

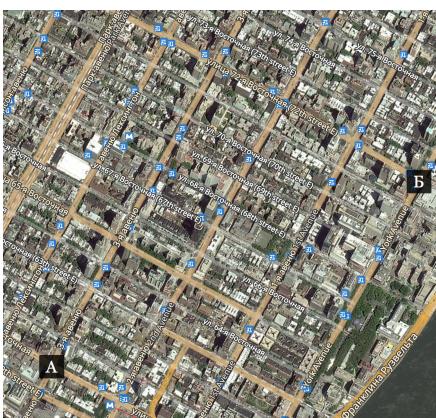
Семинар 2-3: сегментация клиентов и кластеризация

Задача 1

- а) У нас есть точки A(1, 1), B(2, 2), C(3, 0). Нарисуйте их на плоскости и посчитайте между ними расстояние Евклида, Манхэттенское и Чебышева:

$$\begin{aligned}\rho_1(A, B) &= \sqrt{(x_B - x_A)^2 + (y_B - y_A)^2} \\ \rho_2(A, B) &= |x_B - x_A| + |y_B - y_A| \\ \rho_3(A, B) &= \max(|x_B - x_A|, |y_B - y_A|)\end{aligned}\tag{1}$$

- б) Какое расстояние будете использовать для того, чтобы добраться из точки А в точку Б? Почему?



- в) Какое расстояние вы бы использовали для измерения похожести текстов или генома?

В тексте была сделана опечатка

В тексте была сделано очепятка

CTGGG**CT**AAAAGGTCCCTTAGCC..TTTAGAAAAA.GGCCATTAGG**AA**ATTGC
CTGGG**ACT**AAA....CCTTAGC**CT**ATTT**A**AAAAATGGCCATTAGG...TTGC

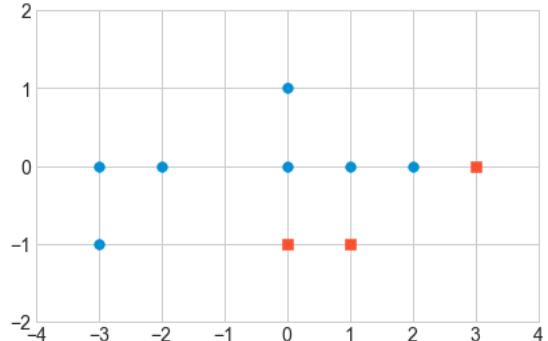
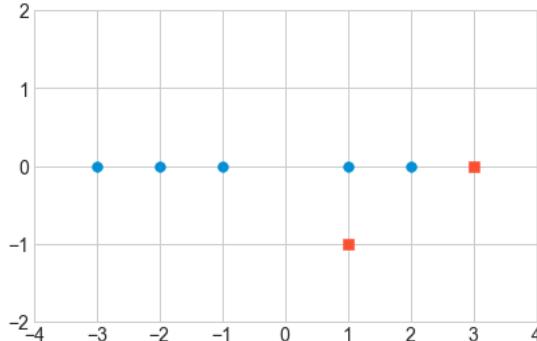
Задача 2

Жокей Святополк решил открыть несколько новых ларьков с шаурмой¹. Перед открытием он подумал о потенциальных покупателях и выяснил, где на районе находятся общежития. На картинках ниже они отмечены синими точками. Святополк понимает, что все общежития, расположенные в районе, можно сегментировать по их географическому положению и, исходя из этого, расположить палатки с шаурмой. Сделать это ему хотелось бы с помощью алгоритма K-means:

1. Ставим ларьки с шаурмой в случайных местах;

¹По мотивам https://vas3k.ru/blog/machine_learning/

2. Смотрим в какой кому ближе идти;
3. Двигаем ларьки ближе к центрам их популярности;
4. Снова смотрим и двигаем;
5. Повторяем так много раз, пока алгоритм не сойдётся и движение не прекратится.



Красными точками отмечены стартовые точки для палаток. В первом районе Святополк ставит две палатки, во втором районе три палатки. Помогите Святополку с сегментацией! Сколько итераций понадобилось сделать до полной сходимости алгоритма? Сколько объектов вошли в каждый из кластеров?

- Используйте для кластеризации Евклидово расстояние.
- Используйте для кластеризации Манхэттенское расстояние.
- 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:

n	x	z
1	8	6
2	6	10
3	2	4
4	4	2

После он отдал блокнот маркетологу Савелию. Аристарх хочет, чтобы Савелий провел агломеративную иерархическую кластеризацию. На совещании 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.

