Dataset1-Regression_output_5

October 19, 2021

1 Dataset 1 - Regression

1.1 Experiment Details

The aim of the experiment is to verify if the: 1. ABC_GAN model corrects model misspecification 2. ABC_GAN model performs better and converges faster than a simple C-GAN model

In the experiment we predict the distribution that represents the real data and simulate realistic fake data points using statistical mode, C-GAN and ABC-GAN model with 3 priors. We analyze and compare their performance using metrics like mean squared error, mean absolute error , manhattan distance and euclidean distance between y_{real} and y_{pred}

The models are as follows:

1. The statistical model assumes the distribution $Y = \beta X + \mu$ where $\mu \sim N(0,1)$
2. The Conditional GAN consists of
   1. Generator with 2 hidden layers with 100 nodes each and ReLu activation.
   2. Discriminator with 2 hidden layers with 25 and 50 nodes and ReLu activation. We use Adam's optimser and BCE Logit Loss to train the model. The input to the Generator of the GAN is (x,e) where x are the features and $e \sim N(0,1)$. The discriminator output is linear.
3. The ABC GAN Model consists of
   1. ABC generator is defined as follows:
      1. $Y = 1 + \beta_1 x_1 + \beta_2 x_2 + \beta_2 x_3 + ... + \beta_n x_n + N(0,\sigma)$ where $\sigma = 0.1$
      2. $\beta_i \sim N(0,\sigma^*)$ when $\mu = 0$ else $\beta_i \sim N(\beta_i^*, \sigma^*)$ where $\beta_i^* s$ are coefficients obtained from statistical model
      3. σ^* takes the values 0.01,0.1 and 1

   2. C-GAN network is as defined above. However the input to the Generator of the GAN is (x, y_{abc}) where y_{abc} is the output of the ABC Generator.

1.2 Import Libraries

[1]:
```
import warnings
warnings.filterwarnings('ignore')
```

[2]:
```
import train_test
import ABC_train_test
import regressionDataset
import network
```

```
import statsModel
import performanceMetrics
import dataset
import sanityChecks
import torch
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import norm
from torch.utils.data import Dataset,DataLoader
from torch import nn
```

1.3 Parameters

General Parameters

1. Number of Samples
2. Number of features

ABC-Generator parameters are as mentioned below: 1. mean : 1 ($\beta \sim N(\beta^*, \sigma)$ where β^* are coefficients of statistical model) or 1 ($\beta \sim N(0, \sigma)$ 2. std : $\sigma = 1, 0.1, 0.01$ (standard deviation)

[3]:
```
n_features = 10
n_samples= 100

#ABC Generator Parameters
mean = 1
variance = 0.001
```

[4]:
```
# Parameters
n_samples = 10
n_features = 10
mean = 0
variance = 0.01
```

1.4 Dataset

Generate a random regression problem

$Y = 1 + \beta_1 x_1 + \beta_2 x_2 + \beta_2 x_3 + ... + \beta_n x_n + N(0, \sigma)$ where $\sigma = 0.1$

[5]:
```
X,Y = regressionDataset.regression_data(n_samples,n_features)
```

```
         X1        X2        X3        X4        X5        X6        X7  \
0 -0.388053  0.981676  0.224700  0.331660 -0.309388  0.377331  0.666731
1 -0.802879 -1.661496  1.194140 -0.229451 -0.236927  2.747723 -0.427769
2 -0.201943  0.458697  0.297090  1.048824 -0.670302  0.134931 -1.442386
3 -0.163737 -0.739434  0.632450 -0.823472  1.169715  0.352610 -0.387504
4  1.029334 -1.384847 -1.268069 -0.239197 -1.025234  0.493804  1.786092

         X8        X9       X10             Y
```

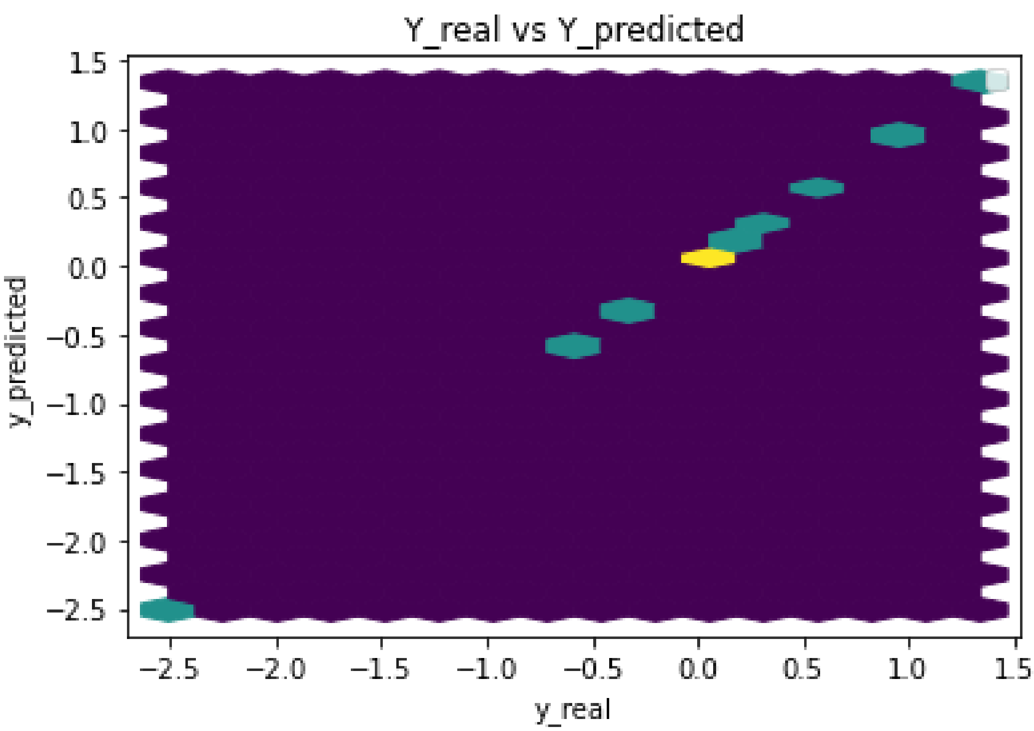
```
0 -0.342893  0.981871 -2.770005   51.554279
1  0.152894 -1.322134 -0.643534   82.842840
2 -1.839216  2.390126 -0.851500   63.135668
3  0.729622  2.659708  0.792772  273.950890
4  0.290161 -0.195772  0.139614  -23.248077
```

1.5 Stats Model

[6]:
```
[coeff,y_pred] = statsModel.statsModel(X,Y)
```

```
No handles with labels found to put in legend.

                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       1.000
Model:                            OLS   Adj. R-squared:                    nan
Method:                 Least Squares   F-statistic:                       nan
Date:                Tue, 19 Oct 2021   Prob (F-statistic):                nan
Time:                        23:18:00   Log-Likelihood:                 328.44
No. Observations:                  10   AIC:                            -636.9
Df Residuals:                       0   BIC:                            -633.9
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       1.665e-16        inf          0        nan         nan         nan
x1             0.2745        inf          0        nan         nan         nan
x2            -0.1266        inf         -0        nan         nan         nan
x3             0.4263        inf          0        nan         nan         nan
x4             0.3390        inf          0        nan         nan         nan
x5             0.3137        inf          0        nan         nan         nan
x6             0.1628        inf          0        nan         nan         nan
x7             0.2131        inf          0        nan         nan         nan
x8             0.2125        inf          0        nan         nan         nan
x9             0.4302        inf          0        nan         nan         nan
x10            0.0466        inf          0        nan         nan         nan
==============================================================================
Omnibus:                       14.916   Durbin-Watson:                   2.306
Prob(Omnibus):                  0.001   Jarque-Bera (JB):                7.625
Skew:                           1.789   Prob(JB):                       0.0221
Kurtosis:                       5.344   Cond. No.                         14.6
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The input rank is higher than the number of observations.
Parameters:  const    1.665335e-16
```

```
x1       2.744951e-01
x2      -1.266294e-01
x3       4.263344e-01
x4       3.390048e-01
x5       3.136960e-01
x6       1.628171e-01
x7       2.131321e-01
x8       2.124772e-01
x9       4.302051e-01
x10      4.659453e-02
dtype: float64
```




```
Performance Metrics
Mean Squared Error: 1.7358262144016326e-30
Mean Absolute Error: 1.0068335054569388e-15
Manhattan distance: 1.0068335054569388e-14
Euclidean distance: 4.1663247765886335e-15
```

1.6 Common Training Parameters (GAN & ABC_GAN)

[7]:
```
n_epochs = 5000
error = 0.001
batch_size = n_samples//2
```

1.7 GAN Model

[8]:
```
real_dataset = dataset.CustomDataset(X,Y)
device = torch.device('cuda' if torch.cuda.is_available() else 'cpu')
```

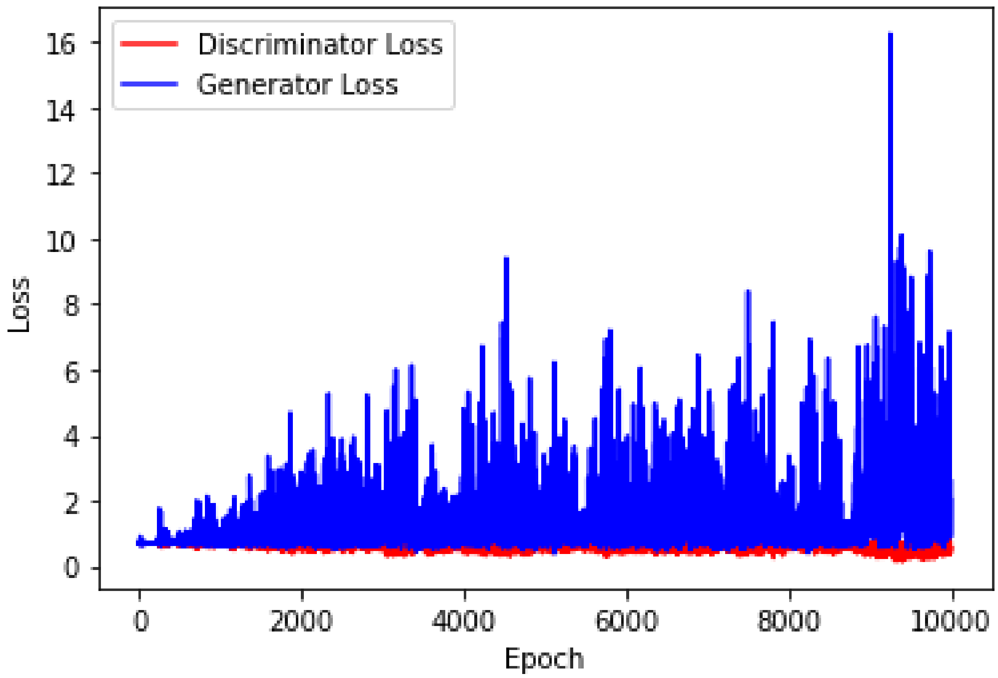
Training GAN for n_epochs number of epochs

[9]:
```
generator = network.Generator(n_features+2)
discriminator = network.Discriminator(n_features+2)

criterion = torch.nn.BCEWithLogitsLoss()
gen_opt = torch.optim.Adam(generator.parameters(), lr=0.01, betas=(0.5, 0.999))
disc_opt = torch.optim.Adam(discriminator.parameters(), lr=0.01, betas=(0.5, 0.
 ↪999))
```

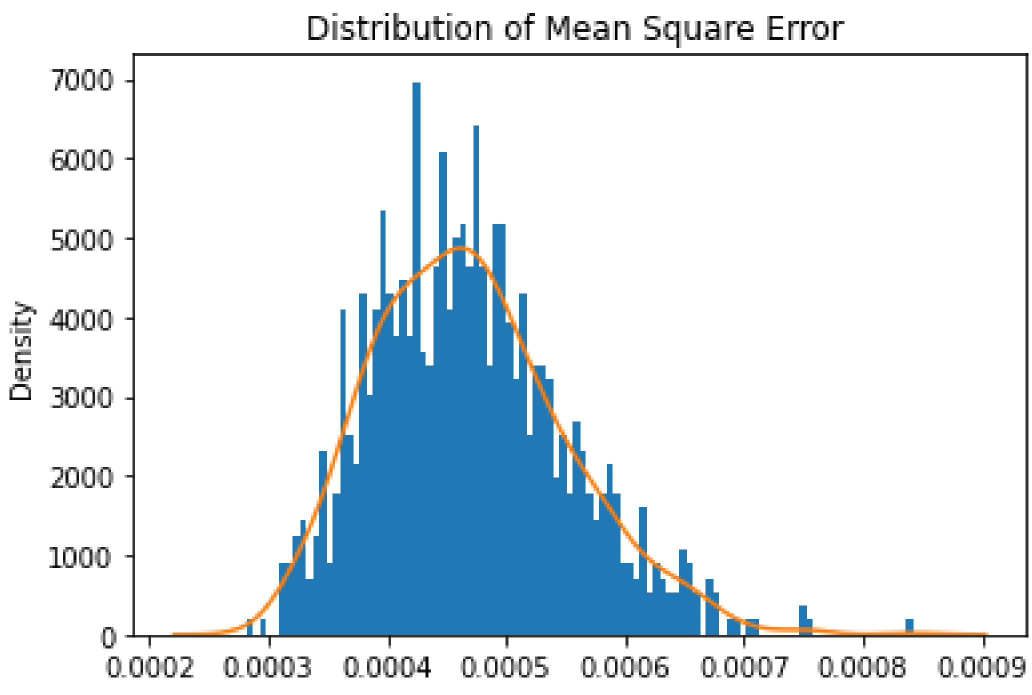
[10]:
```
print(generator)
print(discriminator)
```

```
Generator(
  (hidden1): Linear(in_features=12, out_features=100, bias=True)
  (hidden2): Linear(in_features=100, out_features=100, bias=True)
  (output): Linear(in_features=100, out_features=1, bias=True)
  (relu): ReLU()
)
Discriminator(
  (hidden1): Linear(in_features=12, out_features=25, bias=True)
  (hidden2): Linear(in_features=25, out_features=50, bias=True)
  (output): Linear(in_features=50, out_features=1, bias=True)
  (relu): ReLU()
)
```

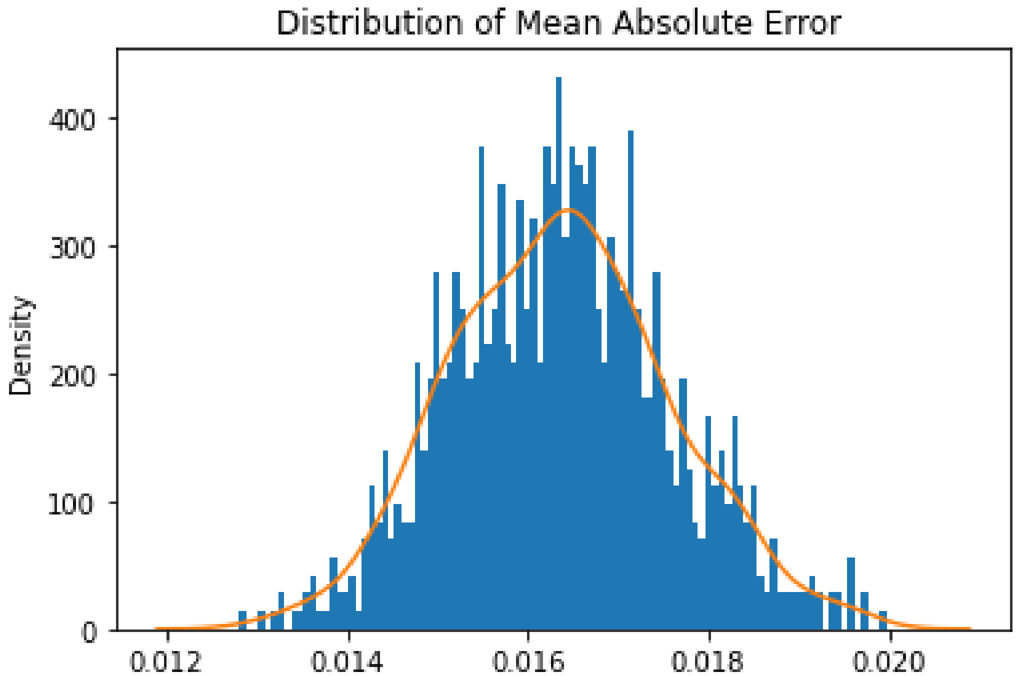
[11]:
```
train_test.
 ↪training_GAN(discriminator,generator,disc_opt,gen_opt,real_dataset,batch_size,␣
 ↪n_epochs,criterion,device)
```



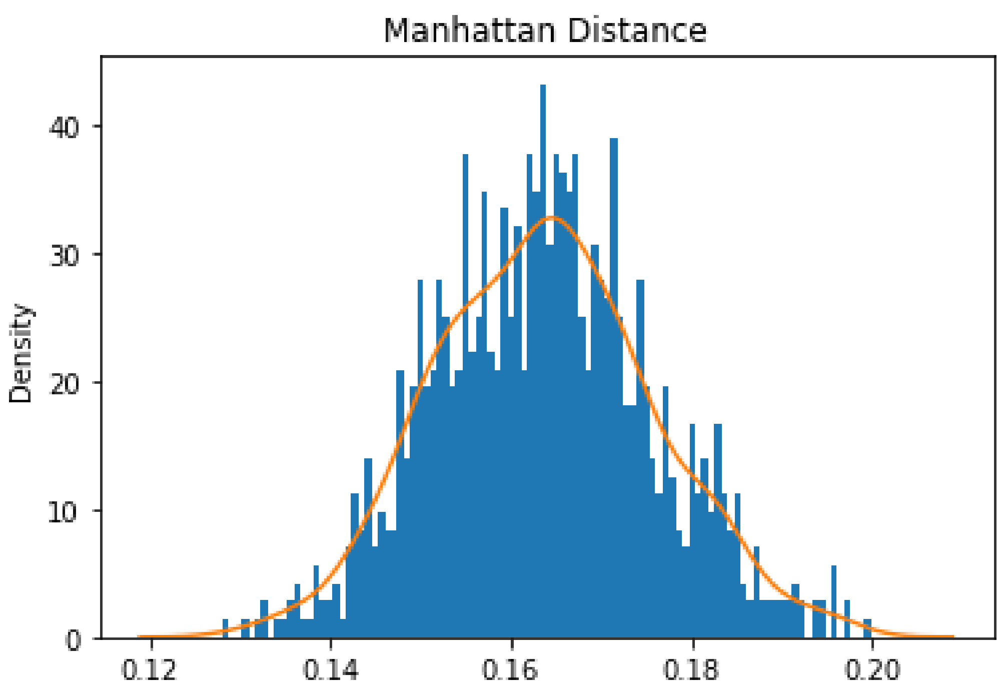
[12]:
```
train_test.test_generator(generator,real_dataset,device)
```



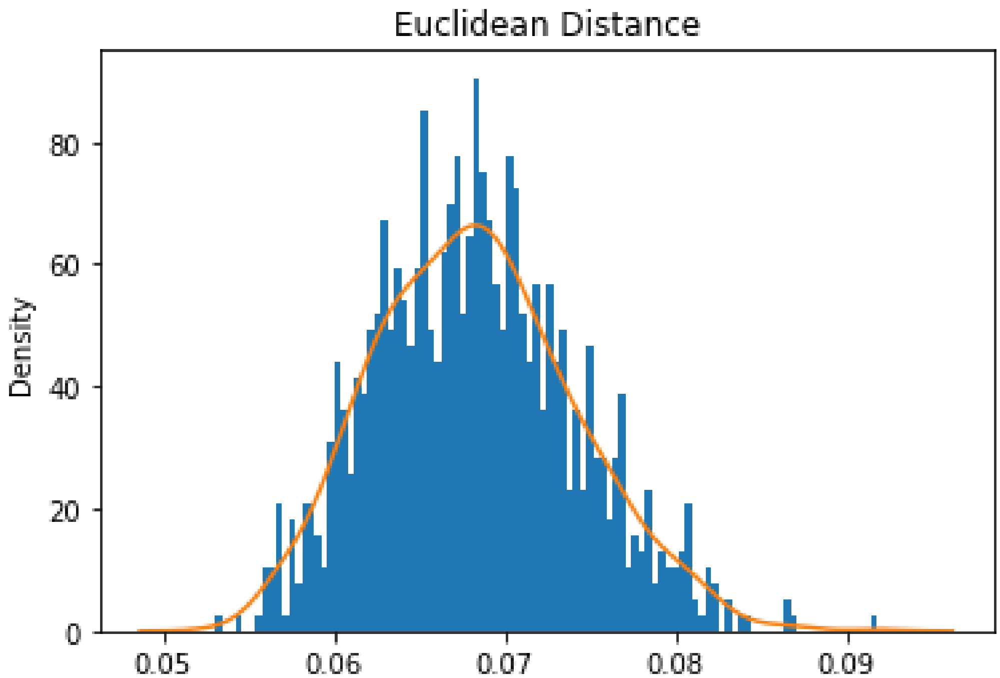
```
Mean Square Error: 0.00046926331946791456
```




```
Mean Absolute Error: 0.016364656138420103
```



