оМинистерство науки и высшего образования Российской Федерации

федеральное государственное автономное образовательное учреждение

высшего образования

«Санкт-Петербургский политехнический университет Петра Великого»

(ФГАОУ ВО СПбПУ)

Институт промышленного менеджмента, экономики и торговли

Высшая инженерно-экономическая школа

**ЛАБОРАТОРНАЯ РАБОТА №4**

По дисциплине «Многомерный статистический анализ»

Кластеризация методом опорных векторов (SVC)

(семестр 2)

|  |  |  |  |
| --- | --- | --- | --- |
| Студент группы |  |  |  |
| 3740105/20101 |  | подпись, дата | К.С. Малышева. |

Оценка выполненной студентом работы:

|  |  |  |
| --- | --- | --- |
| Преподаватель,  Доцент, канд.физ-мат.наук |  |  |
|  | подпись, дата | Л.В. Павлова |

Санкт-Петербург – 2023

**Задача работы:**

Кластеризовать заранее подготовленные модельные данные с помощью метода опорных векторов. Провести исследование зависимости полученного результата от параметров q и p.

**Ход работы:**

1. Подготовить модельные данные объема 100, состоящие из точек в двумерном пространстве, допускающие a) хорошее разделение и b) плохое разделение.

2. Реализовать SVC-алгоритм. Для различных значений гиперпараметров (q и p) провести кластеризацию модельных данных (хорошо разделимых и плохо разделимых). Оценить качество кластеризации (визуально; вычислить внешние/внутренние критерии, |SVs|, |BSVs|, м.б., CPU-time). Выбрать оптимальную кластерзацию.

3. Привести графическую иллюстрацию результатов нескольких кластеризации, также представить результаты в виде таблиц.

4. Сделать выводы.

**Постановка задачи**

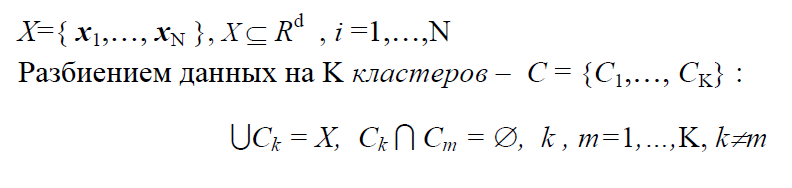
Автоматическая классификация объектов, или кластеризация, – разбиение совокупности объектов, статистически представленных матрицей «объект-свойство», на сравнительно небольшое число однородных групп, или кластеров. Кластерный анализ – это общее название множества вычислительных процедур, используемых при создании классификации. В результате работы с процедурами образуются кластеры, или группы очень похожих объектов.

Для анализа множества объектов достаточно:

– сохранить описание типичного представителя каждого кластера;

– перечислить номера объектов, входящих в данный кластер;

–указать максимальное отклонение каждого свойства от его среднего значения для данного кластера.



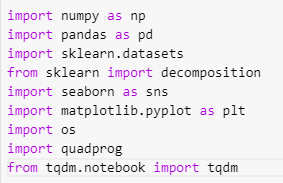
Согласно алгоритму SV-кластеризации (support vector clustering, SVC) исходные векторы, которые необходимо разбить на кластеры с помощью гауссова ядра k(xi,xj ) = exp(−q||xi − xj ||2) с параметром q > 0, отображаются в пространство признаков большей размерности, где отыскивается радиус наименьшей гиперсферы, охватывающей множество исходных векторов. При обратном отображении (в исходное пространство) эта сфера может разделяться на несколько контуров, каждый из которых включает в себя часть исходных векторов. Полученные контуры интерпретируются как границы кластеров.

