Санкт-Петербургский Политехнический Университет Петра Великого

Физико-механический институт

Кафедра прикладной математики и вычислительной физики

Отчёт по лабораторной работе №2 по дисциплине

«Многомерный статистический анализ»

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

Преподаватель: Павлова Л. В.

Санкт-Петербург

2023

# Постановка задачи

Необходимо построить классификатор и протестировать его на сгенерированных данных и на данных из репозитория.

# 1.1 Построение классификатора на модельных данных

Моделируем два класса, каждый объёмом 500 элементов и из нормального трёхмерного распределения с параметрами:

Суммарно выборка состоит из 1000 элементов, возьмём 200 значений на тестовую и 800 на тренировочную. Разбиение на тестовую и тренировочную будем проводить случайным образом ( для первого класса в выборке было 396 элементов, для второго 404)

Строим классификатор по формулам:

Изображение выглядит как текст

Автоматически созданное описаниеИзображение выглядит как текст

Автоматически созданное описание

Изображение выглядит как текст

Автоматически созданное описаниеОценки вероятностей ошибочной классификации считаются двумя способами:

1. С использованием матрицы ошибок, тогда , где
2. *Изображение выглядит как текст

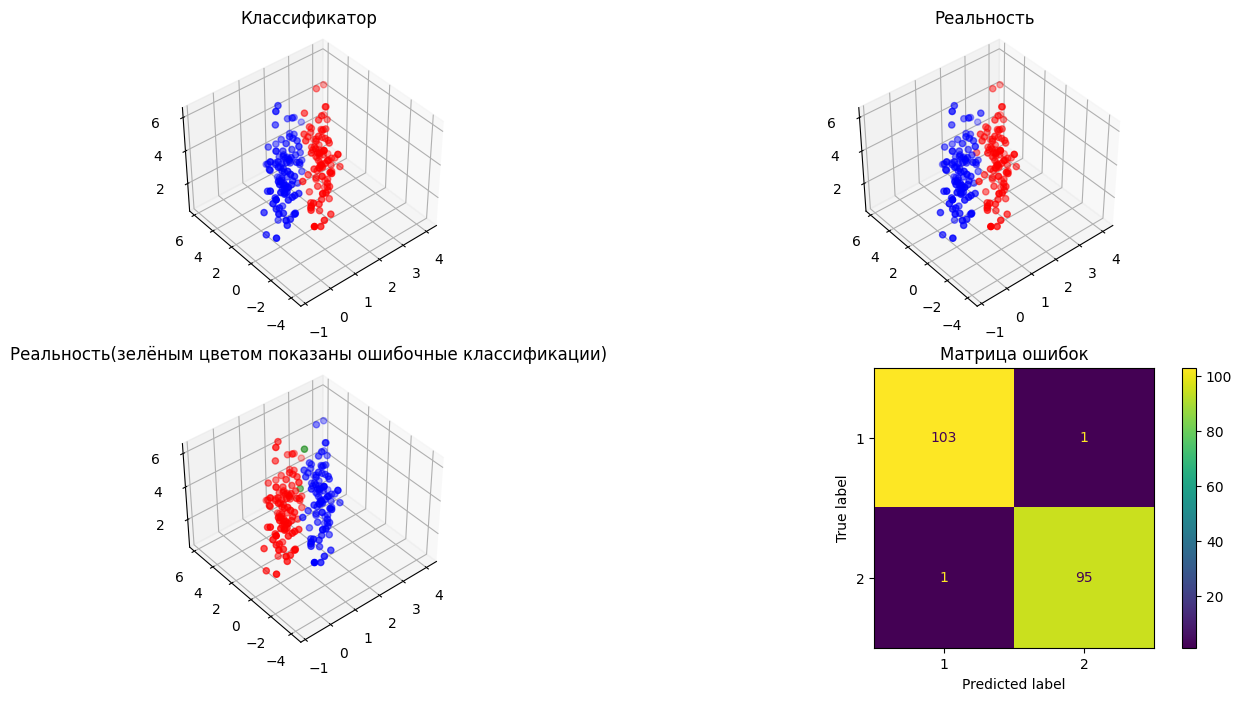
   Автоматически созданное описание*Метрика Махаланобиса ,

*Изображение выглядит как текст, монитор, снимок экрана

Автоматически созданное описаниеИзображение выглядит как текст

Автоматически созданное описание*

# 1.2 Результаты

**

*m/n P(2|1): 0.009615384615384616, P(1|2): 0.010416666666666666*

*Метрика P(2|1): 0.010282984646538642, P(1|2): 0.010049908930866*

Как видим, классификатор довольно хорошо разделил данные, из 200 элементов ошибка была только на 2. Вероятность ошибиться в обоих случаях равна примерно 0.01, что является неплохим результатом. Однако, такие результаты сильно зависят от выборки, которая, в данном случае, хорошо разделима.

