**Лабораторная работа №4**

**Кластеризация данных**

*Цель работы:*

Получить практические навыки решения задачи кластеризации фактографических данных в среде Jupiter Notebook. Научиться проводить настраивать параметры методов и оценивать точность полученного разбиения.

*Задание:*

1. Загрузить выборки согласно варианту задания
2. Отобразить данные на графике в пространстве признаков. Поскольку решается задача кластеризации, то подразумевается, что априорная информация о принадлежности каждого объекта истинному классу неизвестна, соответственно, на данном этапе все объекты на графике должны отображаться одним цветом, без привязки к классу.
3. Провести иерархическую кластеризацию выборки, используя разные способы вычисления расстояния между кластерами: расстояние ближайшего соседа (single), дальнего соседа (complete), Уорда (Ward). Построить дендрограммы для каждого способа. Размер графика должен быть подобран таким образом, чтобы дендрограмма хорошо читалась.
4. Исходя из дендрограмм выбрать лучший способ вычисления расстояния между кластерами.
5. Для выбранного способа, исходя из дендрограммы, определить количество кластеров в имеющейся выборке. Отобразить разбиение на кластеры и центроиды на графике в пространстве признаков (объекты одного кластера должны отображаться одним и тем же цветом, центроиды всех кластеров – также одним цветом, отличным от цвета кластеров)
6. Рассчитать среднюю сумму квадратов расстояний до центроида, среднюю сумму средних внутрикластерных расстояний и среднюю сумму межкластерных расстояний для данного разбиения. Сделать вывод о качестве разбиения.
7. Провести кластеризацию выборки методом k-средних. для k [1, 10].
8. Сформировать три графика: зависимость средней суммы квадратов расстояний до центроида, средней суммы средних внутрикластерных расстояний и средней суммы межкластерных расстояний от количества кластеров. Исходя из результатов, выбрать оптимальное количество кластеров.
9. Составить сравнительную таблицу результатов разбиения иерархическим методом и методом k-средних.

**Варианты заданий:**

Для всех вариантов дополнительно указать параметр n\_samples=100 - объем выборки N.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Вариант** | **1** | **2** | **3** | **4** | **5** | **6** |
| Вид классов | blobs | blobs | blobs | blobs | blobs | blobs |
| Random\_state | 34 | 28 | 41 | 23 | 18 | 68 |
| cluster\_std | 2.1 | 2 | 2 | 2 | 2 | 2 |
| noise | - | - | - | - | - | - |
| Centers | 7 | 7 | 7 | 6 | 8 | 6 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Вариант** | **7** | **8** | **9** | **10** | **11** | **12** |
| Вид классов | classification | classification | classification | classification | classification | classification |
| Random\_state | 55 | 36 | 3 | 68 | 5 | 27 |
| class\_sep | 1.5 | 1.2 | 2 | 1 | 1.5 | 1 |
| Для всех вариантов, использующих для генерации make\_classification, дополнительные параметры: n\_features=2, n\_redundant=0, n\_informative=2, n\_clusters\_per\_class=1, n\_classes = 4 | | | | | | |

**Контрольные вопросы**

1. Постановка задачи кластеризации данных. Отличие от задачи классификации.
2. Какие способы вычисления расстояний между кластерами вам известны?
3. Что такое дендрограмма? Алгоритм построения дендрограммы.
4. Алгоритм метода k-средних.
5. Особенности метода k-средних, его достоинства и недостатки.
6. Способы оценки качества кластеризации данных.