Logistic Regression(Binary Output) - From Scratch

SETTING UP DATA AND DEPENDENCIES

In [4]:

```
# Importing dependencies and dataset
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

from sklearn.datasets import load_breast_cancer
from sklearn.model_selection import train_test_split
from sklearn import preprocessing
```

In [5]:

```
#Creating the dataframe
cancer = load_breast_cancer()
df = pd.DataFrame(data = cancer['data'], columns = cancer['feature_name
s'])
df['class'] = cancer['target']
df.head()
```

Out[5]:

	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry
0	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	0.2419
1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	0.1812
2	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	0.2069
3	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	0.2597
4	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	0.1809

5 rows × 31 columns

In [6]:
```
#Data statistics
df.describe()
```

Out[6]:

	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity
count	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000
mean	14.127292	19.289649	91.969033	654.889104	0.096360	0.104341	0.088799
std	3.524049	4.301036	24.298981	351.914129	0.014064	0.052813	0.079720
min	6.981000	9.710000	43.790000	143.500000	0.052630	0.019380	0.000000
25%	11.700000	16.170000	75.170000	420.300000	0.086370	0.064920	0.029560
50%	13.370000	18.840000	86.240000	551.100000	0.095870	0.092630	0.061540
75%	15.780000	21.800000	104.100000	782.700000	0.105300	0.130400	0.130700
max	28.110000	39.280000	188.500000	2501.000000	0.163400	0.345400	0.426800

8 rows × 31 columns

PREPROCESSING DATA AND SPLITTING DATA INTO TRAINING AND TESTING SETS

In [7]:
```
#Need to scale feature.  Standard Scaling(use with normally distributed
 data)
#Dropping class column to scale data features
df_drop = df.drop(['class'], axis = 1)

df_drop_scaled = preprocessing.scale(df_drop)

#Adding the ones column to training data
X = np.c_[np.ones(df.shape[0]),df_drop_scaled]
y = df['class']
```

In [8]:
```
#Splitting Data Set into Training and Testing Data Sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = .8
0, random_state = 1)
```

CREATING SIGMOID, COST, AND GRADIENT DESCENT FUNCTIONS

In [9]:
```
#creating the sigmoid function for the prediction hypothesis

def sigmoid(z):
    sigmoid = 1/(1+np.exp(-z))
    return sigmoid
```

In [10]:
```
#Calculating the cost function

def cost_function(x_train, y_train, thetas_array, n):
    prediction = sigmoid(np.dot(X_train,np.transpose(thetas_array)))
    cost = np.sum((y_train * np.log(prediction)) + ((1 - y_train) * np.l
og(1-prediction)))/-n
    """error = (-y_train * np.log(prediction)) - ((1-y_train)*np.log(1-p
rediction))
    cost = 1/n * sum(error)"""
    return cost
```

In [11]:
```
#Calculating the gradient descent

def gradient_descent(X_train, y_train, alpha, thetas_array,n):
    thetas = thetas_array - ((alpha/n) * np.dot((sigmoid(np.dot(X_train,
np.transpose(thetas_array))) - y_train), X_train))
    return thetas
```

In [12]:
```
thetas_array = np.zeros(31)
sig = sigmoid(np.dot(X_train,np.transpose(thetas_array))) - y_train
print(sig.shape)
print(np.transpose(X_train).shape)
```

```
(113,)
(31, 113)
```

