UNSUPERVISED

→ Dont have labels
→ it's all about how to group

2 types Unsupervised learning
grouping the data is called Clustering
Reducing the dimensionality is another type

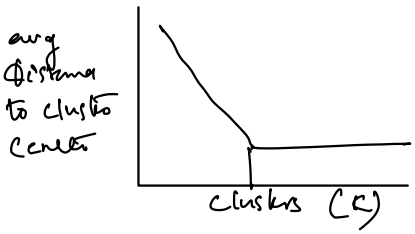
1. Clustering the data
2. Hierarchical and Density based Clustering
3. Gaussian Mixture model and Clustering validation
4. principal Component analysis
5. Random projection and Independent Component analysis

Clustering - K MEANS algorithm

Recommending new items eg
Grouping the data by Closeness of the data

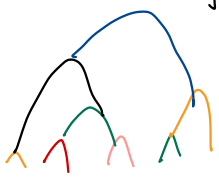
Choosing K value:
1. depends prior knowledge
2. Elbow method (no prior knowledge)

Elbow Method



Hierarchical density based clustering

DBSCAN - Density b



Hierarchical clustering

Hierarchical clustering

Single link clustering
* each point is a cluster
* Smallest d/b/w the points
* Nearest 2 clusters

Complete link:-

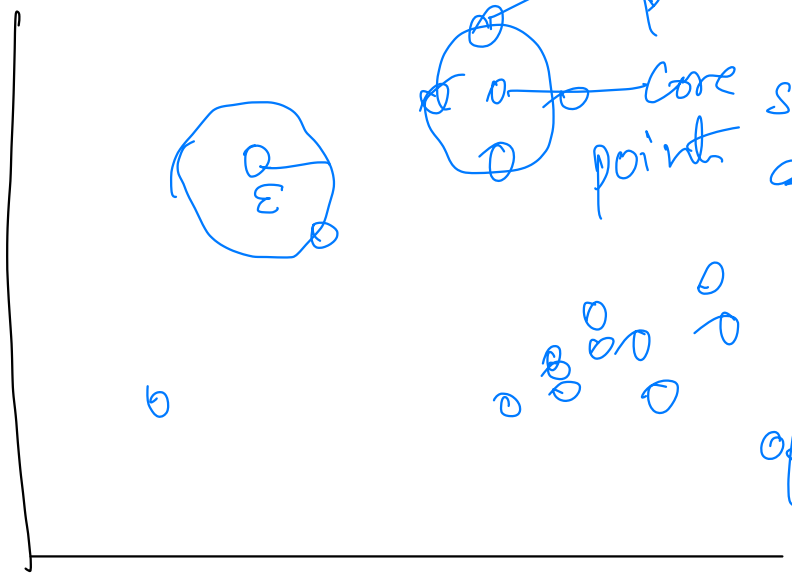
```
from sklearn import datasets, cluster
from scipy.cluster.hierarchy import dendrogram, ward, single
```

```
X =
linkage_matrix = ward(X)
dendrogram(linkage_matrix)
plt.show()
```

```
from sklearn import Agglomerative clustering   Agglomerative clustering ( linkage = 'complete'
                                                                          'average')
  ward = Aggalomerative clustering (n_clusters = 5)
ward_predict = ward.fit_predict(X)

from sklearn.metrics  Adjust_rand_score

    from  Sklearn.cluster  import  linkage
      linkage_type = 'ward'
      linkage_matrix = linkage (X, linkage_type)
    from  scipy.cluster.hierarchy  import  dendrogram
      import matplotlib.pyplot as plt
        plt.figure ('figsize = (22, 18))
      dendrogram (linkage_matrix)
```



