Task-C: Regression outlier effect.

Objective:Visualization best fit linear regression line for different scenarios

In [1]:

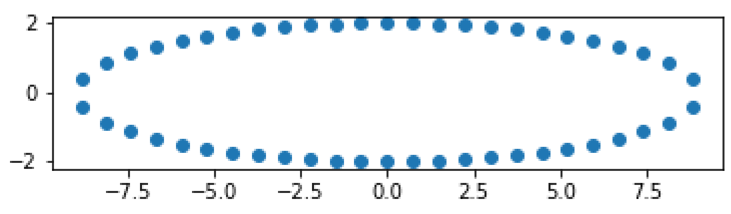
```
# you should not import any other packages
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings("ignore")
import numpy as np
from sklearn.linear_model import SGDRegressor
```

In [2]:

```
import numpy as np
import scipy as sp
import scipy.optimize

def angles_in_ellipse(num,a,b):
    assert(num > 0)
    assert(a < b)
    angles = 2 * np.pi * np.arange(num) / num
    if a != b:
        e = (1.0 - a ** 2.0 / b ** 2.0) ** 0.5
        tot_size = sp.special.ellipeinc(2.0 * np.pi, e)
        arc_size = tot_size / num
        arcs = np.arange(num) * arc_size
        res = sp.optimize.root(
            lambda x: (sp.special.ellipeinc(x, e) - arcs), angles)
        angles = res.x
    return angles
```

In [3]:

```
a = 2
b = 9
n = 50

phi = angles_in_ellipse(n, a, b)
e = (1.0 - a ** 2.0 / b ** 2.0) ** 0.5
arcs = sp.special.ellipeinc(phi, e)

fig = plt.figure()
ax = fig.gca()
ax.axes.set_aspect('equal')
ax.scatter(b * np.sin(phi), a * np.cos(phi))
plt.show()
```



In [4]:

```
X= b * np.sin(phi)
Y= a * np.cos(phi)
```

In [5]:

```
Y.shape
```

Out[5]:

```
(50,)
```

1. As a part of this assignment you will be working the regression problem and how regularization helps to get rid of outliers

2. Use the above created X, Y for this experiment.

3. to do this task you can either implement your own SGDRegression(prefered) excatly similar to "SGD assignment" with mean sequared error or
you can use the SGDRegression of sklearn, for example "SGDRegressor(alpha=0.001, eta0=0.001, learning_rate='constant',random_state=0)"
*note that you have to use the constant learning rate and learning rate **eta0** initialized.*

4. as a part of this experiment you will train your linear regression on the data (X, Y) with different regularizations alpha=[0.0001, 1, 100] and
observe how prediction hyper plan moves with respect to the outliers

5. This the results of one of the experiment we did (title of the plot was not metioned intentionally)

in each iteration we were adding single outlier and observed the movement of the hyper plane.

6. please consider this list of outliers: [(0,2),(21, 13), (-23, -15), (22,14), (23, 14)] in each of tuple the first elemet is the input feature(X) and the second element is the output(Y)

7. for each regularizer, you need to add these outliers one at time to data and then train your model again on the updated data.

*8. you should plot a 3*5 grid of subplots,*
where each row corresponds to results of model with a single regularizer.

9. Algorithm:

for each regularizer:
for each outlier:
#add the outlier to the data
#fit the linear regression to the updated data
#get the hyper plane
#plot the hyperplane along with the data points

