Exercício Clustering

Filipe Assis Mourão

8988914

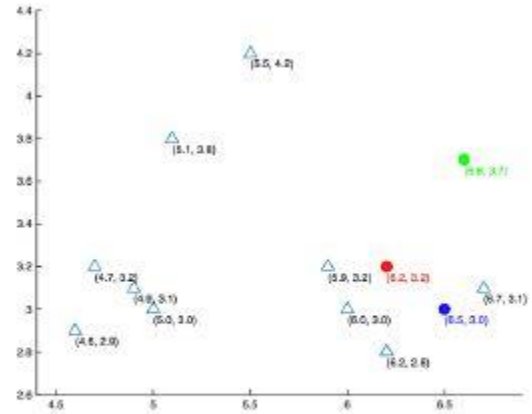


Figure 1: Scatter plot of datasets and the initialized centers of 3 clusters

1.1 Implement k-means manually

Given the matrix $\mathbf{X}$ whose rows represent different data points, you are asked to perform a k-means clustering on this dataset using the Euclidean distance as the distance function. Here k is chosen as 3. The Euclidean distance d between a vector $\mathbf{x}$ and a vector $\mathbf{y}$ both in $\mathcal{R}^p$ is defined as $d = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$. All data in $\mathbf{X}$ were plotted in Figure 1. The centers of 3 clusters were initialized as $\mu_1 = (6.2, 3.2)$ (red), $\mu_2 = (6.6, 3.7)$ (green), $\mu_3 = (6.5, 3.0)$ (blue).

$$\mathbf{X} = \begin{bmatrix} 5.9 & 3.2 \\ 4.6 & 2.9 \\ 6.2 & 2.8 \\ 4.7 & 3.2 \\ 5.5 & 4.2 \\ 5.0 & 3.0 \\ 4.9 & 3.1 \\ 6.7 & 3.1 \\ 5.1 & 3.8 \\ 6.0 & 3.0 \end{bmatrix}$$

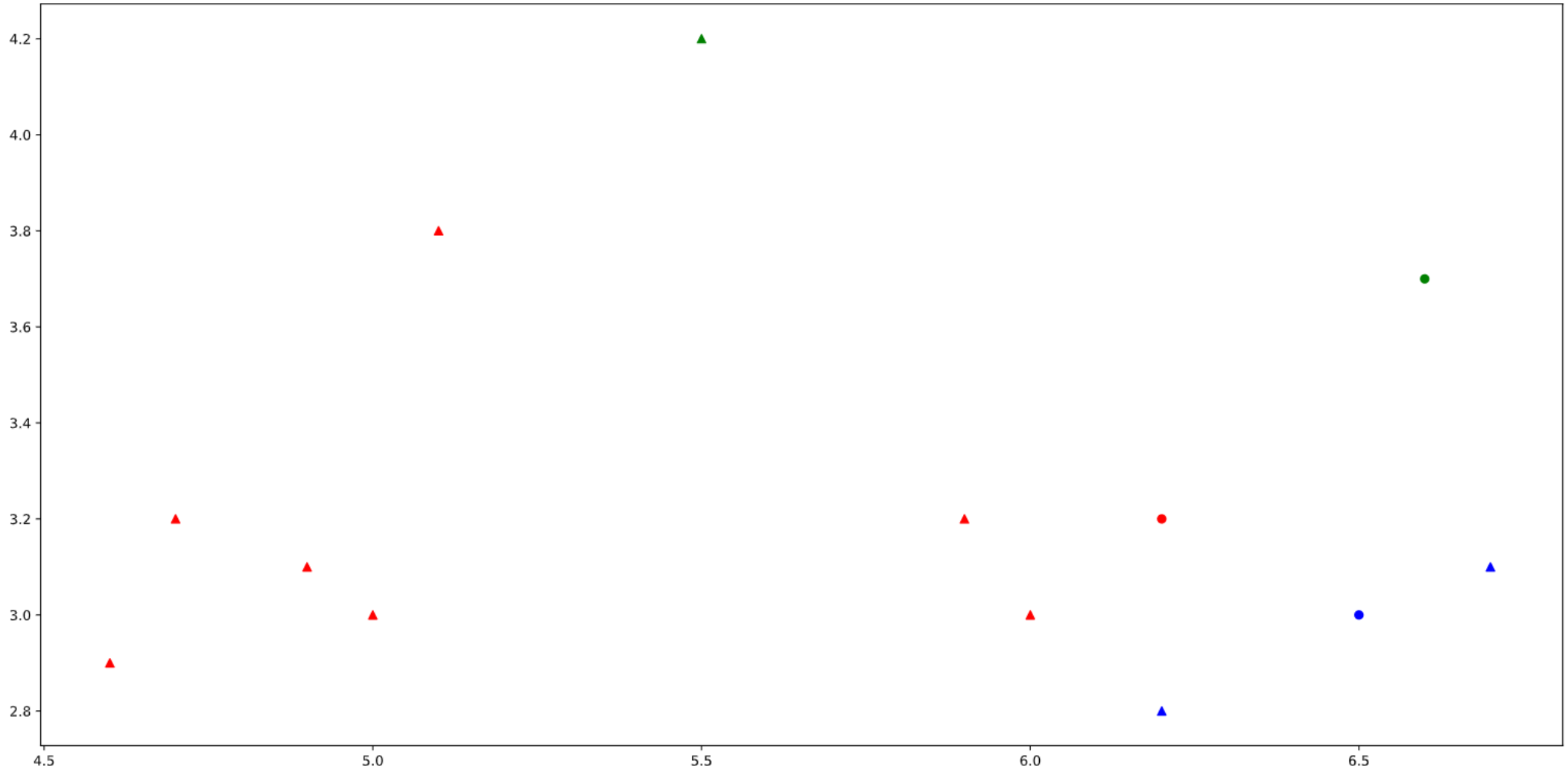
1. What's the center of the first cluster (red) after one iteration? (Answer in the format of **[x1, x2]**, round your results to three decimal places, same as problems 2 and 3)

