Scikit Learn
Essay Discussion

Ishant Nayer
M10669373
Special Topics in BANA (002)-Python

Dataset

I am using Make_blobs dataset from sklearn's in-built datasets library.

Following are its parameters:

n_samples : int, optional (default=100)
The total number of points equally divided among clusters.

n_features : int, optional (default=2)
The number of features for each sample.

centers : int or array of shape [n_centers, n_features], optional
(default=3) The number of centers to generate, or the fixed center locations.

cluster_std : float or sequence of floats, optional (default=1.0)
The standard deviation of the clusters.

center_box : pair of floats (min, max), optional (default=(-10.0, 10.0))
The bounding box for each cluster center when centers are generated at random.

shuffle : boolean, optional (default=True)
Shuffle the samples.

random_state : int, RandomState instance or None, optional (default=None)

If int, random_state is the seed used by the random number generator; If RandomState instance, random_state is the random number generator; If None, the random number generator is the RandomState instance used by np.random.

Following are the Returns that we get:

X : array of shape [n_samples, n_features]
The generated samples.

y : array of shape [n_samples]
The integer labels for cluster membership of each sample.

Creating a model

K-means clustering

Following is the full code that I used for k means clustering:

```
import numpy as np
import matplotlib.pyplot as plt

from sklearn.cluster import KMeans
from sklearn.datasets import make_blobs

#choosing sample size
n_samples = 1500
random_state = 170
x, y = make_blobs(n_samples=n_samples, random_state=random_state)

#filtering x
x_filtered = np.vstack((x[y == 0][:500], x[y == 1][:100], x[y == 2][:10]))

#predicting y
y_pred = KMeans(n_clusters=3,
random_state=random_state).fit_predict(x_filtered)

#Plotting
plt.subplot(224)
plt.scatter(x_filtered[:, 0], x_filtered[:, 1], c=y_pred)
plt.title("K means clustering")

plt.show()
```

Explanation

In our plot, k-means returns intuitive clusters despite unevenly sized blobs.

Plot