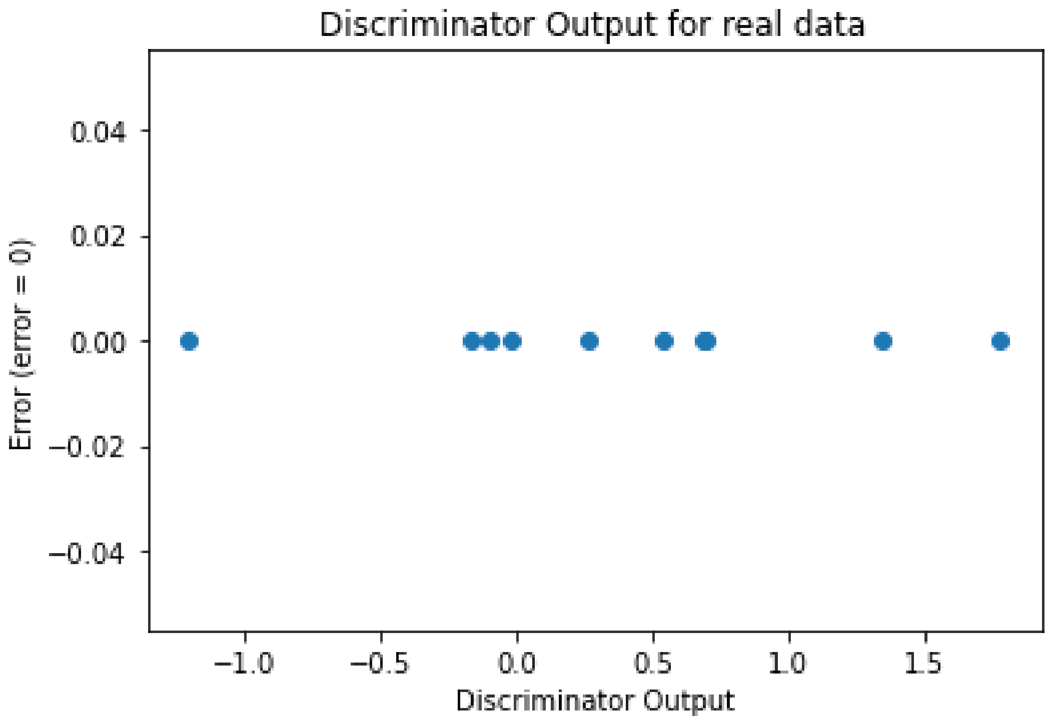
```
Mean Manhattan Distance: 0.16364656138420106
```

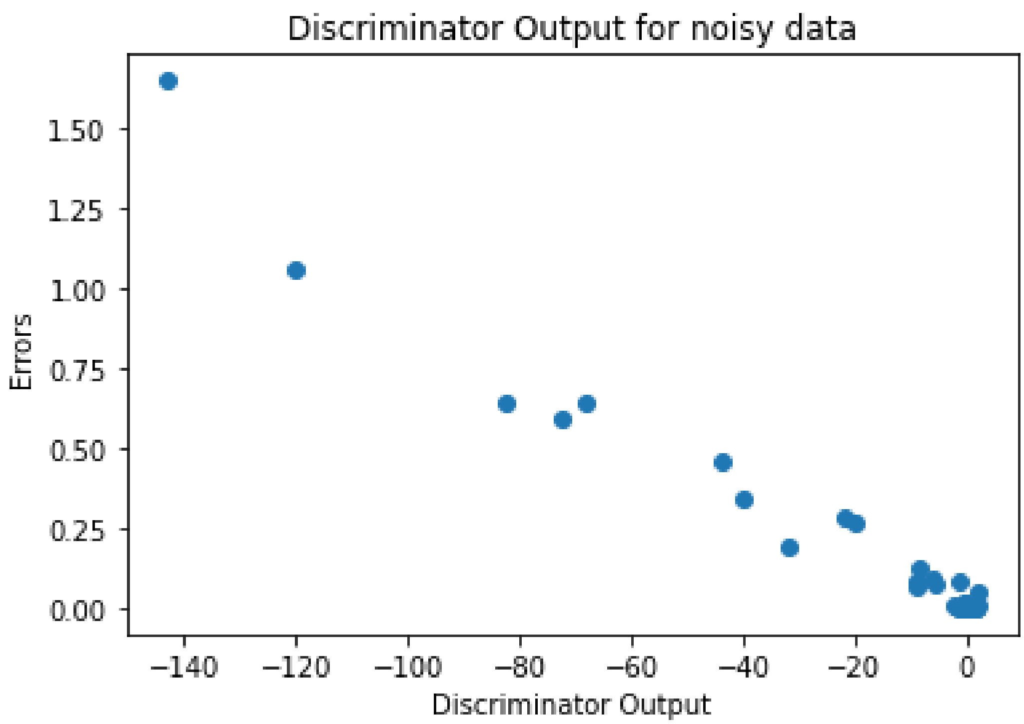



```
Mean Euclidean Distance: 0.06825283924140164
```

[13]:

```
sanityChecks.discProbVsError(real_dataset,discriminator,device)
```


Discriminator Output for real data
0.04
0.02
0.00
−0.02
−0.04
Error (error = 0)
−1.0
−0.5
0.0
0.5
1.0
1.5
Discriminator Output

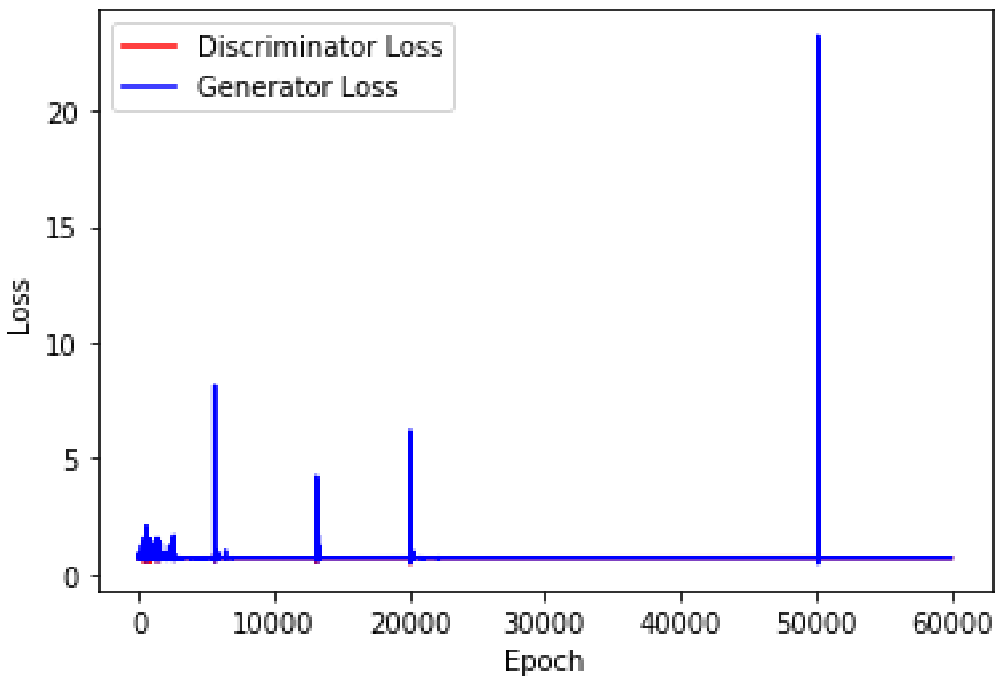

Discriminator Output for noisy data
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Errors
−140
−120
−100
−80
−60
−40
−20
0
Discriminator Output

Training GAN until mse of y_pred is > 0.1 or n_epochs < 30000

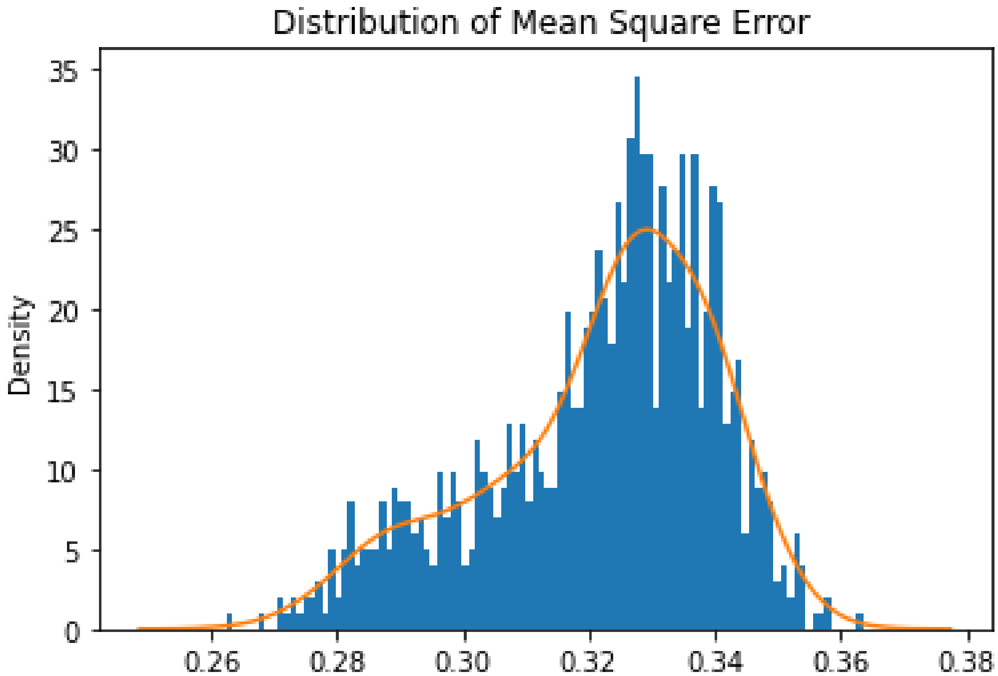
[14]:
```
generator = network.Generator(n_features+2)
discriminator = network.Discriminator(n_features+2)
criterion = torch.nn.BCEWithLogitsLoss()
gen_opt = torch.optim.Adam(generator.parameters(), lr=0.01, betas=(0.5, 0.999))
disc_opt = torch.optim.Adam(discriminator.parameters(), lr=0.01, betas=(0.5, 0.
 ↪999))
```

[15]:
```
train_test.
 ↪training_GAN_2(discriminator,generator,disc_opt,gen_opt,real_dataset,batch_size,error,crite
```

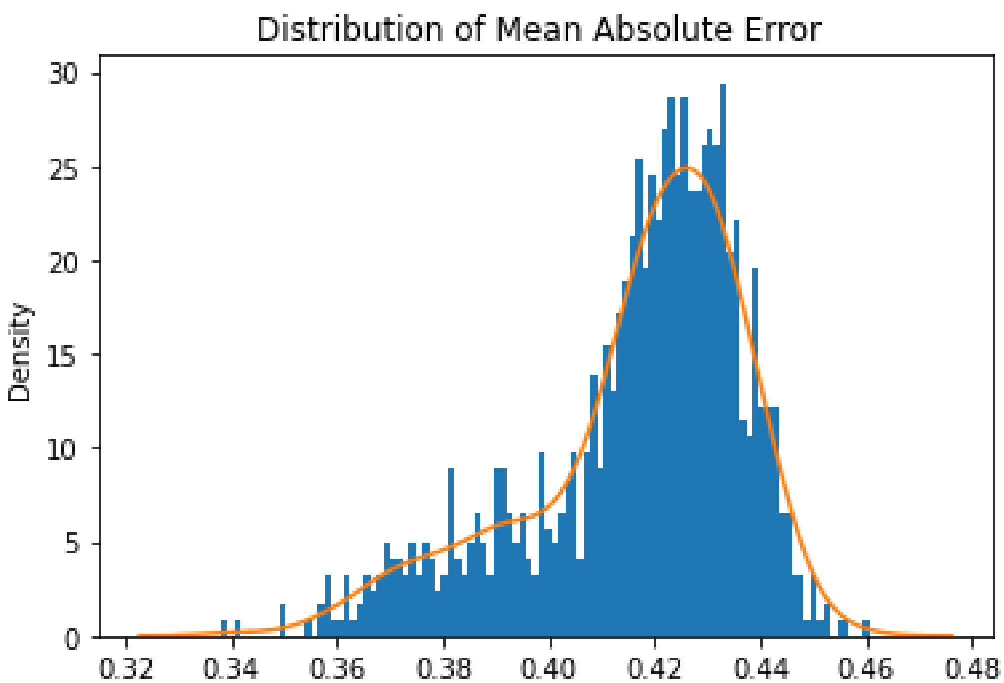
```
Number of epochs needed 30000
```



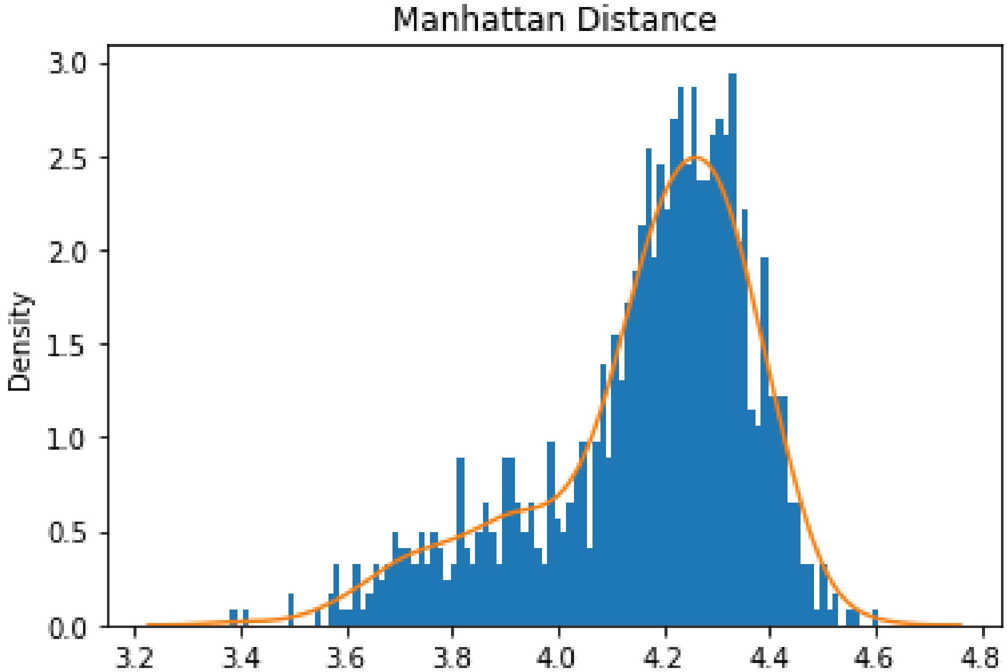
[16]:
```
train_test.test_generator(generator,real_dataset,device)
```



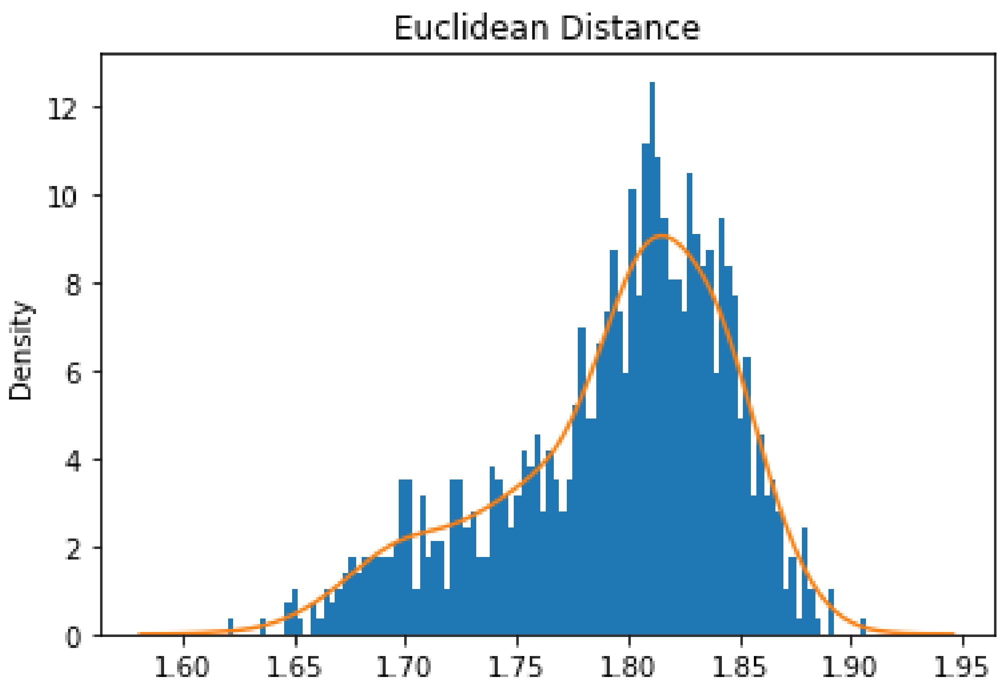
Mean Square Error: 0.3219581482645019



```
Mean Absolute Error: 0.4167684877425432
```




```
Mean Manhattan Distance: 4.167684877425432
```



```
Mean Euclidean Distance: 1.7935713620619622
```

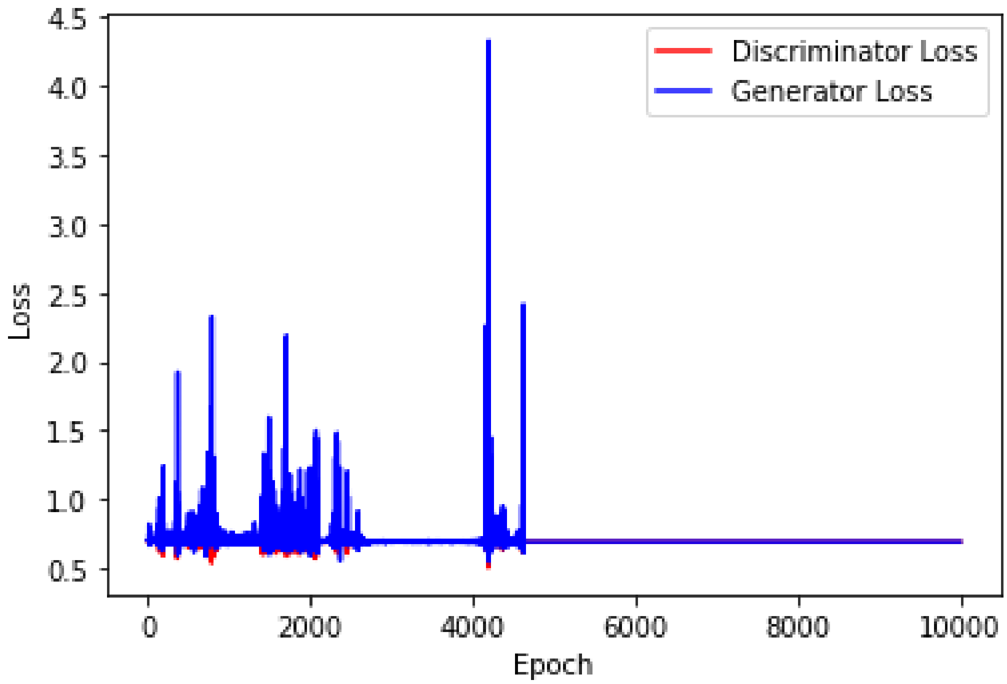
2 ABC GAN Model

2.0.1 Training the network

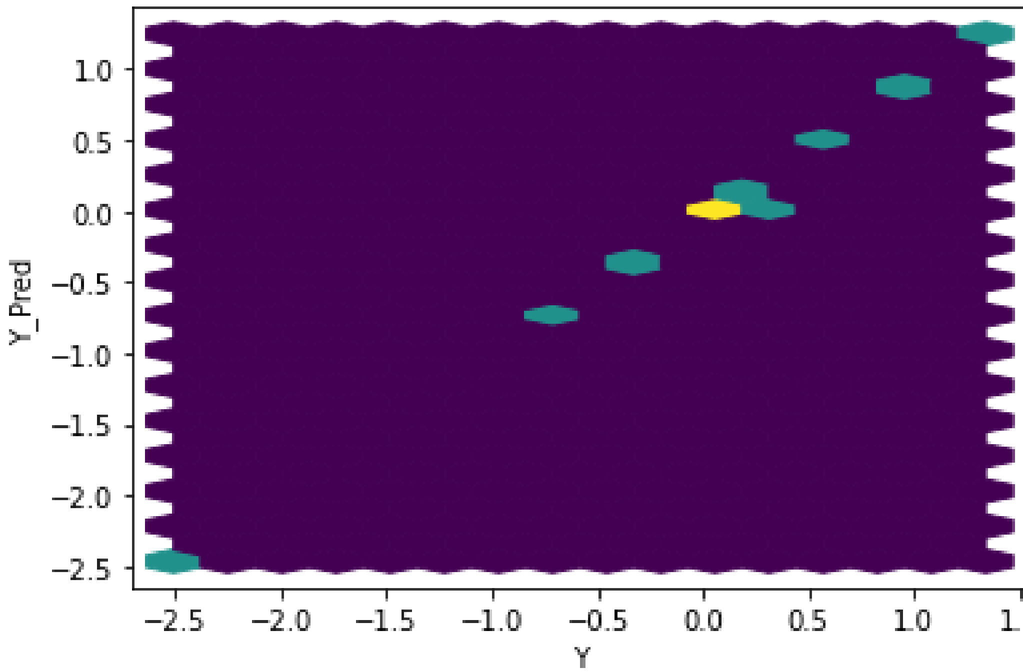
Training ABC-GAN for n_epochs number of epochs