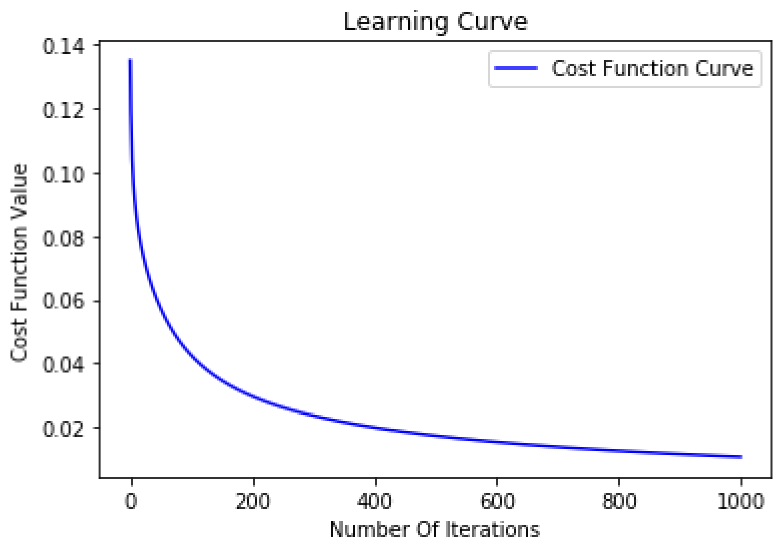
TRAINING THE DATA SET

In [13]:
```
def training(X_train, y_train, alpha, iters):
    n = len(X_train)
    thetas_array = np.zeros(31)
    #thetas_array = np.random.rand(len(cancer['feature_names']) + 1)
    cost = []

    for i in range(iters):
        thetas_array = gradient_descent(X_train, y_train, alpha, thetas_
array,n)
        cost.append(cost_function(X_train, y_train, thetas_array, n))

    #Plot cost function error per iteration
    x = np.arange(0, len(cost), step=1)
    plt.plot(x, cost, "-b", label="Cost Function Curve")
    plt.title("Learning Curve")
    plt.xlabel("Number Of Iterations")
    plt.ylabel("Cost Function Value")
    plt.legend()
    plt.show()
    return thetas_array, cost[-1]
```

In [14]:
```
training(np.array(X_train), np.array(y_train), 1, 1000)
```



Out[14]:
```
(array([-0.22799422, -1.06663335,  1.00608203, -0.95702739, -1.0550155
8,
         0.41521535,  0.87461389, -0.44360275, -1.01100796, -0.5227008
4,
        -1.09043709, -2.69243967, -2.14739907, -2.21045983, -1.7189354
9,
         3.12508893,  1.04422038,  0.16876089, -1.42959538,  1.7102668
5,
         0.70710548, -1.61842906, -1.25537742, -1.58207748, -1.4574248
3,
         0.18746913, -0.55771794, -1.99624321, -1.56301169, -1.1237997
8,
        -0.75522139]), 0.010710413265069925)
```

CALCULATING TRAINING ACCURACY

In [15]:
```
def training_accuracy(X, y, thetas_array):
    y = np.array(y_train)
    z = np.dot(X_train, thetas_array)
    prediction = sigmoid(z)
    total_number_pred = len(prediction)
    TP = 0
    FP = 0
    FN = 0
    TN = 0

    for i in range(len(y_train)):
        if prediction[i] >= 0.5 and y[i] == 1:
            TP += 1
        elif prediction[i] < 0.5 and y[i] == 1:
            FP += 1
        elif prediction[i] >= 0.5 and y[i] == 0:
            FN += 1
        else:
            TN += 1
    accuracy = round((TP + TN)/(TP + TN + FN + FP) * 100, 2)
    print(f'Training Accuracy: {accuracy}%')
```

In [16]:
```
training_accuracy(X_train, y_train, [-0.22799422, -1.06663335,  1.006082
03, -0.95702739, -1.05501558,
         0.41521535,  0.87461389, -0.44360275, -1.01100796, -0.52270084,
        -1.09043709, -2.69243967, -2.14739907, -2.21045983, -1.71893549,
         3.12508893,  1.04422038,  0.16876089, -1.42959538,  1.71026685,
         0.70710548, -1.61842906, -1.25537742, -1.58207748, -1.45742483,
         0.18746913, -0.55771794, -1.99624321, -1.56301169, -1.12379978,
        -0.75522139])

#overfitting occuring due to very good performance on the training data
 set
```

```
Training Accuracy: 100.0%
```

MAKING PREDICTIONS ON NEW DATA POINT

In [17]:
```
#Function to predict probability of developing breast cancer.  Enter a n
ew data point from testing set.
def predict(X_test, thetas_array):
    z = np.dot(X_test, thetas_array)
    prediction = round(sigmoid(z) *100,2)
    return f"You have a {prediction}% change of breast cancer."
```

In [25]:
```
predict(np.array(X_test[30]), [-0.22799422, -1.06663335,  1.00608203, -
0.95702739, -1.05501558,
         0.41521535,  0.87461389, -0.44360275, -1.01100796, -0.52270084,
        -1.09043709, -2.69243967, -2.14739907, -2.21045983, -1.71893549,
         3.12508893,  1.04422038,  0.16876089, -1.42959538,  1.71026685,
         0.70710548, -1.61842906, -1.25537742, -1.58207748, -1.45742483,
         0.18746913, -0.55771794, -1.99624321, -1.56301169, -1.12379978,
        -0.75522139])
```

Out[25]: 'You have a 0.0% change of breast cancer.'

In [ ]: