МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ТЕХНОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ «МИСиС»

Институт ИТАСУ
Группа: **МПИ-20-4-2**

ОТЧЕТ

по лабораторной работе №3
по курсу «Нейронные сети»

Выполнил: Хабибулин М.И.
группа МПИ-20-4-2
Проверил: Курочкин И.И.

Москва 2020

.

Реализация:

В работе были использованы 3 метода кластеризации: K-Means, Spectral clustering и k-medoids clustering.

Расстояние подсчитывалось двумя методами:

1. Евклидово расстояние рассчитывается по формуле

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}.$$

   Где *p, q* – некоторые точки.

2. Манхэттенское расстояние рассчитывается по формуле

Манхэттенская метрика

$$l_{ij} = |x_i - x_j| + |y_i - y_j|$$

Метрики оценки качества кластеризации:

True positive (TP)	False positive (FP)
False negative (FN)	True negative (TN)

В качестве показателей качества разделения были использованы:

1. Folkes and Mallows Index рассчитывается по формуле

$$\mathrm{FMI} = \frac{\mathrm{TP}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})}}$$

2. Rand index рассчитывается по формуле

$$\mathrm{RI} = \frac{a + b}{C_2^{n_{samples}}}$$

3. Adjusted Rand index рассчитывается по формуле

$$\mathrm{ARI} = \frac{\mathrm{RI} - E[\mathrm{RI}]}{\max(\mathrm{RI}) - E[\mathrm{RI}]}$$

Также использовались Adjusted Mutual Information, V-measure и Индекс однородности. https://scikit-learn.org/stable/modules/clustering.html#spectral-clustering

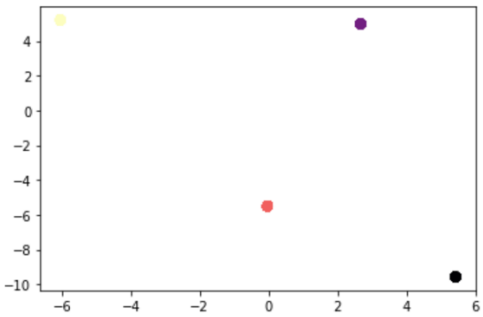
Результат работы:

Пример 1.

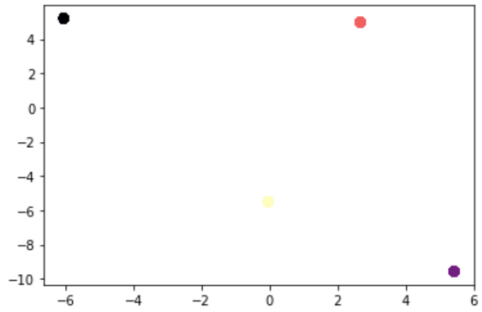
Линейно разделимые множества (с расстоянием между группами в 10^3 раз больше, чем диаметр группы)

Параметры:
n_samples = 1000
centers = 4
n_features = 2
cluster_std = 0,001
random state=10

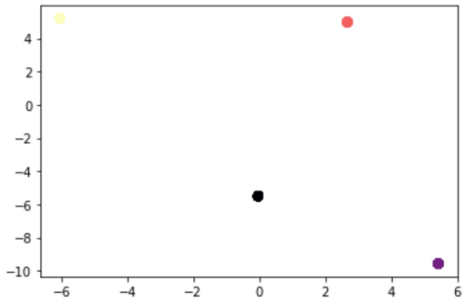
До кластеризации



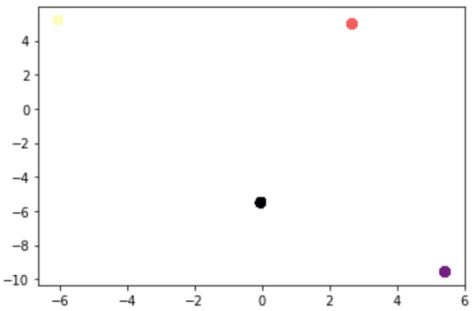
K-means



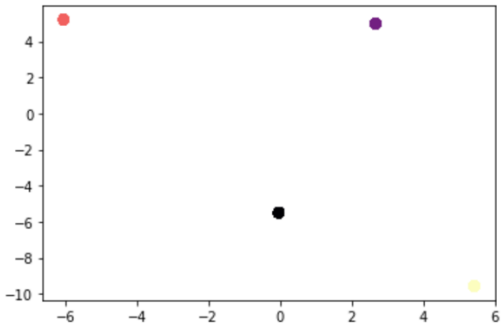
Спектральная кластеризация



KMedoids(eucl)



KMedoids(manh)



Сравнительная таблица 1.

