Министерство науки и высшего образования Российской Федерации

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ

ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ (НИ ТГУ)

Институт прикладной математики и компьютерных наук

Кафедра прикладной информатики

**Отчет по лабораторной работе №1 по дисциплине «Интеллектуальные системы»**

Выполнил

студент группы № 1462

\_\_\_\_\_\_\_\_\_\_\_\_А.В. Шабля

Проверил

доцент, канд. техн. наук

\_\_\_\_\_\_\_\_\_\_\_\_С.В. Аксёнов

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Томск – 2020

**Описание**

Для выполнения работы был выбран набор данных Spatial Location of Bramble Canes (https://vincentarelbundock.github.io/Rdatasets/csv/boot/brambles.csv). Это данные из Diggle, P.J. (1983) Statistical Analysis of Spatial Point Patterns. Academic Press.

Размер выборки: 823 элемента.

Атрибуты:

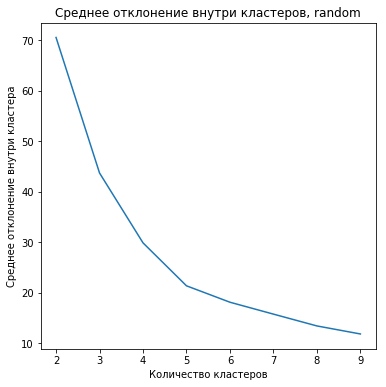
* x – координата ‘x’ положения тростника на графике.
* y – координата ‘y’ положения тростника на графике.
* age – возрастная классификация тростника; 0 обозначает недавно появившийся, 1 обозначает тростник одно- и 2 обозначает тростник двухлетнего возраста.

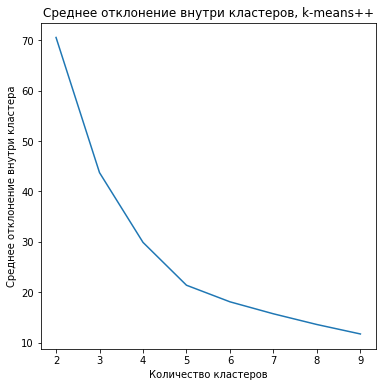
Первая колонка (там содержатся порядковые номера записей), для удобства, была названа «ID». Были убраны кавычки у остальных заголовков, так как данные в кавычках не считываются. В остальном, данные оказались целостными и без пропусков, поэтому дополнительных обработок не потребовалось.

**Кластеризация К-средних**

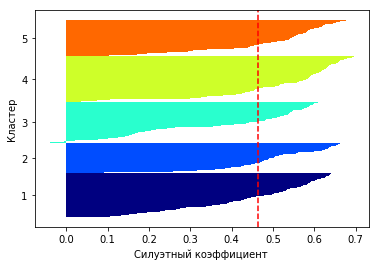
Для кластеризации будем использовать библиотеку sklearn. Библиотечная реализация k-means, по умолчанию, выполняет алгоритм несколько раз с различными центрами, выбирая в качестве конечного результата кластеры с наименьшим внутрикластерным отклонением. Число повторений можно переопределить параметром n\_init, по умолчанию он равен 10. Способ выбора центроидов задаётся параметром init. По умолчанию используется улучшенный метод k-means++, но сначала воспользуется случайным выбором. Выполним кластеризацию с разным желаемым количеством кластеров и посмотрим на изменение среднего внутрикластерного отклонения.

Для начала применим графический инструмент, так называемый метод локтя, для оценки оптимального числа k кластеров для поставленной задачи.



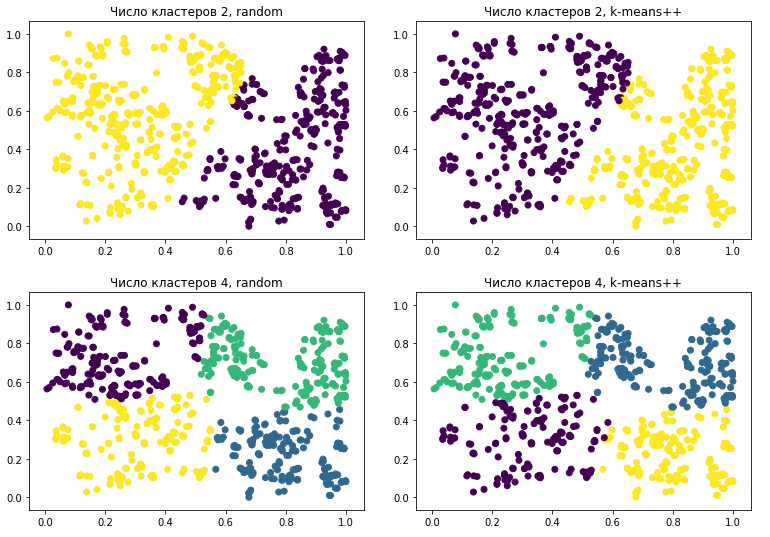


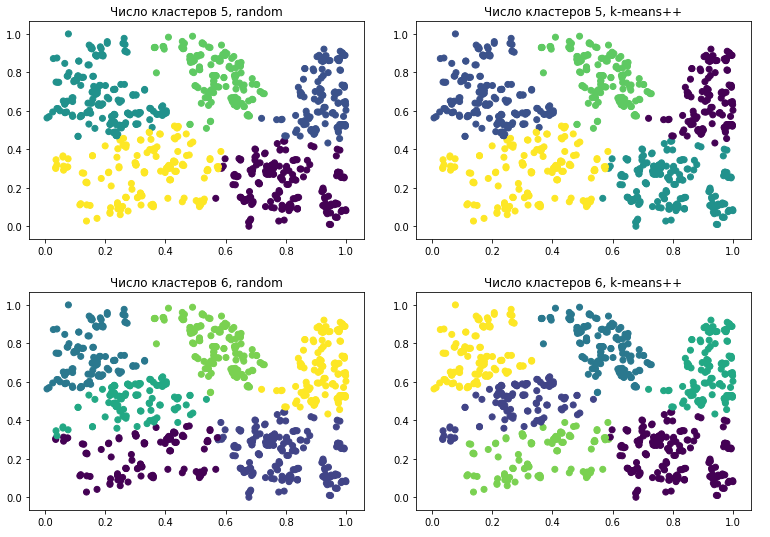
На графике видим, что скорость уменьшения отклонения постепенно падает, а после 5 кластеров практически не меняется. Число 5 является стабильно оптимальным.

Визуальный анализ силуэтного графика позволяет быстро рассмотреть размеры разных кластеров и идентифицировать кластеры, которые содержат выбросы:

Чтобы резюмировать качество нашей кластеризации, мы добавили в график средний силуэтный коэффициент (пунктирная линия).

Оценим визуально кластеры, получившиеся при разном параметре.

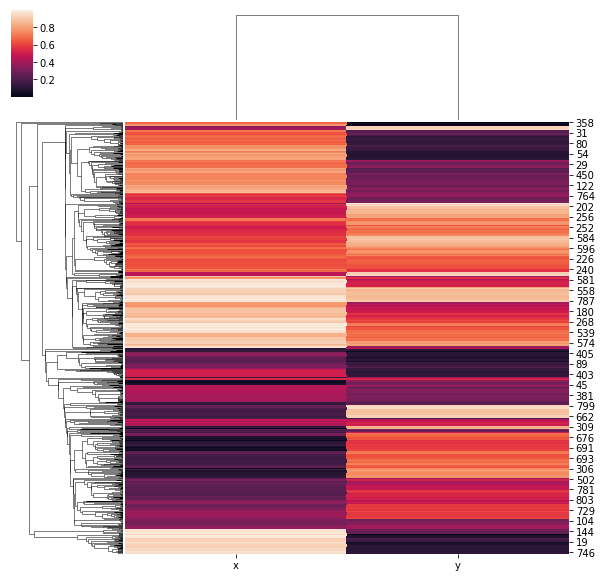
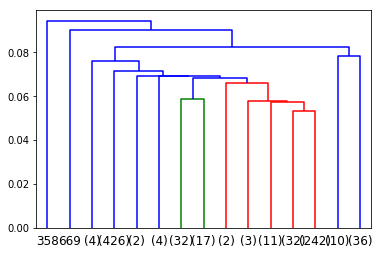




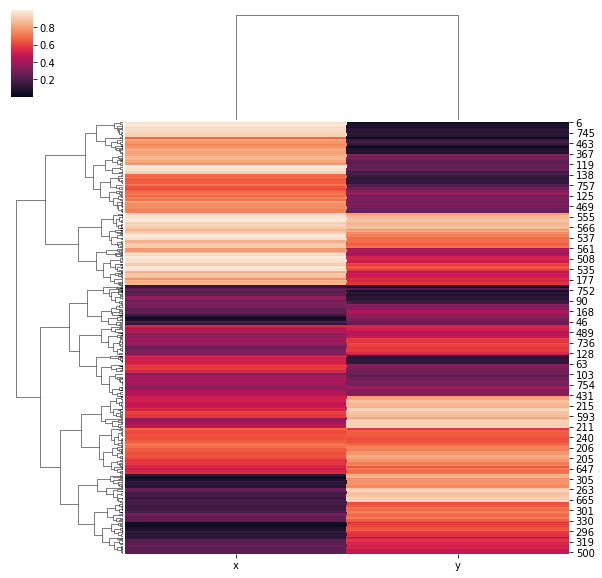
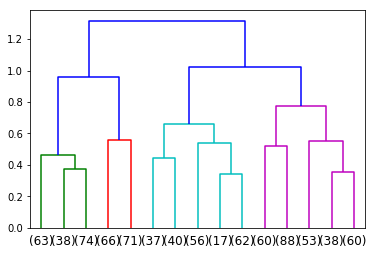
При кластеризации на 5 кластеров оба метода выбора центроидов показали себя отлично.

**Иерархическая кластеризация**

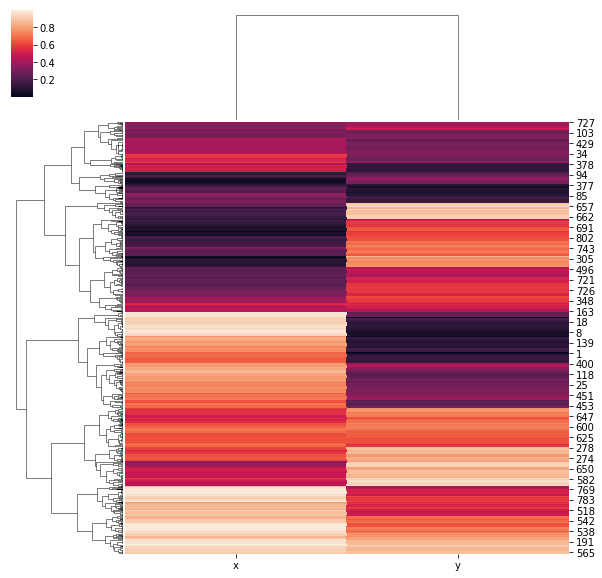
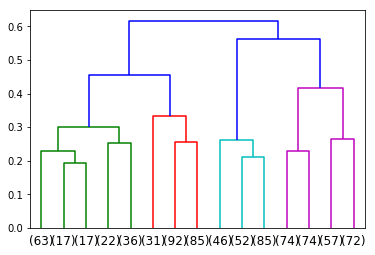
Для кластеризации будем использовать библиотеку scipy и seaborn для тепловой карты. Ниже приведено сравнение разбиения на кластеры при разных способах измерения расстояния между кластерами - single, complete, average, ward.



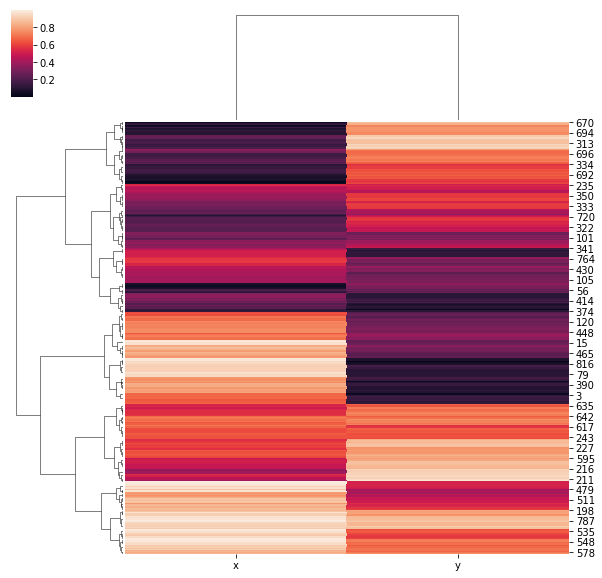
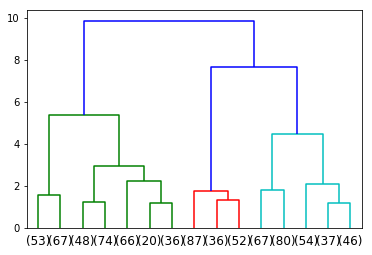
Single - расстояние между кластерами как расстояние между самыми близкими точками из разных кластеров.



Complete - расстояние между кластерами как расстояние между самыми отдаленными точками из разных кластеров.



Average - среднее расстояние по декартовому произведению точек двух кластеров.



Ward - иcпользуется метод Уорда.

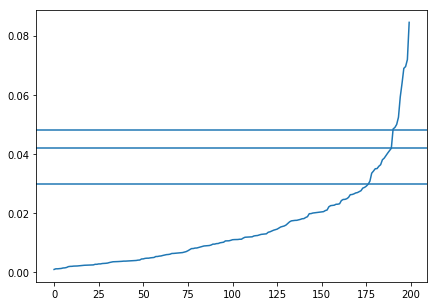
**Кластеризация DBSCAN**

Для кластеризации методом DBSCAN воспользуемся реализацией из библиотеки sklearn. Модель предоставляет 2 основных параметра:

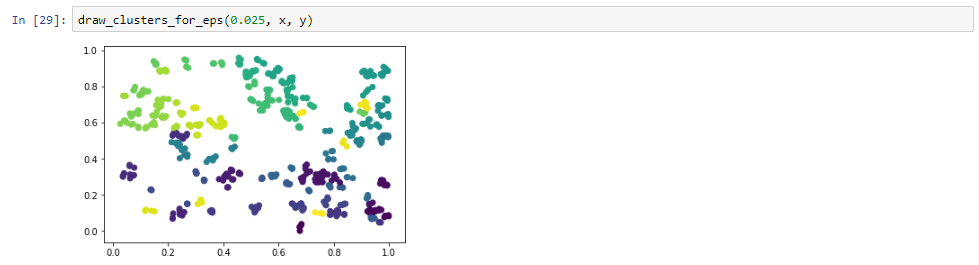
eps - максимальное расстояние между двумя точками, при котором они распознаются как соседи

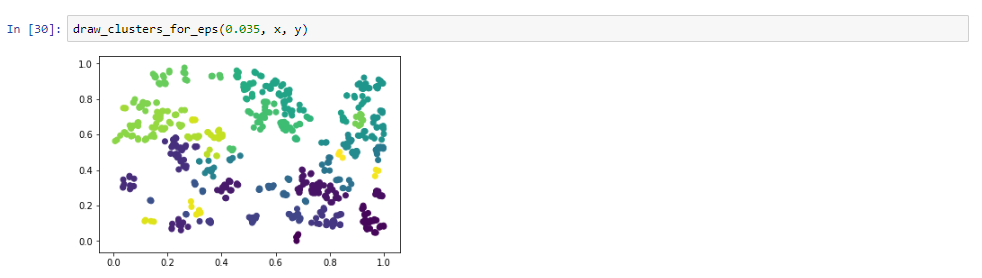
min\_samples - минимальное количество точек, которые могут образовать кластер, возьмём минимальное значение - 3

