COMP9417- Machine Learning & Data Mining

Homework 1

Name: Chongshi Wang

Q1: You have to report the θ parameters in step 3 when you are using TV feature.

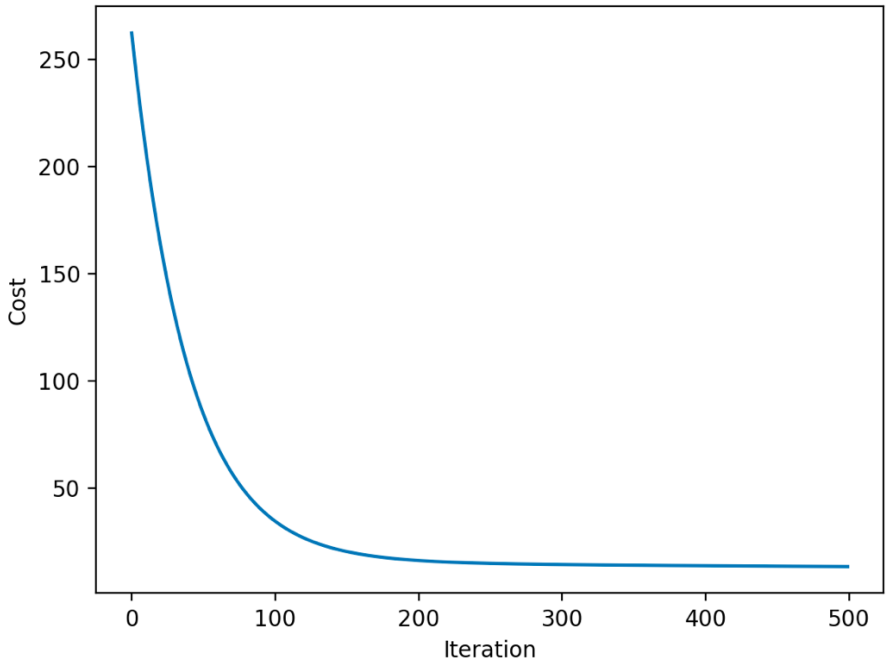
Answer:

When I use TV feature, the final $\theta 0 = 10.113$ and $\theta 1 = 8.272$ in step 3.

```
homework1
/usr/local/bin/python3.7 /Users/edwin/Downloads/homework1.py
TV theta0:  10.11283406777275 TV theta1:  8.271831294479673
```

Q2: A plot, which visualises the change in cost function $J(\theta)$at each iteration.

Answer:



Q3: RMSE for your training set when you use TV feature.

Answer:

When I use TV feature, the RMSE of training set is 3.640.

```
RMSE TV train:  3.640345489368778
```

Q4: RMSE for test set, when you use TV feature.

Answer:

When I use test set of TV feature, the RMSE is 3.909

```
RMSE TV test:  3.9085603448797355
```

Q5: RMSE for test set, when you use Radio feature.

Answer:

When I use Radio feature, the RMSE of test set is 4.200

```
RMSE Radio test:  4.20042579511254
```

Q6: RMSE for test set, when you use newspaper feature.

Answer:

When I use newspaper feature, the RMSE of test set is 5.428

```
RMSE Newspaper test:  5.427909854899054
```

Q7: Compare the performance of your three models and rank them accordingly.

Answer:

Obviously, RMSE of TV < RMSE of Radio < RMSE of Newspaper.

NO.1: TV feature(It has the best performance)

NO.2: Radio feature(It has the medium performance)

NO.3: Newspaper feature(It has the worst perfomance)

My Code:

```
import csv
import math
import numpy
import matplotlib.pyplot as py

file_path = r'/Users/edwin/downloads/Advertising.csv'

def readFunction(file):
    data = []
    with open(file) as f:
        readFile = csv.reader(f)
        for line in readFile:
            data.append(line)
    final_data = numpy.array(data[1:]).astype(float)
    return final_data

def nomalisation(output_data):
    for i in range(1,4):
        max_number = output_data[:,i].max()
        min_number = output_data[:,i].min()
        difference = max_number - min_number
        for j in range(200):
            output_data[j,i] =  (output_data[j,i] - min_number)/difference
    return output_data

def gradient_descent(train_data,theta_0,theta_1,learning_rate,attribute):
    true_y = train_data[:,4]
    train_x = train_data[:,attribute]
    cost = []
```

```
    for i in range(500):
        train_y = theta_0 + theta_1 * train_x
        error = true_y - train_y
        cost_function = numpy.mean((error * error))
        cost_theta_0 = numpy.mean(error)
        cost_theta_1 = numpy.mean(error * train_x)
        cost.append(cost_function)
        theta_0 = theta_0 + learning_rate * cost_theta_0
        theta_1 = theta_1 + learning_rate * cost_theta_1
    return theta_0,theta_1,cost

def test(theta_0,theta_1,attribute,test_data):
    test_x = test_data[:,attribute]
    test_y = test_data[:,4]
    prediction_y = theta_0 + theta_1 * test_x
    rmse = math.sqrt(numpy.mean((test_y - prediction_y) * (test_y -
prediction_y)))
    return rmse

output_data = readFunction(file_path)
nomalisation_data = nomalisation(output_data)
train_data = nomalisation_data[:190,:]
test_data = nomalisation_data[190:,:]
learning_rate = 0.01

attribute1_theta_0, attribute1_theta_1, cost = gradient_descent(train_data,-
1,-0.5,learning_rate,1)
print("TV theta0: ", attribute1_theta_0,"TV theta1: ",attribute1_theta_1)

py.plot(cost)
py.xlabel("Iteration")
py.ylabel("Cost")
py.show()

attribute2_theta_0, attribute2_theta_1, cost = gradient_descent(train_data,-
1,-0.5,learning_rate,2)
print("Radio theta0: ", attribute2_theta_0,"Radio theta1:
",attribute2_theta_1)

attribute3_theta_0, attribute3_theta_1, cost = gradient_descent(train_data,-
1,-0.5,learning_rate,3)
print("Newspaper theta0: ", attribute3_theta_0,"Newspaper theta1:
",attribute3_theta_1)
```

```
rmse0 = test(attribute1_theta_0,attribute1_theta_1,1,train_data)
print("RMSE TV train: ",rmse0)

rmse1 = test(attribute1_theta_0,attribute1_theta_1,1,test_data)
print("RMSE TV test: ",rmse1)

rmse2 = test(attribute2_theta_0,attribute2_theta_1,2,test_data)
print("RMSE Radio test: ", rmse2)

rmse3 = test(attribute3_theta_0,attribute3_theta_1,3,test_data)
print("RMSE Newspaper test: ", rmse3)
```