

Семинар 2-3: сегментация клиентов и кластеризация

Задача 1

- 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:

$$\rho(A, B) = \sqrt{(x_B - x_A)^2 + (y_B - y_A)^2}$$

$$\rho(A, B) = |x_B - x_A| + |y_B - y_A|$$

$$\rho(A, B) = \max(|x_B - x_A|, |y_B - y_A|)$$

Посмотрите на то, какие точки с точки зрения каких расстояний равнодалены друг от друга, а какие нет. Попробуйте врубить интуицию на все 100% и проинтерпретировать это.

- Какое расстояние будете использовать для того, чтобы посчитать расстояние между точками А и Б для случаев ниже? Почему?



- Какую метрику вы бы использовали для измерения похожести текстов или генома?

В тексте была сделана опечатка

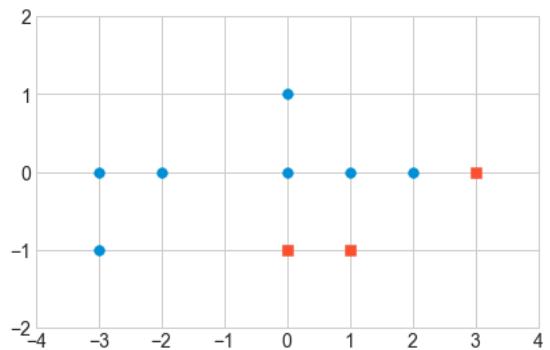
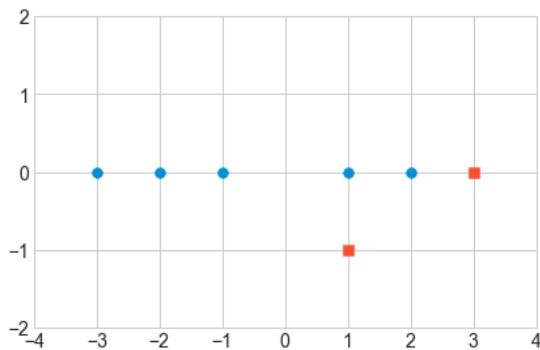
В тексте была сделано очепятка

CTGGG**CT**AAAAGGTCCCTTAGCC..TTTAGAAAAA.GGGCCATTAGG**AA**TTGC
CTGGG**ACT**AAA....CCTTAGCCTATTT**C**AAAAATGGGCCATTAGG...TTGC

Задача 2

Жокей Святополк решил открыть несколько новых ларьков с шаурмой¹. Перед открытием он провёл анализ рынка 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;
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 прекратиться.



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?

- a) Используйте для кластеризации Евклидово расстояние.
- b) Используйте для кластеризации Манхэттенское расстояние.
- c) 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:

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

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 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 на tf idf

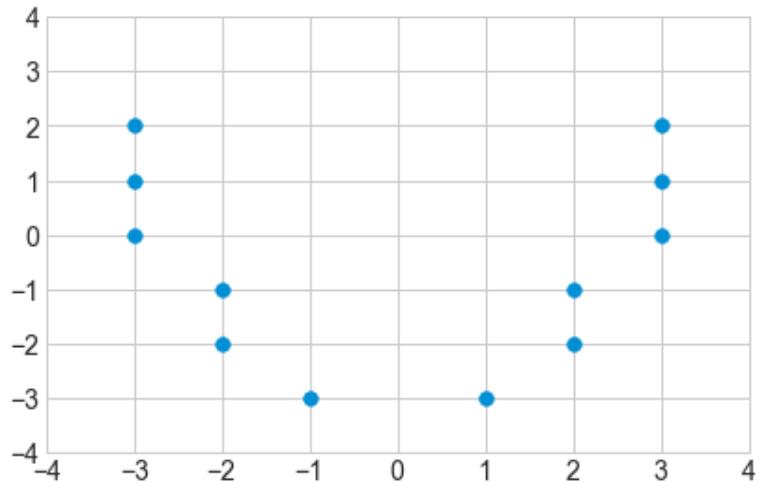
Напомнить про ОНЕ и как предобрабатывать фичи

Ещё задачи!

Задача 4

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 сетки примите за $(0, 0)$.

Задача 5



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 6

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

