|  |  |  |  |
| --- | --- | --- | --- |
| Министерство образования Республики Беларусь  Учреждение образования  Белорусский Государственный Университет Информатики и Радиоэлектроники | | | |
| Факультет | Компьютерных сетей и систем | | |
| Кафедра | Информатики  Дисциплина: Конструирование те технологии электронных вычислительных средств | | |
|  |  | | |
| **лабораторная работа №8**  по курсу Машинное обучение  на тему  **Выявление аномалий** | | | |
| Магистрант:  гр. 858341  Кудин Н.И. | |  | Проверил:  Стержанов М.В. |
| Минск, 2019 | | | |

**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.

**Ход работы**

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

*datafile = 'ex8data1.mat'*

*mat = scipy.io.loadmat( datafile )*

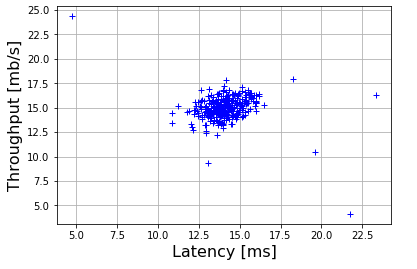
*X = mat['X']*

*ycv = mat['yval']*

*Xcv = mat['Xval']*

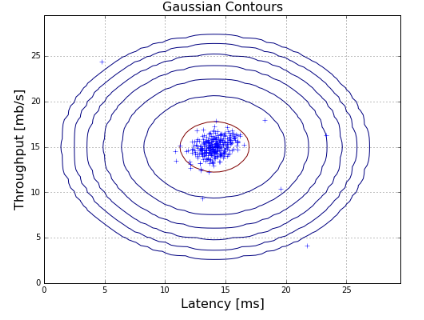
1. Построим график загруженных данных в виде диаграммы рассеяния.

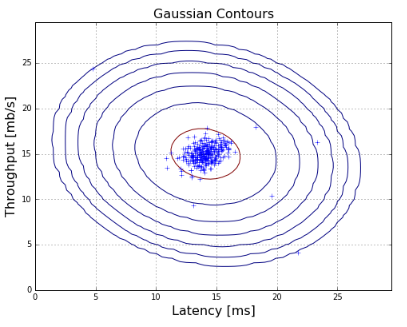
Представим данные в виде двух независимых нормально распределенных случайных величин.



1. Оценим параметры распределений случайных величин.

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





1. Подберем значение порога для обнаружения аномалий на основе валидационной выборки. В качестве метрики используйте F1-меру.

*def computeF1(predVec, trueVec):*

*"""*

*F1 = 2 \* (P\*R)/(P+R)*

*where P is precision, R is recall*

*Precision = "of all predicted y=1, what fraction had true y=1"*

*Recall = "of all true y=1, what fraction predicted y=1?*

*Note predictionVec and trueLabelVec should be boolean vectors.*

*"""*

*#print predVec.shape*

*#print trueVec.shape*

*#assert predVec.shape == trueVec.shape*

*P, R = 0., 0.*

*if float(np.sum(predVec)):*

*P = np.sum([int(trueVec[x]) for x in xrange(predVec.shape[0]) \*

*if predVec[x]]) / float(np.sum(predVec))*

*if float(np.sum(trueVec)):*

*R = np.sum([int(predVec[x]) for x in xrange(trueVec.shape[0]) \*

*if trueVec[x]]) / float(np.sum(trueVec))*

*return 2\*P\*R/(P+R) if (P+R) else 0*

*def selectThreshold(myycv, mypCVs):*

*"""*

*Function to select the best epsilon value from the CV set*

*by looping over possible epsilon values and computing the F1*

*score for each.*

*"""*

*# Make a list of possible epsilon values*

*nsteps = 1000*

*epses = np.linspace(np.min(mypCVs),np.max(mypCVs),nsteps)*

*# Compute the F1 score for each epsilon value, and store the best*

*# F1 score (and corresponding best epsilon)*

*bestF1, bestEps = 0, 0*

*trueVec = (myycv == 1).flatten()*

*for eps in epses:*

*predVec = mypCVs < eps*

*thisF1 = computeF1(predVec, trueVec)*

*if thisF1 > bestF1:*

*bestF1 = thisF1*

*bestEps = eps*

*print "Best F1 is %f, best eps is %0.4g."%(bestF1,bestEps)*

*return bestF1, bestEps*

*# Using the gaussian parameters from the full training set,*

*# figure out the p-value for each point in the CV set*

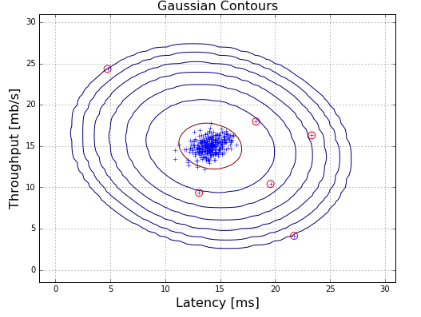
*pCVs = gaus(Xcv, mu, sig2)*

*#You should see a value for epsilon of about 8.99e-05.*

*bestF1, bestEps = selectThreshold(ycv,pCVs)*

**Best F1 is 0.875000, best eps is 9.075e-05.**

1. Выделим аномальные наблюдения на графике из пункта 5 с учетом выбранного порогового значения.



1. Загрузим данные **ex8data2.mat** из файла.

*datafile = 'data/ex8data2.mat'*

*mat = scipy.io.loadmat( datafile )*

*Xpart2 = mat['X']*

*ycvpart2 = mat['yval']*

*Xcvpart2 = mat['Xval']*

*print 'Xpart2 shape is ', Xpart2.shape*

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

**Xpart2 shape is (1000, 11)**

1. Подберем значение порога для обнаружения аномалий на основе валидационной выборки. В качестве метрики используем F1-меру.

Выделим аномальные наблюдения в обучающей выборке.

*mu, sig2 = getGaussianParams(Xpart2, useMultivariate=False)*

*ps = gaus(Xpart2, mu, sig2)*

*psCV = gaus(Xcvpart2, mu, sig2)*

*# Using the gaussian parameters from the full training set,*

*# figure out the p-value for each point in the CV set*

*pCVs = gaus(Xcvpart2, mu, sig2)*

*# You should see a value epsilon of about 1.38e-18, and 117 anomalies found.*

*bestF1, bestEps = selectThreshold(ycvpart2,pCVs)*

*anoms = [Xpart2[x] for x in xrange(Xpart2.shape[0]) if ps[x] < bestEps]*

*print '# of anomalies found: ',len(anoms)*

**Best F1 is 0.615385, best eps is 3.456e-18.**

**# of anomalies found: 117**

**Заключение**

Таким образом, можно сказать, что выявление аномалий — 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.

Были рассмотрены наиболее популярные методы и техники выявления аномалий, приведены примеры практического использования данных техник в проектах, проиллюстрированы примеры работы алгоритмов по выявлению аномалий.

На примере sklearn были детально описаны важные параметры реализации алгоритмов по поиску аномалий. Выделены ключевые моменты при разработке модели обучения для определения выбросов.

Были приведены различия между ключевыми различиями шума и аномалии, определены ключевые параметры, которые помогут определить тип выброса.