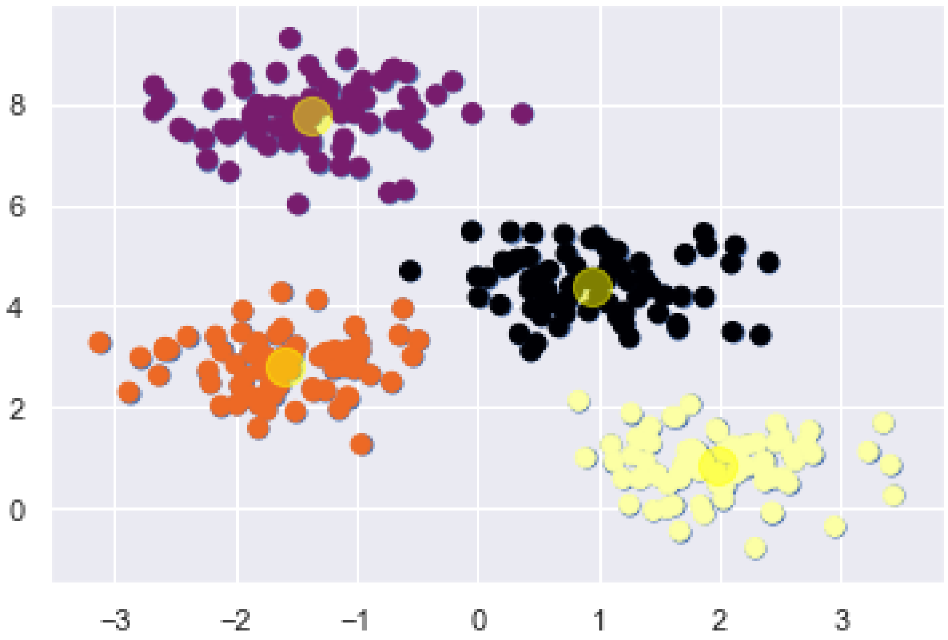
K-Means

November 20, 2019

1 K Means

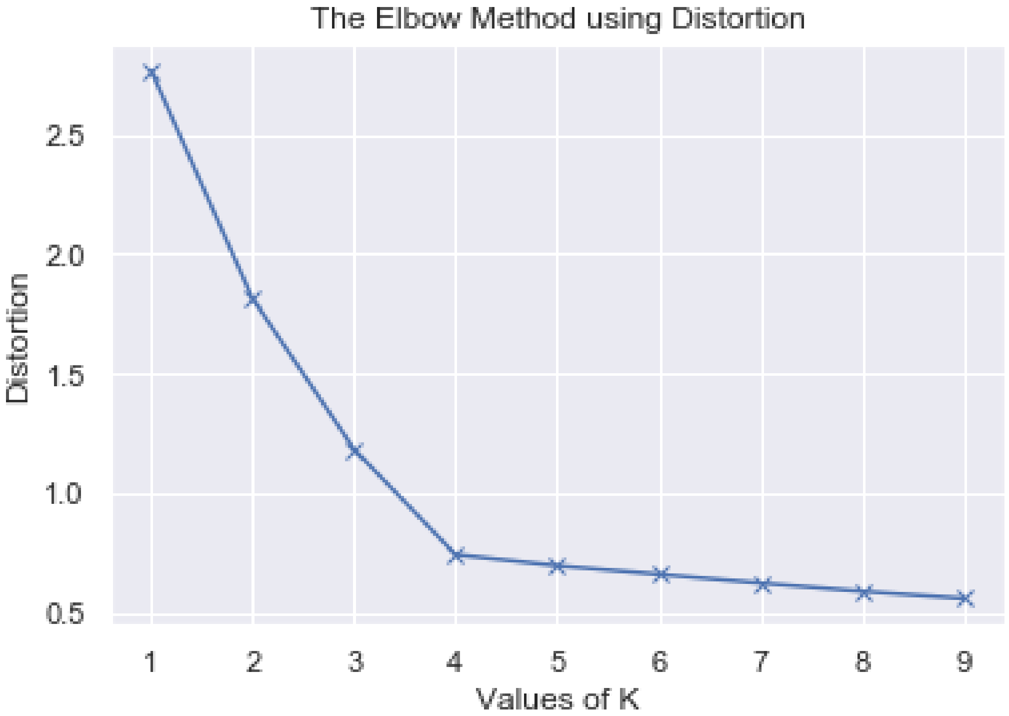
[7]:

```
%matplotlib inline
import matplotlib.pyplot as plt
import seaborn as sns; sns.set()  # for plot styling
import numpy as np
from scipy.spatial.distance import cdist


from sklearn.datasets.samples_generator import make_blobs
X, y_true = make_blobs(n_samples=300, centers=4,
                       cluster_std=0.60, random_state=0)
plt.scatter(X[:, 0], X[:, 1], s=50);


from sklearn.cluster import KMeans
kmeans = KMeans(n_clusters=4)
kmeans.fit(X)
y_kmeans = kmeans.predict(X)

plt.scatter(X[:, 0], X[:, 1], c=y_kmeans, s=50, cmap='inferno')

centers = kmeans.cluster_centers_
plt.scatter(centers[:, 0], centers[:, 1], c='yellow', s=200, alpha=0.5);

distortions = []
inertias = []
mapping1 = {}
mapping2 = {}
K = range(1,10)

for k in K:
    #Building and fitting the model
    kmeanModel = KMeans(n_clusters=k).fit(X)
    kmeanModel.fit(X)

    distortions.append(sum(np.min(cdist(X, kmeanModel.cluster_centers_,
                      'euclidean'),axis=1)) / X.shape[0])
    inertias.append(kmeanModel.inertia_)

    mapping1[k] = sum(np.min(cdist(X, kmeanModel.cluster_centers_,
                 'euclidean'),axis=1)) / X.shape[0]
    mapping2[k] = kmeanModel.inertia_
```



[8]:

```
plt.plot(K, distortions, 'bx-')
plt.xlabel('Values of K')
plt.ylabel('Distortion')
plt.title('The Elbow Method using Distortion')
plt.show()
```



2 Analysis

2.1 **Here K means clustering is studied where k-means algorithm searches for a pre-determined number of clusters within an unlabeled multidimensional dataset.The cluster center is mean of all points belonging to a cluster.By plotting the dataset we can easily observe four distinct blobs belonging to a different cluster in the scatter plot of clusters found by K means clustering model.These are the head clusters.**

2.2 **To determine the optimal number of clusters ‘k’ into which the data may be clustered,the Elbow Method is used to determine this optimal value of k.Here two values can be used for elbow method**

2.3 **(i)Distortion and (ii)Inertia**

2.4 **Distortion is calculated as the average of the squared distances from the cluster centers of the respective clusters.**

2.5 **Inertia s the sum of squared distances of samples to their closest cluster center.**

2.6 **Using distortion we develop the elbow method.**