**МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РФ**

**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ**

**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ТЕХНОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ «МИСиС»**

*ПРОФИЛЬ ИННОВАЦИОННЫЕ IT-ПРОЕКТЫ*

*НАПРАВЛЕНИЕ ПРИКЛАДНАЯ ИНФОРМАТИКА ГРУППА МПИ-20-4-2*

### Отчёт по лабораторной работе №3

*ПО КУРСУ*: «Нейронные сети и машинное обучение»

*СТУДЕНТКА Денисова Е.А.*

*ПРЕПОДАВАТЕЛЬ Курочкин И. И.*

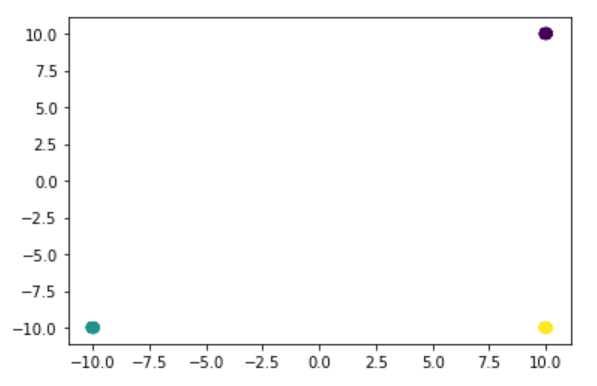
Москва, 2020г.

В ходе данной лабораторной работы были реализованы и протестированы на шести различных датасетах три метода кластеризации с евклидовой и манхеттенской метриками.

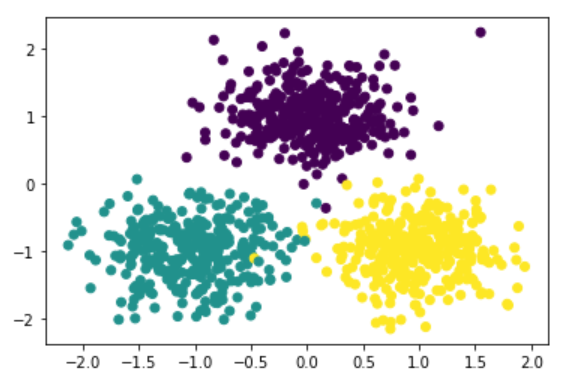
Для тестирования методов использовались датасеты:

1. Сгенерированный датасет с линейно разделимыми множествами, расстояние между группами во много раз превышает диаметр группы;
2. Сгенерированный датасет с линейно разделимыми множествами, группы расположены близко или касаются друг друга;
3. Сгенерированный датасет с линейно неразделимыми множествами, средняя площадь пересечения классов 10-20%;
4. Сгенерированный датасет с линейно неразделимыми множествами, средняя площадь пересечения классов 50-70%;
5. Эталонный датасет «Breast cancer Wiskonsin»;
6. Эталонный датасет «Wine».

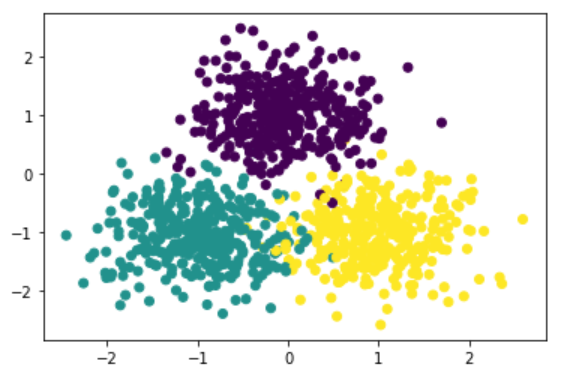
Визуализация сгенерированного датасета с линейно разделимыми множествами, расстояние между группами во много раз превышает диаметр группы. Количество примеров в данном датасете – 1000.



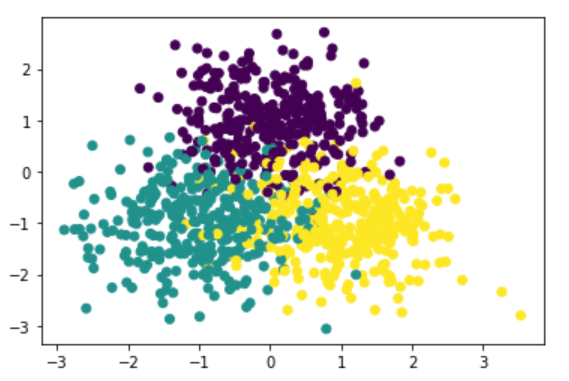
Визуализация сгенерированного датасета с линейно разделимыми множествами, группы расположены близко или касаются друг друга. Количество примеров в данном датасете также равно 1000.



Визуализация сгенерированного датасета с линейно разделимыми множествами, средняя площадь пересечения классов 10-20%. Количество примеров в данном датасете равно 1000.



Визуализация сгенерированного датасета с линейно разделимыми множествами, средняя площадь пересечения классов 50-70%. Количество примеров в данном датасете равно 1000.



Датасет «Breast cancer Wiskonsin» содержит 30 признаков, 569 образцов, 2 класса. Перед применением к датасету методов кластеризации данные в нем проходят препроцессинг с помощью функции sklearn scale(). Данная функция стандартизирует набор данных, центрируя и масштабируя их по среднему значению к единичной дисперсии.

Датасет «Wine» содержит 13 признаков, 178 образцов и 3 класса. Перед применением к датасету методов кластеризации данные в нем также проходят препроцессинг с помощью функции sklearn scale().

Для демонстрации качества разделения данных вычисляются 4 метрики: полнота, однородность, индекс Rand, скорректированная взаимная информация. Все оценки вычисляются с помощью фактических и проставленных меток.

Полнота – это показатель полноты маркировки кластеров с учетом фактической маркировки. Значение метрики уменьшается, если эталонный кластер разделить на части.

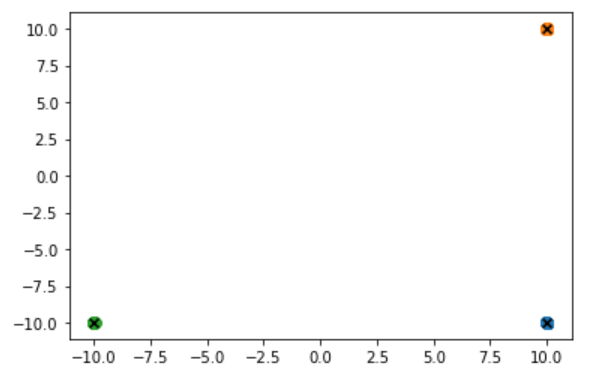
Однородность – это показатель однородности маркировки кластеров с учетом фактической маркировки. Результат кластеризации является однородным, если все его кластеры содержат только те точки данных, которые являются членами одного класса. Значение данной метрики качества уменьшается при объединении в один кластер двух фактических.

Индекс Rand – оценивает, насколько много из тех пар элементов, которые находились в одном классе, и тех пар элементов, которые находились в разных классах, сохранили данное состояние после кластеризации. В данной лабораторной работе вычисляется индекс Rand с поправкой на случайность.

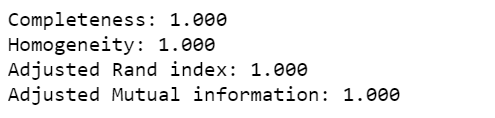
Скорректированная взаимная информация – это скорректированная оценка взаимной информации для учета случайности. Взаимная информация – это функция, которая измеряет согласованность двух наборов меток, игнорируя перестановки.

Первым рассматриваемым методом будет алгоритм k-means с евклидовой и манхеттенской метриками.

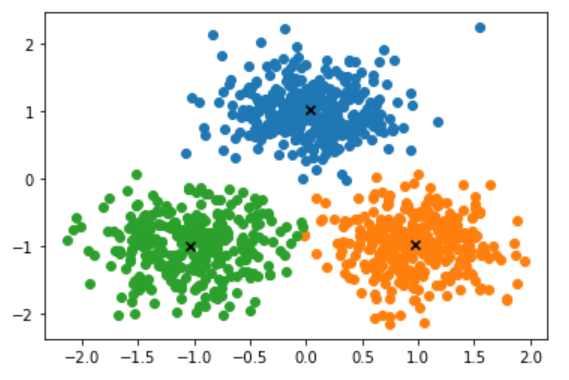
Сначала проверяется работа k-means с евклидовой метрикой. Первый датасет для проверки является сгенерированным датасетом, он представляет собой линейно разделимые множества с расстоянием между группами, во много раз большим, чем диаметр группы. Результат работы k-means с евклидовой метрикой для данного датасета. Крестиками обозначены центры кластеров, точки окрашены в цвета кластеров.



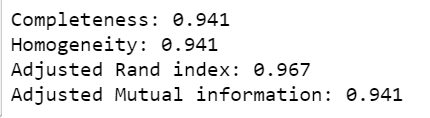
Значения метрик для k-means с евклидовой метрикой для линейно разделимого датасета.



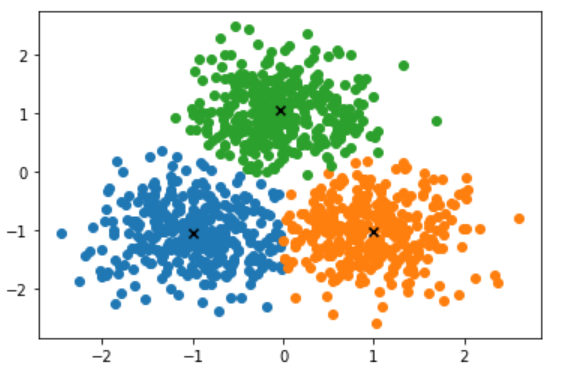
Далее k-means с евклидовой метрикой был протестирован на сгенерированном датасете, который состоит из линейно разделимых множеств, группы расположены близко. Визуализация работы k-means:



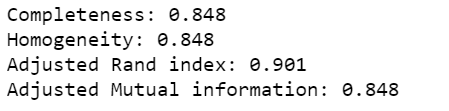
Значения метрик:



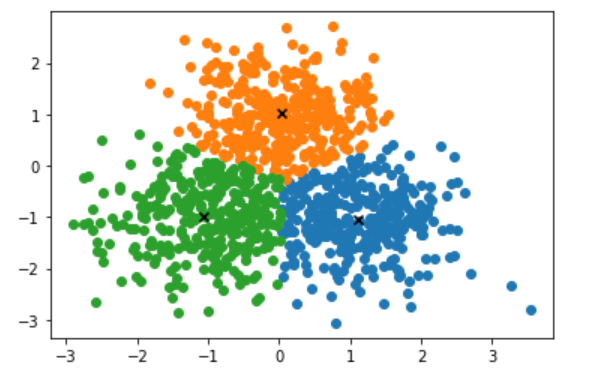
Следующим датасетом для проверки является сгенерированный датасет с линейно неразделимыми множествами, группы пересекаются на 10-20%. Визуализация работы k-means:



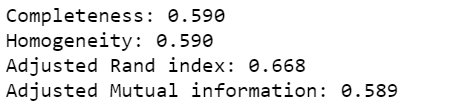
Результат работы k-means:



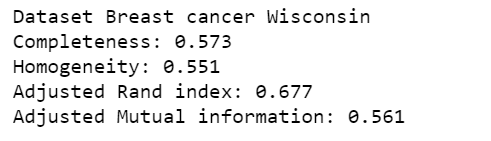
Далее алгоритм k-means был протестирован на линейно неразделимом датасете, со средней площадью пересечения классов 50-70%. Результат работы k-means:



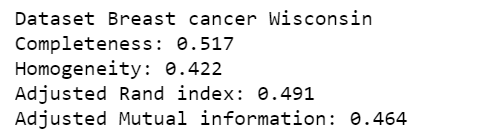
Значения метрик:



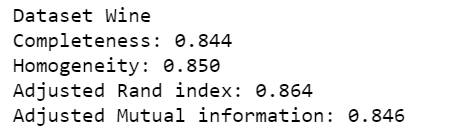
Далее алгоритм был протестирован на 2 эталонных датасетах. В качестве первого датасета был взят датасет «Breast cancer Wiskonsin». Значения метрик для данного датасета:



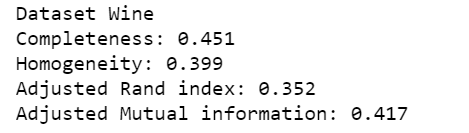
В качестве эксперимента была предпринята попытка применить метод кластеризации к данному датасету, не подготавливая данные. В данном случае значения метрик показывают худшее качество кластеризации:



В качестве второго эталонного датасета был взят датасет «Wine». Значения метрик для данного датасета:



В данном случае также была предпринята попытка не проводить препроцессинг данных. Значения метрик значительно хуже:



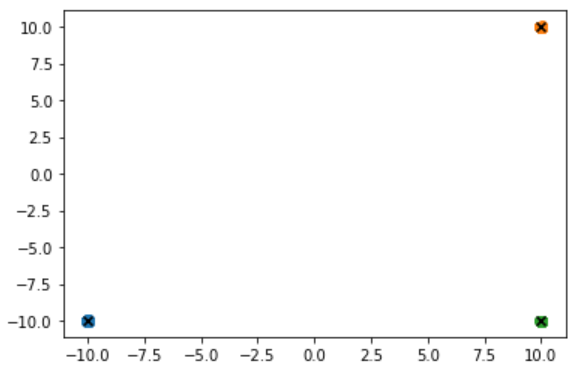
Результаты качества разделения для k-means с евклидовой метрикой для различных датасетов:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Completeness | Homogeneity | Adjusted Rand index | Adjusted Mutual information |
| Большое расстояние между группами | 1.000 | 1.000 | 1.000 | 1.000 |
| Группы расположены близко | 0.941 | 0.941 | 0.967 | 0.941 |
| Группы пересекаются на 10-20% | 0.848 | 0.848 | 0.901 | 0.848 |
| Группы пересекаются на 50-70% | 0.590 | 0.590 | 0.668 | 0.589 |
| «Breast cancer Wiskonsin» | 0.573 | 0.551 | 0.677 | 0.561 |
| «Wine» | 0.844 | 0.850 | 0.864 | 0.846 |

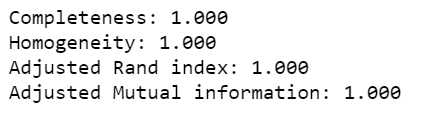
В таблице можно увидеть, что алгоритм работает тем лучше, чем лучше разделяются и чем дальше расположены классы. В случае, если классы пересекаются слишком сильно, значения метрик примерно равны 0.5 – 0.6. Также в ходе проверки работы алгоритма было установлено, что без препроцессинга данных качество кластеризации значительно хуже.

Далее была протестирована работа k-means с манхеттенской метрикой.

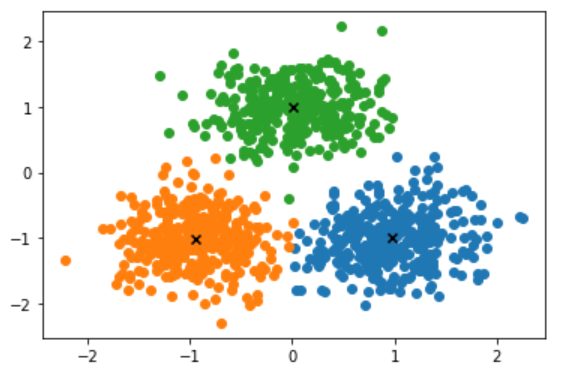
Результат работы k-means с манхеттенской метрикой на сгенерированном линейно разделимом датасете с расстоянием между группами во много раз большим, чем диаметр групп.



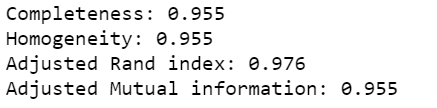
Значения метрик:



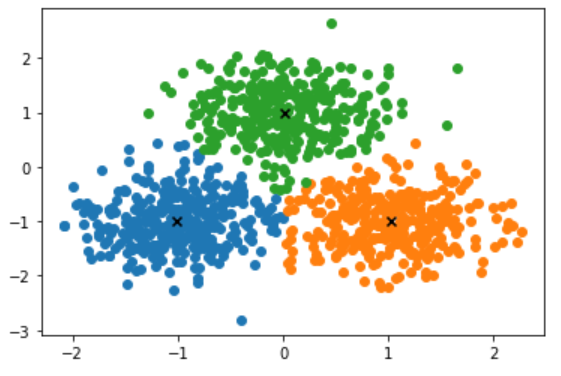
Результат работы k-means с манхеттенской метрикой на сгенерированном линейно разделимом датасете, в котором группы расположены близко или касаются друг друга.



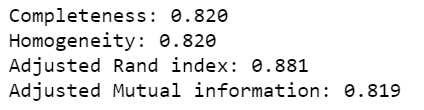
Значения метрик:



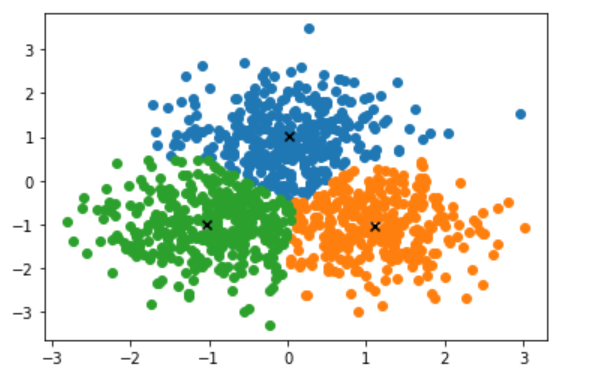
Результат работы k-means с манхеттенской метрикой на сгенерированном линейно неразделимом датасете, где средняя площадь пересечения классов 10-20%:



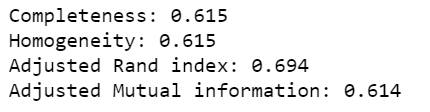
Значения метрик:



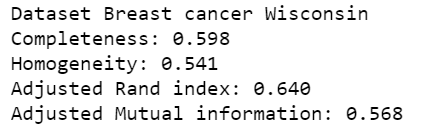
Результат работы k-means с манхеттенской метрикой на сгенерированном линейно неразделимом датасете, где средняя площадь пересечения классов 50-70%:



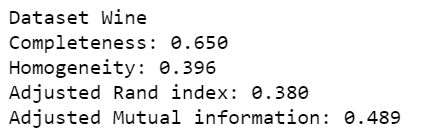
Значения метрик:



Результат работы k-means с манхеттенской метрикой на эталонном датасете «Breast cancer Wiskonsin», значения метрик:



Результат работы k-means с манхеттенской метрикой на эталонном датасете «Wine», значения метрик:



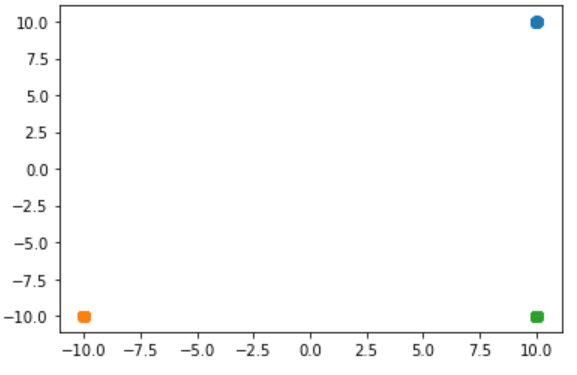
Результаты качества разделения для k-means с манхеттенской метрикой для различных датасетов:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Completeness | Homogeneity | Adjusted Rand index | Adjusted Mutual information |
| Большое расстояние между группами | 1.000 | 1.000 | 1.000 | 1.000 |
| Группы расположены близко | 0.955 | 0.955 | 0.976 | 0.955 |
| Группы пересекаются на 10-20% | 0.820 | 0.820 | 0.881 | 0.819 |
| Группы пересекаются на 50-70% | 0.615 | 0.615 | 0.694 | 0.614 |
| «Breast cancer Wiskonsin» | 0.598 | 0.541 | 0.640 | 0.568 |
| «Wine» | 0.650 | 0.396 | 0.380 | 0.489 |

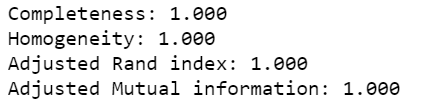
В таблице можно увидеть, что k-means с манхеттенской метрикой показывает примерно те же значения метрик, что и k-means с евклидовой метрикой, кроме датасета «Wine». Это может быть объяснено тем, что алгоритм k-means рассчитан на работу с евклидовой метрикой, а случае, если взять какую-либо иную метрику, алгоритм может не сойтись.

После этого была продемонстрирована работа иерархического агломеративного метода с евклидовой метрикой. В качестве данного метода был выбран метод Уорда. В данном методе в качестве расстояния между кластерами берется прирост суммы квадратов расстояний объектов до центра кластера, получаемого в результате их объединения. На каждом шаге алгоритма объединяются такие два кластера, которые приводят к минимальному увеличению дисперсии.

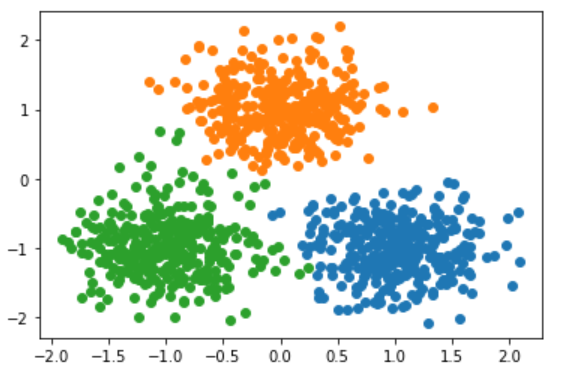
Результат работы метода Уорда на сгенерированном линейно разделимом датасете с расстоянием между группами во много раз большим, чем диаметр групп.



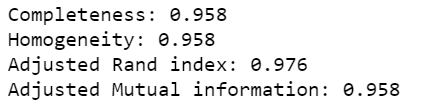
Значения метрик:



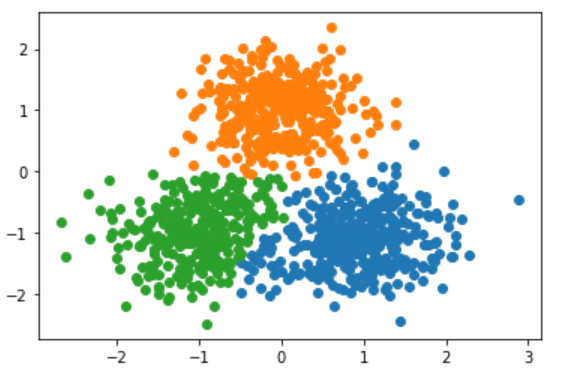
Результат работы метода Уорда на сгенерированном линейно разделимом датасете с группами, расположенными близко.



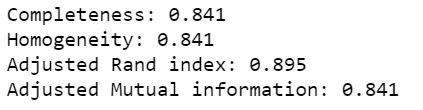
Значения метрик:



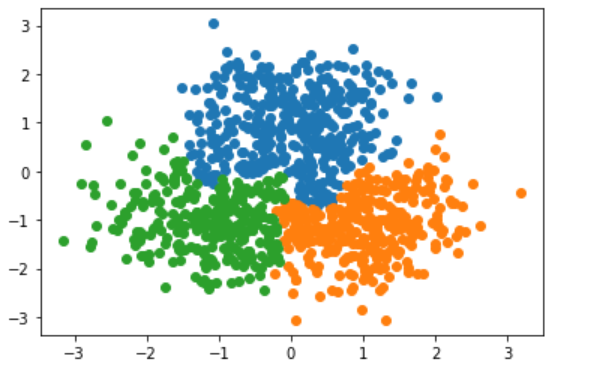
Результат работы метода Уорда на сгенерированном линейно неразделимом датасете с группами, пересекающимися на 10-20%.



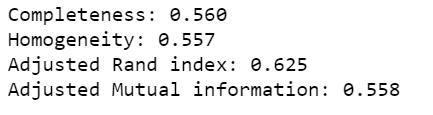
Значения метрик:



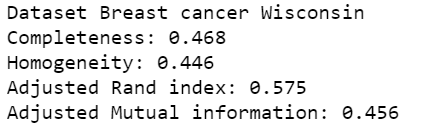
Результат работы метода Уорда на сгенерированном линейно неразделимом датасете с группами, пересекающимися на 50-70%.



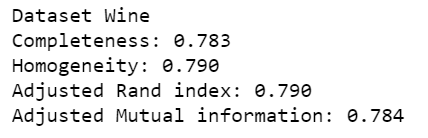
Значения метрик:



Результат работы метода Уорда с евклидовой метрикой на эталонном датасете «Breast cancer Wiskonsin», значения метрик:



Результат работы метода Уорда с евклидовой метрикой на эталонном датасете «Wine», значения метрик:

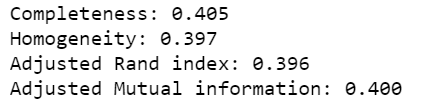


Результаты качества разделения для метода Уорда с евклидовой метрикой для различных датасетов:

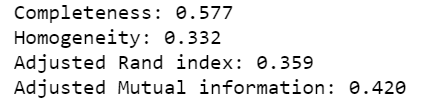
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Completeness | Homogeneity | Adjusted Rand index | Adjusted Mutual information |
| Большое расстояние между группами | 1.000 | 1.000 | 1.000 | 1.000 |
| Группы расположены близко | 0.958 | 0.958 | 0.976 | 0.958 |
| Группы пересекаются на 10-20% | 0.841 | 0.841 | 0.895 | 0.841 |
| Группы пересекаются на 50-70% | 0.560 | 0.557 | 0.625 | 0.558 |
| «Breast cancer Wiskonsin» | 0.468 | 0.446 | 0.575 | 0.456 |
| «Wine» | 0.783 | 0.790 | 0.790 | 0.784 |

Исходя из таблицы, видно, что на разделимых датасетах метод Уорда работает примерно так же, как и предыдущие методы, но на датасетах, классы в которых пересекаются в значительной степени, он работает немного хуже. Также из всех иерархических агломеративных методов метод Уорда наиболее подходит и часто применяется для задач с близко расположенными кластерами. Для демонстрации разницы показателей кластеризации для датасета, в котором средняя площадь пересечения классов 50-70%, были применены другие иерархические агломеративные методы: метод полной связи и метод средней связи.

Значения метрик для метода полной связи:



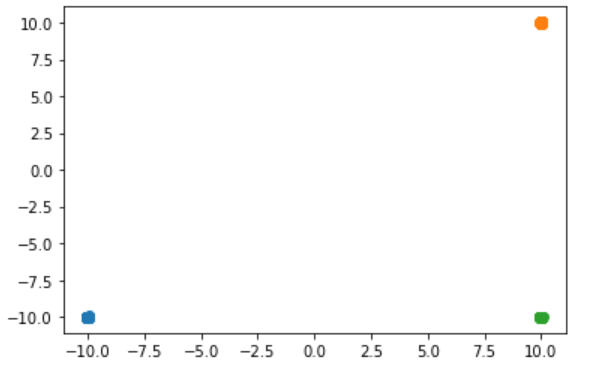
Значения метрик для метода средней связи:



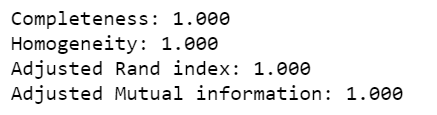
В обоих случаях можно наблюдать, что качество кластеризации хуже, чем при применении метода Уорда.

Далее была продемонстрирована работа иерархического агломеративного метода с манхеттенской метрикой. В качестве данного метода был выбран метод средней связи, поскольку метод Уорда работает только с евклидовой метрикой. Данный метод сводит к минимуму среднее расстояние между всеми образцами пар кластеров.

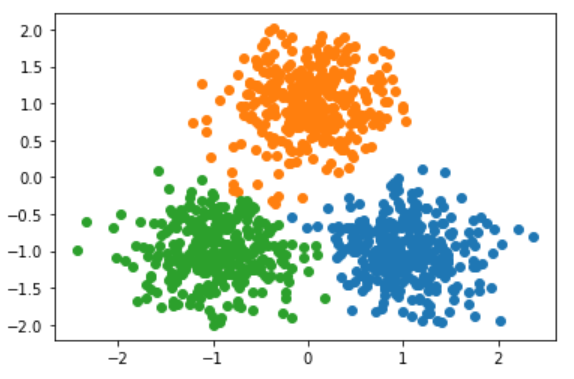
Результат работы метода средней связи на сгенерированном линейно разделимом датасете с расстоянием между группами во много раз большим, чем диаметр групп.



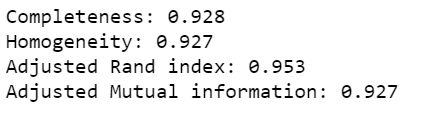
Значения метрик:



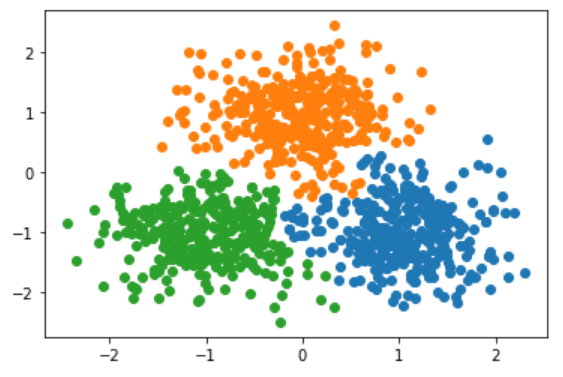
Результат работы метода средней связи на сгенерированном линейно разделимом датасете с группами, расположенными близко.



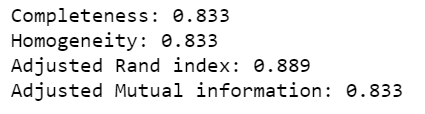
Значения метрик:



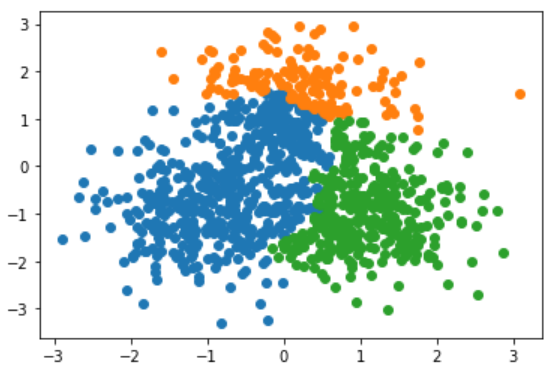
Результат работы метода средней связи на сгенерированном линейно неразделимом датасете с группами, пересекающимися на 10-20%.



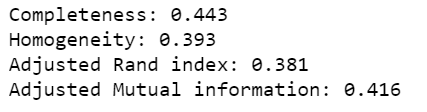
Значения метрик:



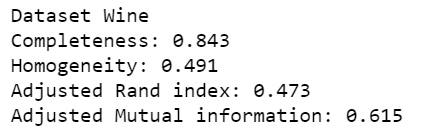
Результат работы метода средней связи на сгенерированном линейно неразделимом датасете с группами, пересекающимися на 50-70%.



Значения метрик:



Результат работы метода средней связи с манхеттенской метрикой на эталонном датасете «Wine», значения метрик:



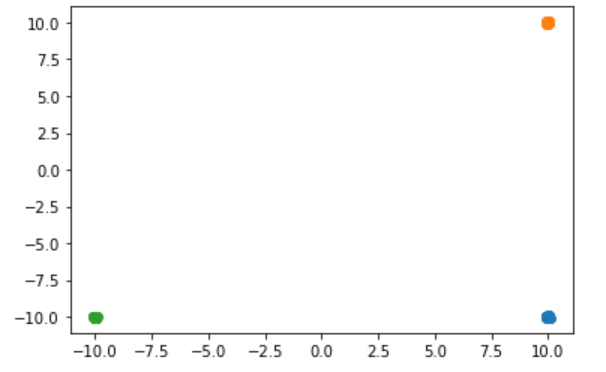
Результаты качества разделения для метода средней связи с манхеттенской метрикой для различных датасетов:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Completeness | Homogeneity | Adjusted Rand index | Adjusted Mutual information |
| Большое расстояние между группами | 1.000 | 1.000 | 1.000 | 1.000 |
| Группы расположены близко | 0.928 | 0.927 | 0.953 | 0.927 |
| Группы пересекаются на 10-20% | 0.833 | 0.833 | 0.889 | 0.833 |
| Группы пересекаются на 50-70% | 0.443 | 0.393 | 0.381 | 0.416 |
| «Wine» | 0.843 | 0.491 | 0.473 | 0.615 |

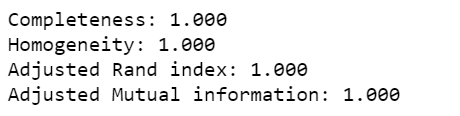
Из данной таблицы видно, что метод средней связи с манхеттенской метрикой на линейно разделимых датасетах или датасетах со слабо пересекающимися группами показал себя примерно так же, как и метод Уорда с евклидовой метрикой, а на датасетах, группы в которых пересекаются в более значительной степени, качество кластеризации с помощью данного метода хуже, чем при кластеризации методом Уорда.

После этого была продемонстрирована работа неиерархического метода с евклидовой метрикой. В качестве данного метода был выбран метод DBSCAN. Данный алгоритм кластеризации основан на плотности – если дан набор точек в пространстве, алгоритм группирует вместе точки, которые тесно расположены, и помечает как выбросы точки, находящиеся в областях с малой плотностью. При применении данного метода из библиотеки sklearn не нужно указывать количество кластеров в качестве входного параметра, но необходимо указать два других входных параметра: и минимальное число точек. Эпсилон () – это максимальное расстояние между двумя точками, чтобы они считались соседними. Минимальное число точек – это минимальное число точек, которые должны образовывать область. В используемом методе из библиотеки sklearn по умолчанию минимальное количество точек равно 5.

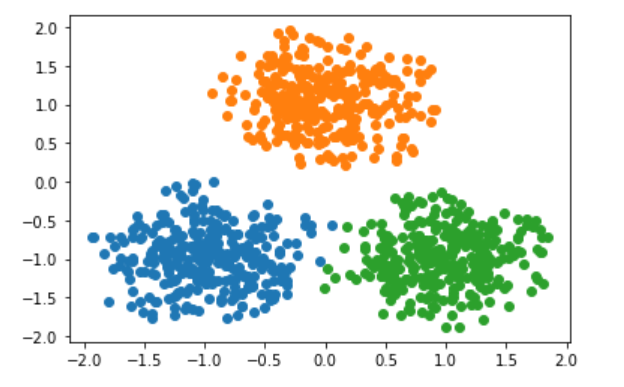
Результат работы метода DBSCAN на сгенерированном линейно разделимом датасете с расстоянием между группами во много раз большим, чем диаметр групп ( = 0.5).



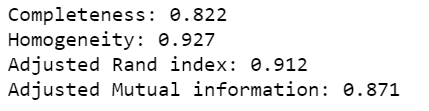
Значения метрик:



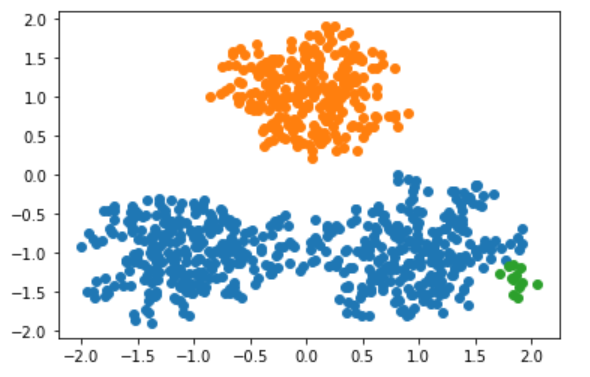
Результат работы метода DBSCAN на сгенерированном линейно разделимом датасете с группами, расположенными близко ( = 0.2).



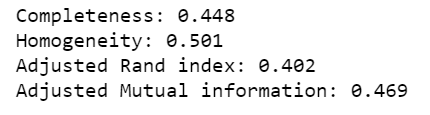
Значения метрик:



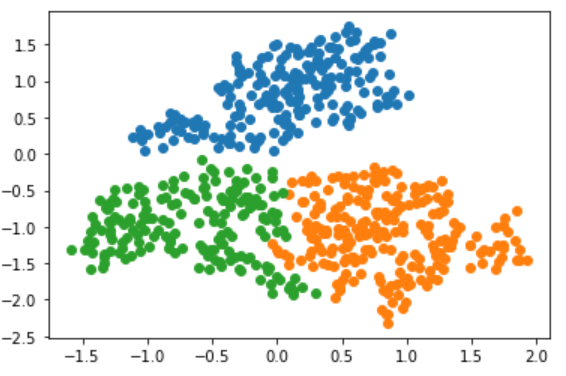
Результат работы DBSCAN на сгенерированном линейно неразделимом датасете с группами, пересекающимися на 10-20% ( = 0.15).



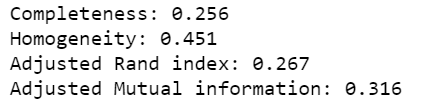
Значения метрик:



Результат работы DBSCAN на сгенерированном линейно неразделимом датасете с группами, пересекающимися на 50-70% ( = 0.15).



Значения метрик:



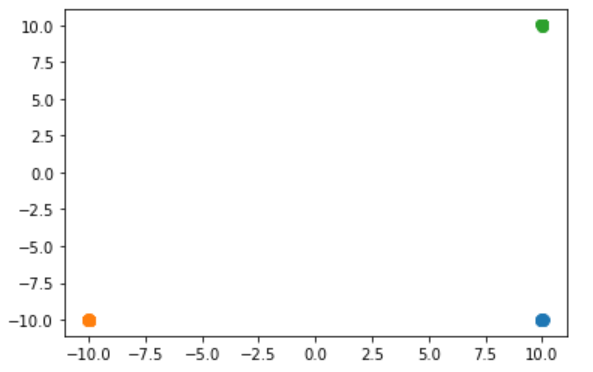
Результаты качества разделения для метода DBSCAN с евклидовой метрикой для различных датасетов:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Completeness | Homogeneity | Adjusted Rand index | Adjusted Mutual information |
| Большое расстояние между группами | 1.000 | 1.000 | 1.000 | 1.000 |
| Группы расположены близко | 0.822 | 0.927 | 0.912 | 0.871 |
| Группы пересекаются на 10-20% | 0.448 | 0.501 | 0.402 | 0.469 |
| Группы пересекаются на 50-70% | 0.256 | 0.451 | 0.267 | 0.316 |

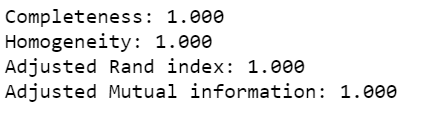
Данный алгоритм показал результат хуже, чем ранее рассмотренные методы. Также согласно данной таблице, качество разделения при использовании данного алгоритма достаточно высокое, когда группы в датасете линейно разделимы, а в случае, когда группы пересекаются, качество разделения резко снижается, поскольку алгоритм основан на плотности. Также в случае, когда группы пересекаются, необходимо достаточно долго эмпирически подбирать значение эпсилон и минимальное число точек для достижения приемлемого результата.

Далее была продемонстрирована работа DBSCAN с манхеттенской метрикой.

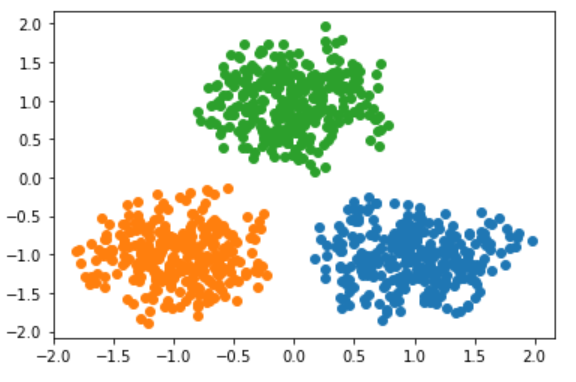
Результат работы метода DBSCAN с манхеттенской метрикой на сгенерированном линейно разделимом датасете с расстоянием между группами во много раз большим, чем диаметр групп ( = 0.5).



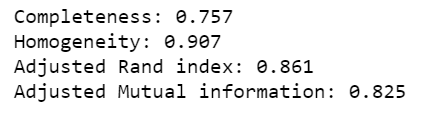
Значения метрик:



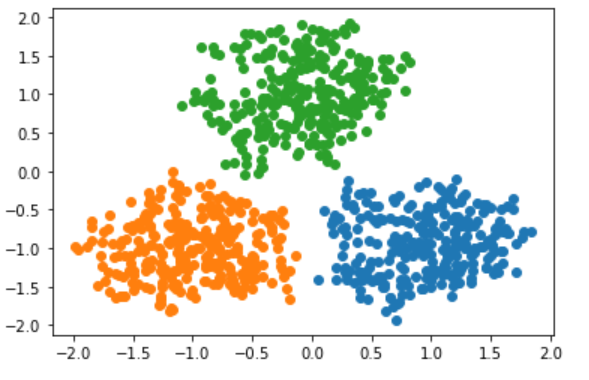
Результат работы метода DBSCAN с манхеттенской метрикой на сгенерированном линейно разделимом датасете с группами, расположенными близко ( = 0.2).



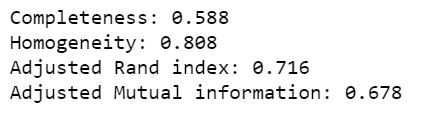
Значения метрик:



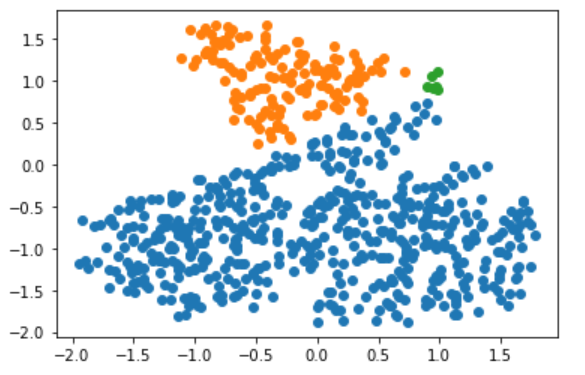
Результат работы DBSCAN с манхеттенской метрикой на сгенерированном линейно неразделимом датасете с группами, пересекающимися на 10-20% ( = 0.2).



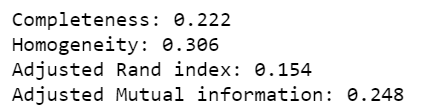
Значения метрик:



Результат работы DBSCAN с манхеттенской метрикой на сгенерированном линейно неразделимом датасете с группами, пересекающимися на 50-70% ( = 0.2).



Значения метрик:



Результаты качества разделения для метода DBSCAN с манхеттенской метрикой для различных датасетов:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Completeness | Homogeneity | Adjusted Rand index | Adjusted Mutual information |
| Большое расстояние между группами | 1.000 | 1.000 | 1.000 | 1.000 |
| Группы расположены близко | 0.757 | 0.907 | 0.861 | 0.825 |
| Группы пересекаются на 10-20% | 0.588 | 0.808 | 0.716 | 0.678 |
| Группы пересекаются на 50-70% | 0.222 | 0.306 | 0.238 | 0.248 |

Данный алгоритм может быть использован с любой функцией расстояния, поэтому DBSCAN с манхеттенской метрикой показал примерно такие же результаты, что и DBSCAN с евклидовой метрикой: на линейно разделимых датасетах хорошие показатели метрик, если датасет линейно неразделим, показатели падают.