Для следующей за этим процедуры маркировки кластеров (cluster labeling) могут применяться различные методики.Простая (с точки зрения понимания) методика маркировки основана на построении матрицы смежности. Далее кластеры определяются как связанные компоненты графа, индуцированного этой матрицей. Гиперпараметры SVC-алгоритма – параметр гауссова ядра q, который управляет масштабом исследуемых данных; параметр регуляризации, или штрафной параметр С, который контролирует долю точек, лежащих вне сферы, и помогает решать проблему перекры-вающихся кластеров и выбросов. Структура данных исследуется в процессе варьирования этих двух параметров при сохранении минимального количества опорных векторов, обеспечивающего гладкость границ кластеров.

a) Данные хорошо разделены

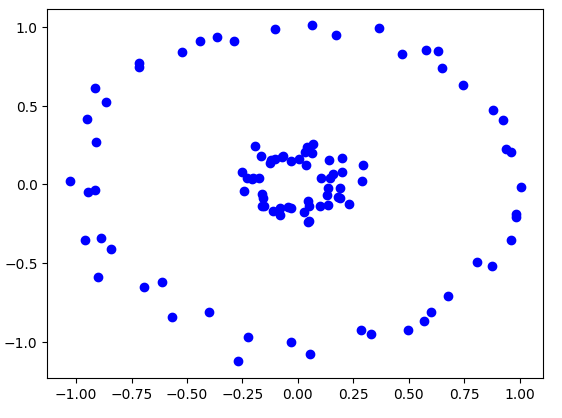
**1. Подготовка данных**

Для реализации алгоритма с помощью языка Python необходимо импортировать следующие библиотеки.



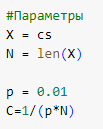
Сгенерируем хорошо разделенные данные из библиотеки sklearn.datasets.



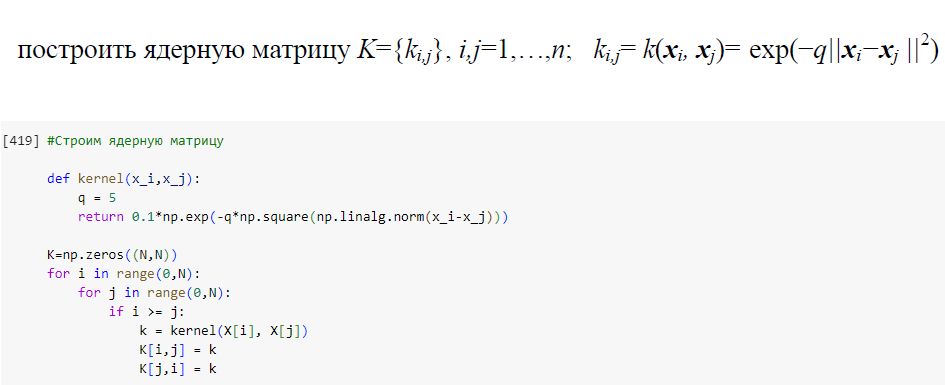


**2. Реализация SVC-алгоритма.**

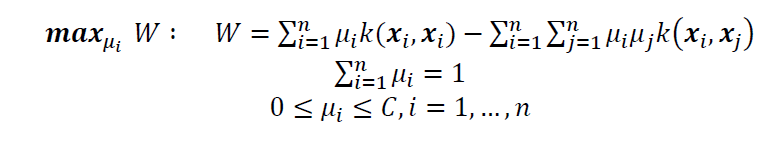
Зададим параметры для алгоритма.

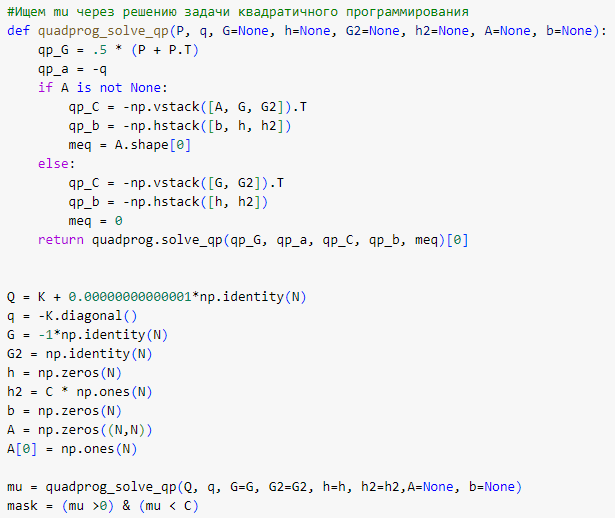


Построим ядерную матрицу с параметром q = 5 и p=0,01 (С=1).