Inputs

Border point

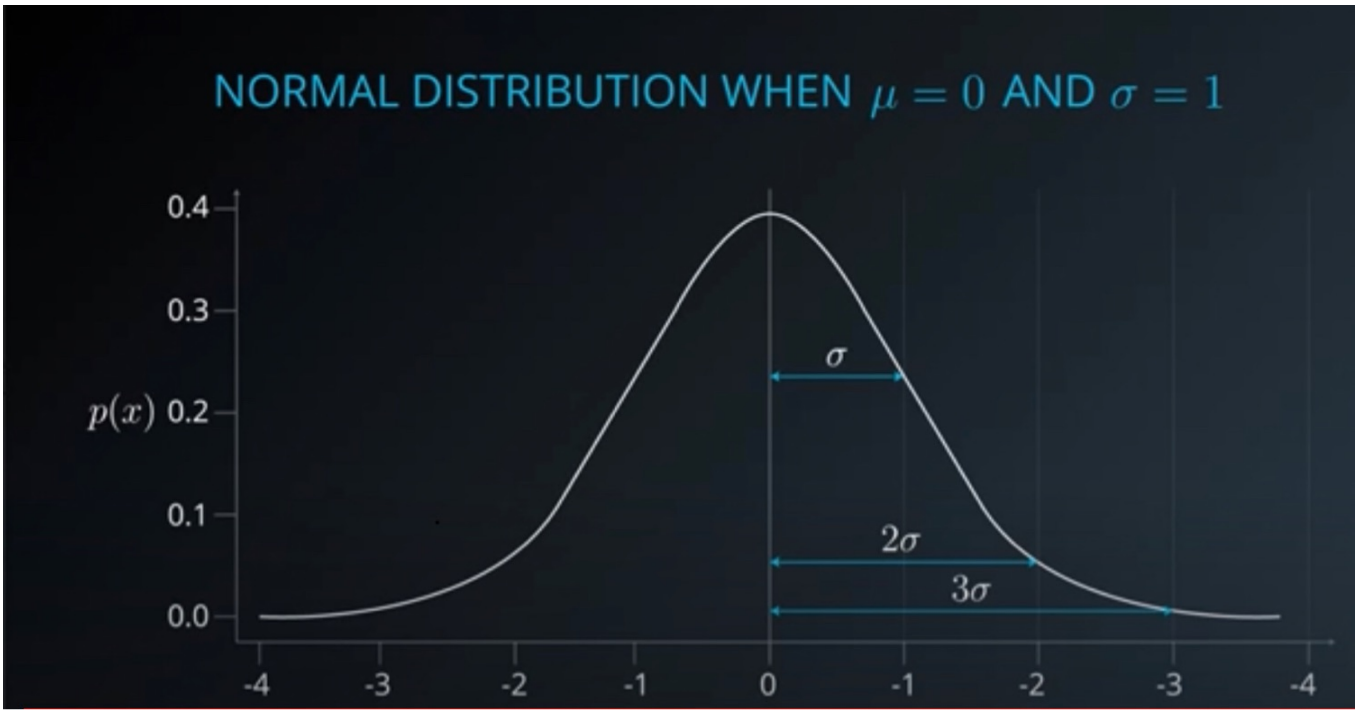
Core point

Epsilon - search distance around point

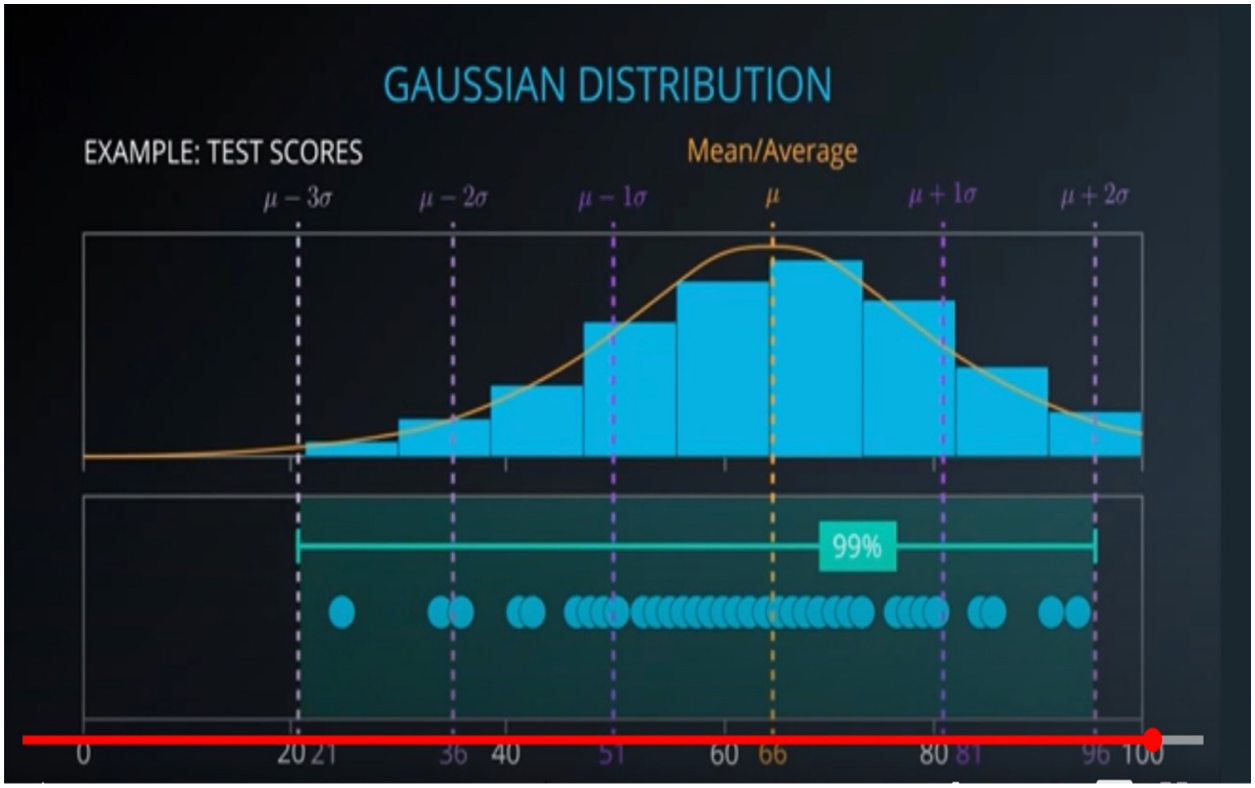
MinPts = 5

Minimum no of required to form density clust

GAUSSIAN MIXTURE MODEL & CLUSTERING



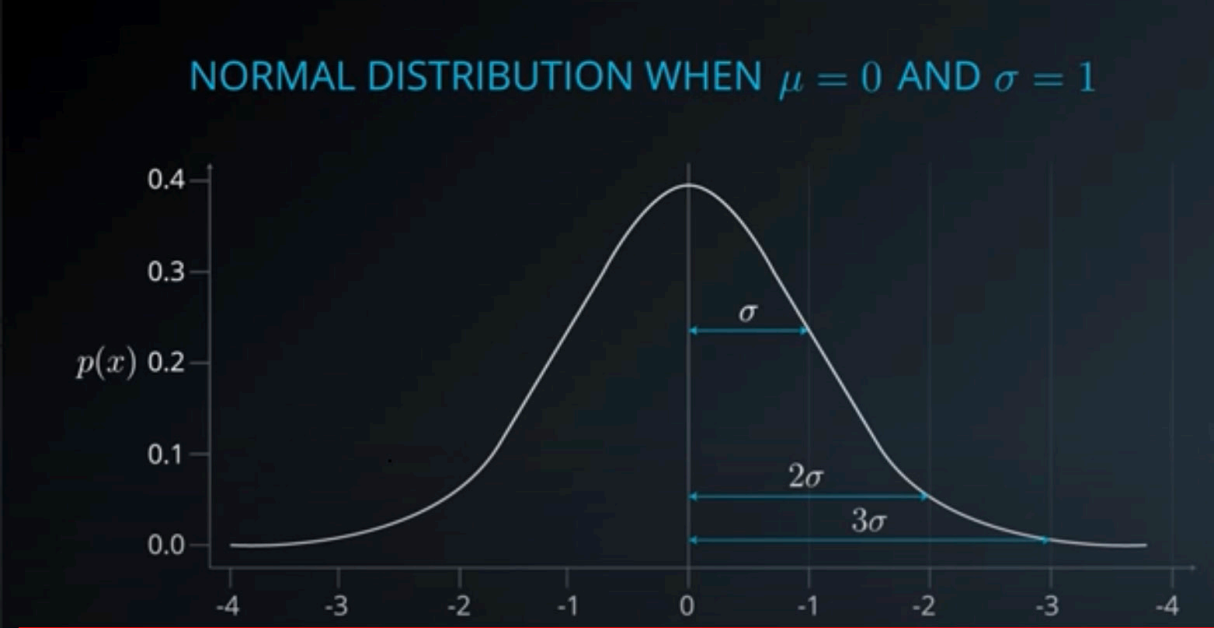
Plotting gaussian distribution

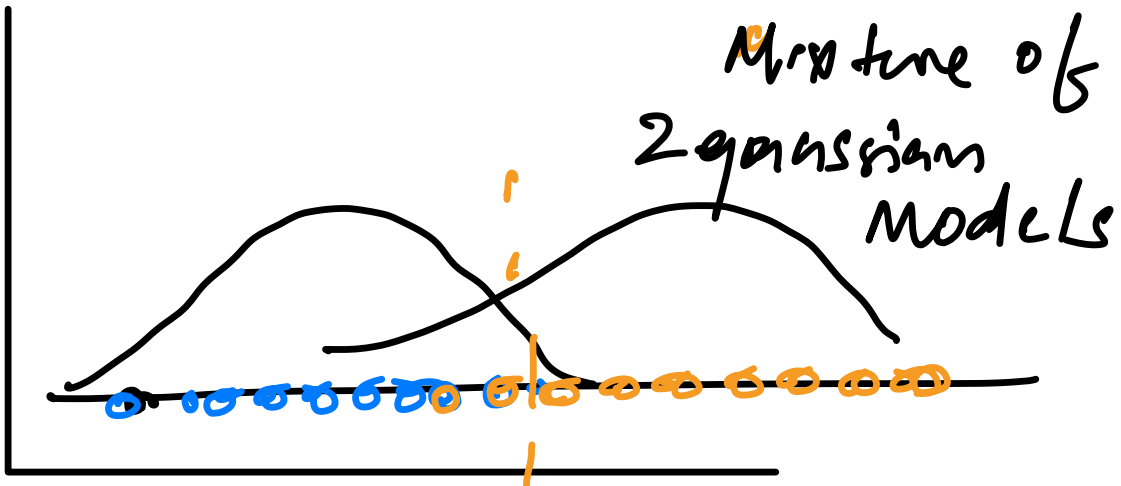


μ = mean
σ - std

$\mu - 3\sigma$ ↓ $\mu + 3\sigma$ — 99%

$\mu - 1\sigma$ — $\mu + 1\sigma$ — 68%
$\mu - 2\sigma$ — $\mu + 2\sigma$ — 95%


NORMAL DISTRIBUTION WHEN $\mu = 0$ AND $\sigma = 1$
0.4
0.3
$p(x)$ 0.2
0.1
0.0
σ
2σ
3σ
-4