2. What's the center of the second cluster (green) after two iteration?

3. What's the center of the third cluster (blue) when the clustering converges?

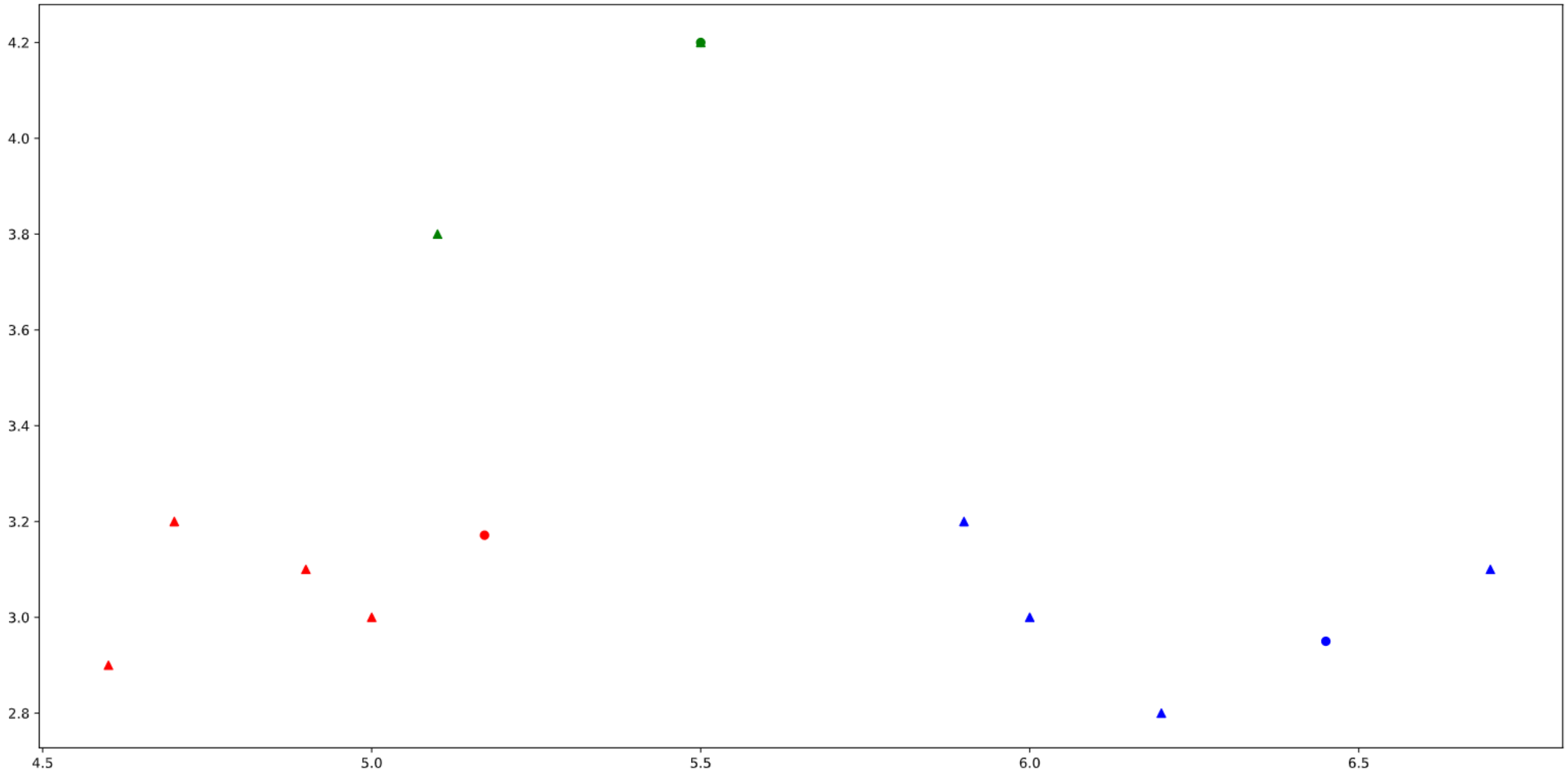
4. How many iterations are required for the clusters to converge?

Configuração inicial



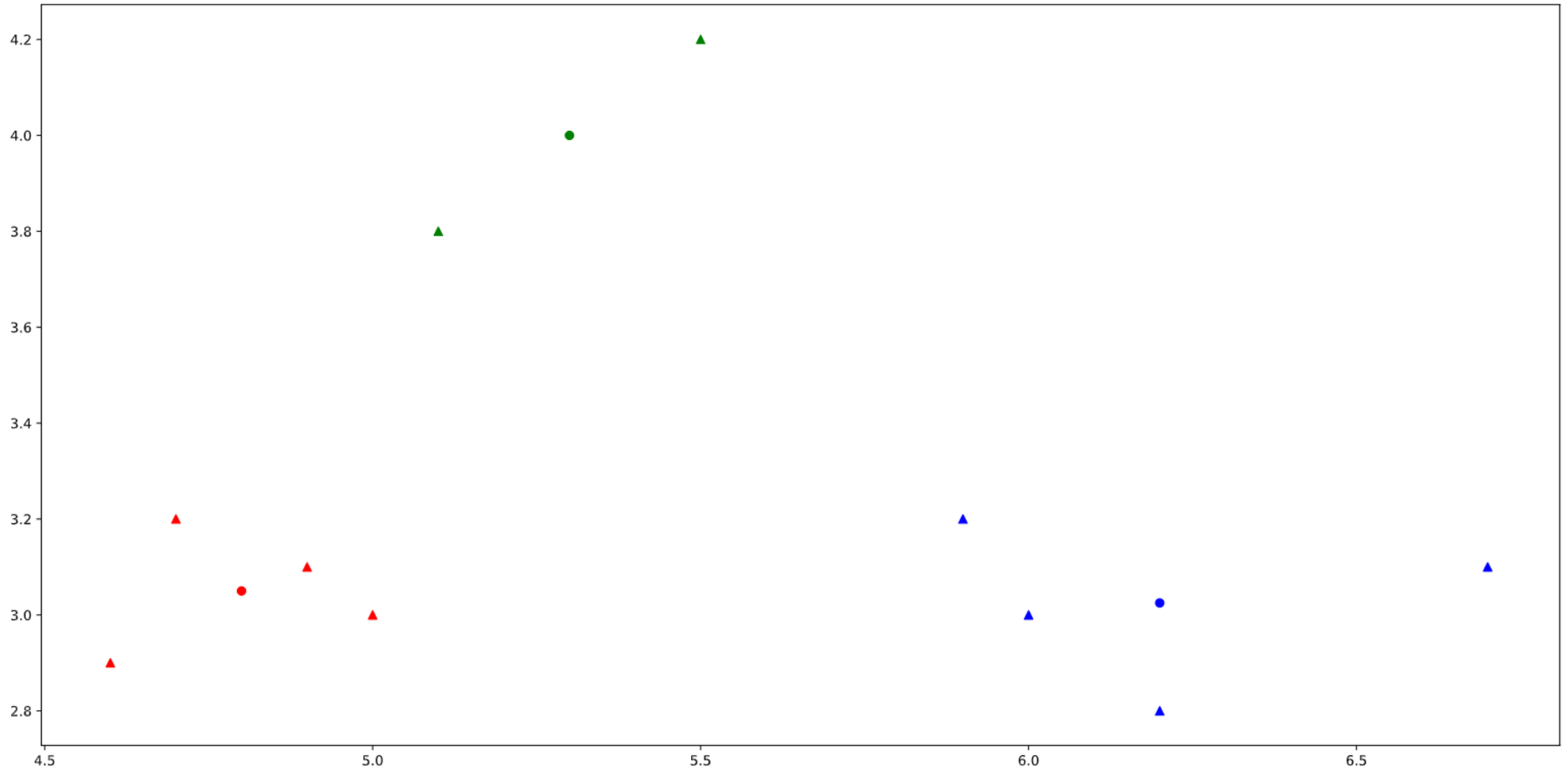
Red cluster = (6.2,3.2) Green cluster = (6.6,3.7) Blue cluster = (6.5,3.0)

Primeira iteração



Red cluster = (5.17,3.17) Green cluster = (5.5,4.2) Blue cluster = (6.45,2.95)

Segunda iteração



Red cluster = (4.8,3.05)

Green cluster = (5.3,4.0)

Blue cluster = (6.2,3.025)

A partir da terceira iteração os clusters não mudam de lugar, dessa forma podemos concluir que são necessárias duas iterações para convergência

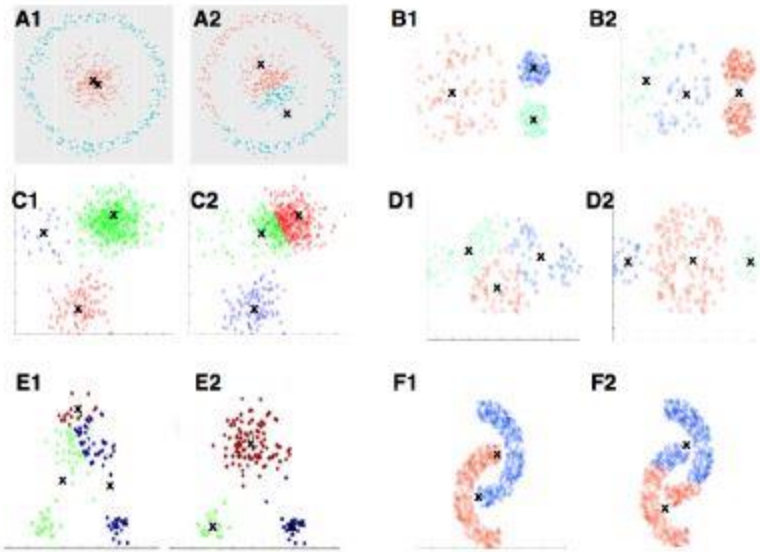


Figure 2: Clustered results for 6 datasets

is more likely to be generated by K-means method. (Hint: check the state when K-means converges; Centers for each cluster have been noted as **X**; Since x and y axis are scaled proportionally, you can determine the distance to centers geometrically). The distance measure used here is the Euclidean distance.

1. Dataset A (write A1 or A2, [ ] same in the following question);
2. Dataset B
3. Dataset C
4. Dataset D
5. Dataset E
6. Dataset F

Devemos descartar todos os casos em que há uma separação não linear, logo:

A – A2

B – B2

C – C2

D – D1

E – E2

F – F2

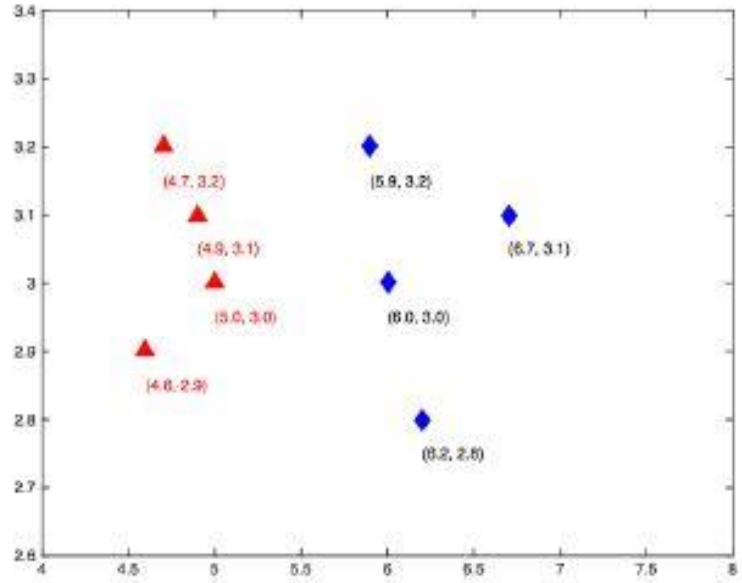


Figure 3: Scatter plot samples in two clusters

1.3 Hierarchical clustering

In Figure 3 there are two clusters A (red) and B (blue), each has four members and plotted in Figure 3. The coordinates of each member are labeled in the figure. Compute the distance between two clusters using Euclidean distance.

1. What is the distance between the two farthest members? (complete link) (round to **four** decimal places here, and next 2 problems);

2. What is the distance between the two closest members? (single link)

3. What is the average distance between all pairs?

4. Among all three distances above, which one is robust to noise? Answer either "complete", "single", or "average".

1. Distância dos dois membros mais distantes:

euclidianDistante([4.6,2.9],[6.7,3.1]) =2.1095

2. Distância dos dois membros mais próximos :

euclidianDistante([5.0,3.0],[5.9,3.2]) =0.9219

3. Distância média de todos os pares

avgEuclidianDistance = 1.4128

4. O método “average” é mais robusto à ruídos, tendo em vista que mais pontos são considerados e os “outliers” que poderiam estar muito próximos ou distantes tem seu efeito reduzido