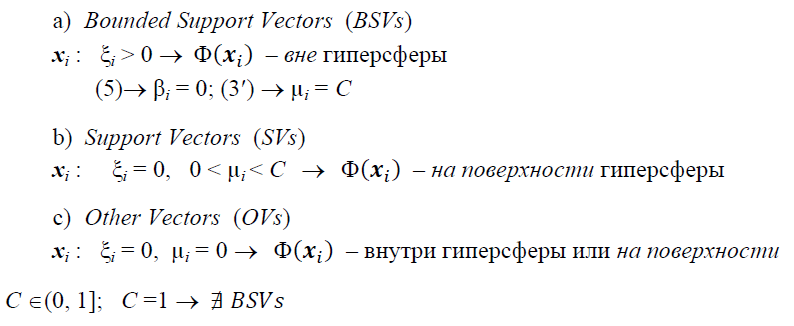


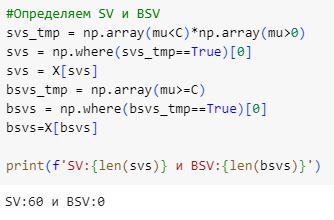
Далее решим задачу квадратического программирования и найдем множители Лагранжа µ.





Находим опорных вектора (SV) и связанные опорные вектора (BSV).

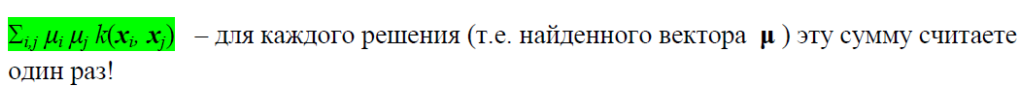


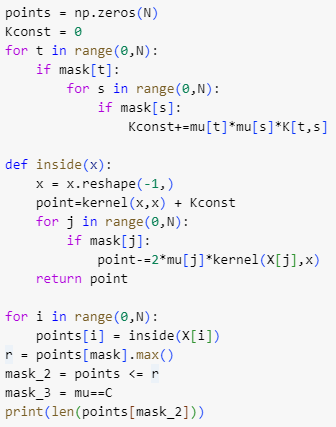


В данном случае количество опорных векторов составило 60 векторов из 100, а связанных векторов = 0.

Воспользуемся гауссовым ядром. В этом случае сложность кластерных контуров будет регулироваться единственным параметром q, управляющим «пропускной способностью», или шириной ядра. Для любого значения x∈ Х определим расстояние от точки Φ(x)до центра сферы a: r 2(x) = ||Φ(x) − a||2. Учитывая разложение a(34), получим

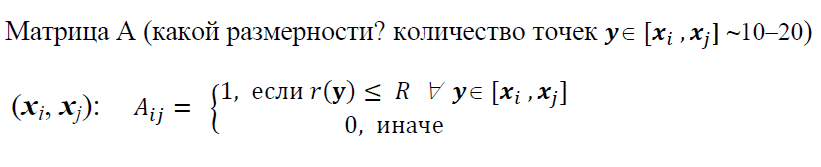




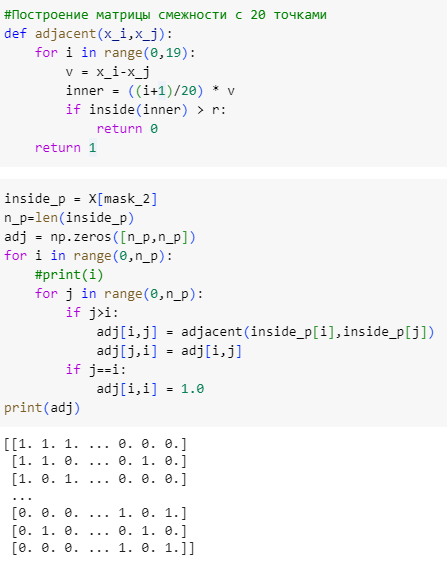


После определения границ кластеров требуется провести маркировку кластеров. Для этого используют геометрический подход с участием r (x), основанный на следующем наблюдении: если пара точек принадлежит разным кластерам, любой путь, соединяющий эти точки, должен выходить за границы сферы радиуса R с центром в точке a в пространстве признаков.

