Учреждение образования
«Белорусский государственный университет
информатики и радиоэлектроники»
Кафедра информатики

Лабораторная работа №8
«Выявление аномалий»

Выполнил: Чёрный Родион Павлович
магистрант кафедры информатики
группа №858642

Проверил: доцент кафедры информатики
Стержанов Максим Валерьевич

Минск 2019

Постановка задачи

Набор данных **ex8data1.mat** представляет собой файл формата *.mat (т.е. сохраненного из Matlab). Набор содержит две переменные X_1 и X_2 - задержка в мс и пропускная способность в мб/с серверов. Среди серверов необходимо выделить те, характеристики которых аномальные. Набор разделен на обучающую выборку (X), которая не содержит меток классов, а также валидационную (Xval, yval), на которой необходимо оценить качество алгоритма выявления аномалий. В метках классов 0 обозначает отсутствие аномалии, а 1, соответственно, ее наличие.

Набор данных **ex8data2.mat** представляет собой файл формата *.mat (т.е. сохраненного из Matlab). Набор содержит 11-мерную переменную X - координаты точек, среди которых необходимо выделить аномальные. Набор разделен на обучающую выборку (X), которая не содержит меток классов, а также валидационную (Xval, yval), на которой необходимо оценить качество алгоритма выявления аномалий.

1. Загрузите данные **ex8data1.mat** из файла.
2. Постройте график загруженных данных в виде диаграммы рассеяния.
3. Представьте данные в виде двух независимых нормально распределенных случайных величин.
4. Оцените параметры распределений случайных величин.
5. Постройте график плотности распределения получившейся случайной величины в виде изолиний, совместив его с графиком из пункта 2.
6. Подберите значение порога для обнаружения аномалий на основе валидационной выборки. В качестве метрики используйте F1-меру.
7. Выделите аномальные наблюдения на графике из пункта 5 с учетом выбранного порогового значения.
8. Загрузите данные **ex8data2.mat** из файла.
9. Представьте данные в виде 11-мерной нормально распределенной случайной величины.
10. Оцените параметры распределения случайной величины.
11. Подберите значение порога для обнаружения аномалий на основе валидационной выборки. В качестве метрики используйте F1-меру.
12. Выделите аномальные наблюдения в обучающей выборке. Сколько их было обнаружено? Какой был подобран порог?

Описание реализации

1. Загрузим данные из файла ex8data1.mat:

```
mat = loadmat("ex8data1")
x = mat['X']
x_val, y_val = mat['Xval'], mat['yval'].flatten()
```

2. График исходных данных изображен на рисунке 1.

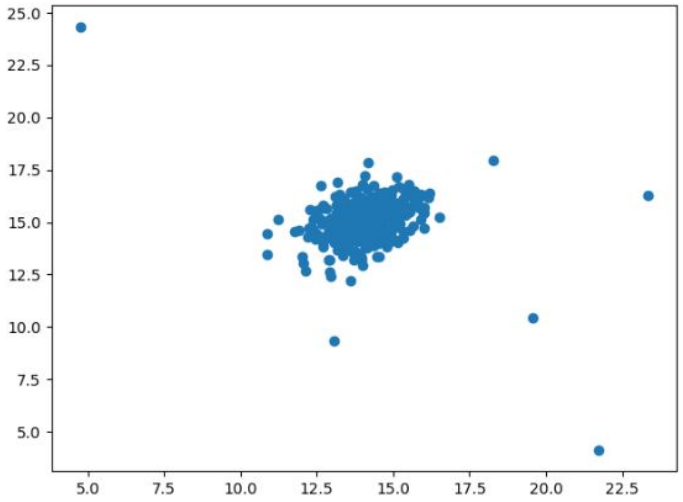

Рисунок 1 - График исходных данных

3. Построим гистограмму частот обеих случайных величин. Графики гистограмм случайных величин изображены на рисунке 2.

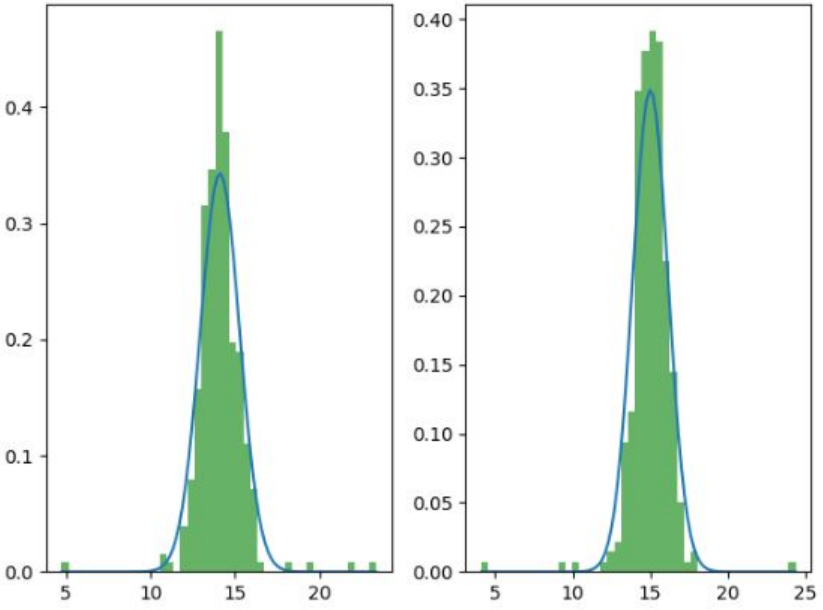

Рисунок 2 - Гистограммы случайных величин

4. Параметры распределения случайных величин:

mu = 14.112225783945592. sigma = 1.163506413768144.
mu = 14.99771050813621. sigma = 1.1434912786047284

5. Совместим график плотности распределения двумерной СВ в виде изолиний и совместим его с графиком исходных данных. Результат изображен на рисунке 3.