# 2.1 Построение классификатора с использованием данных из репозитория

Выполним вышеописанные шаги, заменив выборку на скачанную из репозитория

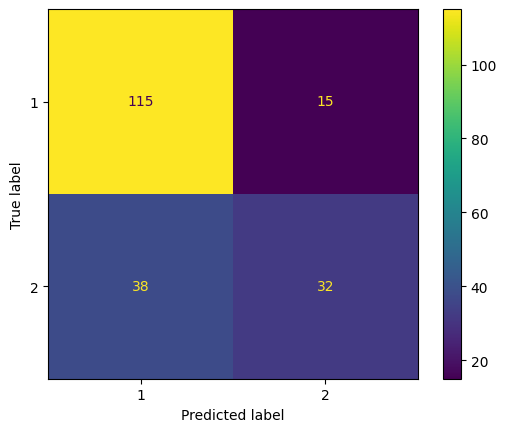
Изображение выглядит как текст

Автоматически созданное описание

# 2.2 Результаты

*m/n P(2|1): 0.11538461538461539, P(1|2): 0.5428571428571428*

*С помощью метрики и P(2|1): 0.0866706969583475, P(1|2): 0.5779697445528251*

**

На этих данных классификатор даёт менее хорошие результаты, поскольку это уже в какой-то степени плохо разделимая выборка. Можно заметить интересную особенность, что одна вероятность в несколько раз больше другой. Это объясняется тем, что в данных превалирует число элементов 1 класса, нежели 2.

# 3 Классификация данных (модельных и из репозитория) после перехода к главным компонентам (PCA)

Сократим число признаков с помощью метода главных компонент. Для отбора наиболее значимых признаков будем руководствоваться правилом

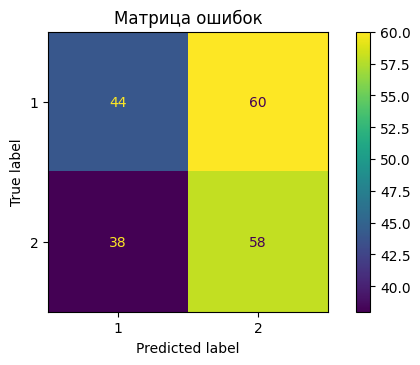
# Изображение выглядит как текст Автоматически созданное описание

# 3.1 На модельных данных

**Число признаков = 1 (выбрано с помощью правила, описанного выше)**

*m/n P(2|1): 0.5769230769230769, P(1|2): 0.3958333333333333*

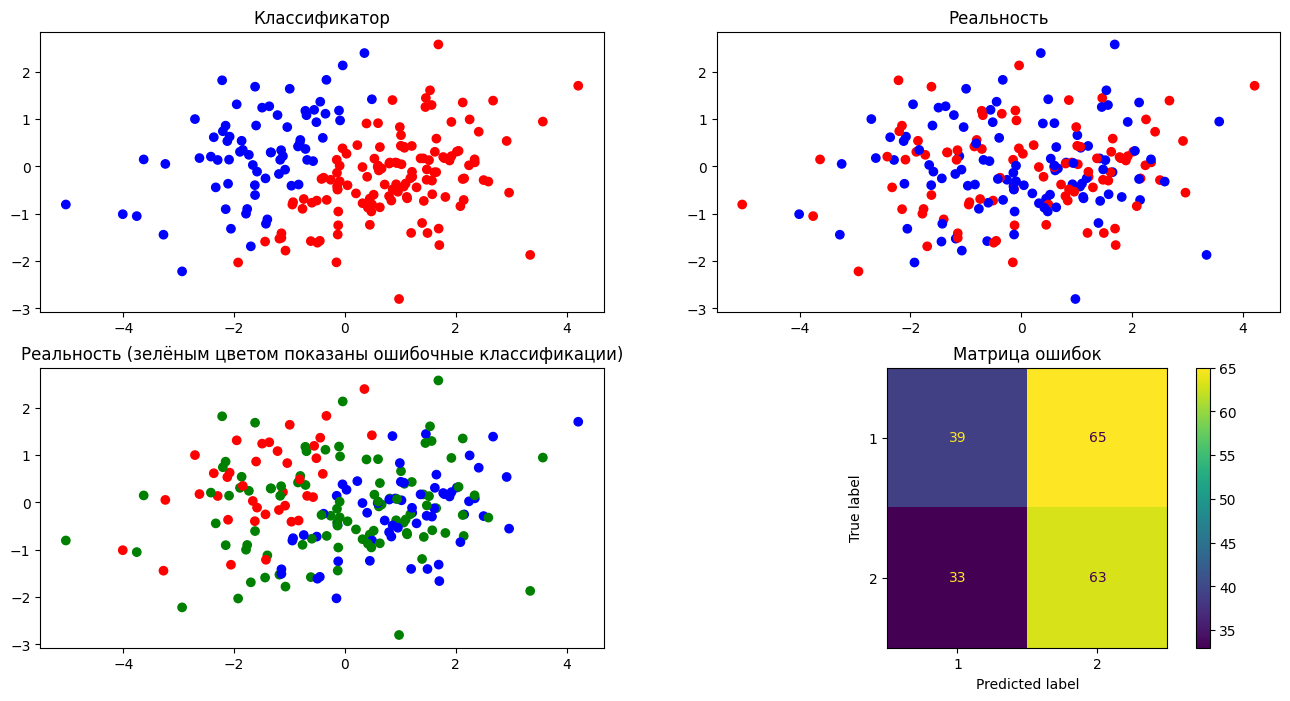
*Метрика P(2|1): 0.6185774395116671, P(1|2): 0.35867165994896477*

**

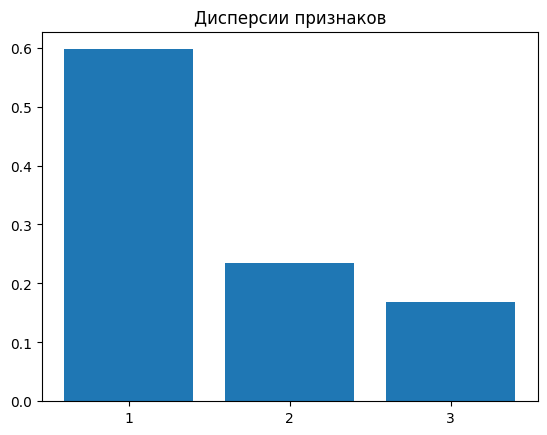
**Число признаков = 2 (выбрано вручную)**

*m/n P(2|1): 0.625, P(1|2): 0.34375*

*Метрика P(2|1): 0.8801368531697024, P(1|2): 0.1165204804628786*

**

По сравнению со случаем, когда мы не уменьшали число признаков, результаты значительно ухудшились. Такое результат можно обосновать тем, что у признаков примерно одинаковые дисперсии, поэтому убирая любую из компонент, мы сильно теряем в точности. Это как раз и показывает график ниже (дисперсии отсортированы в порядке убывания)

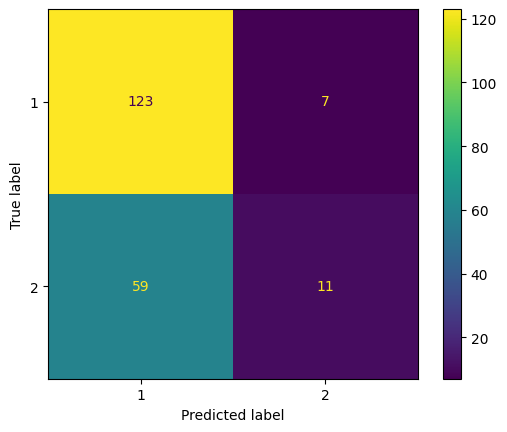


# 3.2 На данных из репозитория

**Число признаков = 3 (выбрано с помощью правила, описанного выше)**

*m/n P(2|1): 0.05384615384615385, P(1|2): 0.8428571428571429*

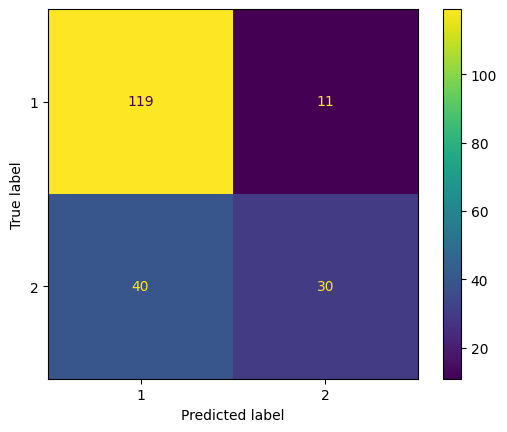
*С помощью метрики и P(2|1): 0.01986233307149921, P(1|2): 0.9398915064805747*

**

**Число признаков = 15 (выбрано вручную)**

*m/n P(2|1): 0.08461538461538462, P(1|2): 0.5714285714285714*

*С помощью метрики и P(2|1): 0.08405153172326751, P(1|2): 0.6144637270061754*

**

Как и ожидается, с уменьшением числа признаков (24 против 3 и 15) возрастает и ошибка. Если сравнить результат, полученный при 24 признаках и 15, то видим, что ошибка хоть и стала побольше, однако не настолько критичной, учитывая сокращение числа признаков на 9, что чуть больше 1/3 от общего их числа. На графике ниже видно, что дисперсии 3 признаков как раз выделяются на фоне остальных, что согласуется с “автоматической” оценкой.



# 4 Выводы

Уменьшая число признаков, получается выигрыш в вычислительной сложности, но падает точность, однако, учитывая тот факт, что возможен случай, когда счёт занимает много времени, то преимущества такого отсечения признаков очевидны. Стоит находить “золотую середину”, поскольку можно получить классификацию очень быстро и очень неточно, что не будет иметь никакой практической ценности