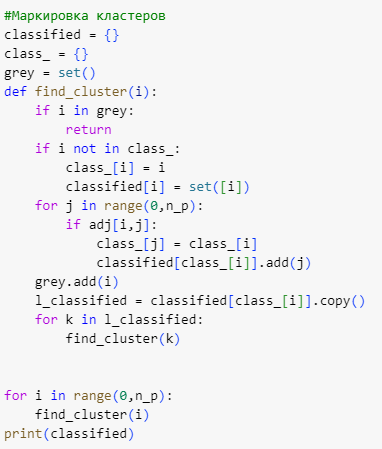
Исходя из этого, для пар точек xi и xj , чьи образы находятся внутри или на поверхности сферы, определяют элементы Aij матрицы смежности A:

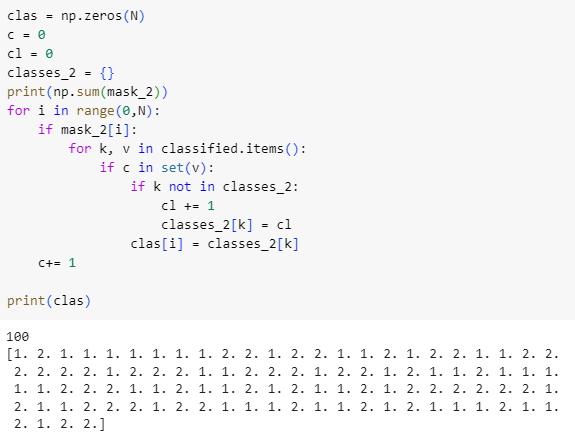


Таким образом, построим матрицу смежности с количеством точек 20.

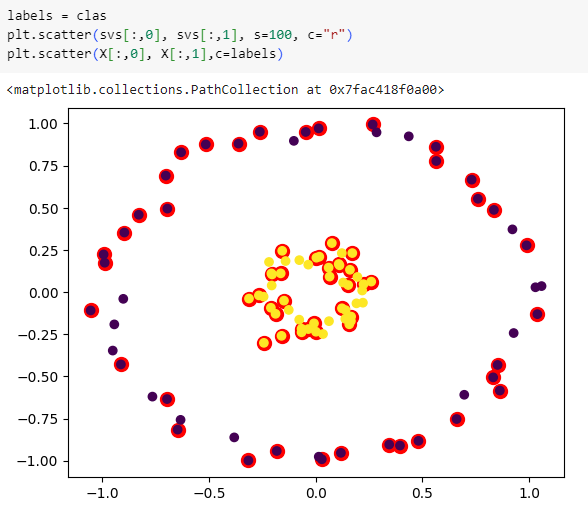


Далее на основании построенной матрицы промаркируем существующие значения и определим классы нашей выборки.





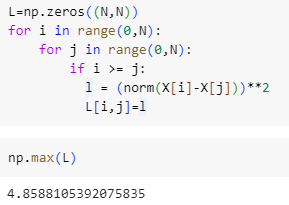
Таким образом, визуализируем получившиеся данные и отметим красным опорные вектора.



Другие варианты кластеризации и их графики представлены в таблице.

