Министерство образования Республики Беларусь

Учреждение образования

“Белорусский государственный университет информатики и радиоэлектроники”

Факультет компьютерных систем и сетей

Специальность “Обработка больших объемов информации”

Лабораторная работа №8

**Выявление аномалий**

Выполнил:

магистрант гр. 858641 Кальман В.А.

Проверил:

Стержанов М. В.

Минск 2019

**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.

**1. Загрузите данные ex8data1.mat из файла.**

ex8data1 = io.loadmat('Data/Lab 8/ex8data1.mat')

X, X\_val, y\_val = ex8data1['X'], ex8data1['Xval'], ex8data1['yval'].ravel()

**2. Постройте график загруженных данных в виде диаграммы рассеяния.**

plt.figure(figsize=(16,10))

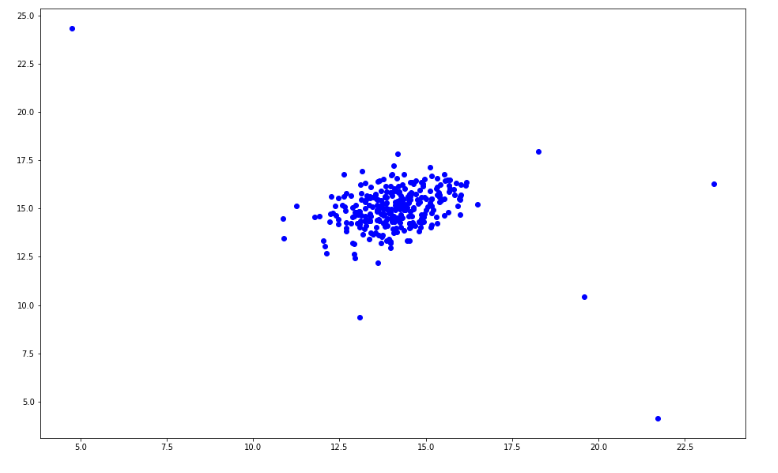
plt.scatter(X[:,0], X[:,1], color='b')

Рисунок 1 – визуализация исходных данных

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

plt.figure(figsize=(16,10))

std1 = np.std(X[:, 0])

mean1 = np.mean(X[:, 0])

plt.plot(norm.pdf(np.linspace(mean1-3\*std1, mean1+3\*std1, 100), mean1, std1))

std2 = np.std(X[:, 1])

mean2 = np.mean(X[:, 1])

plt.plot(norm.pdf(np.linspace(mean2-3\*std2, mean2+3\*std2, 100), mean2, std2))

**4. Оцените параметры распределений случайных величин.**

print('X\_1: mean = **{}**, std = **{}**'.format(mean1, std1))

print('X\_2: mean = **{}**, std = **{}**'.format(mean2, std2))

Результат работы:

X\_1: mean = 14.112225783945592, std = 1.353747174879607

X\_2: mean = 14.99771050813621, std = 1.3075723042450764

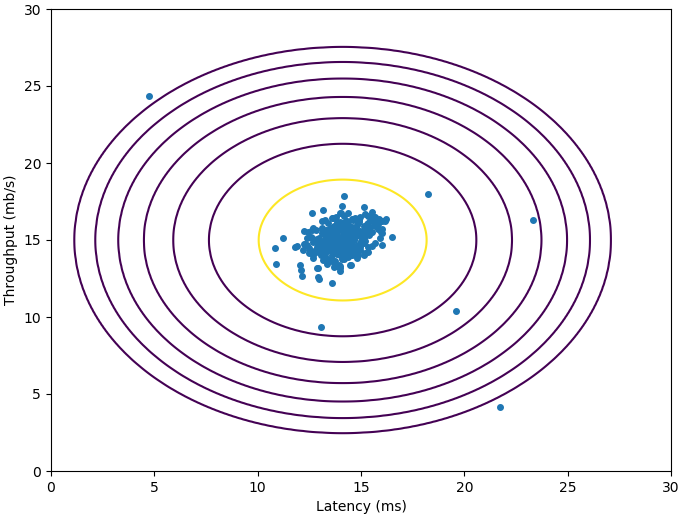
  
**5. Постройте график плотности распределения получившейся случайной величины в виде изолиний, совместив его с графиком из пункта 2.**

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.**

x, y = np.mgrid[np.min(X[:, 0])-1:np.max(X[:, 0])+1:.1, np.min(X[:, 1])-1:np.max(X[:, 1])+1:.1]

pos = np.dstack((x, y))

rv = multivariate\_normal([mean1, mean2], [[std1\*\*2, 0.], [0., std2\*\*2]])

plt.figure(figsize=(16,10))

plt.contour(x, y, rv.pdf(pos), levels=np.linspace(0, rv.pdf([mean1, mean2]), 50))

mask\_anom = rv.pdf(X) < np.mean(potential\_thr)

plt.scatter(X[mask\_anom][:, 0], X[mask\_anom][:, 1], color='r')

plt.scatter(X[~mask\_anom][:, 0], X[~mask\_anom][:, 1], color='b')

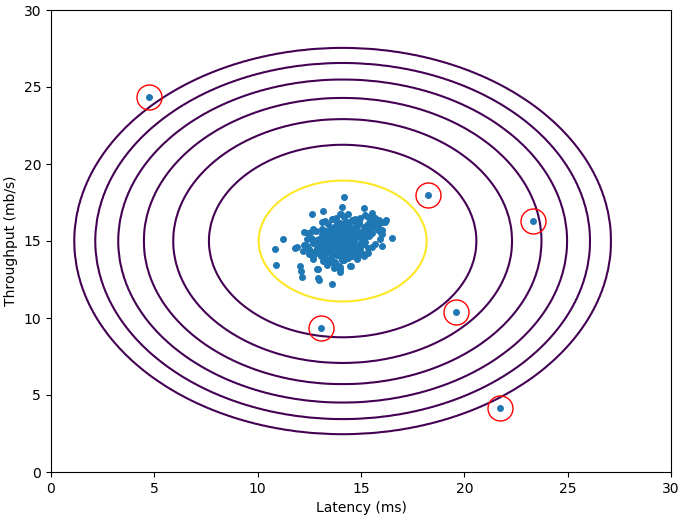


Рисунок 3 – аномальные значения, с учетом выбранного порового значения

**8. Загрузите данные ex8data2.mat из файла.**

ex8data2 = io.loadmat('Data/Lab 8/ex8data2.mat')

X, X\_val, y\_val = ex8data2['X'], ex8data2['Xval'], ex8data2['yval'].ravel()

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

mean\_vec = np.mean(X, axis=0)

cov\_mtx = np.cov(X.T)

rv = multivariate\_normal(mean\_vec, cov\_mtx)

**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

**Вывод**

Выявление аномалий (также обнаружение выбросов) — это опознавание во время интеллектуального анализа данных редких данных, событий или наблюдений, которые вызывают подозрения ввиду существенного отличия от большей части данных. Обычно аномальные данные превращаются в некоторый вид проблемы, такой как мошенничество в банке, структурный дефект, медицинские проблемы или ошибки в тексте. Аномалии также упоминаются как выбросы, необычности, шум, отклонения или исключения.