CS150A Homework3 -- Coding

Instructions / Notes:

Read these carefully

- You may need to install the `sklearn` module to run the scripts.
- You **may** create new Jupyter notebook cells to use for e.g. testing, debugging, exploring, etc.- this is encouraged in fact!
- There will be deductions from your grades if you don't have ouputs when the question requires you to do so.

Submission Instructions:

- Do *NOT* submit your iPython notebook directly.
- Instead, upload your answers in PDF version of the HW3.ipynb with your outputs to Gradescope.

If you hava any confusion, please ask TA team in Piazza.

Have fun!

In [2]:
```
!pip install scikit-learn
```

```
Looking in indexes: https://pypi.tuna.tsinghua.edu.cn/simple
Requirement already satisfied: scikit-learn in c:\users\y\appdata\local\programs\python\python312\lib\site-packages (1.5.0)
Requirement already satisfied: numpy>=1.19.5 in c:\users\y\appdata\local\programs\python\python312\lib\site-packages (from scikit-learn) (1.26.4)
Requirement already satisfied: scipy>=1.6.0 in c:\users\y\appdata\local\programs\python\python312\lib\site-packages (from scikit-learn) (1.13.1)
Requirement already satisfied: joblib>=1.2.0 in c:\users\y\appdata\local\programs\python\python312\lib\site-packages (from scikit-learn) (1.4.2)
Requirement already satisfied: threadpoolctl>=3.1.0 in c:\users\y\appdata\local\programs\python\python312\lib\site-packages (from scikit-learn) (3.5.0)
```

In [26]:
```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.datasets import make_blobs
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import DBSCAN
from sklearn.datasets import make_moons
from sklearn.metrics import adjusted_rand_score
from sklearn import preprocessing
import pandas as pd
```

Part 1: Implementation of K-means (70 points)

In this part, you should finish question 1 to 7.

In the lecture we have learnt about the K-means algorithm, now we need to finish the sectional functions of k-means method, and use it in the following tasks.

The k-means clustering method is an unsupervised machine learning technique used to identify clusters of data objects in a dataset. There are many different types of clustering methods, but k-means is one of the oldest and most approachable.

Conventional k-means requires only a few steps. The first step is to randomly select k centroids, where k is equal to the number of clusters you choose. Centroids are data points representing the center of a cluster.The main element of the algorithm works by a two-step process called expectation-maximization. The expectation step assigns each data point to its nearest centroid. Then, the maximization step computes the mean of all the points for each cluster and sets the new centroid.

Hint: It will be helpful to refer to the pseudocode of K-means in the lecture.

In [27]:
```
X = np.array([[0,2],[0,0],[1,0],[5,0],[5,2]])
```

Question 1: Definition of Euclidean Distance (10 points)

In this question, you should finish the "euclidean_distance" function, where we compute the euclidean distance of two given points.

In [28]:
```
def euclidean_distance(x1, x2):
# ----- Write You Code Here ----- #
    return np.sqrt(np.sum((x1 - x2) ** 2))
```

Question 2: Initialization of Centroids (10 points)

In this question, you should finish the "centroids_init" function. Centroid initialization is class-center initialization, that is, k samples of the dataset are randomly selected for each category for class-center initialization. This process is also the starting point of K-means clustering algorithm.

input:
X: dataset
k: num of centroids
output:
return `k * n` matrix of centroids for `m * n` dataset X

In [50]:
```
def centroids_init(k, X):
# ----- Write You Code Here ----- #
    indices = np.random.choice(X.shape[0], k, replace=False)
    return X[indices]
```

Question 3: Determination of Belonging Centroids (10 points)

In this question, you should finish the "closest_centroid" function. This function defines the class index to which the nearest centroid of the sample belongs according to euclidean distances.

input:
sample: a single sample of X
centroids: matrix of centroids
output:
return the index of center that the sample belongs to

In [31]:
```
def closest_centroid(sample, centroids):
# ----- Write You Code Here ----- #
    distances = np.array([euclidean_distance(sample, centroid) for centroid in centroids])
    return np.argmin(distances)
```

Question 4: Assignment of Clusters (10 points)

In this question, you should finish the "build_clusters" function, where we should assign clusters to each data point. This step is actually the clustering process, that is, each sample is assigned to the nearest class cluster.

input:
centroids: matrix of centroids k: num of centroids
X: dataset
output:
return the matrix of clusters with index of each assigned sample x_i as item

In [32]:
```
def build_clusters(centroids, k, X):
# ----- Write You Code Here ----- #
```

```
    clusters = [[] for _ in range(k)]
    for idx, sample in enumerate(X):
        centroid_idx = closest_centroid(sample, centroids)
        clusters[centroid_idx].append(idx)
    return clusters
```

Question 5: Recalculation of Centroids (10 points)

In this question, you should finish the "calculate_centroids" function. The core idea of K-means clustering algorithm is to continuously dynamically adjust, recalculate the centroid according to the class cluster generated in the previous step, and then perform the clustering process.

