Семинар 2-3: сегментация клиентов и кластеризация

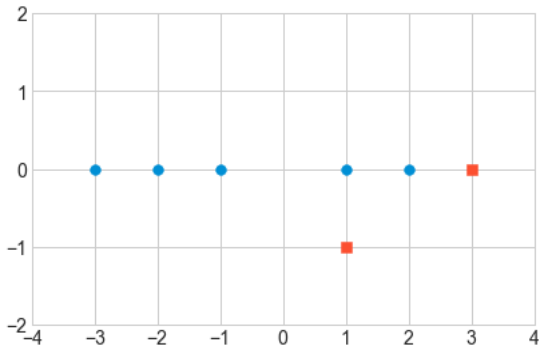
Задача 1

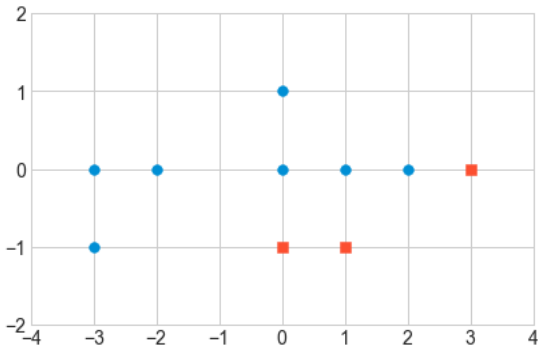
Тут будет задача про расстояния. В первом пункте между несколькими точками эти расстояния надо будет посчитать. Во втором пункте будут рисунки и вопросы какое расстояние лучше использовать.

Идеи рисунков: две точки в поле — евклидово, между небоскребами по улицам — манхеттенское и тп.

Задача 2

На картинках ниже синими точками отмечены наблюдения. Красными точками отмечены стартовые центроиды для алгоритма K-means.





Кластеризуйте данные на первой картинке на два кластера, на второй картинке на три кластера. Сколько итераций понадобилось сделать до полной сходимости алгоритма? Сколько объектов вошли в каждый из кластеров?

а) Используйте для кластеризации Евклидово расстояние.

б) Используйте для кластеризации Манхеттенское расстояние.

в) В этой задачке мы сами предложили вам для кластеризации начальные точки (красные квадраты). На практике начальное приближение центроидов обычно генерирует компьютер. Изменится ли разбиение на кластеры, если изменить стартовые точки?

Задача 3

Начальник Аристарх был в командировке. Там он услышал про иерархическую агломеративную кластеризацию. По приезду, находясь в состоянии восторга, он записал в

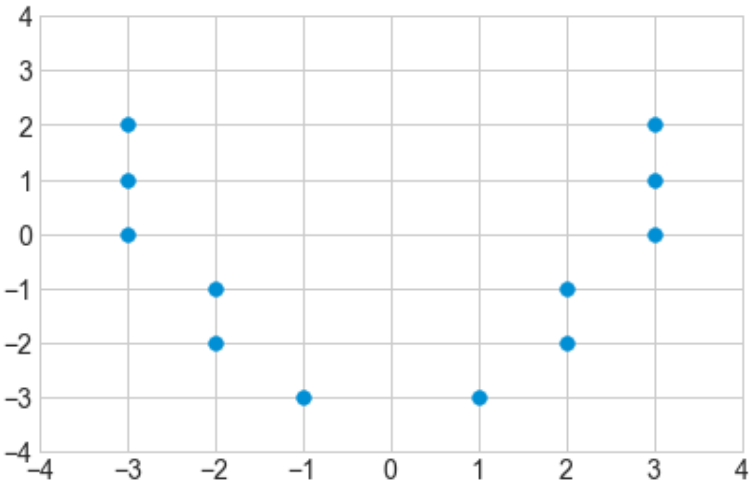
свой блокнот следующие четыре наблюдения:

x	z
8	6
6	10
2	4
4	2

После он отдал блокнот маркетологу Савелию. Аристарх хочет, чтобы Савелий провел агломеративную иерархическую кластеризацию. На совещаний было решено использовать в качестве расстояния между объектами обычное Евклидово расстояние. Расстояние между кластерами решено определять по принципу дальнего соседа. Помогите Савелию с агломеративной иерархической кластеризацией. И не забудьте нарисовать дендрограмму. Начальники любят красивые картинки.

Ещё задачи!

Задача 4



1. Примените метод K-means с $K = 2$, $K = 3$, $K = 4$ и $K = 5$. Начальные точки каждый раз выбирайте случайно. Для всех ли начальных точек кластеризация каждый раз будет выдавать один и тот же результат?
2. Примените метод агломеративной иерархической кластеризации. Нарисуйте дендрограмму. Руководствуясь дендрограммой выберете оптимальное количество кластеров. Обоснуйте свой выбор.
3. Правда ли, что для всех рассмотренных K оба метода разбивают выборку на одинаковые кластеры? Всегда ли так происходит? Приведите контр-пример.
4. Сюда вопрос про выбор оптимального k по межкластерному расстоянию и силуэту.

Задача 5

Обозначьте расположение центроидов и границ кластеров после применения метода K-means с $K = 2$ на следущих данных: