Министерство 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 №2**

По дисциплине «Многомерный статистический анализ»

Классификация многомерных объектов при наличии обучающей выборки (Классификация объектов на основе дискриминантного анализа)

(семестр 2)

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

Оценка выполненной студентом работы:

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

Санкт-Петербург – 2023

**Ход работы:**

1. Моделирование ОВ1 и ОВ2 – выборки заданного объема n1, n2 из нормального трехмерного распределения, т.е. моделирование последовательность векторов.

2. Моделирование тестовой выборки (ТВ) (или тестовой последовательности) заданного объема n, представленную смесью (других) объектов из двух предыдущих нормальных трехмерных распределений.

3. Вычисление выборочных оценок параметров распределения и построение классификатора, используя вместо значений параметров распределения – их оценки.

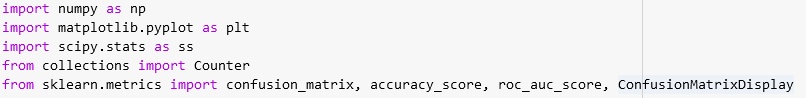
4. Работа с ТВ: классификация векторов ТВ, приведение результатов в виде Таблицы сопряженности.

5. Анализ полученных результатов.

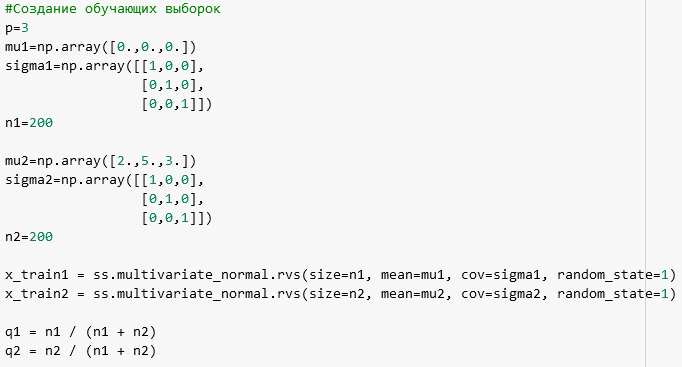
**1. Моделирование ОВ1 и ОВ2 – выборки заданного объема n1, n2 из нормального трехмерного распределения, т.е. моделирование последовательность векторов.**

Классификация объектов на основе дискриминантного анализа будет реализовываться на языке программирования Python.

Для начала необходимо импортировать следующие библиотеки.



Для создания и обучения классификатора на определенных данных необходимо смоделировать обучающие выборки. Моделирование обучающих выборок происходит из трехмерного нормального распределения с указанными параметрами. Для начала произведем анализ с хорошо разделенными данными.

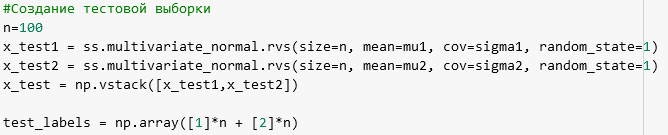


Объемы обучающих выборок возьмем 200 значений.

Рассчитаем также параметры q1 и q2 как долю объектов соответствующих выборок. В данном случае они будут одинаковы.

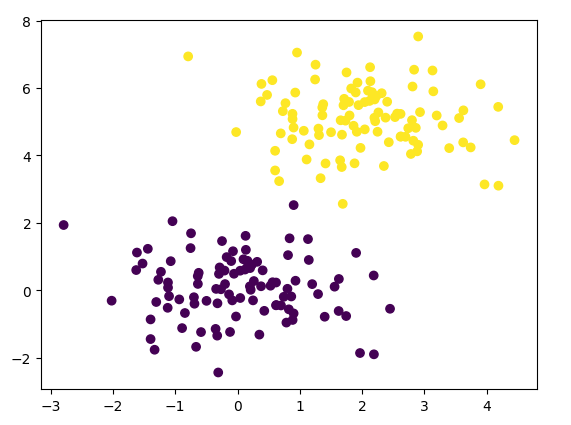
**2. Моделирование тестовой выборки (ТВ) заданного объема n, представленную смесью (других) объектов из двух предыдущих нормальных трехмерных распределений.**

На основе параметров распределения ОВ1 и ОВ2 смоделируем тестовую выборку как смесь объектов из двух предыдущих распределений. Размер тестовой выборки также будет составлять 200 значений (100 значений из распределения ОВ1 и 100 значений из распределения ОВ2).



Также по тестовой выборке присвоены метки и сохранены в пtременную test\_labels.

Отобразим тестовую выборку на графике.

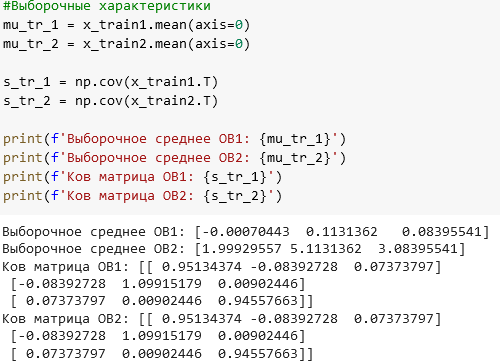


На графике ясно видно разделение на две группы.

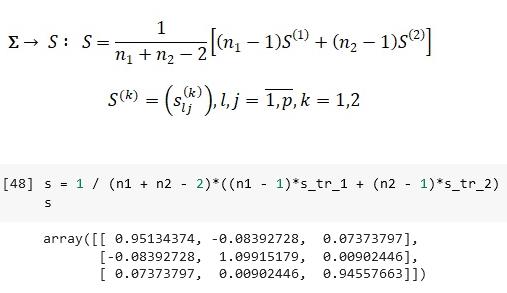
**3. Вычисление выборочных оценок параметров распределения и построение классификатора, используя вместо значений параметров распределения – их оценки.**

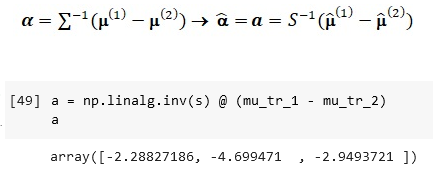
Классификатор будет строиться на основе выборочных оценок, для этого необходимо их рассчитать.

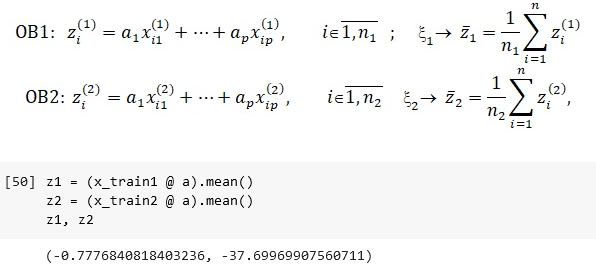
Рассчитаем выборочные характеристики ОВ1 и ОВ2.

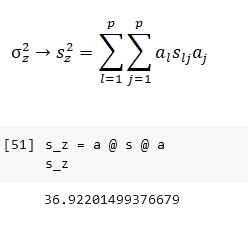


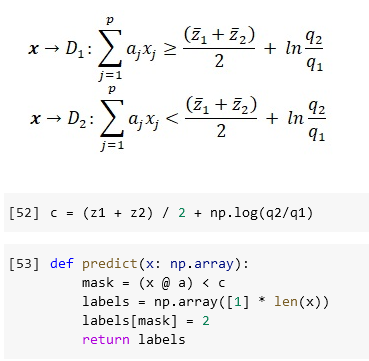
Далее на основании полученных оценок рассчитаем параметры и построим дискриминантные функции.





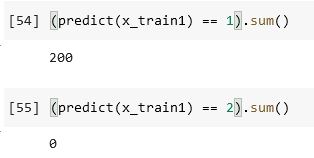




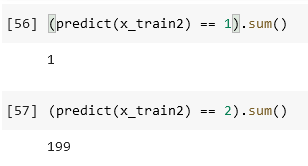


Проверим работу данного классификатора на уже разделенные ОВ1 и ОВ2.

Получили следующие результаты по ОВ1: к 1 классу принадлежат 200 значений из 200, ко 2 классу – 0 значений из 200. Так и должно быть объекты ОВ1 принадлежат первому распределению.

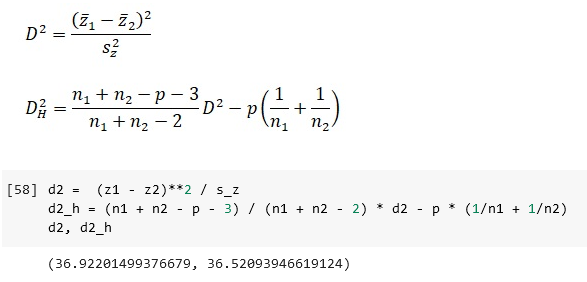


Получили следующие результаты по ОВ2: к 1 классу принадлежит 1 значение из 200, ко 2 классу – 199 значений из 200. В данном случае есть небольшая погрешность, но ситуация почти соответствует реальной.

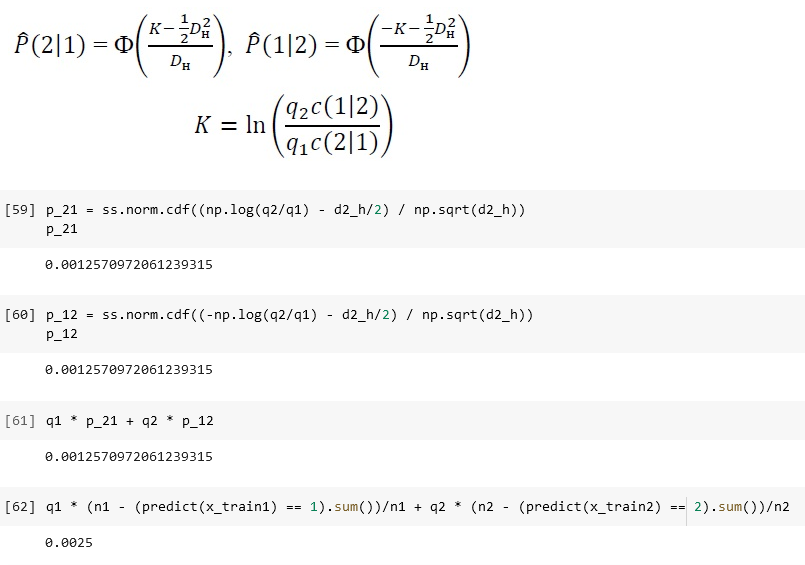


Далее оценим вероятности ошибочной классификации и ожидаемую вероятность ошибочной классификации .

Оценки расстояния Махаланобиса (смещенная и несмещенная):

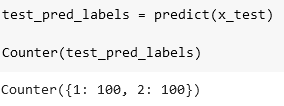


Вероятности ошибочной классификации и ожидаемая вероятность ошибочной классификации (одинаковые, так как размеры выборок были взяты одинаковые).



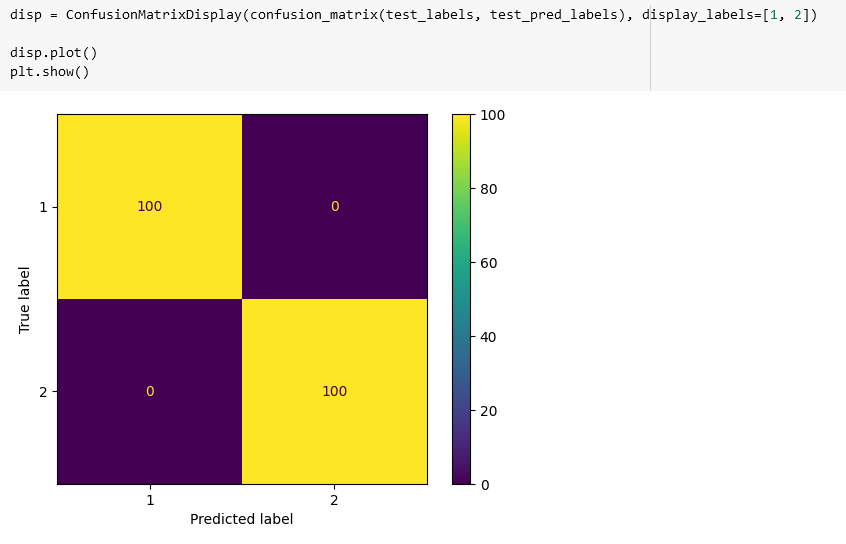
**4. Работа с ТВ: классификация векторов ТВ, приведение результатов в виде Таблицы сопряженности.**

Далее опробуем построенный классификатор на смоделированной тестовой выборке (объем ТВ = 200).

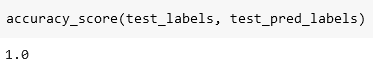


Таким образом, вышло, что к 1 классу нормального распределения относится 100 значений, ко 2 классу – также 100 значений. То есть тестовая выборка была поделена поровну, как и должно было выйти.

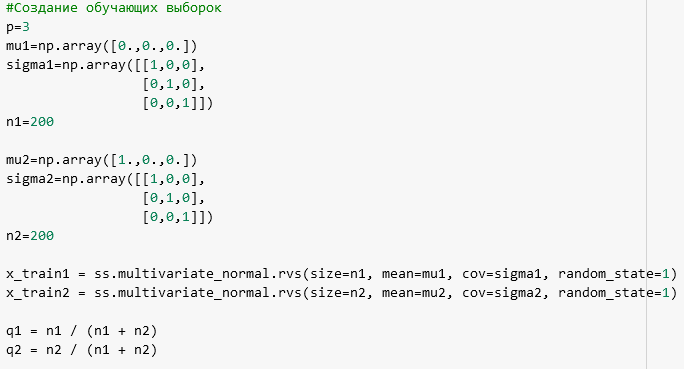
Построим таблицу сопряженности на основании полученного результата.



Таким образом, вышло следующее: все значения тестовой выборки были верно предсказаны. И в таком случае процент точности будет составлять 100%.

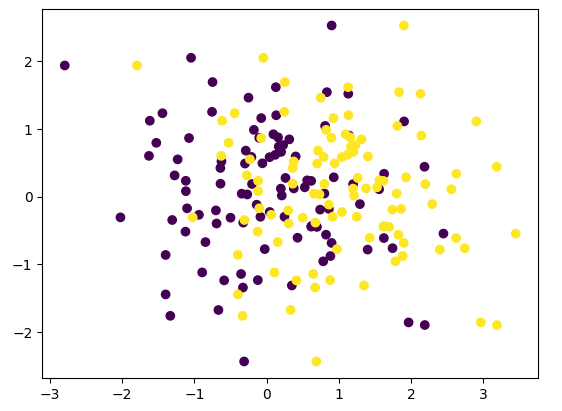


В данном случае данные изначально были хорошо разделены по классам, имея четко разные характеристики распределения. Повторим анализ для плохо разделенных данных.



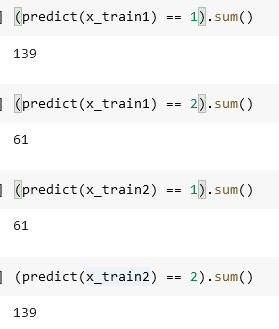
Процесс построения классификатора аналогичен предыдущему.

Отобразим на графике значения тестовой выборки.



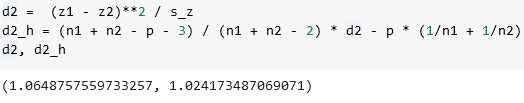
На графике видно, что данные плохо разделены.

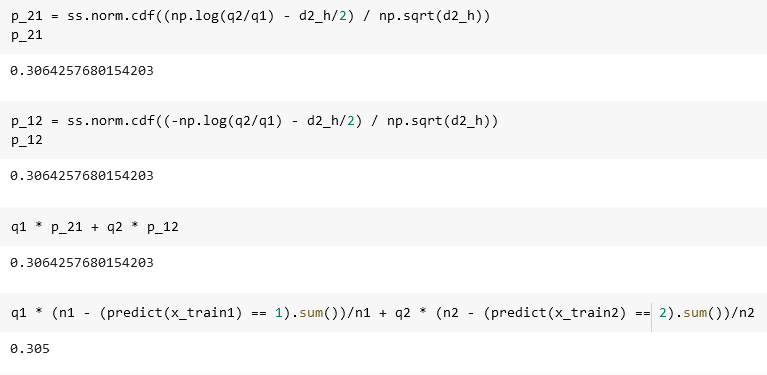
Проверим работу данного классификатора на уже разделенные ОВ1 и ОВ2.



