Задача 3

Постановка задачи

Провести кластерный анализ точек на плоскости.

Решение

С помощью иерархической кластеризации на малой выборке получил приближенные кластеры. Затем запустил k-means на полученных центрах.
Полученные данные:
Как можно заметить ниже, K-means плохо справляется с задачей, когда кластеры распложены слишком близко, либо перекрываются, поэтому нужно использовать другой способ кластеризации. Чем сильнее разграничены кластеры, тем для K-means лучше.

Начальные данные/Иерархическая кластеризация/K-means

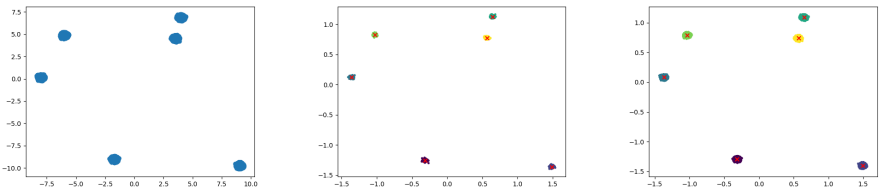

Рис. 1:
Так как кластеры расположены далеко друг от друга, то мы можем удачно использовать K-means.

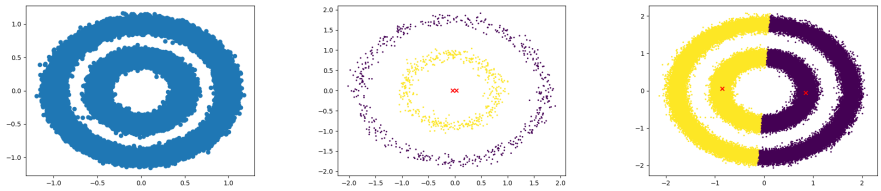

Рис. 2:
Иерархическая кластеризация дает понять, что центр одного кластера лежит "внутри" другого кластера, что говорит о невозможности использовать K-means.

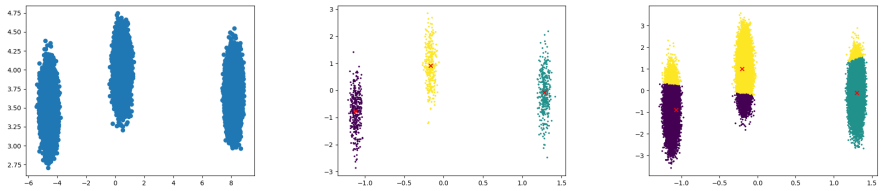

Рис. 3:
Несмотря на то, что иерархическая кластеризация правильно выделила кластеры, которые находятся на заметном расстоянии, K-means все равно не справляется полностью с задачей из-за вытянутой формы кластеров.

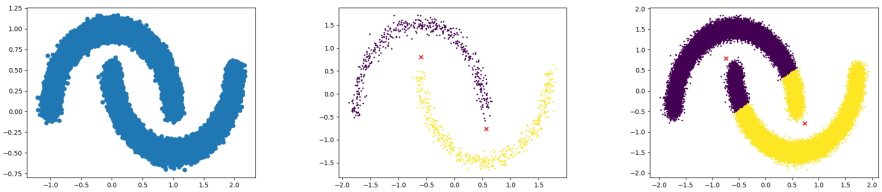

Рис. 4:
Иерархическая кластеризация дала правильное разбиение на кластеры, но из-за того, что центры каждого кластера расположены близко к точкам другого, K-means не может полностью правильно разделить данные.

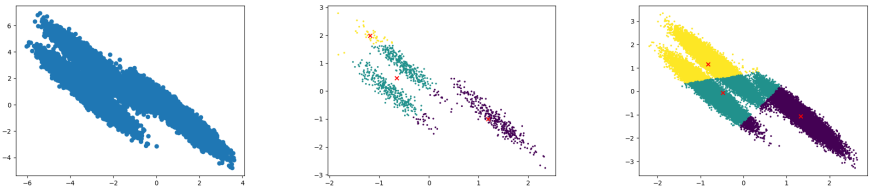

Рис. 5:
Из-за сильной близости кластеров и их формы не справляется правильно с задачей уже иерарахическая.

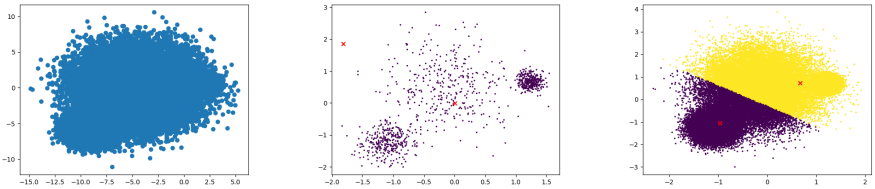

Рис. 6:
Кластеры сильно пересекаются, что делает кластеризацию невозможной.

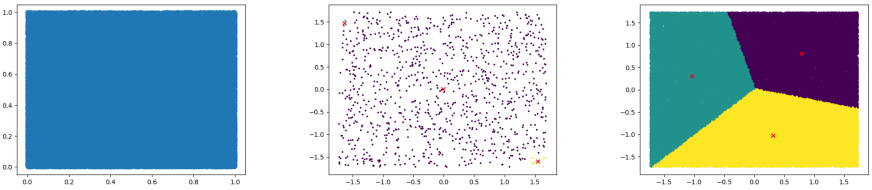

Рис. 7:
Кластеризация изначально не имеет смысла, так как данные неразделимы.