input:
clusters: the matrix of clusters with index of each assigned sample x_i as item k: num of centroids
X: dataset
output:
return the updated matrix of centroids

In [33]:
```
def calculate_centroids(clusters, k, X):
# ----- Write You Code Here ----- #
    centroids = np.zeros((k, X.shape[1]))
    for cluster_idx, cluster in enumerate(clusters):
        if cluster:  # Avoid empty clusters
            centroids[cluster_idx] = np.mean(X[cluster], axis=0)
    return centroids
```

Question 6: Get Labels (10 points)

In this question, you should finish the "get_cluster_labels" function, where we should get the cluster category to which each sample belongs, a.k.a. the labels of each data point.

input:
clusters: the matrix of clusters with index of each assigned sample x_i as item
X: dataset
output:
return the predicted labels of X

In [34]:
```
def get_cluster_labels(clusters, X):
# ----- Write You Code Here ----- #
    labels = np.zeros(X.shape[0])
    for cluster_idx, cluster in enumerate(clusters):
        for sample_idx in cluster:
            labels[sample_idx] = cluster_idx
    return labels
```

Question 7: K-means Clustering (10 points)

In this question, you should finish the "mykmeans" function. Define the K-means clustering algorithm flow based on the above functions.

input:
X: dataset
k: num of centroids
max_iterations: max time of iterations
output:
return the predicted labels of X

Note: you should exactly finish the following 5 serial parts of the function.

In [48]:
```
def mykmeans(X, k, max_iterations):
    # 1. Initialize the centroids
    centroids = centroids_init(k, X)
```

```
    for _ in range(max_iterations):
        # 2. Build clusters according to the current centroids
        clusters = build_clusters(centroids, k, X)

        # 3. Compute the new centroids according to the new clustering results
        new_centroids = calculate_centroids(clusters, k, X)

        # 4. Set the convergence condition as whether the centroids change
        if np.all(centroids == new_centroids):
            break
        centroids = new_centroids

    # 5. Return final labels of each point
    return get_cluster_labels(clusters, X)
```

In [49]:
```
# comparision for check
from sklearn.cluster import KMeans
pred1 = mykmeans(X, 2, 10)
pred2 = KMeans(n_clusters=2).fit_predict(X)
print(pred1, pred2)
```

```
[0. 0. 0. 1. 1.] [0 0 0 1 1]
```

In [51]:
```
ari = adjusted_rand_score(pred1, pred2)
print("Adjusted Rand Index:", ari)
```

```
Adjusted Rand Index: 1.0
```

Good for you! Now you can use your K-means functions to do more wonderful things!

Part 2: Plots and ARI (10 points)

In this part, you should finish question 8.

Now you may wonder given the prediction of K-means clustering method, then how to evaluate the results properly. If the ground truth labels are known, it's possible to use a clustering metric that considers labels in its evaluation. You can use the `scikit-learn` implementation of a common metric called the **adjusted rand index (ARI)**, which uses true cluster assignments to measure the similarity between true and predicted labels. The ARI output values range between -1 and 1. A score close to 0.0 indicates random assignments, and a score close to 1 indicates perfectly labeled clusters. Besides, it would be helpful if we can see the assignment of clusters with **plots**.

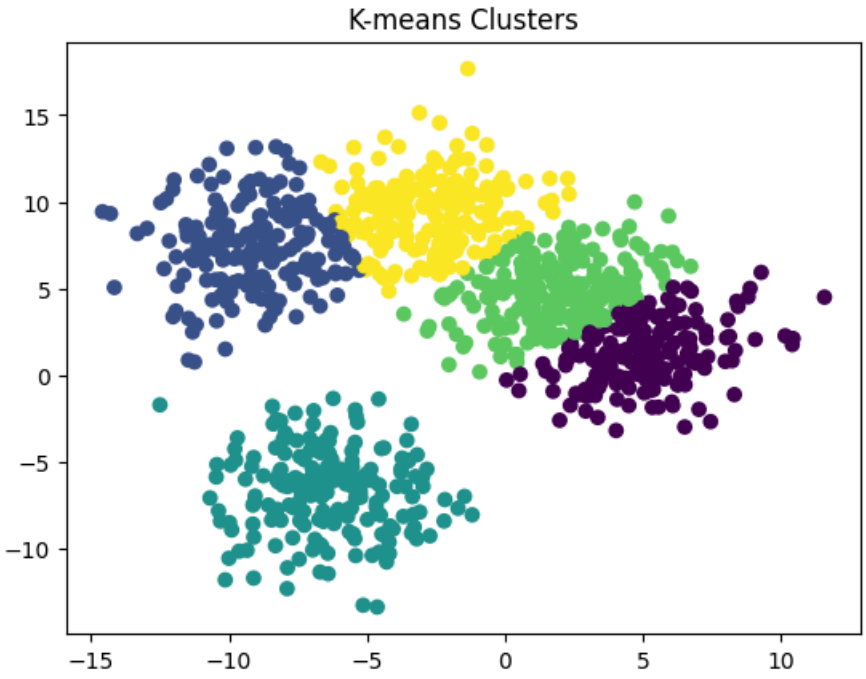
Question 8: Reach higher ARI (10 points)

In this question, you should tune the parameters `n_clusters` and `max_iters` to see better results of your K-means method, the output ARI should be higher than **0.7** with clear plots.

