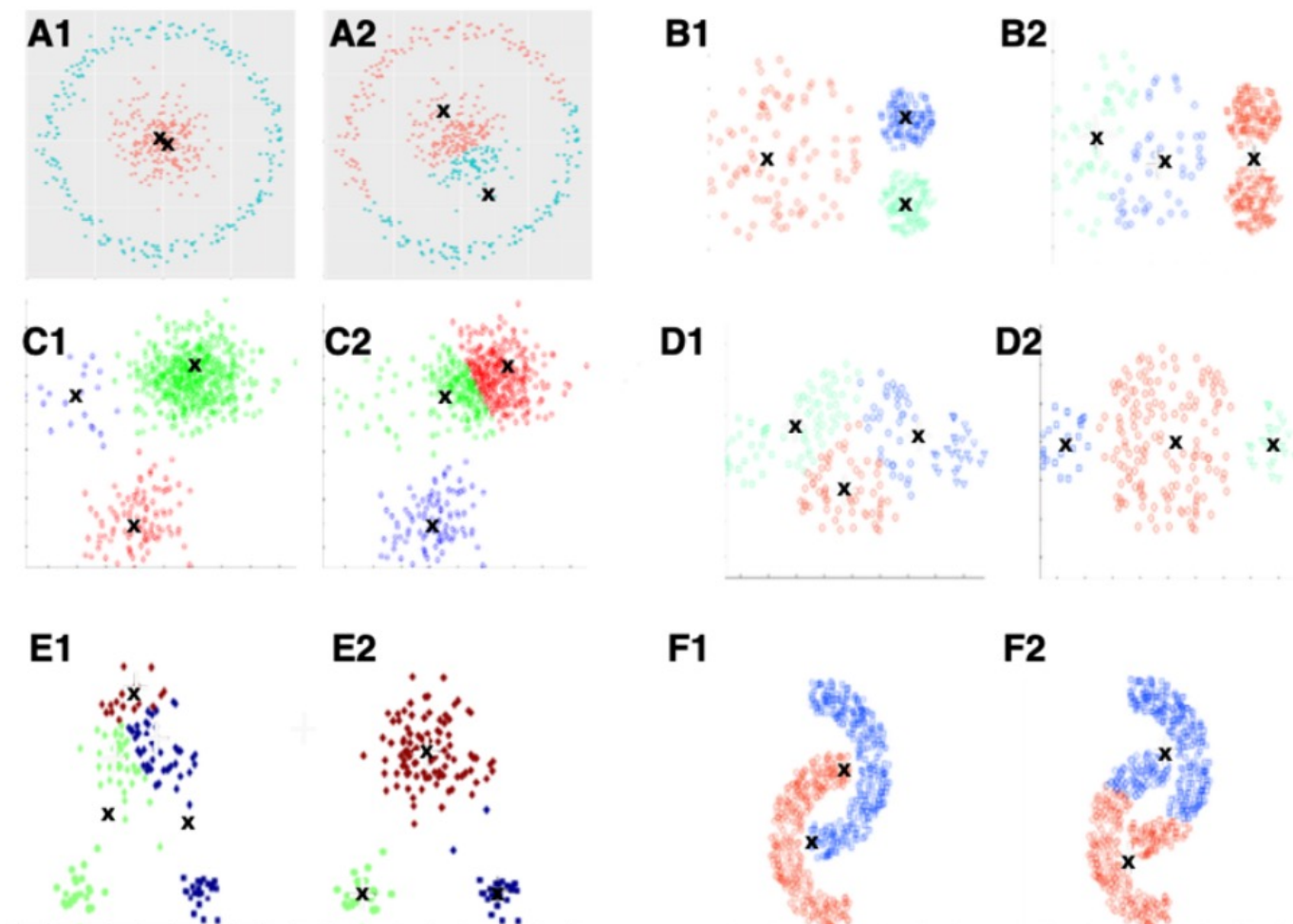




作业

1. 下图给出6个数据集A-F分别用两种算法得到的聚类结果，其中一种是K均值聚类。请问哪些最可能是K均值聚类的结果？如果K均值聚类结果不够理想，建议采用哪种聚类算法？