-3
-2
-1
0
-1
-2
-3
-4



1. Initialize gaussian distributio
2. Soft data clustering
3. Re estimate gaussian Maximizatio
4. Evaluate log likelihood to check
the Con

Initialize gaussian distribution

1. Apply KMEANS to the dataset identify the clusters
   pick Random Mean and Std Deviation
2. Soft Clustering find the Membership of the clusters
   use the probability density function of the normal distribution function to identify the membership of the cluster
3. Re estimate the parameters for the gaussians
   find the new mean that comes from the weighted avg of the points and do find the membership of the point
   apply this process
4. Evaluate log likelihood

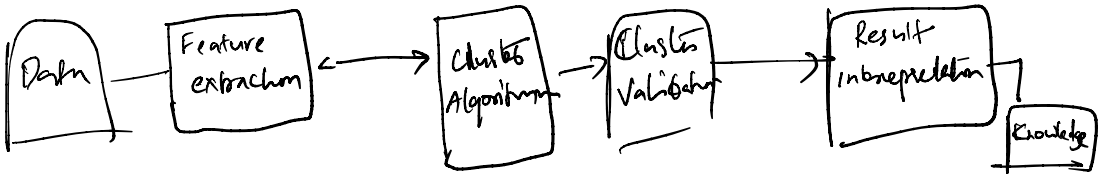
from sklearn.mixture import GaussianMixture

GaussianMixture(n_components=5)

gmm clustering

Advantages
1. Soft clustering
2. cluster shape flexibility

Disadvantages
1. Sensitive to initialization values
2. possible to converge local optimum



Adjusted Rand index — External indices

Ranges from −1 to 1

Silhouette coefficient | internal indices

Ranges from −1 to 1

Silhouette coefficient helps to identify the best values for
for suitable k value silhouette coefficient
will be nigh