Министерство образования Республики Беларусь

Учреждение образования

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ

Факультет компьютерных систем и сетей

Кафедра информатики

Дисциплина: Машинное обучение (МО)

Отчет о лабораторной работе №8

«Выявление аномалий»

|  |
| --- |
| **Выполнил:** |
| **Студент гр. 858341** |
| **Немкович А. В.** |

|  |
| --- |
| **Проверил:** |
| **Стержанов М. В.** |

**Минск 2019**

Содержание

[Постановка задачи 3](#_Toc25275643)

[Ход выполнения 4](#_Toc25275644)

[Выводы 9](#_Toc25275645)

# Постановка задачи

Набор данных **ex8data1.mat** представляет собой файл формата \*.mat (т.е. сохраненного из Matlab). Набор содержит две переменные X1 и X2 - задержка в мс и пропускная способность в мб/c 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.

**Задание.**

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?
13. Ответы на вопросы представьте в виде отчета.

# Ход выполнения

**Задания 1-2**: Загрузка данных ex8data1.mat из файла. Построение графика загруженных данных в виде диаграммы рассеяния.

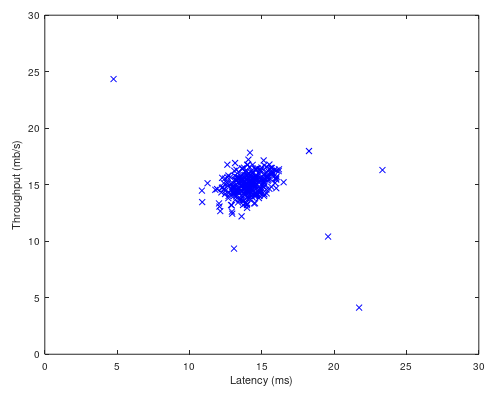
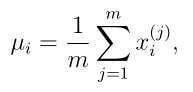
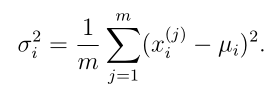


Рис 1. – График загруженных данных (ex8data1.mat)

**Задания 4-5**: Оценка параметров распределения случайных величин. Построение графика плотности распределения получившейся случайной величины в виде изолиний, совместив его с графиком из пункта 2.

Формулы для оценки параметров распределения случайных величин:





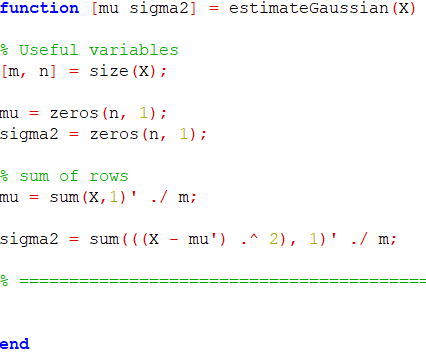


Рис 2. – Код функции оценки параметров распределения случайных величин

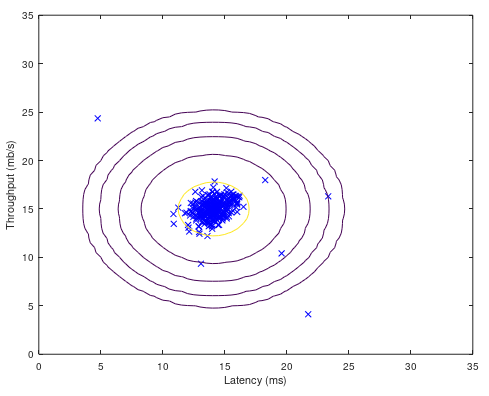
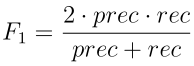


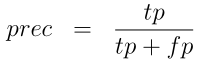
Рис 3. – График плотности распределения получившейся случайной величины в виде изолиний совместно с исходными данными

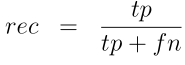
**Задание 6**: Подбор значения порога для обнаружения аномалий на основе валидационной выборки. В качестве метрики используя F1-меру.

Формула F1-меры:



где





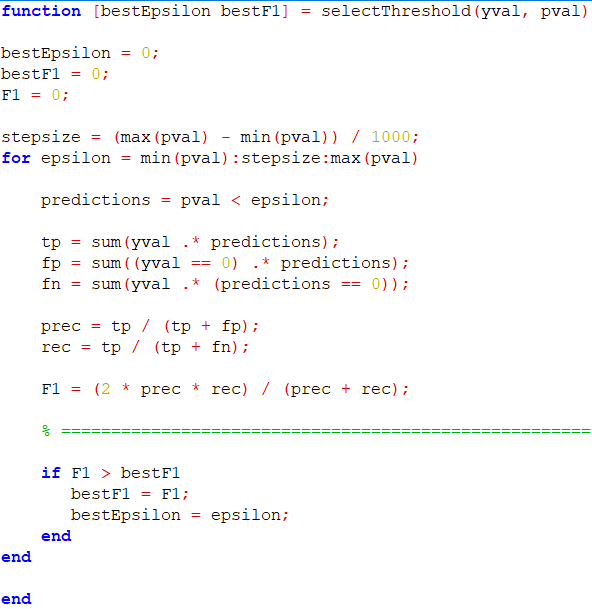


Рис 4. – Код функции подбора значения порога для обнаружения аномалий



Рис 5. – Результаты подбора значения порога для обнаружения аномалий

**Задание 7**: Выделение аномальных наблюдений на графике из пункта 5 с учетом выбранного порогового значения.

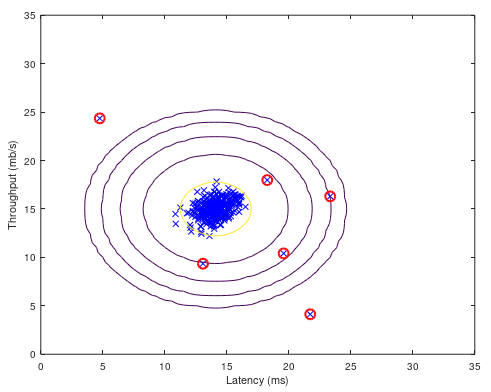


Рис 5. – График с выделением аномальных наблюдений

**Задания 8-11**: Загрузка данных ex8data2.mat из файла. Представление данных в виде 11-мерной нормально распределенной случайной величины. Представление данных в виде 11-мерной нормально распределенной случайной величины. Оценка параметров распределения случайной величины. Подбор значение порога для обнаружения аномалий на основе валидационной выборки. В качестве метрики используя F1-меру.

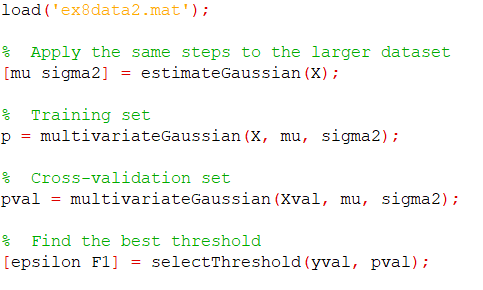


Рис 6. – Код подбора значений порога для обнаружения аномалий для данных из файла ex8data2.mat

**Задание 12**: Выделение аномальных наблюдений в обучающей выборке. Сколько их было обнаружено? Какой был подобран порог?

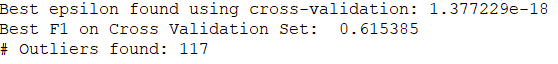


Рис 7. – Результаты подбора значения порога для обнаружения аномалий для данных из файла ex8data2.mat

# Выводы

Мы научились выявлять аномалии с помощью Гауссового распределения. А также подбирать значения порога для обнаружения аномалий на основе валидационной выборки. Метод выявления аномалий достаточно похож на обучение с учителем, но они имеют разные области применения. Сравнительная характеристика областей применения представлена следующей таблицей.

Таблица 1. Сравнение областей применения метода выявления аномалий и обучения с учителем

|  |  |
| --- | --- |
| **Выявление аномалий** | **Обучение с учителем** |
| Малое количество положительные примеры (y=1) | Большое количество позитивных и негативных примеров |
| Большое количество отрицательных примеров (y=0) |  |
| Множество различных типов аномалий |  |
| Будущие положительные примеры могут быть не похожи на примеры из обучающей выборки | Будущие положительные примеры могут похожи на примеры из обучающей выборки |