Санкт-Петербургский Политехнический Университет Петра Великого

Физико-механический институт

Кафедра прикладной математики и вычислительной физики

Отчёт по лабораторной работе №3 по дисциплине

«Многомерный статистический анализ»

Выполнил студент гр. 5030102/90401: Веселый Д. В.

Преподаватель: Павлова Л. В.

Санкт-Петербург

2023

# Постановка задачи

Необходимо реализовать SVC-алгоритм, и для хорошо и плохо разделённых данных провести кластеризацию. Провести исследование влияния параметров , на результат кластеризации и сделать выводы

# 1. Реализация SVC-алгоритма

1) Задаём параметры

2) Решаем задачу:

Изображение выглядит как текст

Автоматически созданное описание

где гауссово ядро,

3) Вычисляем:

Изображение выглядит как текст

Автоматически созданное описание

Конкретно в данной реализации считалось как

4) Строим матрицу смежности:

Изображение выглядит как текст

Автоматически созданное описание

Проверка условия считалась путём деления отрезка на 10 частей

() и для каждой проверялось

5) На последнем шаге по матрице смежности возвращаются кластеры, полученные в результате работы алгоритма

**Опционально:** можно считать начальное приближение для по таким формулам:



Изображение выглядит как текст, монитор, экран, снимок экрана

Автоматически созданное описание Изображение выглядит как текст

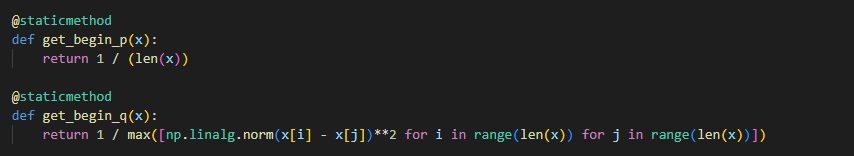
Автоматически созданное описание

Изображение выглядит как текст

Автоматически созданное описание

Изображение выглядит как текст

Автоматически созданное описание



# 2. Исследования

Дополнительно для оценки результатов будем использоваться два критерия:  
1) Внешний улучшенный критерий Рэнда

Изображение выглядит как текст, письмо

Автоматически созданное описание

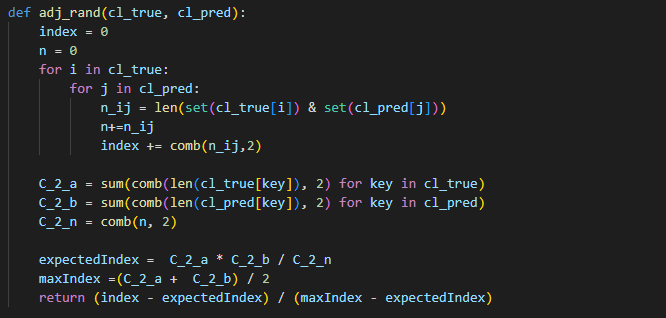
2) Внутренний – критерий “силуэт”

Изображение выглядит как текст, письмо

Автоматически созданное описание

Где ,

**Для случая, когда число кластеров = 1, определим значение критерия “силуэт” = 0**



Изображение выглядит как текст, в помещении, снимок экрана

Автоматически созданное описание

## 

## 2.1.1 Хорошо разделимые данные

Берём датасет из библиотеки sklearn, установим размерность выборки равной 50 элементам, число кластеров возьмём равным трём:

Изображение выглядит как диаграмма

Автоматически созданное описание

Изображение выглядит как текст

Автоматически созданное описание

Далее крестиками будут отмечаться SVs, плюсиками BVs, остальные – точками.

Проводим кластеризацию для начальных , считаемых по формулам, описанных в пункте 1:

*svs: 4 bsv: 0*

*p = 0.02, q = 0.002822982965231981*

*clusters: 1*

*adj\_rand: 0.0, silhouette: 0*Изображение выглядит как диаграмма

Автоматически созданное описание

Оставим p таким же, но увеличим q:

*svs: 12 bsv: 0*

*p = 0.02, q = 0.1*

*clusters: 3*

*adj\_rand: 1.0, silhouette: 0.8094228634022077*

*Изображение выглядит как диаграмма

Автоматически созданное описание*

Не будем останавливаться и пойдём дальше в сторону увеличения q:

*svs: 21 bsv: 0*

*p = 0.02, q = 0.3*

*clusters: 6*

*adj\_rand: 0.8968951832542215, silhouette: 0.6367775434764981*Изображение выглядит как диаграмма

Автоматически созданное описание

Сделаем ещё большим q:

*svs: 47 bsv: 0*

*p = 0.02, q = 4*

*clusters: 43*

*adj\_rand: 0.06143344709897611, silhouette: 0.018745396039117672*Изображение выглядит как диаграмма

Автоматически созданное описание

Попробуем увеличить p:

*svs: 16 bsv: 6*

*p = 0.3, q = 0.3*

*clusters: 11*

*adj\_rand: 0.7639205470530772, silhouette: 0.16022531065892476*Изображение выглядит как диаграмма

Автоматически созданное описание

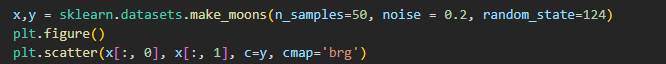
## 2.1.2 Итоговая таблица и выводы

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | Кол-во кластеров | Кол-во SVs | Кол-во BSVs | Критерий Рэнда | Критерий “силуэт” |
| 0.02 | 0.0028 | 1 | 4 | 0 | 0 | 0 |
| 0.02 | 0.1 | 3 | 12 | 0 | 1 | 0.809 |
| 0.02 | 0.3 | 6 | 21 | 0 | 0.89 | 0.63 |
| 0.02 | 4 | 43 | 47 | 0 | 0.061 | 0.018 |
| 0.3 | 0.3 | 11 | 16 | 6 | 0.763 | 0.16 |

Видно зависимость от параметра , чем больше он становится, тем граница как бы ложится плотнее на данные и мы получаем бОльшее число кластеров. Кроме того, удалось получить необходимое разбиение, не прибегая к BSVs. Сравнивая значения критериев, замечаем, что они достаточно хорошо показывают успешность кластеризации. Так, например, критерий Рэнда явно показывает, насколько сильно мы удаляемся/приближаемся к исходной кластеризации, что ожидаемо, ибо критерий внешний, а значит использует истинное разбиение. “Силуэт” хоть и является внутренним критерием, но тоже демонстрирует хорошие результаты по оценке кластеризации: довольно большое значение при правильной кластеризации, уменьшается по мере удалённости от нужного разбиения.

## 2.2.1 Плохо разделимые данные

Берём очередной датасет из sklearn, так же число элементов равно 50, и добавим шум, чтобы данные были “плохие”



Изображение выглядит как диаграмма

Автоматически созданное описание

Начальные значения:

svs: 3 bsv: 0

p = 0.02, q = 0.07484727768150307

clusters: 1

adj\_rand: 0.0, silhouette: 0

Изображение выглядит как диаграмма

Автоматически созданное описание

Пробуем увеличивать :

*svs: 21 bsv: 0*

*p = 0.02, q = 3*

*clusters: 1*

*adj\_rand: 0.0, silhouette: 0*

Изображение выглядит как диаграмма

Автоматически созданное описание

И ещё:

*svs: 27 bsv: 0*

*p = 0.02, q = 4*

*clusters: 5*

*adj\_rand: 0.1050747455997878, silhouette: -0.3030338291107546*Изображение выглядит как диаграмма

Автоматически созданное описание

Видим, что уже лучше, но пошли как будто не в ту сторону(получается очень большой кластер). Попробуем, оставить , поскольку там видно, что области сжимаются довольно близко к исходным. Теперь будем увеличивать долю BSVs, т. е. параметр .

*svs: 19 bsv: 6*

*p = 0.3, q = 3*

*clusters: 8*

*adj\_rand: 0.6357731199357128, silhouette: -0.33189901108052844*Изображение выглядит как диаграмма

Автоматически созданное описание

Здесь уже результат визуально похож на истинный, однако существуют некие “выбросы”. Но, в целом, примерную картину того, как должна быть кластеризована исходная выборка, мы видим.

## 2.2.2 Итоговая таблица и выводы

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | Кол-во кластеров | Кол-во SVs | Кол-во BSVs | Критерий  Рэнда | Критерий  “Силуэт” |
| 0.02 | 0.074 | 1 | 3 | 0 | 0 | 0 |
| 0.02 | 3 | 1 | 21 | 0 | 0 | 0 |
| 0.02 | 4 | 5 | 27 | 0 | 0.105 | -0.303 |
| 0.3 | 3 | 8 | 19 | 6 | 0.635 | -0.331 |

Наблюдается та же закономерность, что и с прошлыми данными, при увеличении параметра области сжимаются и становится больше кластеров. Параметр , представляющий долю BSVs, как раз и показывает увеличение числа BSVs. Критерии отрабатывают так же хорошо, как и в прошлом случае, однако здесь можно заметить, что при приближении к истинному разбиению “Силуэт” немного ухудшился. Такой эффект можно объяснить тем, что оказывают влияние два существенных фактора:  
1) Выборка плохо разделена  
2) Критерий внутренний, а значит использует только информации о составе кластеров и их взаимном расположении. Учитывая тот факт, что значение “силуэта” является мерой того, насколько похож на объект на другие объекты из своего кластера в сравнении с объектами из других кластеров, можно объяснить ухудшение появлением “выбросов” (p = 0.3, q = 3)

# 3. Выводы

В результате исследований видно, что данным алгоритмом можно хорошо кластеризовать различные данные. Путём подбора параметров можно даже на плохо разделённых довести результат до визуально различимой кластеризации, которая очень близка к исходной. Использование критериев так же является удобным способом определения “правильности” кластеризации