Учреждение образования

«Белорусский государственный университет

информатики и радиоэлектроники»

Кафедра информатики

ЛабоРАТОРНАЯ РАБОТА №2

«Логистическая регрессия. Многоклассовая классификация»

Выполнил: Яловчук Валерий Валерьевич

магистрант кафедры информатики

группа №858641

Проверил: доцент, кандидат технических наук Стержанов Максим Валерьевич

Минск 2019

ХОД РАБОТЫ

**Данные.**

Набор данных ex2data1.txt представляет собой текстовый файл, содержащий информацию об оценке студента по первому экзамену (первое число в строке), оценке по второму экзамену (второе число в строке) и поступлении в университет (0 - не поступил, 1 - поступил).

Набор данных ex2data2.txt представляет собой текстовый файл, содержащий информацию о результате первого теста (первое число в строке) и результате второго теста (второе число в строке) изделий и результате прохождения контроля (0 - контроль не пройден, 1 - контроль пройден).

Набор данных ex2data3.mat представляет собой файл формата \*.mat (т.е. сохраненного из Matlab). Набор содержит 5000 изображений 20x20 в оттенках серого. Каждый пиксель представляет собой значение яркости (вещественное число). Каждое изображение сохранено в виде вектора из 400 элементов. В результате загрузки набора данных должна быть получена матрица 5000x400. Далее расположены метки классов изображений от 1 до 9 (соответствуют цифрам от 1 до 9), а также 10 (соответствует цифре 0).

**Выполнение:**

1. Загрузите данные ex2data1.txt из текстового файла:

file\_path = 'ex2data1.txt'

data = pd.read\_csv(file\_path, header=None)

X = data.iloc[:, :-1] # first 2 column

y = data.iloc[:, 2] # last column

data.head()

2. Постройте график, где по осям откладываются оценки по предметам, а точки обозначаются двумя разными маркерами в зависимости от того, поступил ли данный студент в университет или нет:

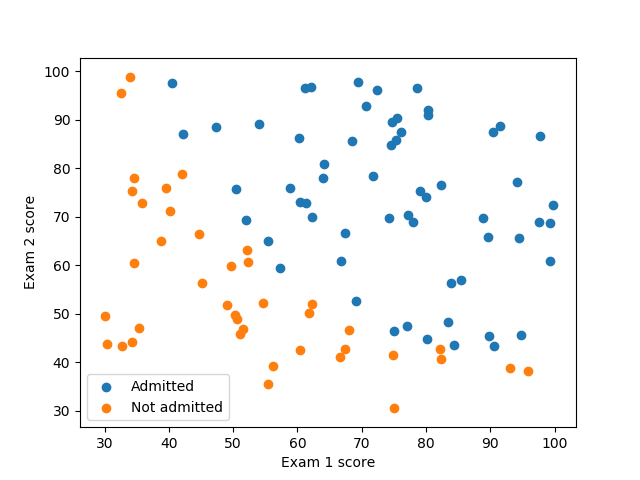
****

Рисунок 1 – график поступления студентов в университет

3. Реализуйте функции потерь J(θ) и градиентного спуска для логистической регрессии с использованием векторизации:

def sigmoid(x):

return 1 / (1 + np.exp(-x))

def cost\_function(theta, X, y):

m = len(y)

h\_theta = sigmoid(np.dot(X, theta))

# J = (1 / m) \* ((-y' \* log(h\_theta)) - (1 - y)' \* log(1 - h\_theta));

J = (1 / m) \* ((np.dot(-y.T, np.log(h\_theta))) - np.dot((1 - y).T, np.log(1 - h\_theta)))

return J

def gradient(theta, X, y):

# grad = (1 / m) \* (h\_theta - y)' \* X;

m = len(y)

h\_theta = sigmoid(np.dot(X, theta))

return (1 / m) \* np.dot((h\_theta - y).T, X)

4. Реализуйте другие методы (как минимум 2) оптимизации для реализованной функции стоимости (например, Метод Нелдера — Мида, Алгоритм Бройдена — Флетчера — Гольдфарба — Шанно, генетические методы и т.п.). Разрешается использовать библиотечные реализации методов оптимизации (например, из библиотеки scipy):

temp = optimize.fmin\_tnc(

func=cost\_function,

x0=theta.flatten(),

fprime=gradient,

args=(X, y.flatten())