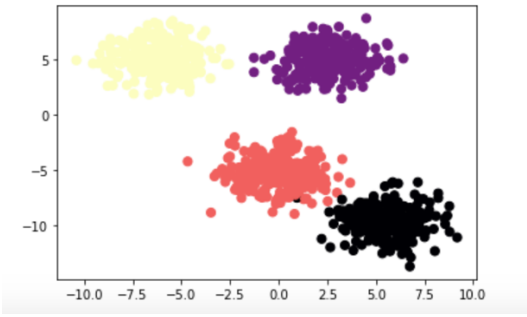
Метрика	K-means	SpectralClustering	KMedoids(euc)	KMedoids(manh)
Fowlkes-Mallows Index	1.0	1.0	1.0	1.0
Rand Index	1.0	1.0	1.0	1.0
Adjusted Rand index	1.0	1.0	1.0	1.0
V-measure	1.0	1.0	1.0	1.0

Пример 2.

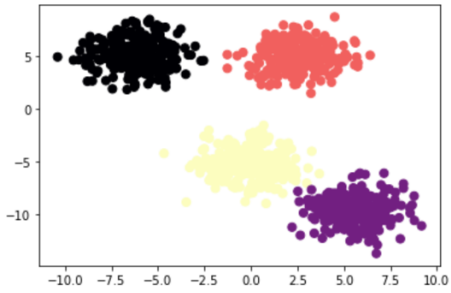
30 линейно разделимых класса, находящихся далеко друг от друга

Параметры:
n_samples = 1000
centers = 4
n_features = 2
cluster_std = 1.4
random state=10

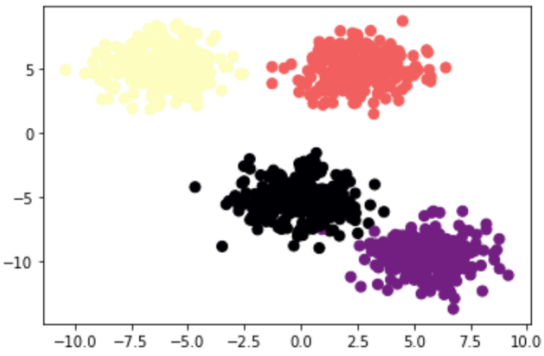
До кластеризации



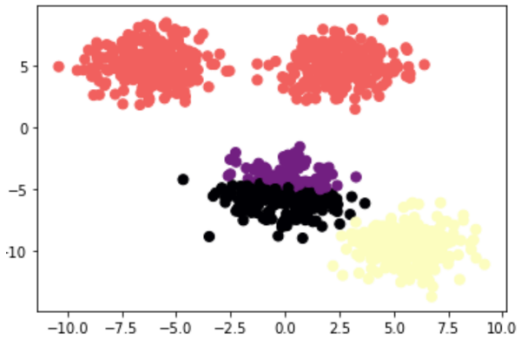
K-means



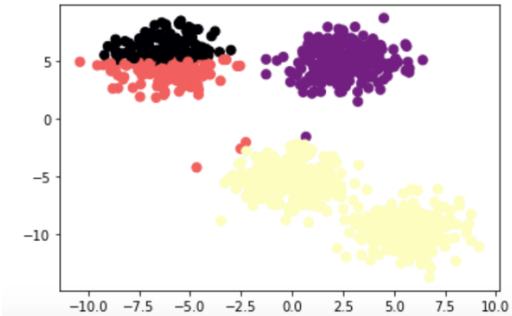
Спектральная кластеризация



KMedoids(euc)



KMedoids(manh)



Сравнительная таблица 2.

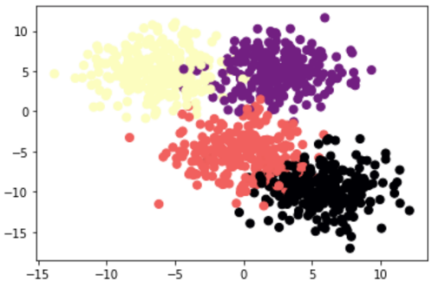
Метрика	K-means	SpectralClustering	KMedoids(euc)	KMedoids(manh)
Fowlkes-Mallows Index	0.996	1.0	0.754	0.741
Rand Index	0.990	1.0	0.745	0.737
Adjusted Rand index	0.994	1.0	0.639	0.624
V-measure	0.990	1.0	0.800	0.784

Пример 3.

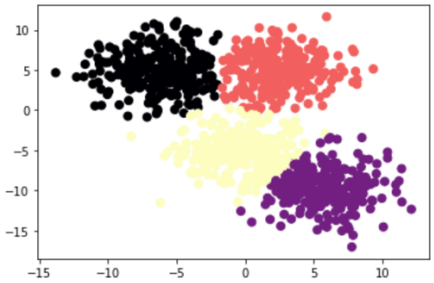
линейно неразделимое множество (средняя площадь пересечения классов 10-20%)

Параметры:
n_samples = 1000
centers = 4
n_features = 2
cluster_std = 2.5
random state=10

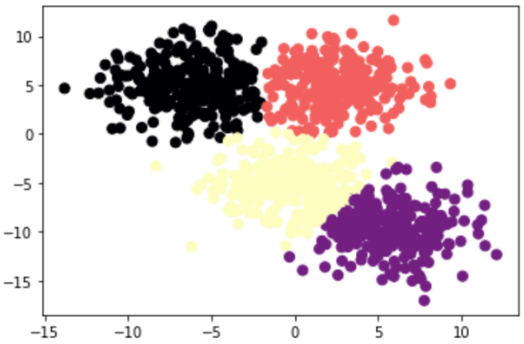
До кластеризации



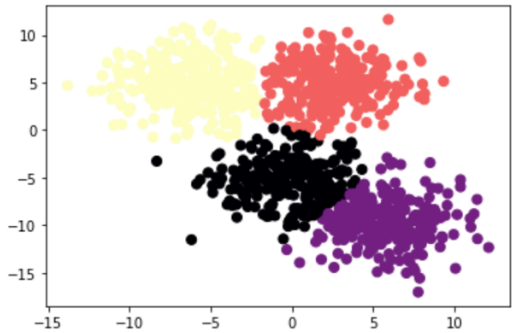
K-means



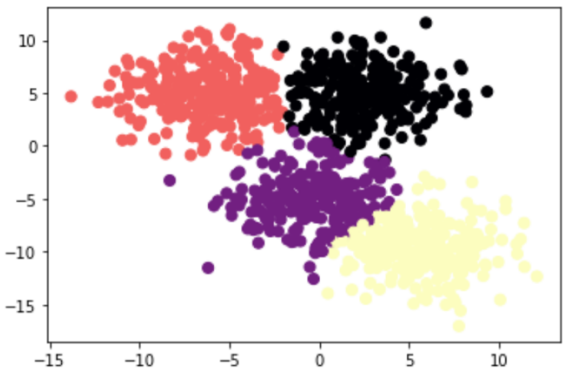
Спектральная кластеризация



KMedoids



KMedoids(manh)



Сравнительная таблица 3.

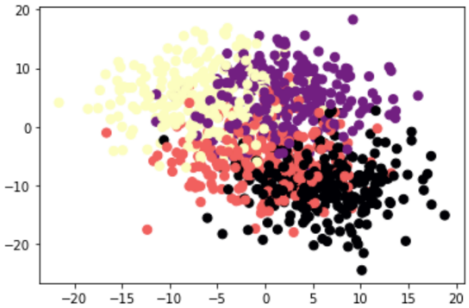
Метрика	K-means	SpectralClustering	KMedoids	KMedoids(manh)
Fowlkes-Mallows Index	0.893	0.869	0.888	0.886
Rand Index	0.828	0.809	0.822	0.819
Adjusted Rand index	0.858	0.825	0.851	0.849
V-measure	0.828	0.811	0.822	0.819

Пример 4.