Для исследуемых данных наилучшей кластеризацией является кластеризация с параметром около q = 5, согласно следующей формуле

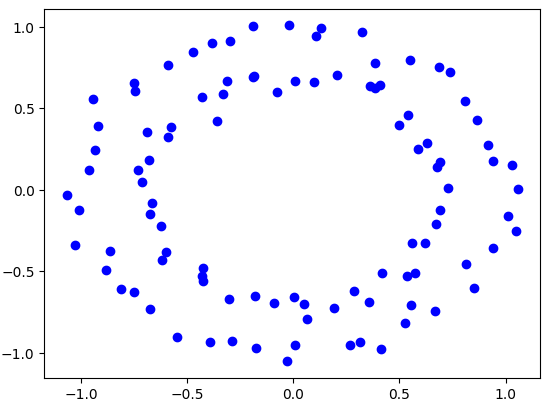




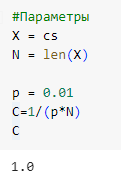
b) Плохо разделимые данные

Смоделируем также плохо разделимые данные в объеме 100 точек.

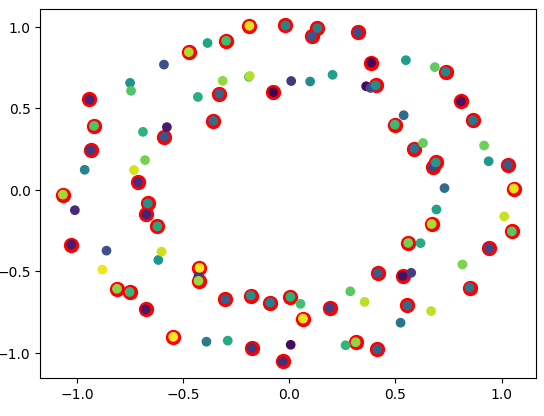




Повторим аналогичный прошлому случаю алгоритм SVC.

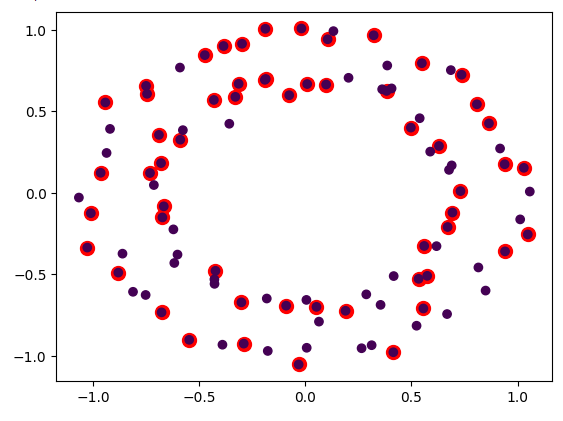


Таким образом, при параметре q=4 получается данный результат



Алгоритму сложно работать в условиях плохо разделимых данных.

Также продемонстрируем результат при q=1



4. Сделать выводы.

Наилучшая кластеризация достигается за счет подбора параметров q и p. В первую очередь необходимо увеличивать параметр q для достижения наименьшего количества SV векторов и при отсутствии BSV векторов. Если хорошего разбиения за счет увеличения q достичь не удается, увеличивают значение параметра p. При этом, так как часть опорных векторов трансформируется в связанные, а границы кластеров становятся более гладкими, количество опорных векторов уменьшается.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **p** | **q** | **C** | **Кол-во кластеров** | **Кол-во SVs** | **Кол-во BSVs** | **CPU-time** |  |
| 0,01 | 1 | 1 | 1 | 56 | 0 | 1 мин 2 сек |  |
| 0,01 | 2 | 1 | 100 | 58 | 0 | 5 сек |  |
| 0,01 | 5 | 1 | 2 | 65 | 0 | 36 сек |  |
| 0,01 | 10 | 1 | 2 | 74 | 0 | 43 сек |  |
| 0,01 | 50 | 1 | 100 | 84 | 0 | 8 сек |  |
| 0,001 | 1 | 10 | 1 | 60 | 2 | 1 мин 7 сек |  |
| 0,001 | 5 | 10 | 2 | 57 | 5 | 32 сек |  |
| 0,001 | 10 | 10 | 2 | 77 | 8 | 44 сек |  |