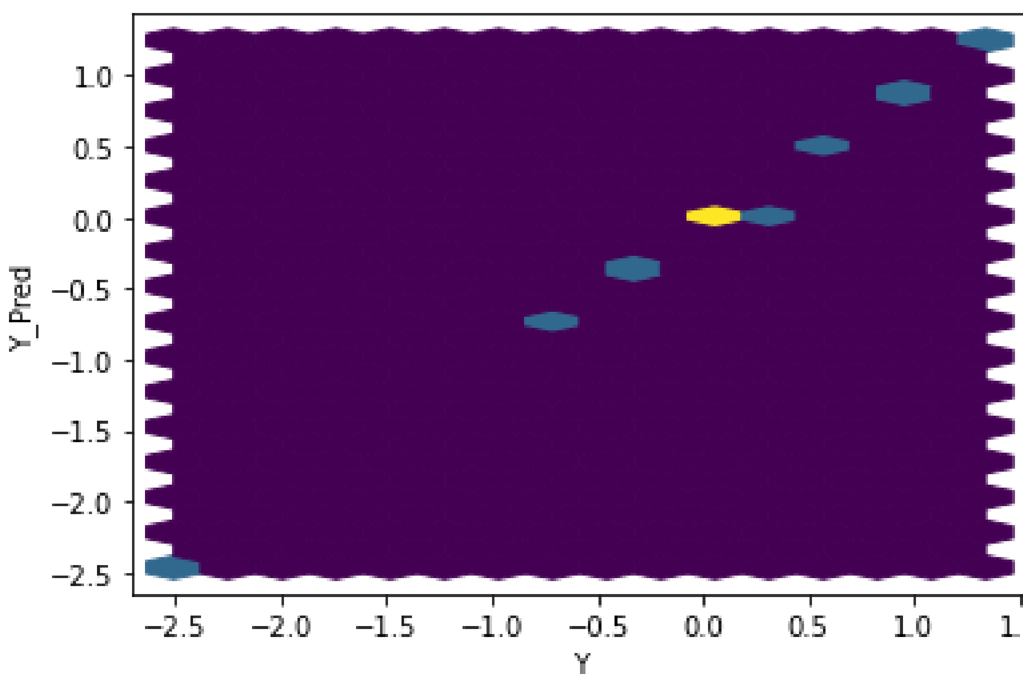
[17]:
```
gen = network.Generator(n_features+2)
disc = network.Discriminator(n_features+2)

criterion = torch.nn.BCEWithLogitsLoss()
gen_opt = torch.optim.Adam(gen.parameters(), lr=0.01, betas=(0.5, 0.999))
disc_opt = torch.optim.Adam(disc.parameters(), lr=0.01, betas=(0.5, 0.999))
```

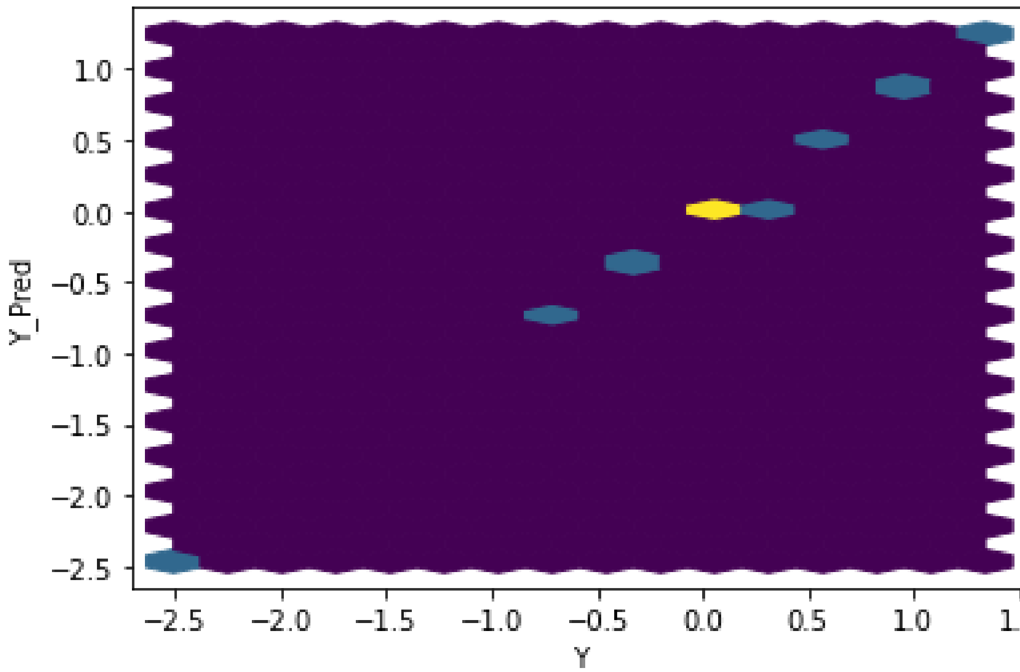
[18]:
```
ABC_train_test.training_GAN(disc, gen,disc_opt,gen_opt,real_dataset,␣
 ↪batch_size, n_epochs,criterion,coeff,mean,variance,device)
```

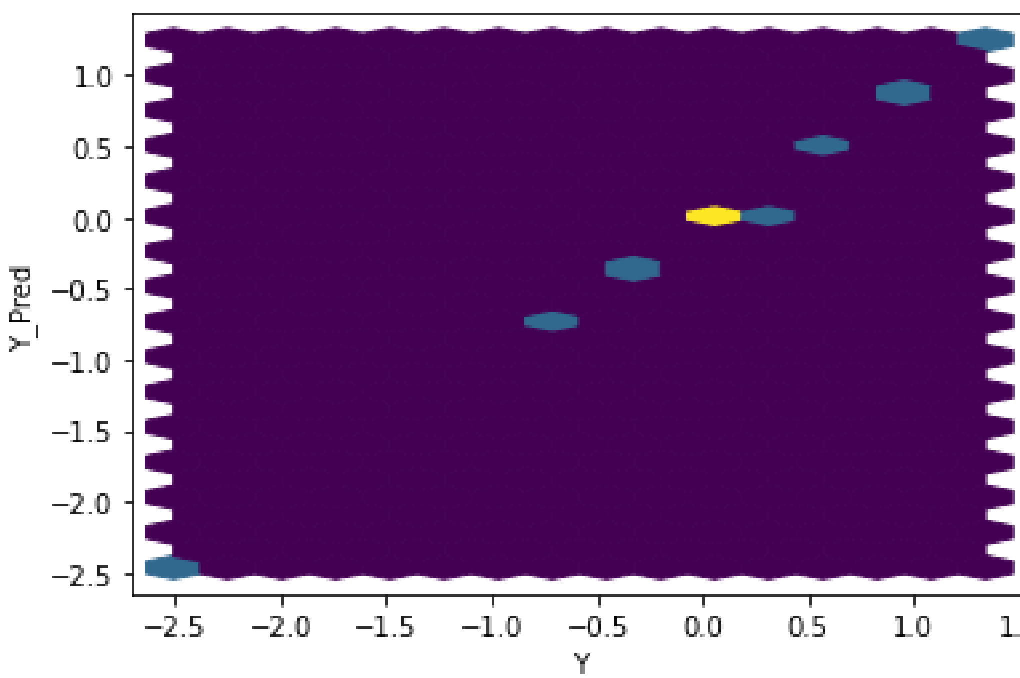


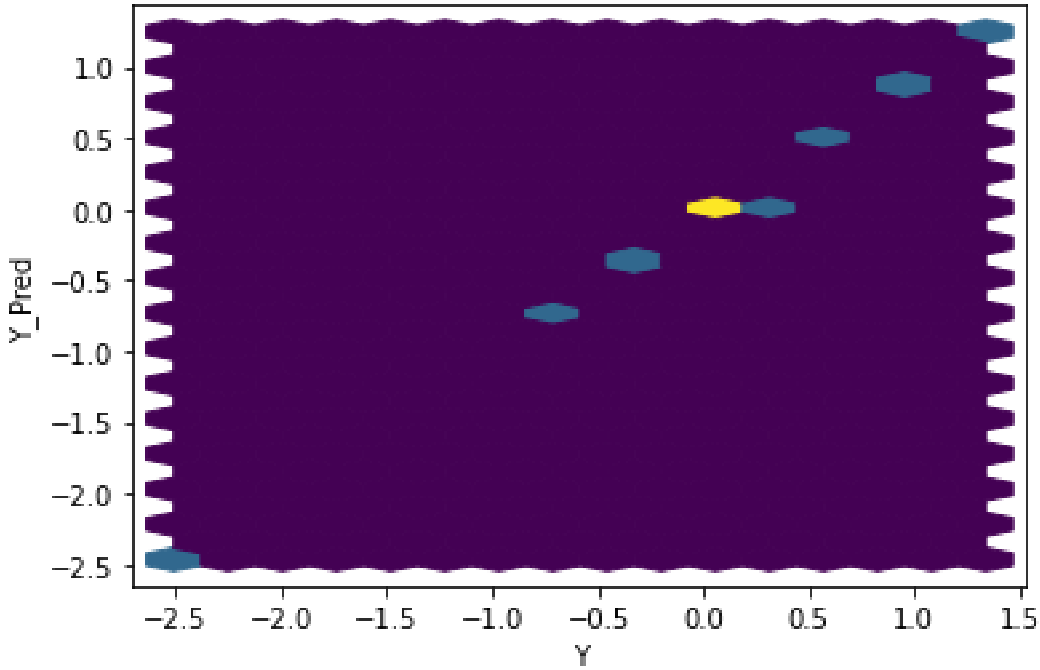
[19]:
```
ABC_train_test.test_generator(gen,real_dataset,coeff,mean,variance,device)
```

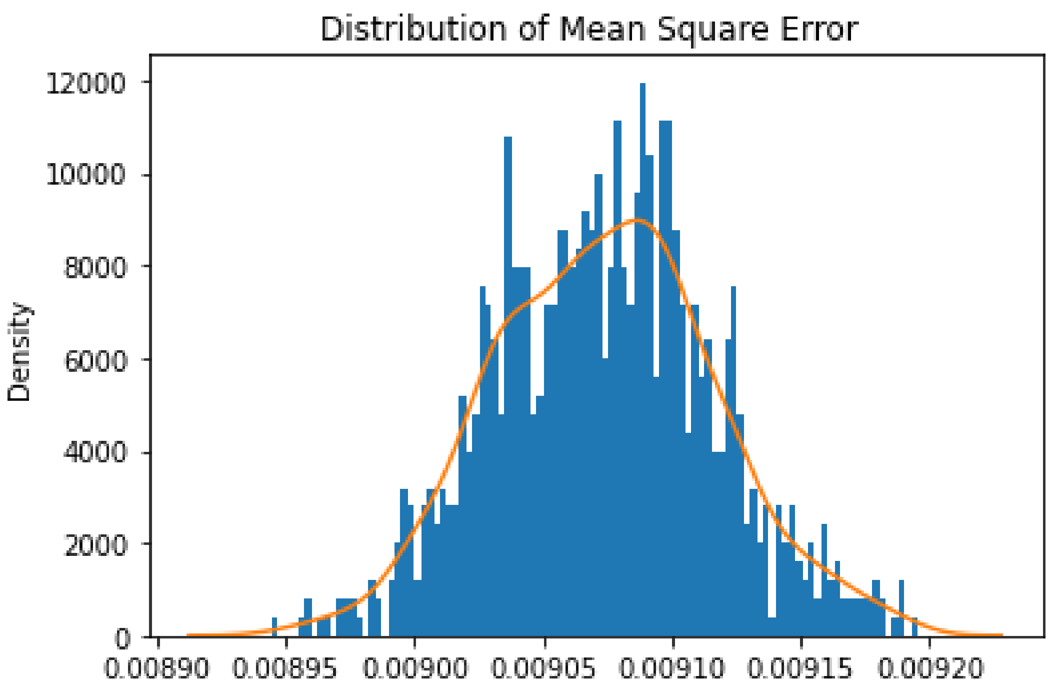

1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
Y_Pred
−2.5 −2.0 −1.5 −1.0 −0.5 0.0 0.5 1.0 1.5
Y


