Учреждение образования

«Белорусский государственный университет

информатики и радиоэлектроники»

Кафедра информатики

ЛабоРАТОРНАЯ РАБОТА №8

«Выявление аномалий»

Выполнил: Яловчук Валерий Валерьевич

магистрант кафедры информатики

группа №858641

Проверил: доцент, кандидат технических наук Стержанов Максим Валерьевич

Минск 2019

ХОД РАБОТЫ

**Данные.**

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.

**Выполнение:**

1. Загрузите данные ex8data1.mat из файла:

file\_path = 'ex8data1.mat'

dataset = sio.loadmat(file\_path)

X = dataset["X"]

Xval = dataset["Xval"]

yval = dataset["yval"]

1. Постройте график загруженных данных в виде диаграммы рассеяния:

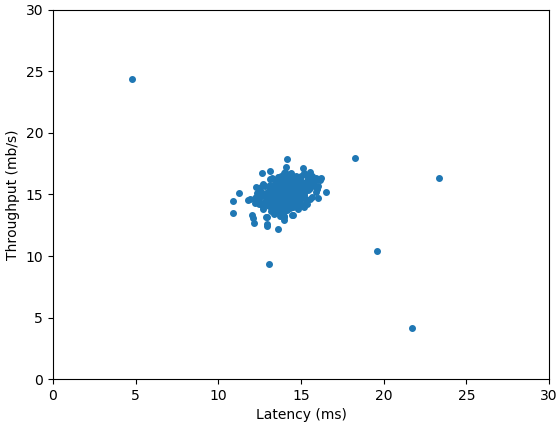


Рисунок 1 – визуализация исходных данных

3. Представьте данные в виде двух независимых нормально распределенных случайных величин:

def estimate\_gaussian(X):

mu = np.mean(X, axis=0)

sigma2 = np.var(X, axis=0)

return mu, sigma2

4. Оцените параметры распределений случайных величин:

p = multivariate\_normal(mu, np.diag(sigma2))

print(mu, sigma2)

Результат выполнения:

(array([14.11222578, 14.99771051]), array([1.83263141, 1.70974533]))

5. Постройте график плотности распределения получившейся случайной величины в виде изолиний, совместив его с графиком из пункта 2:

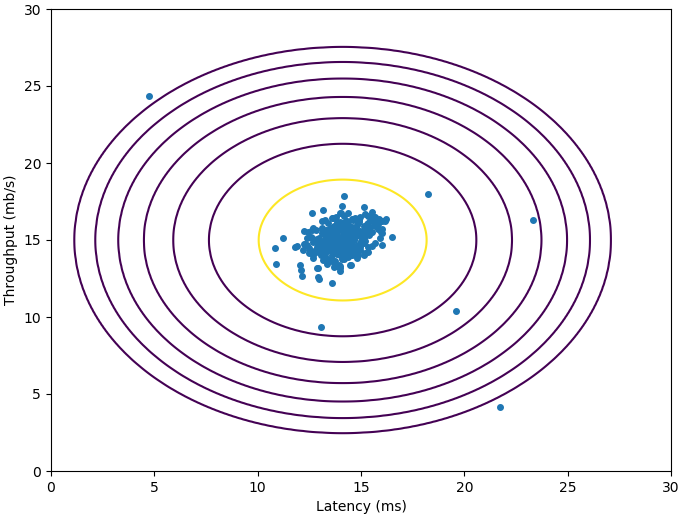


Рисунок 2 – собственные векторы матрицы ковариации

6. Подберите значение порога для обнаружения аномалий на основе валидационной выборки. В качестве метрики используйте F1-меру:

def select\_threshold(y\_val, p\_val):

best\_epsilon, best\_F1 = 0, 0

step\_size = (max(p\_val) - min(p\_val)) / 1000

for epsilon in np.arange(p\_val.min(), p\_val.max(), step\_size):

predictions = (p\_val < epsilon)[:, np.newaxis]

tp = np.sum(predictions[y\_val == 1] == 1)

fp = np.sum(predictions[y\_val == 0] == 1)

fn = np.sum(predictions[y\_val == 1] == 0)

prec = tp / (tp + fp)

rec = tp / (tp + fn)

F1 = 2 \* prec \* rec / (prec + rec)

if F1 > best\_F1:

best\_epsilon = epsilon

best\_F1 = F1

return best\_epsilon, best\_F1

p\_val = p.pdf(Xval)

epsilon, F1 = select\_threshold(yval, p\_val)

print("Best epsilon found using cross-validation:", epsilon)

print("Best F1 on Cross Validation Set:", F1)

Результат выполнения:

('Best epsilon found using cross-validation:', 8.990852779269493e-05)

('Best F1 on Cross Validation Set:', 0.8750000000000001)

7. Выделите аномальные наблюдения на графике из пункта 5 с учетом выбранного порогового значения:

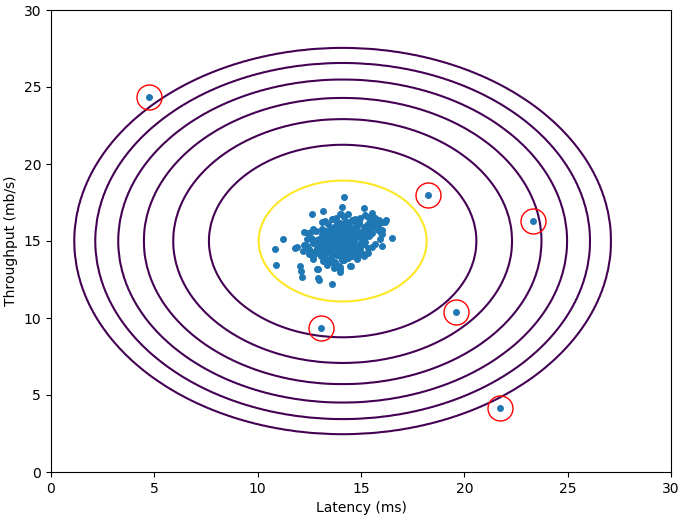


Рисунок 3 – аномальные значения, с учетом выбранного порового значения

8. Загрузите данные ex8data2.mat из файла:

file\_path = 'ex8data2.mat'

dataset = sio.loadmat(file\_path)

X, Xval, yval = dataset['X'], dataset['Xval'], dataset['yval'][:, 0]

9. Представьте данные в виде 11-мерной нормально распределенной случайной величины:

mu, sigma2 = estimate\_gaussian(X)

10. Оцените параметры распределения случайной величины:

p = multivariate\_normal(mu, np.diag(sigma2))

11. Подберите значение порога для обнаружения аномалий на основе валидационной выборки. В качестве метрики используйте F1-меру:

p\_val = p.pdf(Xval)

epsilon, F1 = select\_threshold(yval, p\_val)

print('Best epsilon found using cross-validation: %.2e' % epsilon)

print('Best F1 on Cross Validation Set : %f\n' % F1)

print(' (you should see a value epsilon of about 1.38e-18)')

print(' (you should see a Best F1 value of 0.615385)')

Результат выполнения:

Best epsilon found using cross-validation: 1.38e-18

Best F1 on Cross Validation Set : 0.615385

(you should see a value epsilon of about 1.38e-18)

(you should see a Best F1 value of 0.615385)

12. Выделите аномальные наблюдения в обучающей выборке. Сколько их было обнаружено? Какой был подобран порог?

print('\nAnomalies found: %d' % np.sum(p.pdf(X) < epsilon))

Результат выполнения:

Anomalies found: 117