Задача 4

Маркетолог с аналитическим складом ума Оля (она кстати говоря ещё и фрилансер) занимается заказом от туристической фирмы. Ей нужно сделать с помощью методов машинного обучения сегментацию клиентов. На этапе предобработки фичей Оля столкнулась с двумя проблемами: категориальной переменной x , в которой записаны курорты и текстовой переменной z , в которой дан отзыв о курорте:

n	x	z
1	Испания	Нежился на пляже
2	Крым	Копали яму на пляже
3	Дача	Копал картошку
4	Крым	Ел картошки и картошку

- а) Что такое категориальная переменная? Почему её надо как-то предобрабатывать?
- б) Почему нельзя сделать замену Крым = 1, Дача = 2, Испания = 3? Что такое ОНЕ-кодирование?
Как будет выглядеть наша табличка с переменными после ОНЕ?
- в) Почему нельзя сделать ОНЕ для текстовой переменной?
- г) Какие этапы предобработки тестов вы знаете? На самом деле вы их скорее всего не знаете и мы их сейчас обсудим.
- д) Сделайте для корпуса текстов из задачки tf-idf.

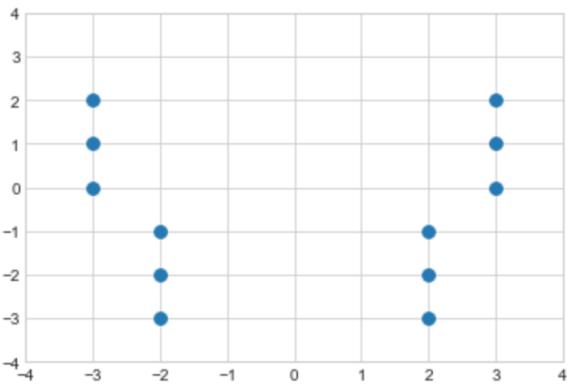
Ещё задачи!

Задача 5

Представьте себе, что вы шахматная фигура. Какое расстояние вы бы использовали, чтобы измерить насколько далёкий путь вам предстоит пройти по шахматной доске. Придумайте метрику для каждой шахматной фигуры: ладья, слон, ферзь, король, слон, конь. Попытайтесь формализовать эту метрику в виде формулы.

Подсказка: попробуйте нарисовать шахматную доску на бумажке, поставить в какую-нибудь клетку фигуру, а после для каждой клетки выписать цифру: сколько до этой клетки идти фигуре. Это поможет придумать для формулы общий вид. Клетку $a1$ в координатной сетке примите за $(0, 0)$.

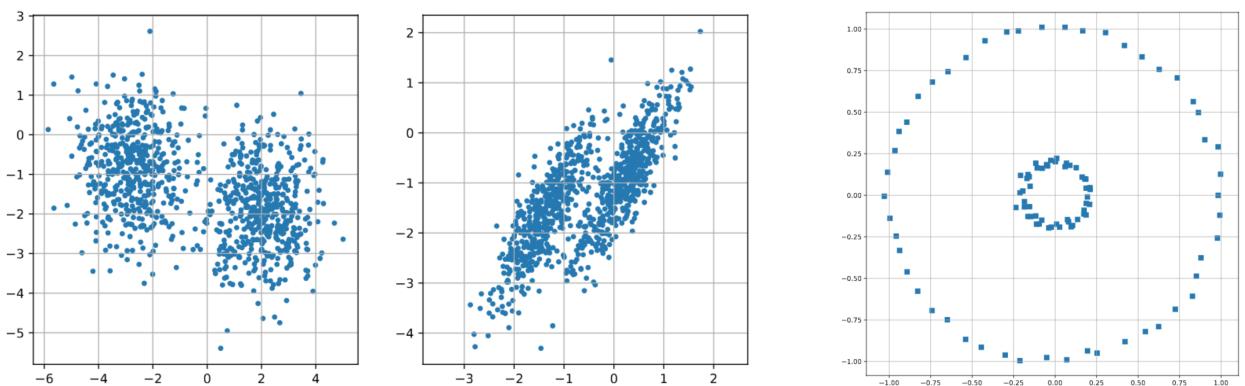
Задача 6



- Примените метод K-means с $K = 2$, $K = 3$, $K = 4$ и $K = 5$. Начальные точки каждый раз выбирайте случайно. Спишитесь с другими людьми и выясните какие точки они выбрали для инициализации. Для каких K у вас получились одинаковые результаты? Почему?
- Правда ли, что для всех рассмотренных K оба метода разбивают выборку на одинаковые кластеры? Всегда ли так происходит? Докажите это или приведите контр-пример.
- В первом пункте вы попробовали провести кластеризацию для разных K . Какое из K является оптимальным? Для того, чтобы определить это используйте сумму квадратов расстояний от точек до центров кластеров.
- Примените метод агломеративной иерархической кластеризации. Нарисуйте дендрограмму. Руководствуясь дендрограммой выберете оптимальное количество кластеров. Обоснуйте свой выбор.

Задача 7

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



Для каких из этих ситуаций имеет смысл попробовать отличные от k – means алгоритмы? Если вы подумали, что это изи-задание и на третьей картинке сходу взяли в один кластер внутреннюю окружность, а во второй внешнюю - вы балбес. Это неправильно. Идите и подумайте ещё.