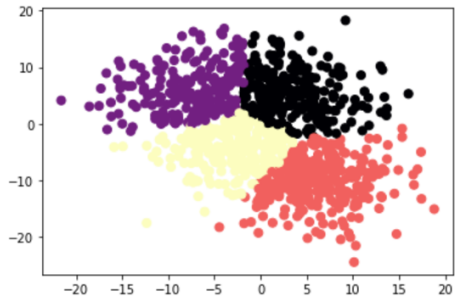
линейно неразделимое множество (средняя площадь пересечения классов 50-70%)

Параметры:
n_samples = 1000
centers = 4
n_features = 2
cluster_std = 5
random state=10

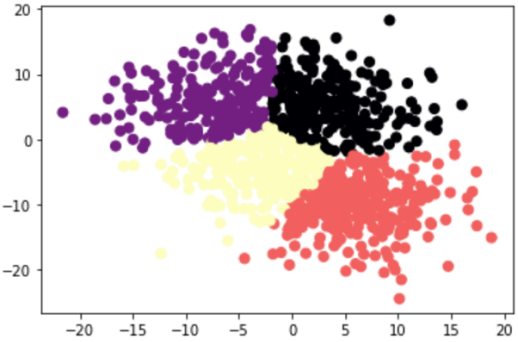
До кластеризации



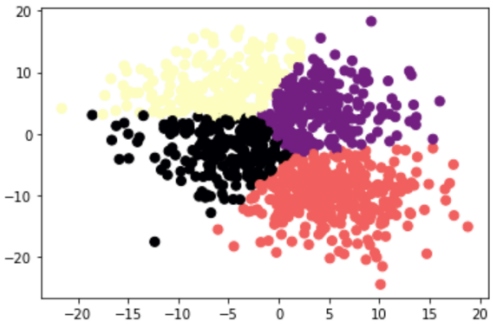
K-means



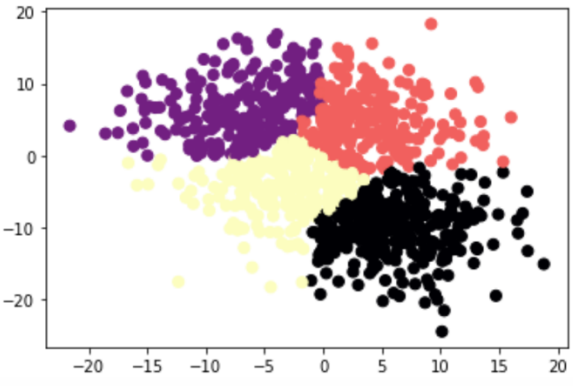
Спектральная кластеризация



KMedoids



KMedoids(manh)



Сравнительная таблица 4.

Метрика	K-means	SpectralClustering	KMedoids	KMedoids(manh)

Fowlkes-Mallows Index	0.536	0.576	0.520	0.533
Rand Index	0.397	0.284	0.388	0.393
Adjusted Rand index	0.381	0.314	0.386	0.376
V-measure	0.398	0.375	0.390	0.394

Пример 5.

Car Evaluation Data Set https://archive.ics.uci.edu/ml/datasets/Car+Evaluation

Attribute Information:

Class Values:

unacc, acc, good, vgood

Attributes:

- buying: vhigh, high, med, low
- maint: vhigh, high, med, low.
- doors: 2, 3, 4, 5more.
- persons: 2, 4, more.
- lug_boot: small, med, big.
- safety: low, med, high.

- **Сравнительная таблица 4.**

Метрика	K-means	SpectralClustering	KMedoids	KMedoids(manh)
Fowlkes-Mallows Index	0.372	0.369	0.374	0.388
Rand Index	0.011	0.010	0.012	0.013
Adjusted Rand index	0.014	0.011	0.013	0.014
V-measure	0.008	0.007	0.009	0.010

Во всех случая K-means показывает лучшие результаты. KMedoids показывает более низкие результаты, но догоняет K-means при большом смешение кластеров. Также стоит отметить, что алгоритм KMedoids обладает большей вычислительной сложностью: $O(k(n-k)^2)$, против O(nkl), где k – число кластеров, l – число итераций у K-means. Кластеризации методов на эталонном датасете Car Evaluation Data Set показали неудовлетворительный результат. Предположительно это связано с большим количеством признаков.