**Student’s Name: ISHAAN GUPTA**

**Roll Number: B20292**

**Mobile No: 9179242114**

**Branch:M.E.**

Chart, line chart

Description automatically generated

Figure Eigenvalue vs. components

**Inferences:**

1. he eigenvalue decreases corresponding to each component increase.
2. Eigen values tell the variation along each principal component and the component with highest eigen values carry the most data.

# a.

Chart, scatter chart

Description automatically generated

Figure 2 K-means (K=3) clustering on Iris flower dataset

**Inferences:**

* 1. As we specified K=3, the K-means model has given 3 distinct clusters.
  2. The boundary seem circular (non linear) around the centroids although there is some distortion.

1. The value for distortion measure is 63.87.
2. The purity score after examples are assigned to the clusters is 0.89.

Chart, line chart

Description automatically generated

Figure Number of clusters(K) vs. distortion measure

**Inferences:**

1. Distortion decreases with increasing K.
2. Justify the observed trend.
3. The elbow is formed at K = 3, this shows that 3 is the optimal number of clusters. It follows our data as there were three varieties in the label column.

Table Purity score for K value = 2,3,4,5,6 & 7

|  |  |
| --- | --- |
| **K value** | **Purity score** |
| 2 | 0.667 |
| 3 | 0.886 |
| 4 | 0.686 |
| 5 | 0.680 |
| 6 | 0.513 |
| 7 | 0.513 |

**Inferences**:

1. The highest purity score is obtained with K =3.
2. The purity score increase upto k=3, then decreases with increasing k.
3. The data had 3 class originally and so the purity is highest at 3 cluster.
4. The purity is highest at elbow of distortion plot.

# a.

Chart, scatter chart

Description automatically generated

Figure 2 GMM (K=3) clustering on Iris flower dataset

**Inferences:**

* 1. 3 clusters (k=3) are formed by GMM clustering. GMM performs soft clustering as it gives the probability of sample point to belong to a particular cluster.
  2. The clusters are elliptical with a little distortion.
  3. There is an observable difference in shape and boundary of class 1 and class 2 in K-means and GMM clustering.

1. Total Log Likelihood for K=3 clusters = -280.87.
2. Purity score for data = 0.98.

Chart, line chart

Description automatically generated

Figure Number of clusters(K) vs. distortion measure

**Inferences:**

1. Does the distortion measure increase with an increase in K.
2. Justify the observed trend.
3. From the number of species in the given dataset, intuitively what should be the number of optimum clusters? Does the elbow and distortion measure plot follow the intuition?

Note: The plot above is for illustration purposes. Replace it with the plot obtained by you. Label x-axis as distortion measure and y-axis as number of clusters (K).

Table Purity score for K value = 2,3,4,5,6 & 7

|  |  |
| --- | --- |
| **K value** | **Purity score** |
| 2 | 0.66 |
| 3 | 0.98 |
| 4 | 0.82 |
| 5 | 0.60 |
| 6 | 0.73 |
| 7 | 0.54 |

**Inferences**:

1. e highest purity score is obtained with K =3.
2. The purity score increase upto k=3, then decreases with increasing k.
3. The original data had 3 clusters and hence purity is highest at k =3.
4. The maximum purity score for GMM is 0.98 whereas for KMeans is 0.886.

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Figure 6 DBSCAN clustering on Iris flower dataset

**Inferences:**

1. We cannot specify the number of clusters in DBSCAN. It is robust to outliers and can also form arbitrarily shaped clusters. It is not suitable for too dense or too sparse data with connected boundaries.
2. There are 3 clusters in gmm (as specified), but 2 clusters in eps 1 and only 1 cluster in eps 5.

**b.**

|  |  |  |
| --- | --- | --- |
| **Eps** | **Min\_samples** | **Purity Score** |
| 1 | 5 | 0.667 |
| 10 | 0.667 |
| 4 | 5 | 0.333 |
| 10 | 0.333 |

**Inferences:**

1. For the same eps value, the Purity score remains the same on increasing the min samples.
2. For the same min\_samples, increasing the value of eps decreases the value of purity score for the sample.