1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
Y_Pred
−2.5 −2.0 −1.5 −1.0 −0.5 0.0 0.5 1.0 1.5
Y

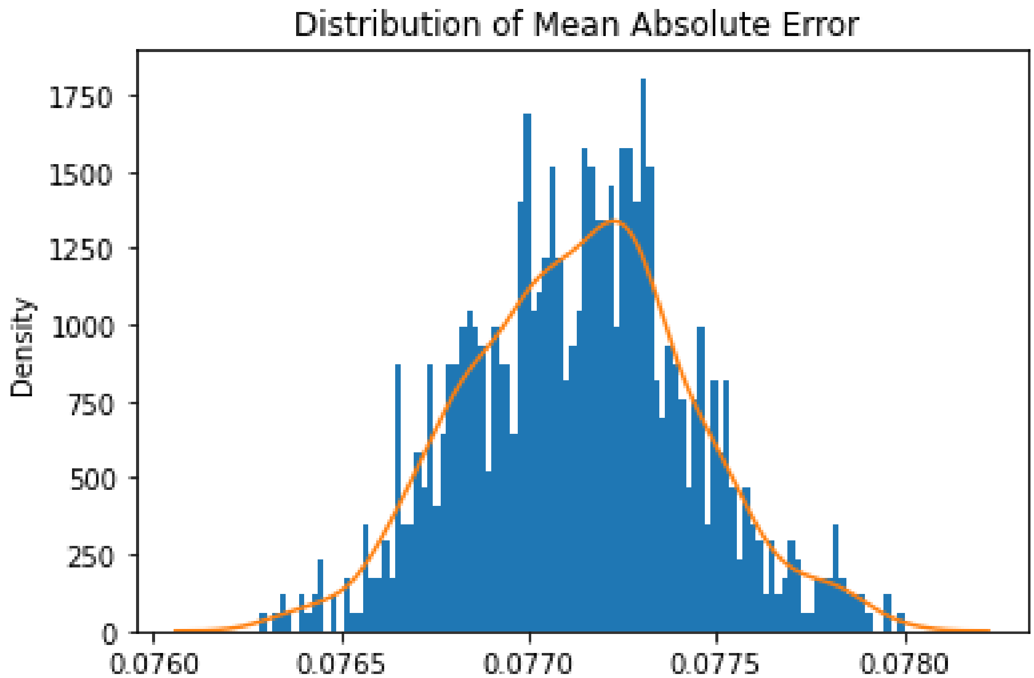

Y_Pred
Y
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5


Y_Pred
Y
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

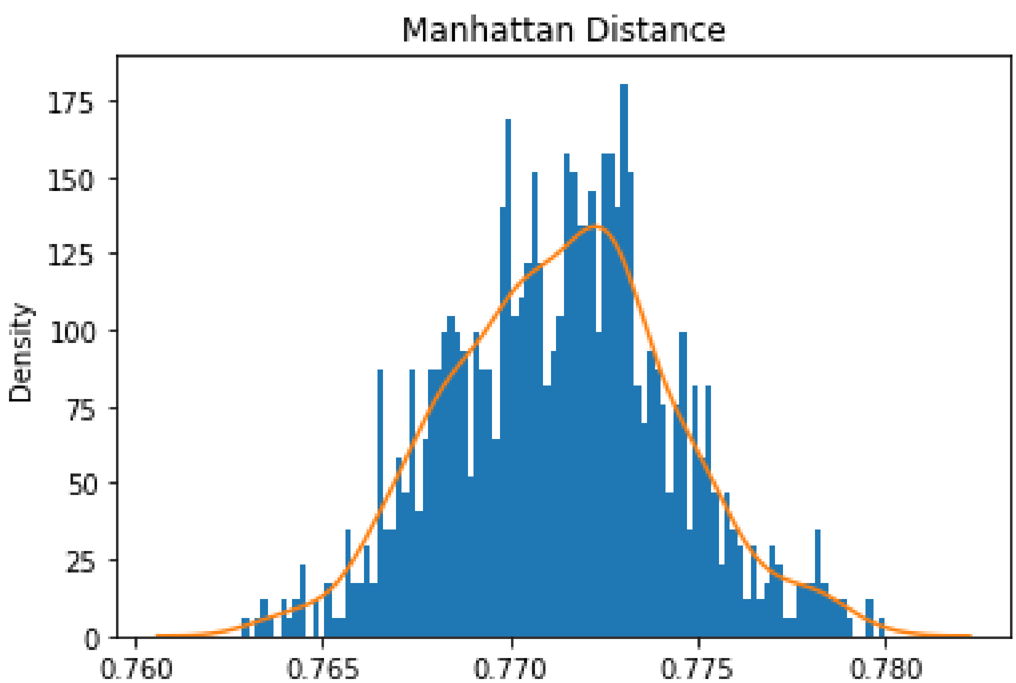




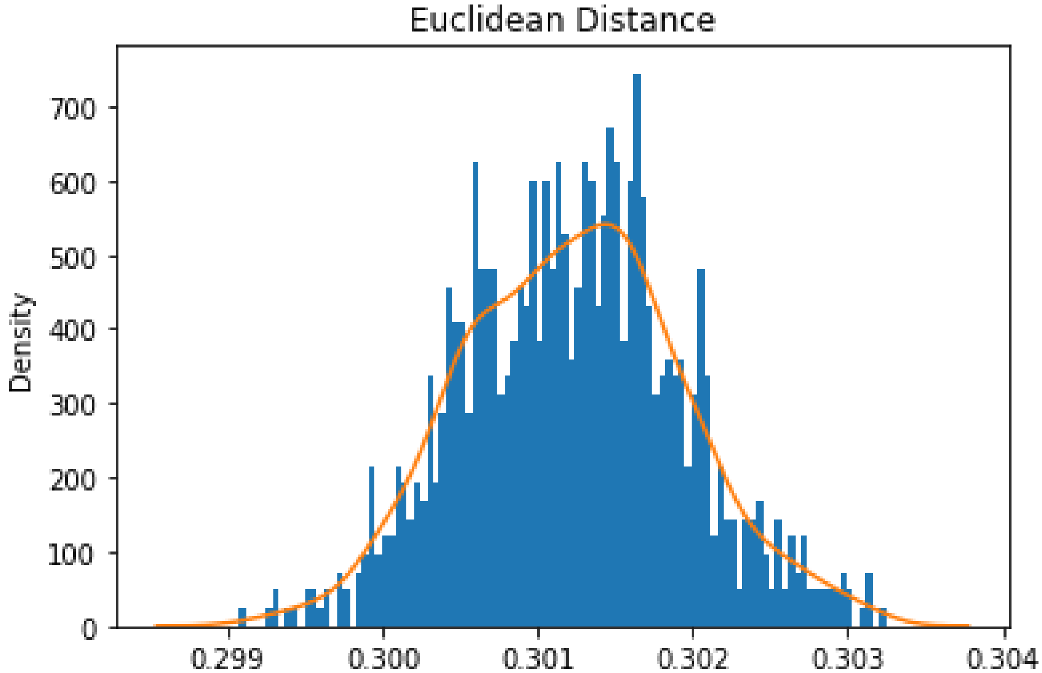

```
Mean Square Error: 0.009073815352499336
```




```
Mean Absolute Error: 0.0771349802929908
Mean Manhattan Distance: 0.7713498029299081
```