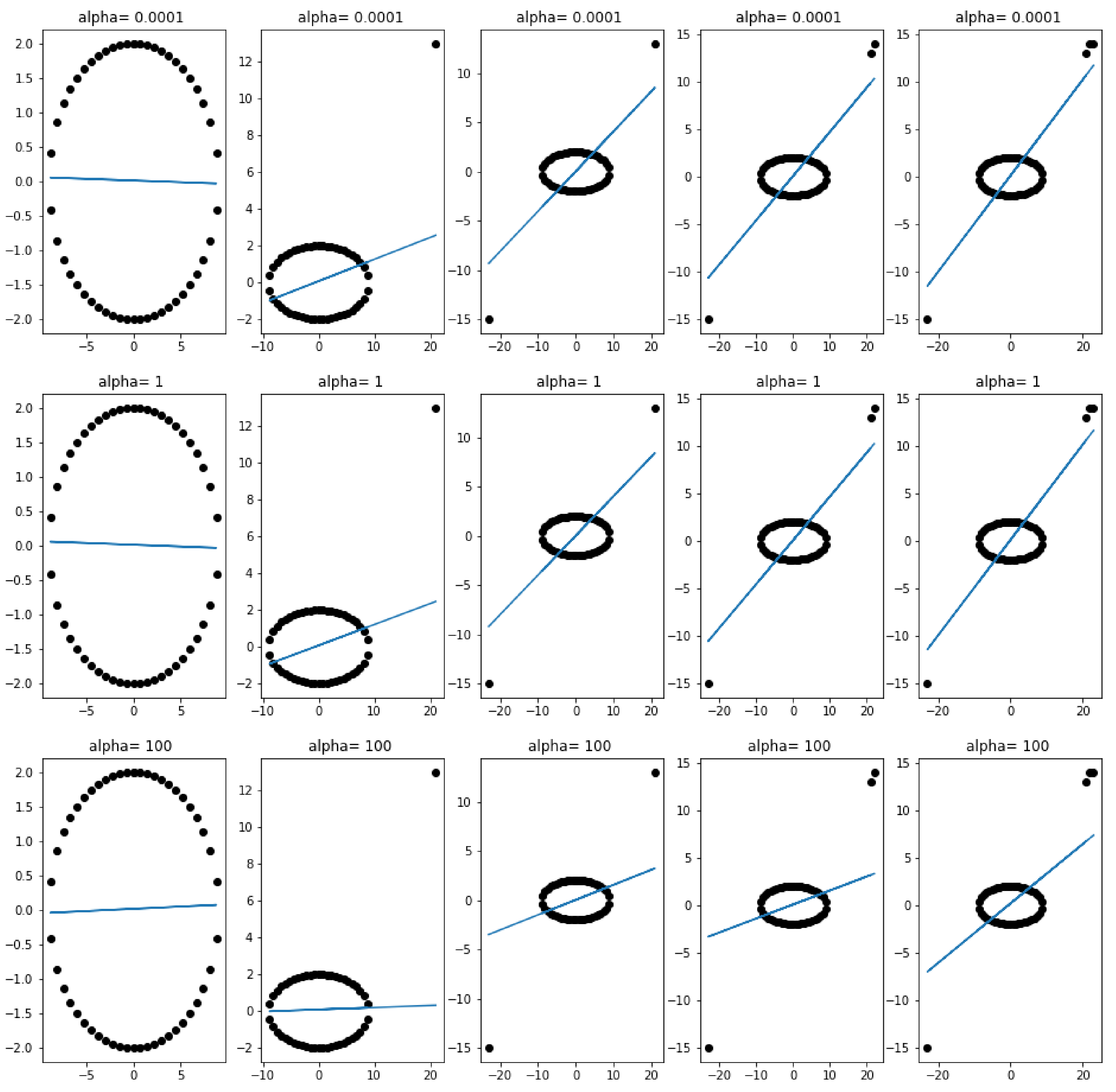
10. MAKE SURE YOU WRITE THE DETAILED OBSERVATIONS, PLEASE CHECK THE LOSS FUNCTION IN THE SKLEARN DOCUMENTATION
(please do search for it).

In [6]:

```
outliers = [(0,2),(21, 13), (-23, -15), (22,14), (23, 14)]
alpha = [0.0001, 1, 100]
```

In [7]:

```
p=1
plt.figure(figsize=(16, 16))
for i in alpha:
    X_with_outliers = X
    Y_with_outliers = Y
    for j in outliers:
        X_with_outliers = np.append(X_with_outliers, j[0]).reshape(-1,1)
        Y_with_outliers = np.append(Y_with_outliers, j[1]).reshape(-1,1)
        sgd_regressor = SGDRegressor(alpha=i, eta0=0.001, learning_rate='constant',random_state=0)
        sgd_regressor.fit(X_with_outliers,Y_with_outliers)
        predicted_value = sgd_regressor.predict(X_with_outliers)
        plt.subplot(3,5,p)
        plt.title("alpha= {} ".format(i))
        plt.scatter(X_with_outliers, Y_with_outliers, color = 'black')
        plt.plot(X_with_outliers, predicted_value)
        p=p+1
```



Observations:

1. Regularization hyperparameter 'alpha' determines how much the SGD Regressor model will try to reduce the loss caused by introduction of outliers.
2. At low values of alpha, SGD Regression will overfit which makes outliers to easily affect the model and the model will try to reduce the effect of outlier by moving the hyperplane.
3. At alpha = 0.0001, adding a single outlier causes the hyperplane to move significantly
4. The performance of SGD regressor is nearly the same at alpha = 0.0001 and alpha = 1. After adding few more outliers, the hyperplane completely changes to overfit the outlier data (which causes high loss in our original data).
5. At high values of alpha, SGD Regression will reduce overfitting making SGD Regression to not immediately respond to outliers and the hyperplane starts to move as more outliers are added.
6. At alpha = 100, adding a single outlier does not affect the hyperplane and the hyperplane is nearly in the same position as when there were no outliers in our data.
7. Even after 4 outliers for alpha = 100, the model does not change drastically like it did with alpha values 0.0001 and 1.