1 Валидация кластеризации

На лекции была рассмотрена задача *валидации кластеризации* и описаны различные подходы для ее решения. Эти методы делятся на два основных типа: внутренние и внешние критерии. В частности, были рассмотрены внутренние критерии *Davies-Bouldin criteria* и *Calinski-Harabasz criteria* и внешние критерии *Rand Index* и *Fowlkes-Mallows Index*. Описания формул для вычисления статистик представлены на слайдах лекции.

2 Задание

Задание состоит из двух частей: реализации и применения внутреннего и внешнего критериев для нахождения оптимального количества кластеров.

В предыдущем задании было предложено реализовать алгоритм k-means и применить его для "сжатия" изображения. В этом задании необходимо реализовать четыре критерия качества кластеризации, и, применив его в алгоритму кластеризации из прошлого задания, выбрать оптимальное количество кластеров в данных.

Внутренние критерии

В этой части задания необходимо реализовать два внутренних критерия качества кластеризации: *Davies-Bouldin* и *Calinski-Harabasz*. Затем с помощью каждого из них определить оптимальное количество кластеров для реализованного в предыдущем задании алгоритма k-means, примененного к изображению "policemen.jpg"

Внешние критерии

В этой части задания необходимо реализовать два внешних критерия качества кластеризации: *Rand Index* и *Fowlkes-Mallows Index*, а затем с помощью каждого из них определить оптимальное количество кластеров в данных. В файле "task2_data_7.txt" содержатся данные в следующем формате: в каждой строке находится 3 числа, разделенные запятыми: индекс кластера (y_i), значения наблюдения ($x_i \in \mathbb{R}^2$). Индекс кластера y_i является правильным разбиением данных на кластеры, которое используется для построения внешних критериев.

3 Содержание ответа

Для получения зачета по этому заданию необходимо на адрес natalia.kizhaeva@gmail.com прислать следующее

- Графики зависимостей значений каждого из 4 критериев от количества кластеров

- Код реализации критериев *Davies-Bouldin* и *Calinski-Harabasz* и поиска оптимального количества кластеров для алгоритма k-means, реализованного в предыдущем задании, примененного к изображению “policemen.jpg”
- Изображение “policemen.jpg”, сжатое с оптимальным количеством кластеров
- Код реализации критериев *Rand Index* и *Fowlkes-Mallows Index* и поиска оптимального количества кластеров для алгоритма k-means, реализованного в предыдущем задании, примененного к данным “task2_data_7.txt”