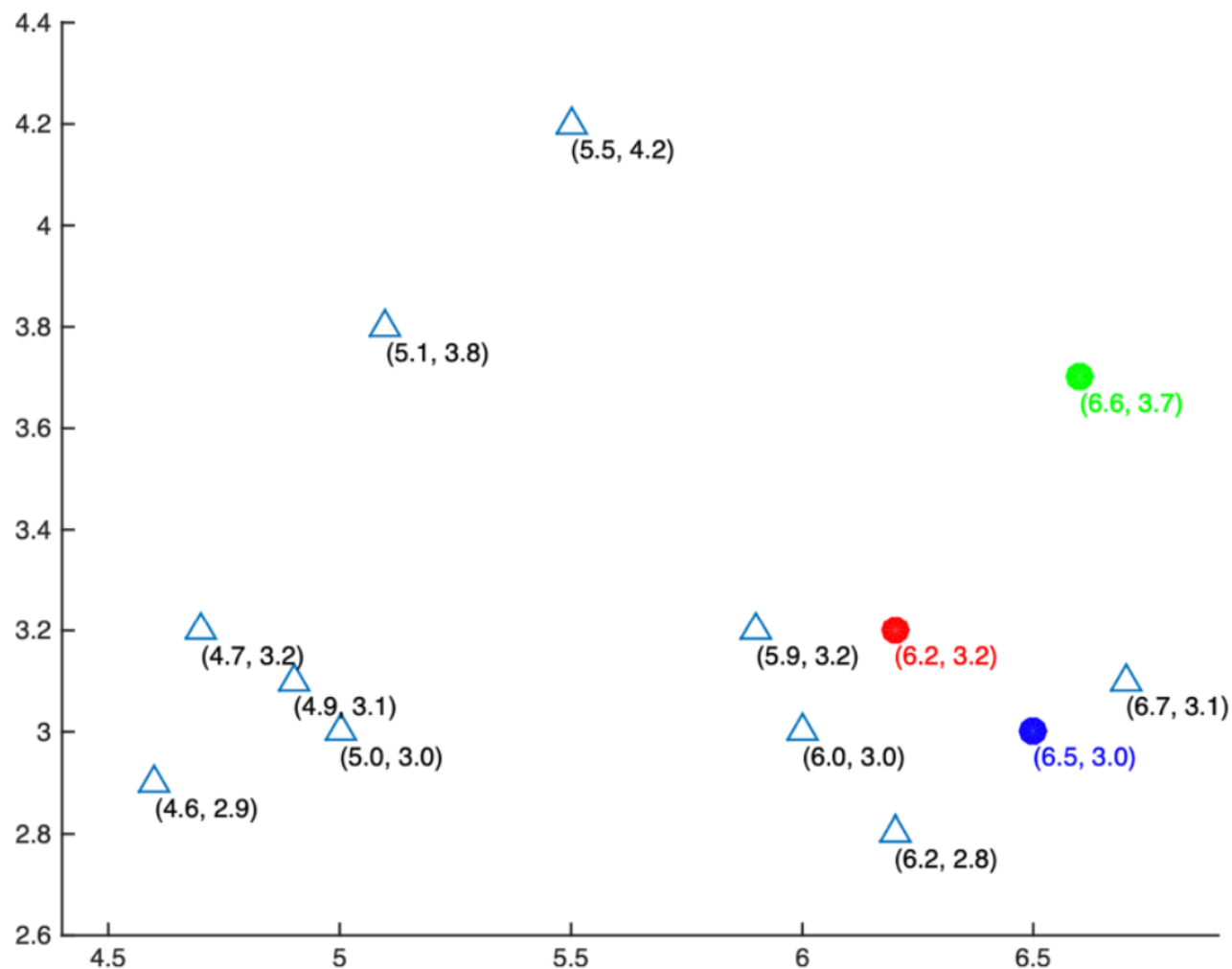




2. 对如图所示的数据集，采用 K 均值聚类。设 $K = 3$ ，3 个聚类中心分别为

$\mu_1 = (6.2, 3.2)^T$ (红色), $\mu_2 = (6.6, 3.7)^T$ (绿色), $\mu_3 = (6.5, 3.0)^T$ (蓝色)

请给出一次迭代后属于第一簇的样本及更新后的簇中心(保留两位小数)



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



• 3. 简答题

1	K均值聚类算法隐含了对簇的什么假设？
2	K均值迭代一定会收敛吗？
3	K均值算法的初始化参数怎么选择？
4	K均值算法的超参K怎么选择？
5	K均值算法有什么局限性？可能用什么方法改进？
6	基于GMM的EM聚类算法为什么要引入隐变量 z ？
7	K均值和EM算法有什么共通之处？
8	如何度量簇的相似性？不同度量方法的优缺点？
9	层次聚类的限制有哪些？
10	请简述PCA流程。