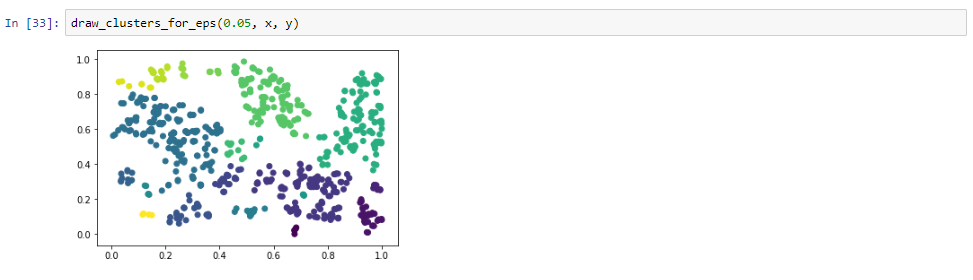
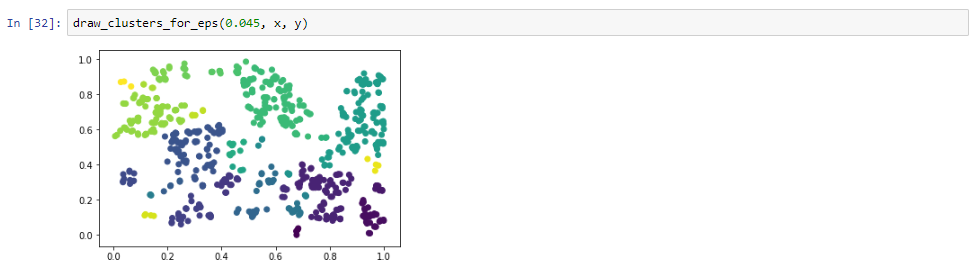
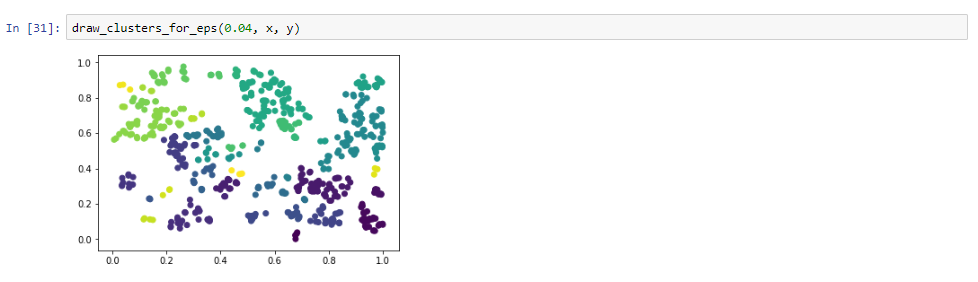
Посмотрим, как изменяется среднее попарное расстояние между точками при различном eps.

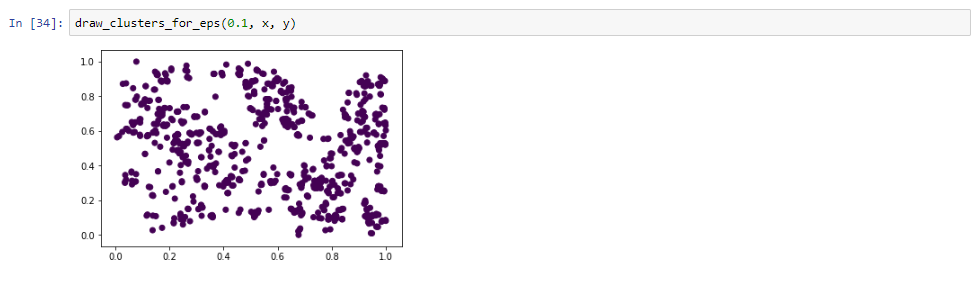


Видим, что большинство точек имеет среднее расстояние < 0.03, после чего видны два небольших скачка, и, в конце концов, наблюдается резкий скачок. Рассмотрим полученные кластеры при eps = (0,025, 0.035, 0.04, 0.045, 0.05). Алгоритм помечает специальным лейблом -1 точки, которые он посчитал шумом. Их будем убирать с графика.









**Выводы:**

На практике были изучены методы преобразования данных, а так же три алгоритма кластеризации.

Лучше всего на взятых данных себя показал алгоритм К-средних при 5 центроидах. Учитывая, что кластеризация является методом обучения без учителя, мы не могли себе позволить такую роскошь, как использование определенных данных о метках классов с целью оценки качества модели. Поэтому мы обратились к полезным внутренним метрикам оценки качества, таким как метод локтя или силуэтный анализ, в качестве попытки количественно определить качество кластеризации.

Затем мы проанализировали другой подход к кластеризации: агломеративную иерархическую кластеризацию. Иерархическая кластеризация не требует предварительно указывать число кластеров, и ее результат может визуализироваться в виде дендограммы, которая помогает интерпретировать результаты. Последним алгоритмом кластеризации, который мы рассмотрели в этой главе, был алгоритм DBSCAN, который группирует точки, основываясь на локальных плотностях, и способен обрабатывать выбросы и идентифицировать нешаровидные фигуры.

DBScan при маленьком eps выделяет слишком много кластеров, при большом наоброт, все данные соединяются в один кластер.

Иерархическая кластеризация неэффективно разделила данные на кластеры при методе одиночной связи (single), все остальные методы сработали вполне неплохо.