In [60]:
```
features = np.genfromtxt('data_x.csv')
true_labels = np.genfromtxt('data_y.csv')
n_clusters = 5
max_iters = 500

pred = mykmeans(features, n_clusters, max_iters)
plt.plot()
plt.scatter(features[:, 0], features[:, 1], c=pred)
plt.title("K-means Clusters")
ari_kmeans = adjusted_rand_score(true_labels, pred)
print('ARI of K-means with {} clusters and {} max iterations:'.format(n_clusters, max_iters),ari_kmeans)
```

```
ARI of K-means with 5 clusters and 500 max iterations: 0.7090321013591384
```



Part 3: Prediction of Asian Football Teams (20 points)

In this part, you should finish question 8.

Did you watch the recent 2022 World Cup in Qatar? WOW, that would be super exciting! Maybe you are thinking of which football team to celebrate or even pay closer attention to Asian teams. Now imagine that you could have the chance to choose which Asian football teams to be participated in the next round of World Cup with your K-means methods. What a surpise!

In [4]:
```
# data = pd.read_csv('soccer.csv')
# train_x = data[["2019Global","2018WorldCup","2015AsianCup"]]
# df = pd.DataFrame(train_x)
```

In [58]:
```
import pandas as pd
import numpy as np
from sklearn.preprocessing import MinMaxScaler
from sklearn.cluster import KMeans

data = pd.read_csv('soccer.csv')
print(data.columns)
identifier_column = 'Country'

train_x = data[["2019Global", "2018WorldCup", "2015AsianCup"]]

# Normalize the data
scaler = MinMaxScaler()
train_x_scaled = scaler.fit_transform(train_x)

def select_top_teams(kmeans, data, n_teams):
    centroids = kmeans.cluster_centers_
    selected_teams = []

    for i in range(len(centroids)):
        cluster_data = data[data['Cluster'] == i]
        if not cluster_data.empty:
            distances = cluster_data.apply(lambda row: np.linalg.norm(row[["2019Global", "2018WorldCup",
            closest_idx = distances.idxmin()
            selected_teams.append(data.loc[closest_idx, identifier_column])

    return selected_teams[:n_teams]

k_5 = 5
k_9 = 9

kmeans_5 = KMeans(n_clusters=k_5, random_state=42)
```

```
clusters_5 = kmeans_5.fit_predict(train_x_scaled)
data['Cluster_5'] = clusters_5

kmeans_9 = KMeans(n_clusters=k_9, random_state=42)
clusters_9 = kmeans_9.fit_predict(train_x_scaled)
data['Cluster_9'] = clusters_9

# Select 5 teams
data['Cluster'] = clusters_5
top_5_teams = select_top_teams(kmeans_5, data, 5)

# Select 9 teams
data['Cluster'] = clusters_9
top_9_teams = select_top_teams(kmeans_9, data, 9)
```

```
Index(['Country', '2019Global', '2018WorldCup', '2015AsianCup'], dtype='object')
```

Question 9: Choose 5 Teams (10 points)

If you are about to choose **5** Asian teams from the pool, which five teams would you choose and why? Please briefly explain your reasons with your clustering results as output.

Hint: `preprocessing.MinMaxScaler()` **function can help for dataset scaling**

In [56]:
```
print("Top 5 teams:", top_5_teams)
```

```
Top 5 teams: ['China', 'Syria', 'Iran', 'Bahrain', 'Kuwait']
```

The top 5 teams are selected based on their proximity to the centroids of the clusters formed with k=5. These teams are considered representative of different strengths and characteristics across the clusters.

Question 10: Choose 9 Teams (10 points)

It's announced that the 2026 World Cup wil allow 48 teams in total. Now, maybe we could choose up to 9 Asian teams to participate in the contest. So if you are about to choose **9** Asian teams from the pool, which nine teams would you choose and why? Please briefly explain your reasons with your clustering results as output.

Hint: `preprocessing.MinMaxScaler()` **function can help for dataset scaling**

In [57]:
```
print("Top 9 teams:", top_9_teams)
```

```
Top 9 teams: ['China', 'Palestine', 'Iran', 'Kuwait', 'Syria', 'Australia', 'Saudi', 'Bahrain', 'Oman']
```

Similarly, the top 9 teams are selected using k=9. This ensures a broader representation of the diversity in the dataset, which is especially important given the increase in the number of teams for the 2026 World Cup.

The selected teams are chosen to represent different clusters identified by the K-means algorithm, ensuring a diverse set of strengths and capabilities. The normalization step ensures that each feature contributes equally to the clustering, leading to a fair selection of teams. This approach helps in identifying teams that are strong in various aspects of international and regional performance metrics.