)

# the output of above function is a tuple whose first element contains the optimized values of theta

theta\_optimized = temp[0]

print(theta\_optimized)

temp = optimize.minimize(cost\_function, theta.flatten(), (X, y.flatten()), method='Nelder-Mead')

print(temp.x)

# Brovden Fletcher Goldfarb Shanno alghoritm

theta\_optimized = optimize.fmin\_bfgs(

cost\_function,

theta.flatten(),

gradient,

(X, y.flatten())

)

print(theta\_optimized)

Результат выполнения:

[-25.16131856 0.20623159 0.20147149]

[-25.16130062 0.20623142 0.20147143]

[-25.16133284 0.2062317 0.2014716 ]

5. Реализуйте функцию предсказания вероятности поступления студента в зависимости от значений оценок по экзаменам:

def h0x(X, theta):

return sigmoid(np.dot(X.T, theta))

6. Постройте разделяющую прямую, полученную в результате обучения модели. Совместите прямую с графиком из пункта 2:

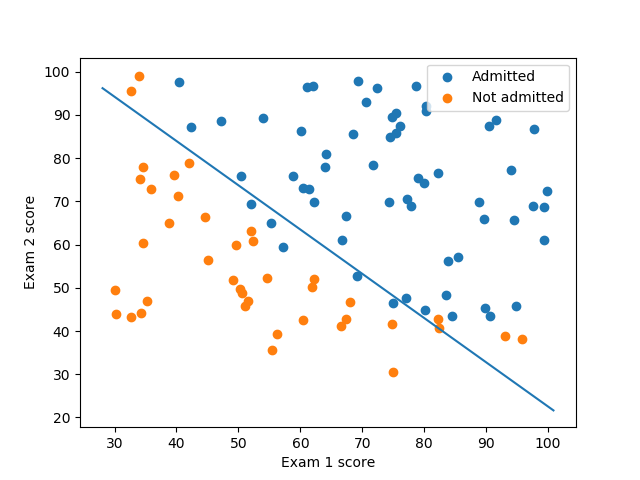


Рисунок 2 – график разделяющей прямой совмещенный с исходными данными

7. Загрузите данные ex2data2.txt из текстового файла:

file\_path = 'ex2data2.txt'

data = pd.read\_csv(file\_path, header=None)

X = data.iloc[:, :-1] # first 2 column

y = data.iloc[:, 2] # last column

data.head()

8. Постройте график, где по осям откладываются результаты тестов, а точки обозначаются двумя разными маркерами в зависимости от того, прошло ли изделие контроль или нет:

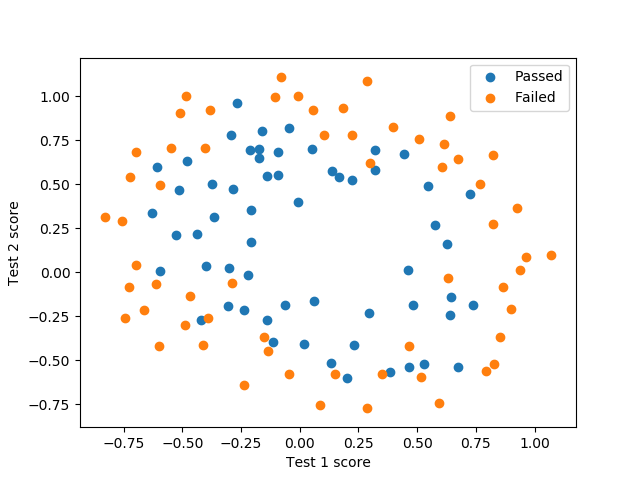


Рисунок 3 – график результатов тестов, где точки обозначаются двумя разными маркерами в зависимости от того, прошло ли изделие контроль или нет

9. Постройте все возможные комбинации признаков x1 (результат первого теста) и x2 (результат второго теста), в которых степень полинома не превышает 6, т.е. 1, x1, x2, x12, x1x2, x22, …, x1x25, x26 (всего 28 комбинаций):

def polynom\_multi\_var(p1, p2):

def multiply(x): # 6 combination

return (x[0] \*\* p1) \* (x[1] \*\* p2)

return ['(x1^%s)\*(x2^%s)' % (p1, p2), multiply]

# 9

map = {}

for i in range(0, 7):

for j in range(0, 7):

if i + j <= 6:

[key, fn] = polynom\_multi\_var(i, j)

map[key] = fn

# len(map.keys()) == 28

XX = []

for i in X.values:

a = []

for key in map.keys():

a.append(map[key](i))

XX.append(np.array(a))

X = np.array(XX)

10. Реализуйте L2-регуляризацию для логистической регрессии и обучите ее на расширенном наборе признаков методом градиентного спуска:

def cost\_function\_regularized(theta, X, y, lambda\_=0):

m = len(y)

h\_theta = sigmoid(np.dot(X, theta))

J = (1 / m) \* ((np.dot(-y.T, np.log(h\_theta))) - np.dot((1 - y).T, np.log(1 - h\_theta))) + (lambda\_ / (2 \* m)) \* np.sum(theta[1:]\*\*2)

return J

def gradient\_regularized(theta, X, y, lambda\_=0):

m = len(y)

grad = np.zeros([m, 1])

grad = (1 / m) \* np.dot(X.T, (sigmoid(np.dot(X, theta)) - y))

grad[1:] = grad[1:] + (lambda\_ / m) \* theta[1:]

return grad

# Set regularization parameter lambda to 1

lambda\_ = 0.1

(m, n) = X.shape

theta = np.zeros((n + 1, 1))

X = np.hstack((np.ones((m, 1)), X))

y = y[:, np.newaxis]

print('Cost at initial theta (zeros): %s', cost\_function\_regularized(theta, X, y, lambda\_)[0][0])

print('Expected cost (approx): 0.693')

output = optimize.fmin\_tnc(

func=cost\_function\_regularized,

x0=theta.flatten(),

fprime=gradient\_regularized,

args=(X, y.flatten(), lambda\_)

)

temp = output[0]

print('Reg fmin\_tnc: %s' % temp) # theta contains the optimized values

11. Реализуйте другие методы оптимизации:

temp = optimize.minimize(cost\_function\_regularized, theta.flatten(), (X, y.flatten(), lambda\_), method='Nelder-Mead')

print('Nelder-Mead: %s' % temp.x)

theta\_optimized = optimize.fmin\_bfgs(

cost\_function\_regularized,

theta.flatten(),

gradient\_regularized,

(X, y.flatten(), lambda\_)

)

print('Brovden Fletcher Goldfarb Shanno alghoritm: %s' % theta\_optimized)

12. Реализуйте функцию предсказания вероятности прохождения контроля изделием в зависимости от результатов тестов.

Код реализован в пункте номер 5.

13. Постройте разделяющую кривую, полученную в результате обучения модели. Совместите прямую с графиком из пункта 7:

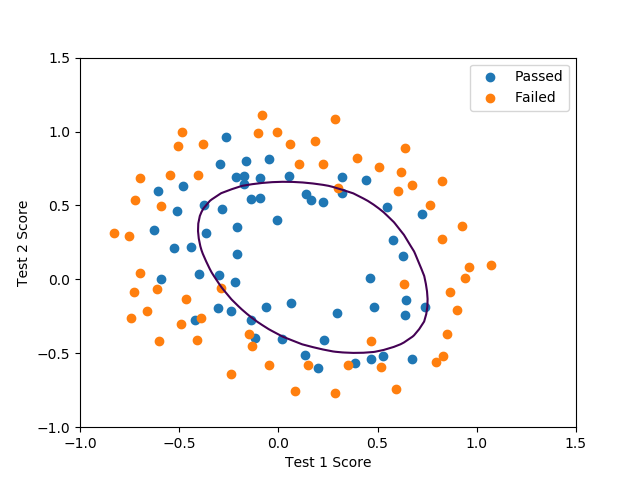


Рисунок 4 – график разделяющей совмещенный с исходными данными (lambda=0.1)

14. Попробуйте различные значения параметра регуляризации λ. Как выбор данного значения влияет на вид разделяющей кривой? Ответ дайте в виде графиков:

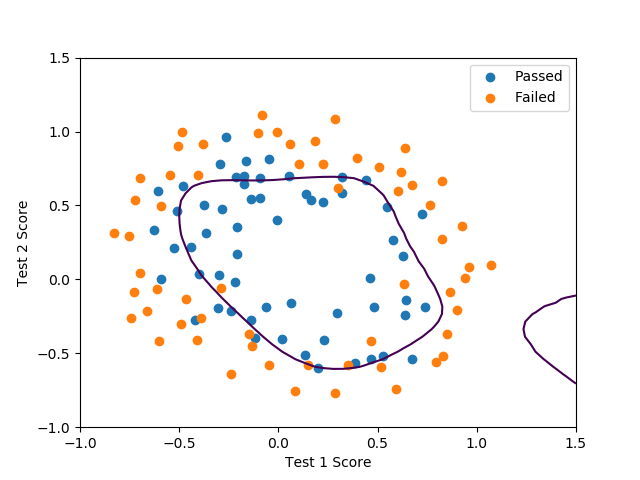


Рисунок 5 – Разделяющая кривая при lambda=0.001

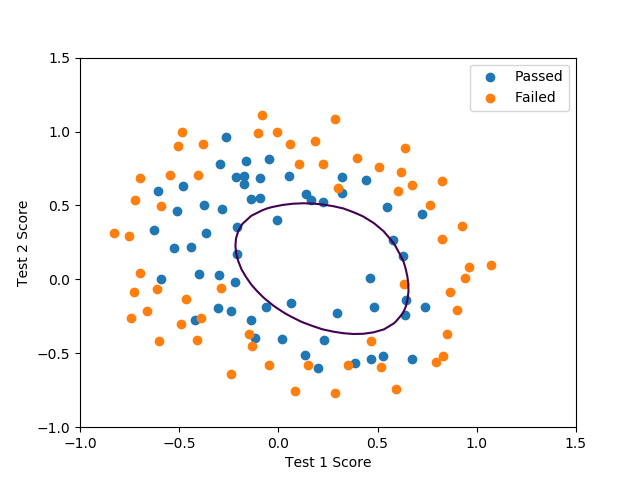


Рисунок 6 – разделяющая кривая при lambda=0.5

15. Загрузите данные ex2data3.mat из файла:

file\_path = 'ex2data3.mat'

data = sio.loadmat(file\_path)

X = data.get('X')

y = data.get('y')

16. Визуализируйте несколько случайных изображений из набора данных. Визуализация должна содержать каждую цифру как минимум один раз:



Рисунок 7 – визуализация чисел

17. Реализуйте бинарный классификатор с помощью логистической регрессии с использованием векторизации (функции потерь и градиентного спуска):

m = len(y)

X = np.hstack((np.ones((m, 1)), X))

(m, n) = X.shape

lmbda = 0.1

k = 10

theta = np.zeros((k, n)) # initial parameters

print("Cost with zeros theta: ", cost\_function\_regularized(theta[0], X, y))

print("Gradient with zeros theta: ", gradient\_regularized(theta.T, X, y))

Результат выполнения (theta-вектор обрезан для отчёта):

'Cost with zeros theta: ', array([-17.05142064])

18. Добавьте L2-регуляризацию к модели:

print("Cost with zeros theta: ", cost\_function\_regularized(theta[0], X, y, lambda\_))

print("Gradient with zeros theta: ", gradient\_regularized(theta.T, X, y, lambda\_))

Результат выполнения (указан только один результат, потому что тета-вектор слишком большой, чтобы вставлять его в отчёт):

'Cost with zeros theta: ', array([-17.05142064])

19. Реализуйте многоклассовую классификацию по методу “один против всех”:

for i in range(k):

digit\_class = i if i else 10

theta[i] = optimize.fmin\_cg(

f=cost\_function\_regularized,

x0=theta[i],

fprime=gradient\_regularized,

args=(X, (y == digit\_class).flatten().astype(np.int), lmbda),

maxiter=50

)

20. Реализуйте функцию предсказания класса по изображению с использованием обученных классификаторов:

def predict\_number(X, theta):

return np.argmax(np.dot(X, theta.T))

print("Predicted number: ", predict\_number(X[1490], theta), "Real: ", y[1490][0])

Результат выполнения:

'Predicted number: ', 2, 'Real: ', 2

21. Процент правильных классификаций на обучающей выборке должен составлять около 95%:

pred = np.argmax(np.dot(X, theta.T), axis=1)

pred = [e if e else 10 for e in pred] # convert 0 to 10

predictions = 0

for i in range(len(pred)):

if pred[i] == y[i][0]:

predictions += 1

print("Accuracy: ", (predictions / len(y)) \* 100)

Результат выполнения:

Accuracy: 95.12