Yash Shah J060

Yash Shah J060

```
%matplotlib inline
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

Univariate Linear regression

```
data=pd.read_csv("C:/Users/riyav/OneDrive/Desktop/SEM 5/ML/dataset1.txt", header=None)
data.head()
```

	0	1
0	6.1101	17.5920
1	5.5277	9.1302
2	8.5186	13.6620
3	7.0032	11.8540
4	5.8598	6.8233

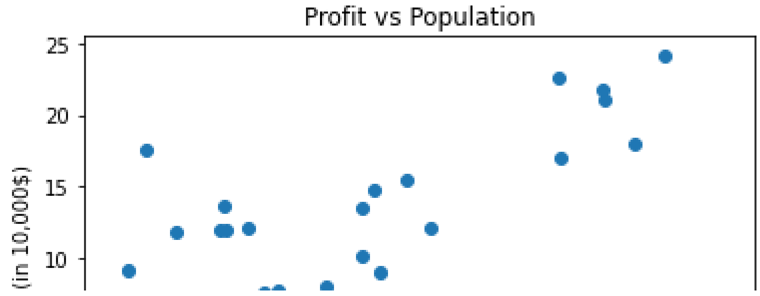
```
data.describe()
```

	0	1
count	97.000000	97.000000
mean	8.159800	5.839135
std	3.869884	5.510262
min	5.026900	-2.680700
25%	5.707700	1.986900
50%	6.589400	4.562300
75%	8.578100	7.046700
max	22.203000	24.147000

```
data.columns = ['Population','Profit']
```

```
plt.scatter(data['Population'],data['Profit'])
plt.xticks(np.arange(5,30,step=5))
plt.yticks(np.arange(-5,30,step=5))
plt.xlabel('Population (in 10,000s)')
plt.ylabel('Profit (in 10,000$)')
plt.title('Profit vs Population')
```

```
Text(0.5, 1.0, 'Profit vs Population')
```



Cost function $J(\Theta)$

```
def computeCost(X,y,theta):
    """
    Take in a numpy array X,y,theta and get cost function using theta as parameter in a linea
    """
    m=len(y)
    prediction =X.dot(theta)
    square_err = (prediction -y)**2

    return 1/(2*m)*np.sum(square_err)
```

```
data['x0'] =1
```

```
data_val= data.values
m = len(data_val[:-1])
X =data[['x0','Population']].iloc[:-1].values
y = data['Profit'][:-1].values.reshape(m,1)
theta = np.zeros((2,1))

m, X.shape, y.shape, theta.shape
```

```
(96, (96, 2), (96, 1), (2, 1))
```

h(θ) = x0θ0 +x1θ1 ....(x0 = 1)

```
computeCost(X,y,theta)
```

```
32.40484177877031
```

```
data.tail()
```

	Population	Profit	x0
92	5.8707	7.20290	1
93	5.3054	1.98690	1
94	8.2934	0.14454	1
95	13.3940	9.05510	1
96	5.4369	0.61705	1

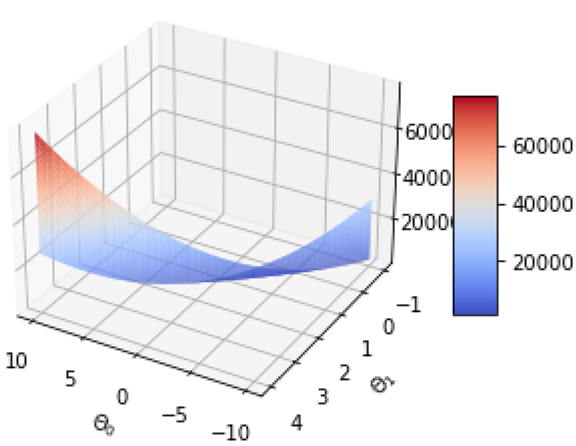
Gradient Descent

```
def gradientDescent(X,y,theta,alpha,num_iters):
    """
    Take numpy aarray for X,y,theta and update theta for every iteration of gradient steps
    return theta adn the list of cost of theta during each iteration
```

```
    """
    m=len(y)
    J_history=[]
    for i in range(num_iters):
        predictions= X.dot(theta)
        error =np.dot(X.transpose(),(predictions - y))
        descent= alpha * 1/m *error
        theta-= descent
        J_history.append(computeCost(X,y,theta))

    return theta,J_history


theta, J_history = gradientDescent(X,y,theta,0.001,2000)


print(f"h(x) = {str(round(theta[0,0],2))} + {str(round(theta[1,0],2))}x1")
```

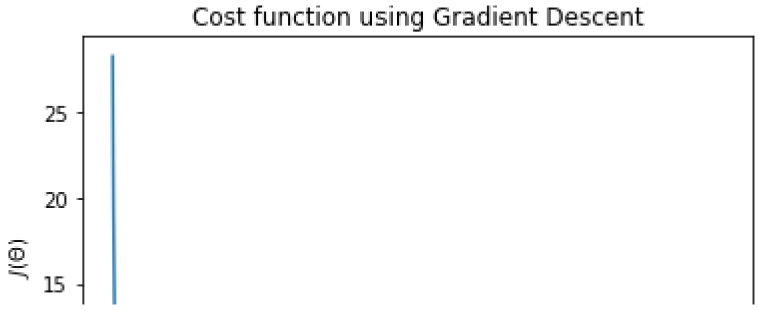
```
h(x) = -1.11 + 0.92x1
```

```
from mpl_toolkits.mplot3d import Axes3D
#Generating values for theta0, theta1 and the resulting cost value
theta0_vals=np.linspace(-10,10,100)
theta1_vals=np.linspace(-1,4,100)
J_vals=np.zeros((len(theta0_vals),len(theta1_vals)))
for i in range(len(theta0_vals)):
    for j in range(len(theta1_vals)):
        t=np.array([theta0_vals[i],theta1_vals[j]])
        J_vals[i,j]=computeCost(X,y,t)
#Generating the surface plot
fig = plt.figure()
ax = fig.add_subplot(111, projection='3d')
surf=ax.plot_surface(theta0_vals,theta1_vals,J_vals,cmap="coolwarm")
fig.colorbar(surf, shrink=0.5, aspect=5)
ax.set_xlabel("$\Theta_0$")
ax.set_ylabel("$\Theta_1$")
ax.set_zlabel("$J(\Theta)$")
#rotate for better angle
ax.view_init(30,120)
```



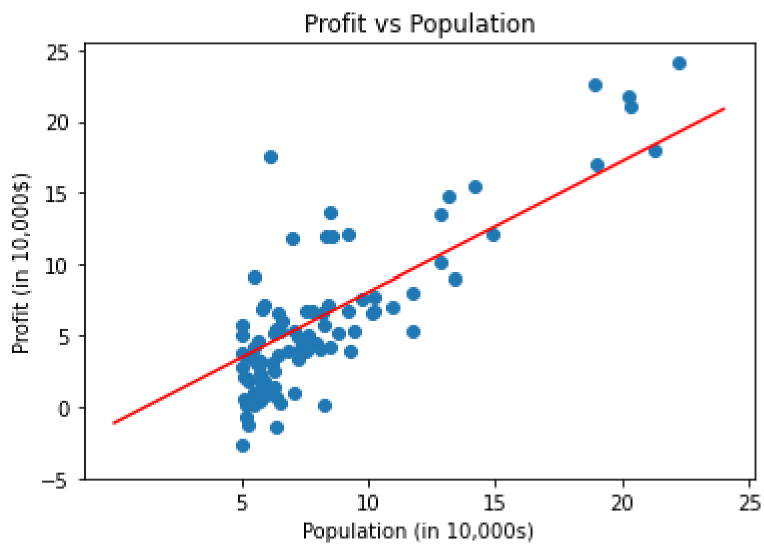

```
plt.plot(J_history)
plt.xlabel("Iteration")
plt.ylabel("$J(\Theta)$")
plt.title("Cost function using Gradient Descent")
```

```
Text(0.5, 1.0, 'Cost function using Gradient Descent')
```




```
plt.scatter(data['Population'], data['Profit'])
x_value = [x for x in range(25)]
y_value = [x*theta[1] + theta[0] for x in x_value]
plt.plot(x_value, y_value, color = 'r')
plt.xticks(np.arange(5,30,step=5))
plt.yticks(np.arange(-5,30,step=5))
plt.xlabel('Population (in 10,000s)')
plt.ylabel('Profit (in 10,000$)')
plt.title('Profit vs Population')
```

```
Text(0.5, 1.0, 'Profit vs Population')
```




```
def predict(x,theta):
    """
    takes in numpy array x and theta and returns predicted value of y
    """
    predictions = np.dot(theta.transpose(),x)
    return predictions[0]
```

```
data.tail(1)
```

	Population	Profit	x0
96	5.4369	0.61705	1

Multivariate Linear Regression

Imports

```
import statsmodels.api as sm
from sklearn.linear_model import LinearRegression
np.random.seed(123)
```

```
data = pd.read_csv('C:/Users/Priya/OneDrive/Desktop/SEM 5/ML/dataset2.csv')
data.head()
```

	Size of the house (in square feet)	Number of bedrooms	Price of the house
0	2104	3	399900
1	1600	3	329900
2	2400	3	369000
3	1416	2	232000
4	3000	4	539900

