.029531

```
fig, ax =plt.subplots(1,1,figsize =(15,10))
for name, group in groups:
    ax.plot(group.PC1,group.PC2,color =col_code[name],label =label[name],marker='o',linestyle='',ms=10)
ax.legend()
plt.show()
```

AIML Assignment5.ipynb - Colab

colab.research.google.com/drive/1aUYjO74QwyIFk6UmZzJlR1X-IfVFwNx4

AIML Assignment5.ipynb

File Edit View Insert Runtime Tools Help Last edited on January 10

Comment Share

+ Code + Text

RAM Disk Editing

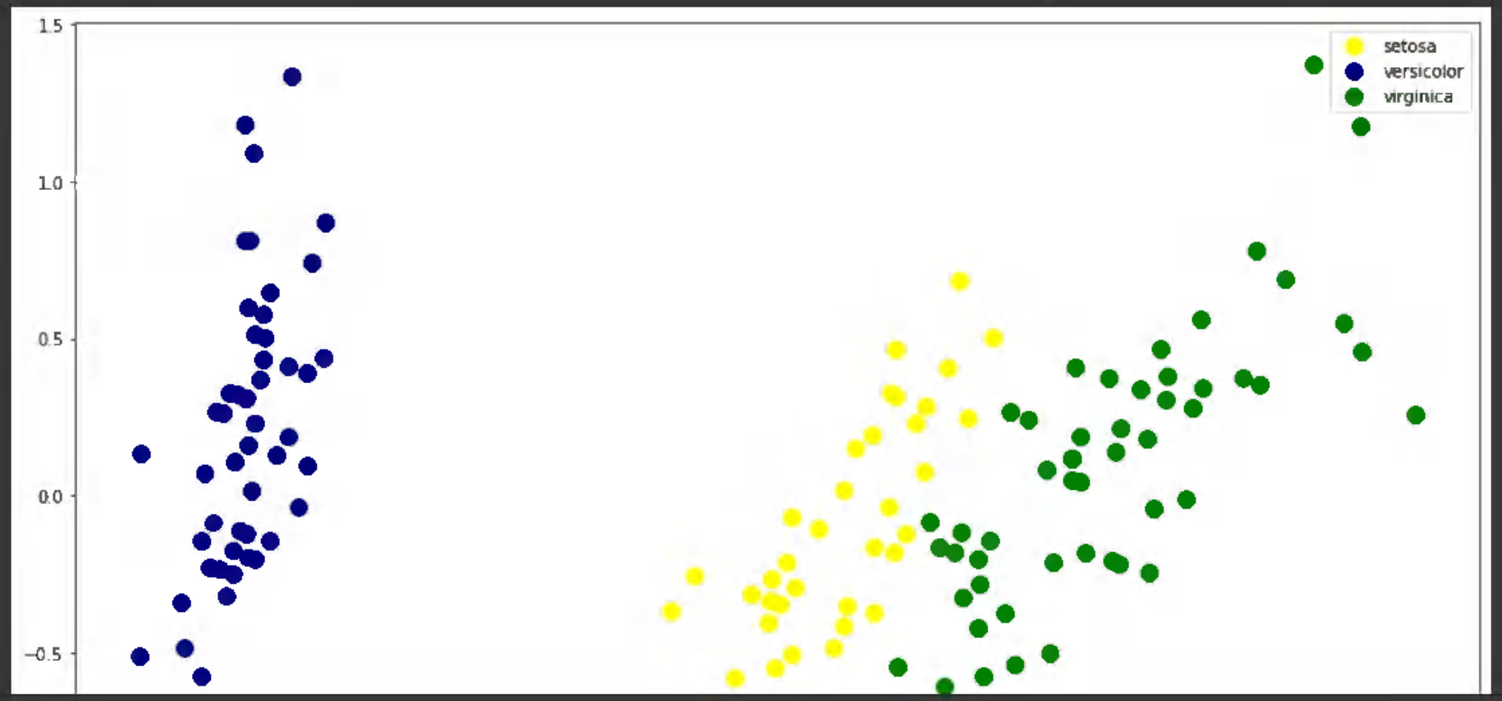
Files

..

sample_data

Disk 65.98 GB available

```
[ ] fig, ax =plt.subplots(1,1,figsize =(15,10))
    for name, group in groups:
        ax.plot(group.PC1,group.PC2,color =col_code[name],label =label[name],marker='o',linestyle='',ms=10)
    ax.legend()
    plt.show()
```



TCOB41 AIML Assi....pdf TCOB41 AIML Assi....pdf TCOB41 AIML Assi....pdf Show all

Type here to search

24°C Light rain ENG 13:48 13-01-2022

AIML Assignment5.ipynb - Colab

Combine your Documents and In

colab.research.google.com/drive/1aUYjO74QwyIFk6UmZzJlR1X-IfVFwNx4

Apps HackerRank W3Schools Elearn Codeforces Django ERDPlus LeetCode Tinkercad Coursera My Captain Google Tech Dev G... Reading list

AIML Assignment5.ipynb

File Edit View Insert Runtime Tools Help Last edited on January 10

Comment Share

Files

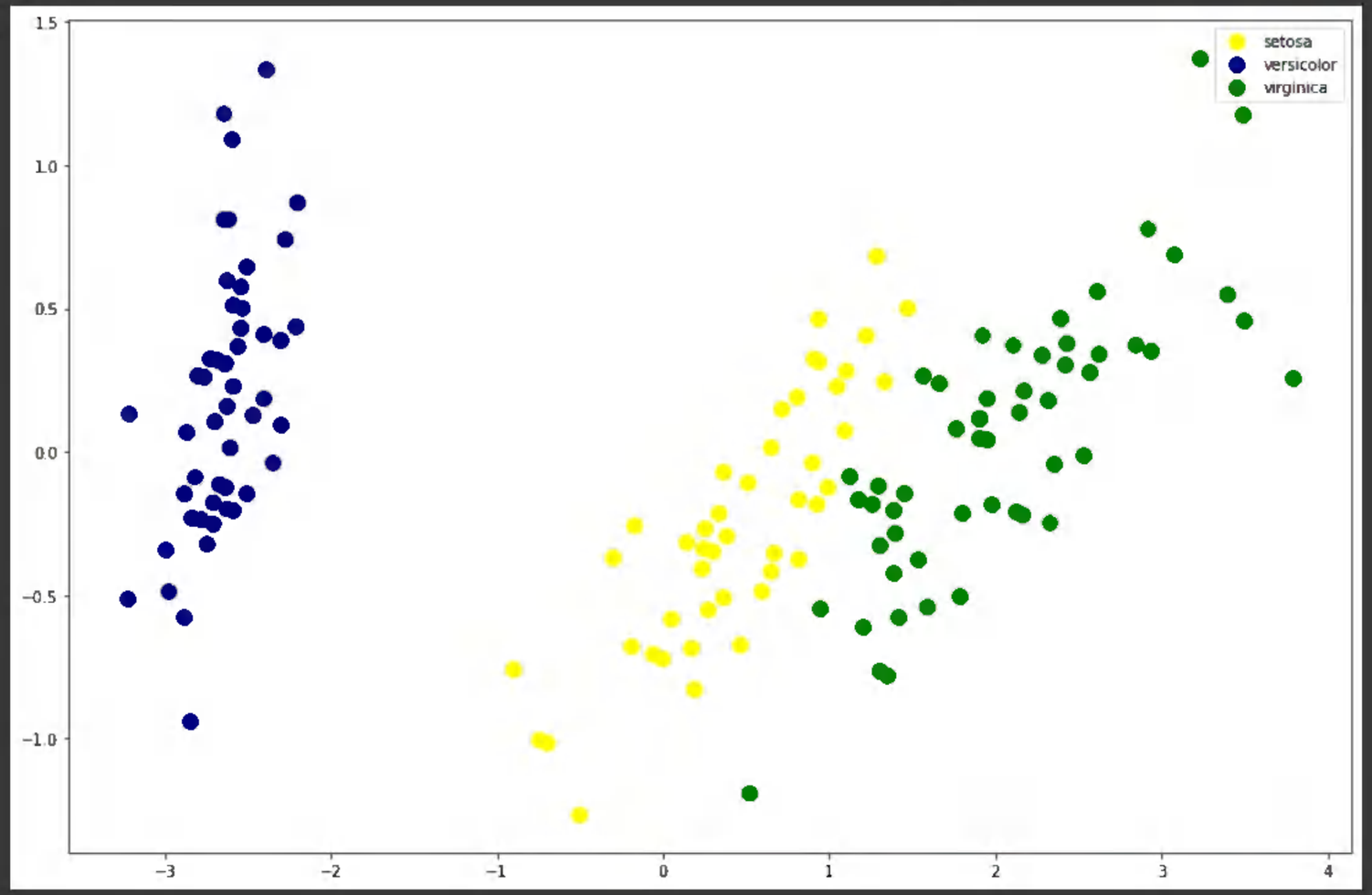
..

sample_data

+ Code + Text

RAM Disk Editing

[ ]



Disk 65.98 GB available

Type here to search

24°C Light rain ENG 13:48 13-01-2022