В данном случае выходит, что верно было определено 139 значений, а ошибочно классифицированы – 61 значение.

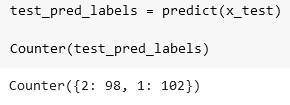
Далее оценим вероятности ошибочной классификации и ожидаемую вероятность ошибочной классификации .



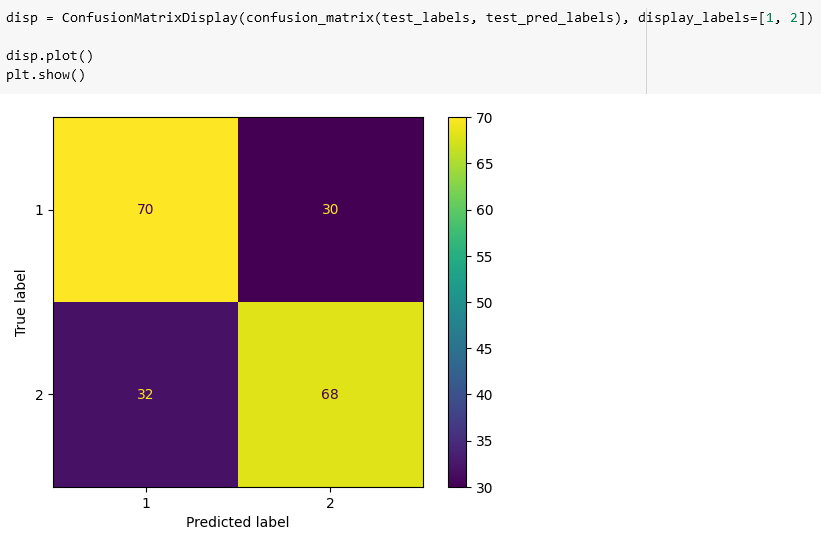


В данном случае значение ошибки будет гораздо больше.

Используем классификатор на тестовой выборке.

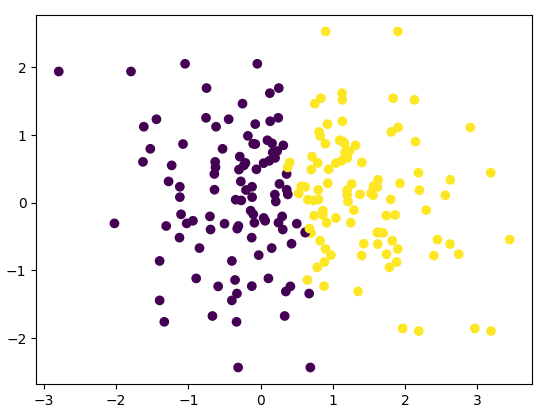


Значения были распределены неровно, что не совсем соответствует действительности. Построим таблицу сопряженности.



Таким образом, по 1 классу верно были определены 70 значений и неверно 30 значений; по 2 классу – 68 верных значений и 32 неверных значений.

Тестовую выборку классификатор разделил следующим образом.



Получается, точность классификатора составляет 69%.



**5. Анализ полученных результатов.**

В результате анализа можно сделать следующие выводы на тему классификации на основе дискриминантного анализа:

1) для обучения классификатора необходимо предварительно подготовить обучающие выборки с уже имеющимися метками данных;

2) для повышения точности классификатора обучающие данные должны быть хорошо разделены по классам и хорошо отличаться друг от друга, иначе точность классификатора будет понижаться как в смоделированном примере (100% при хорошо разделенных данных, 69% при плохо разделенных данных);

3) линейный дискриминантный анализ четко по линии разделяет имеющиеся данные на две зоны, согласно классовой принадлежности.