```
data.isnull().sum()
```

```
Size of the house (in square feet)    0
Number of bedrooms                    0
Price of the house                    0
dtype: int64
```

```
def normalize(dataframe):
    df = dataframe.copy()
    for col in df.columns:
       df[col] = (df[col]-df[col].mean())/df[col].std()
    return df
```

```
normallized_data = normalize(data)
normallized_data.head()
```

	Size of the house (in square feet)	Number of bedrooms	Price of the house
0	0.130010	-0.223675	0.475747
1	-0.504190	-0.223675	-0.084074
2	0.502476	-0.223675	0.228626
3	-0.735723	-1.537767	-0.867025
4	1.257476	1.090417	1.595389

```
X = normallized_data.iloc[:,:-1].values
y = normallized_data.iloc[:,-1].values
```

```
m = y.size
n = data.shape[1]
```

```
y.shape
```

```
(47,)
```

```
y = y.reshape(m,1)
y.shape
```

```
(47, 1)
```

```
ones = np.ones((m,1))
X1 = np.concatenate((ones,X),axis=1)
X1[:5]
```

```
array([[ 1.        ,  0.13000987, -0.22367519],
       [ 1.        , -0.50418984, -0.22367519],
       [ 1.        ,  0.50247636, -0.22367519],
```

```
            [ 1.        , -0.73572306, -1.53776691],
            [ 1.        ,  1.25747602,  1.09041654]])
```

```
alpha = 0.01
theta = np.random.rand(n,1)
epoch = 10000
```

```
def GD(X1,y,theta,epoch,alpha,decimals=5):
    past_cost = []
    past_theta = [theta]
    m = y.size
    n = X1.shape[1]
    for i in range(epoch):
        h_theta = np.dot(X1,theta)
        error = h_theta-y
        cost = np.dot(error.T, error)/(2*m)
        past_cost.append(cost[0][0])
        diff = np.dot(X1.T, error)/m
        theta = theta - (alpha*diff)
        past_theta.append(theta)
        # Task 4 - do early stopping (I have considered 5 decimal places, you can change the
        if np.equal(np.round(past_theta[i],decimals=decimals),np.round(past_theta[i+1],decima
            break
    return past_cost, past_theta, i+1
```

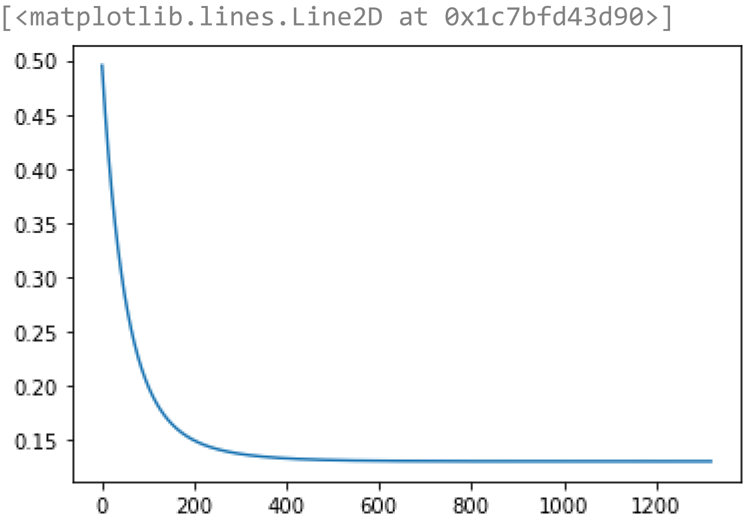
```
pastCost, pastTheta,stop_epoch = GD(X1=X1, y=y, theta=theta, epoch=epoch,alpha=alpha)
```

```
print(f'Our model performed {stop_epoch} epochs out of {epoch} epochs before converging')
```

```
Our model performed 1320 epochs out of 10000 epochs before converging
```

```
plt.plot(pastCost)
```

```
[<matplotlib.lines.Line2D at 0x1c7bfd43d90>]
```




```
best_theta = np.array(pastTheta[-1]).reshape(n,)
print(best_theta)
```

```
[ 1.20603184e-06  8.83291779e-01 -5.17046112e-02]
```

```
print(f'Parameters from StatsModels -> {sm.OLS(y,X1).fit().params}')
print(f'Parameters from SciKitLearn -> {LinearRegression().fit(X1,y).coef_}')
```

```
Parameters from StatsModels -> [-9.71445147e-17  8.84765988e-01 -5.31788197e-02]
Parameters from SciKitLearn -> [[ 0.          0.88476599 -0.05317882]]
```