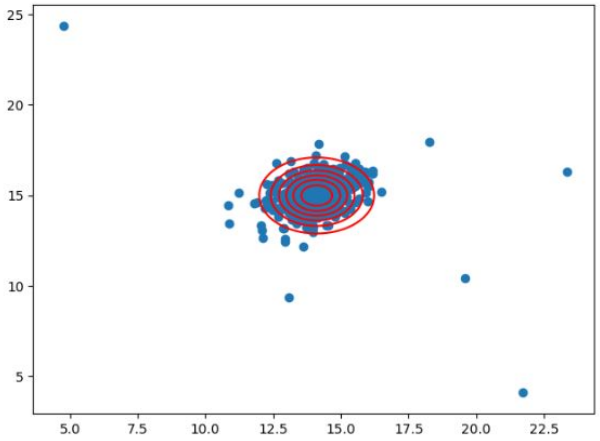

Рисунок 3 - Изолинии графика плотности распределения и исходные данные

6. Будем искать оптимальное значение порога для определения аномалий среди диапазона значений от 0.00001 до 0.001 с шагом 0.00001 при помощи валидационной выборки.

В результате получаем, что лучшее значение порога: 0.00001. При нем значение метрики f1 на валидационной выборке равняется 0.875.

7. Выделим аномальные наблюдения на графике из пункта 5, с учетом значения порога равному 0.00001. Результат изображен на рисунке 4.

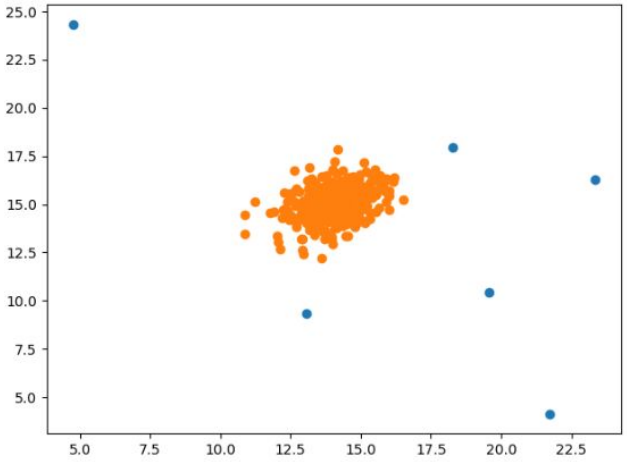

Рисунок 4 - Исходный график с отмеченными аномальными точками

8. Загрузим данные из файла ex8data2.mat:

```
mat = loadmat("ex8data2")
x = mat['X']
x_val, y_val = mat['Xval'], mat['yval'].flatten()
```

9. Данные, представленные в виде 11-мерной нормально распределенной случайной величины, изображены на рисунке 5.

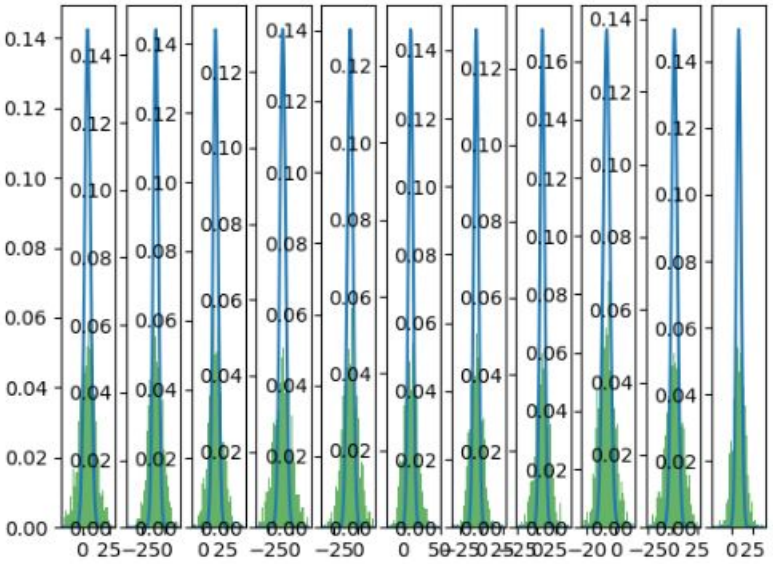

Рисунок 5 - Гистограмма частот и кривая плотности распределения 11-ти случайных величин

10. Их параметры распределения:

1. mu = 4.939400340737879. sigma = 2.7943947906576825
2. mu = -9.637268193926154. sigma = 2.7007823386651695
3. mu = 13.814707488470157. sigma = 2.7657801852276873
4. mu = -10.464488797333269. sigma = 3.02923683178236575
5. mu = -7.956229223945922. sigma = 2.8423402207249726
6. mu = 10.199503723381111. sigma = 3.076426586047039
7. mu = -6.019407545086613. sigma = 2.7310779382440424
8. mu = 7.969828957140624. sigma = 3.05549950395384339
9. mu = -6.253181900763459. sigma = 2.3330832279536344
10. mu = 2.3245128949467495. sigma = 2.900585181027919
11. mu = 8.473722523654011. sigma = 2.6658181078698453

11. Будем искать значение порога среди очень малых значений: 1e-67, 1e-44, 1.6e-44, 1.65e-44. Получим оптимальное значение порога равное 1e-44, которое дает f1 на валидационной выборке равный 0.67.

12. При значении порога 1e-44 на тренировочной выборке была обнаружена 51 аномалия из 1000, а на валидационной выборке - 11 аномалий из 100.