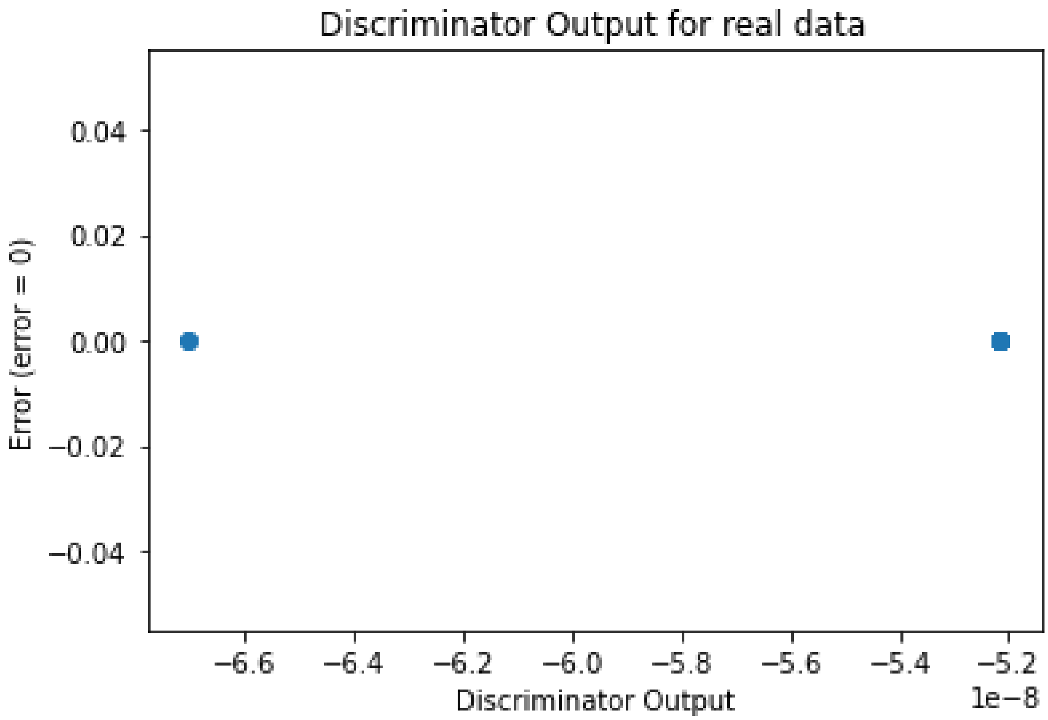


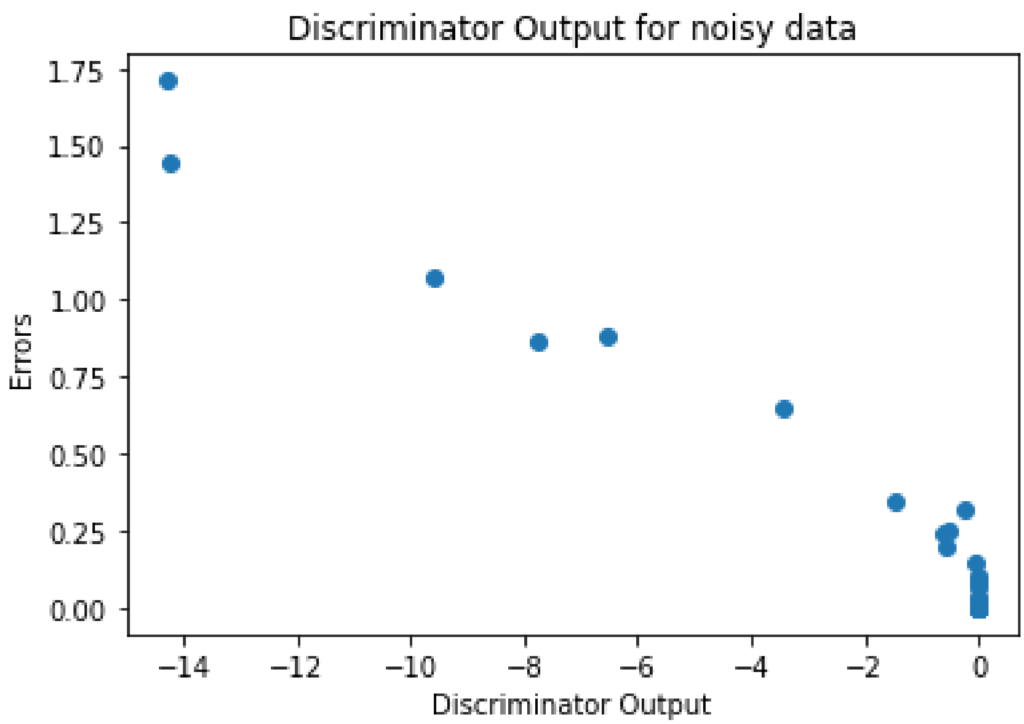
```
Mean Euclidean Distance: 0.3012269074844144
```



Sanity Checks

[20]:
```
sanityChecks.discProbVsError(real_dataset,disc,device)
```


Discriminator Output for real data
0.04
0.02
0.00
−0.02
−0.04
Error (error = 0)
−6.6
−6.4
−6.2
−6.0
−5.8
−5.6
−5.4
−5.2
1e−8
Discriminator Output

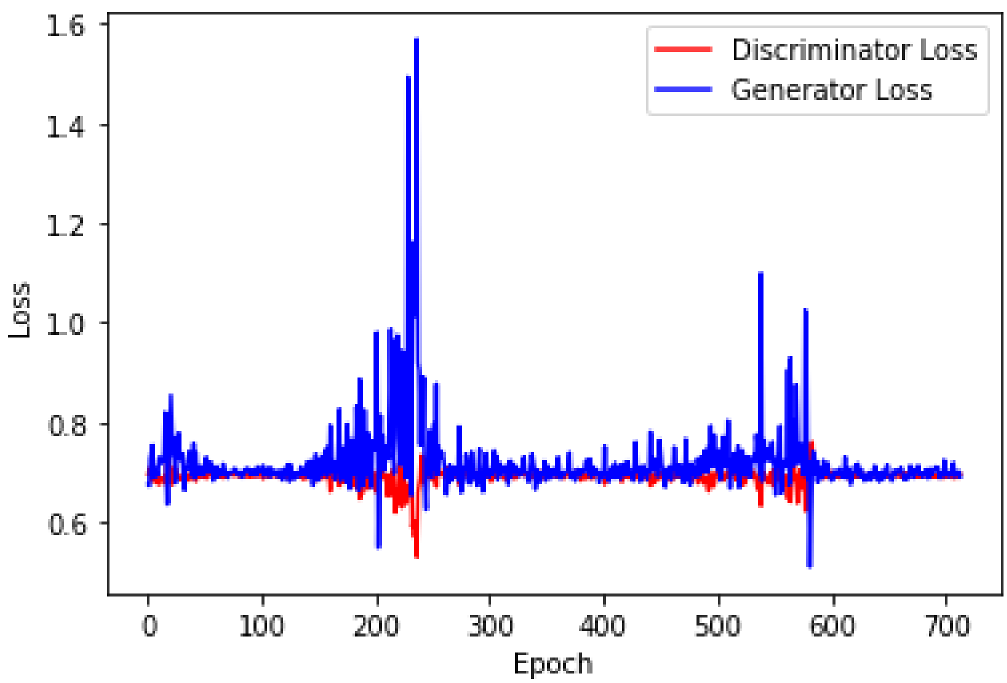

Discriminator Output for noisy data
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Errors
−14
−12
−10
−8
−6
−4
−2
0
Discriminator Output

Training GAN until mse of y_pred is > 0.1 or n_epochs < 30000

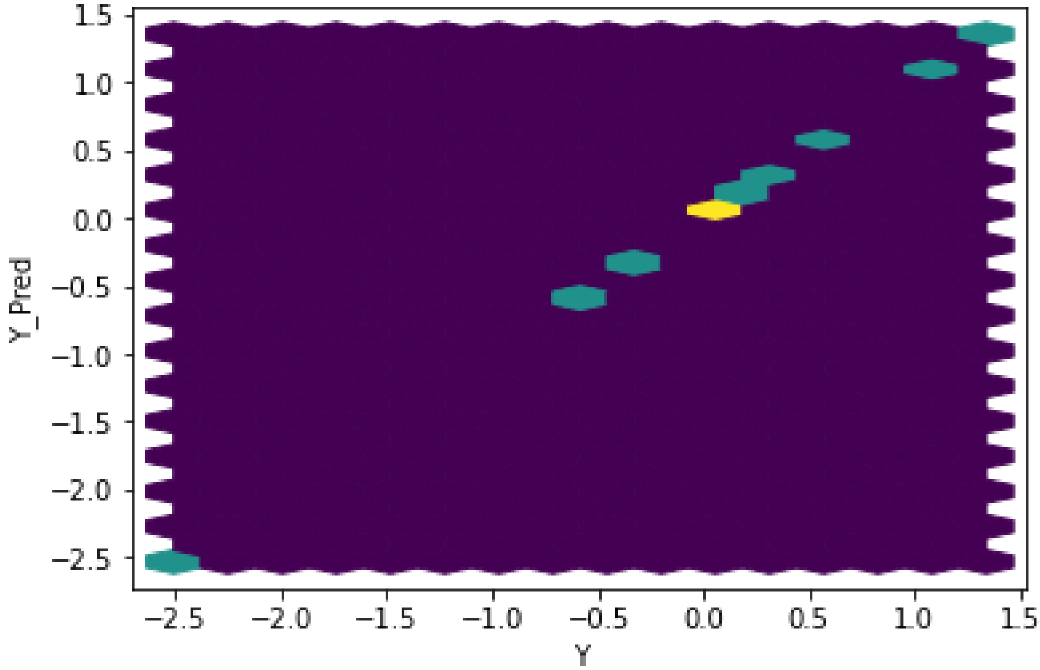
[21]:
```
gen = network.Generator(n_features+2)
disc = network.Discriminator(n_features+2)

criterion = torch.nn.BCEWithLogitsLoss()
gen_opt = torch.optim.Adam(gen.parameters(), lr=0.01, betas=(0.5, 0.999))
disc_opt = torch.optim.Adam(disc.parameters(), lr=0.01, betas=(0.5, 0.999))
```

[22]:
```
ABC_train_test.
↪training_GAN_2(disc,gen,disc_opt,gen_opt,real_dataset,batch_size,␣
↪error,criterion,coeff,mean,variance,device)
```

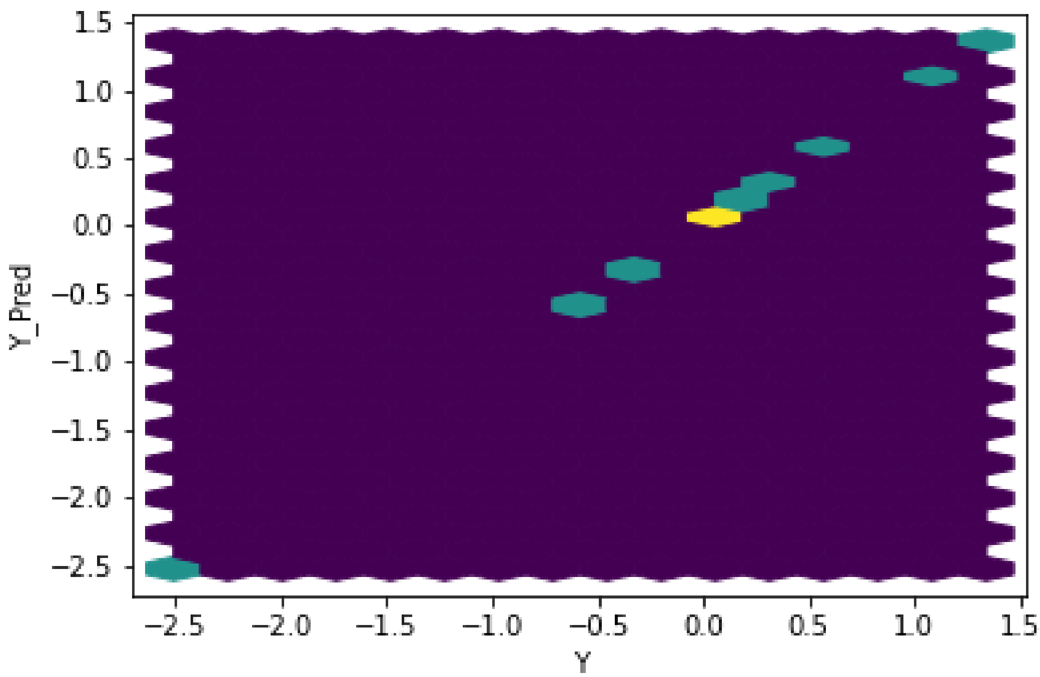
```
Number of epochs 357
```



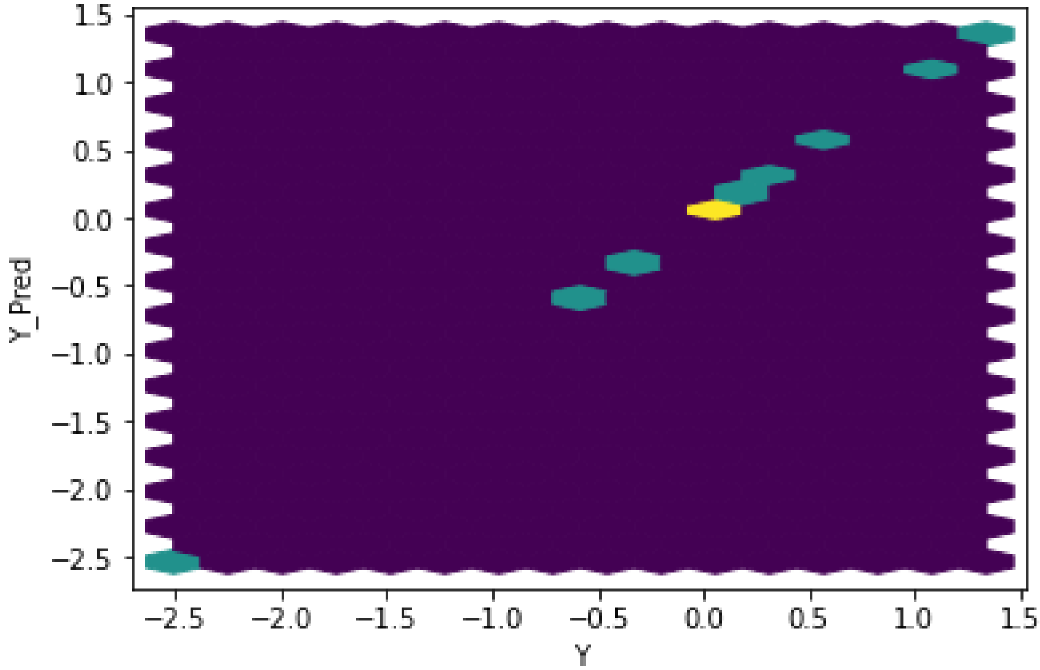
[23]:
```
ABC_train_test.test_generator(gen,real_dataset,coeff,mean,variance,device)
```



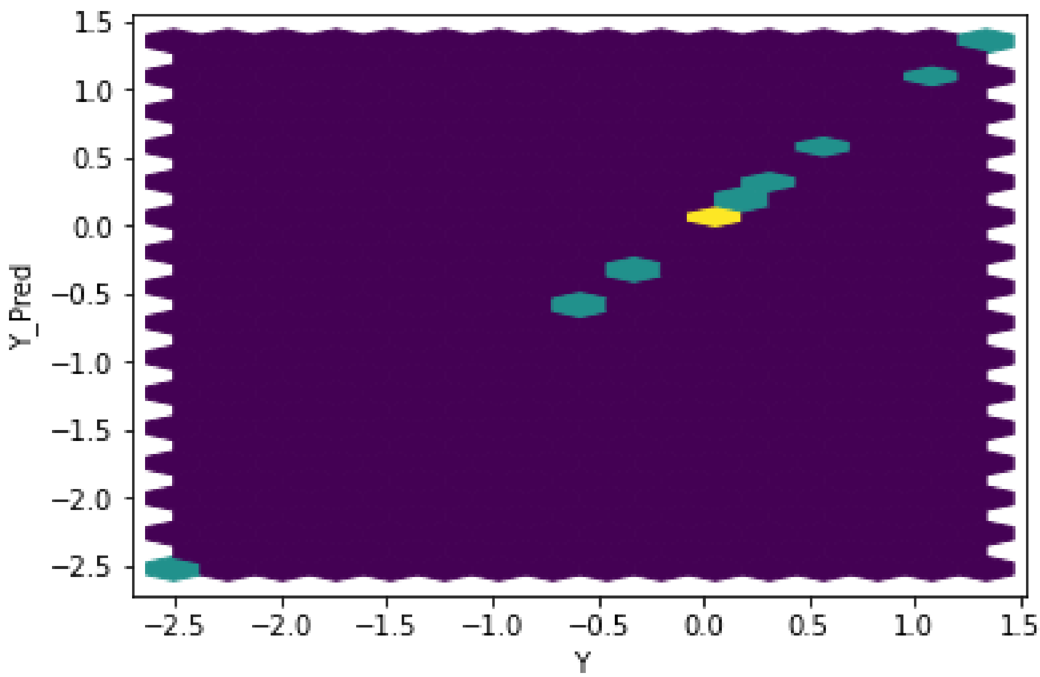
Y_Pred
Y
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.5



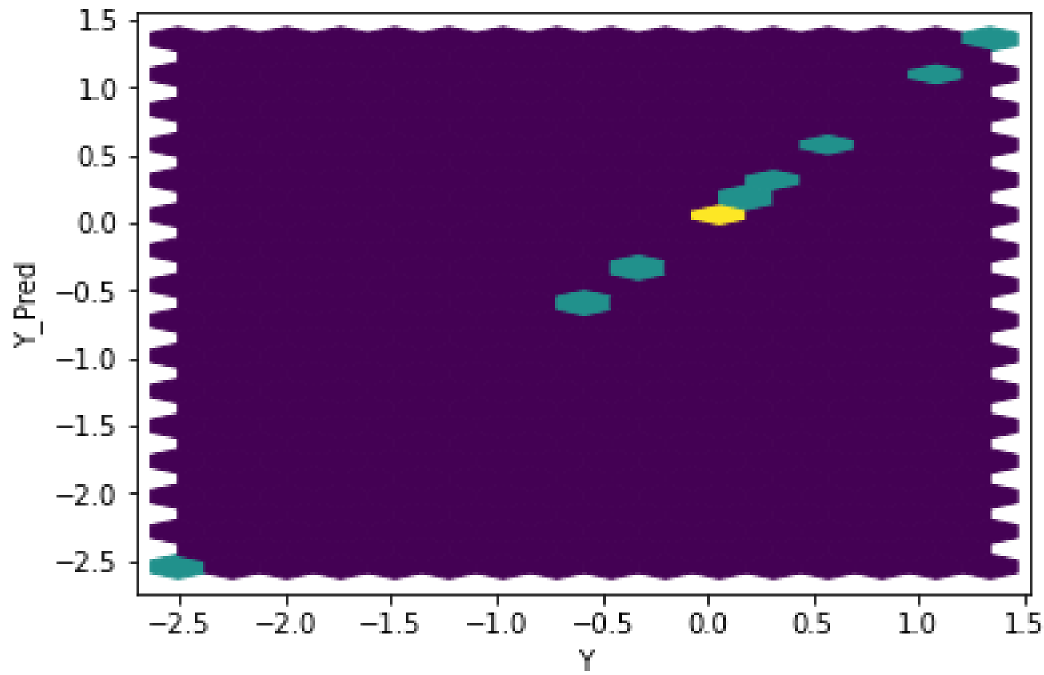
Y_Pred
Y
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.5

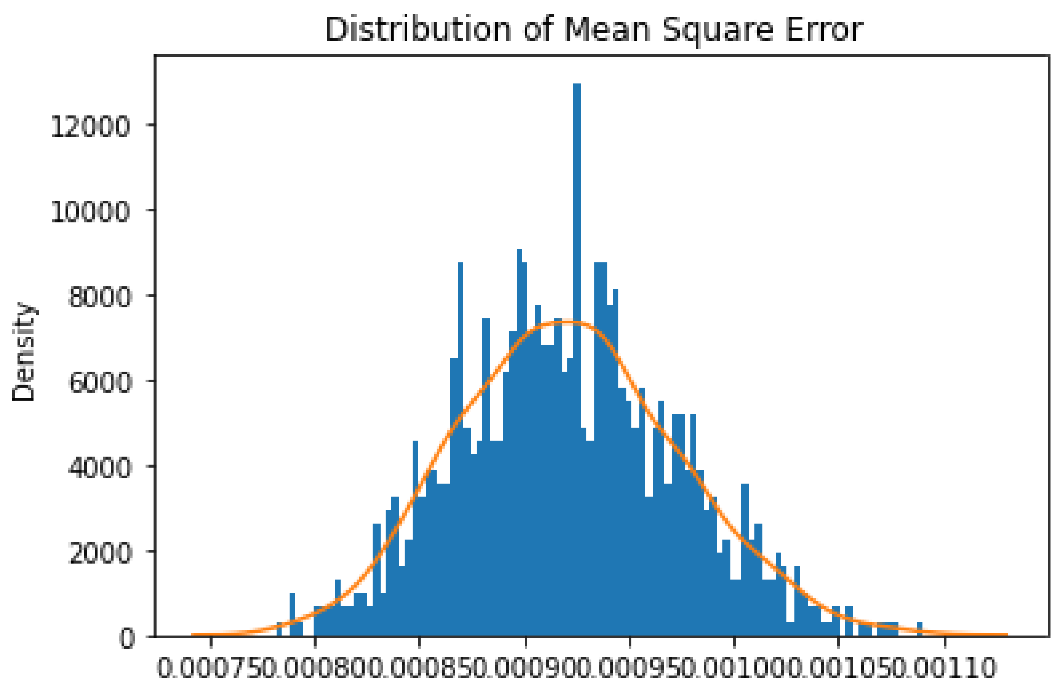


1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
Y_Pred
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Y

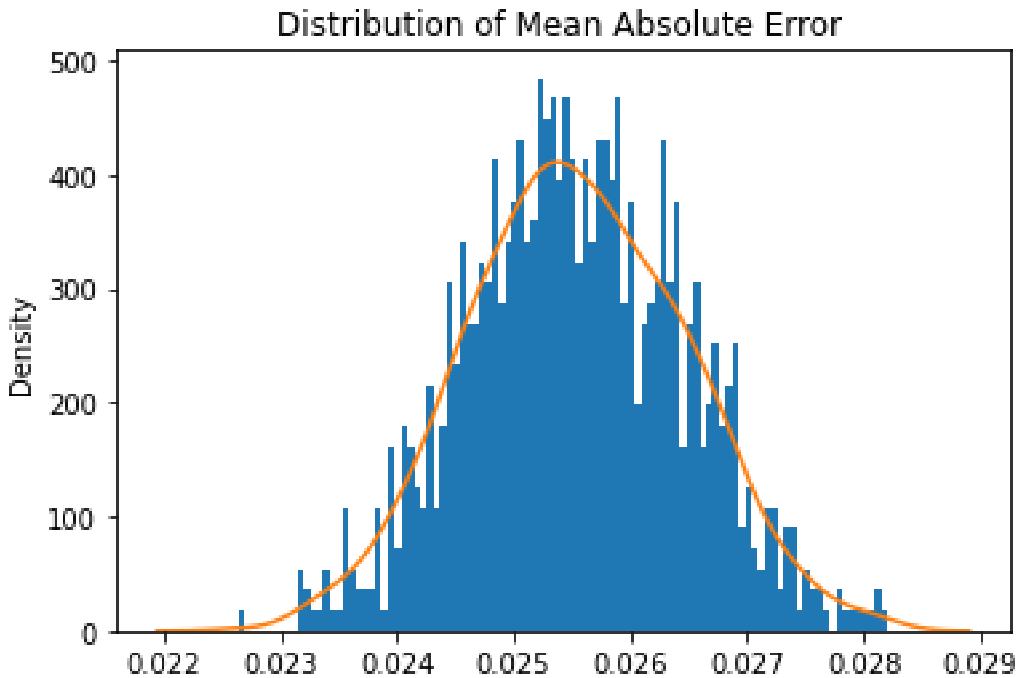


1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
Y_Pred
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Y

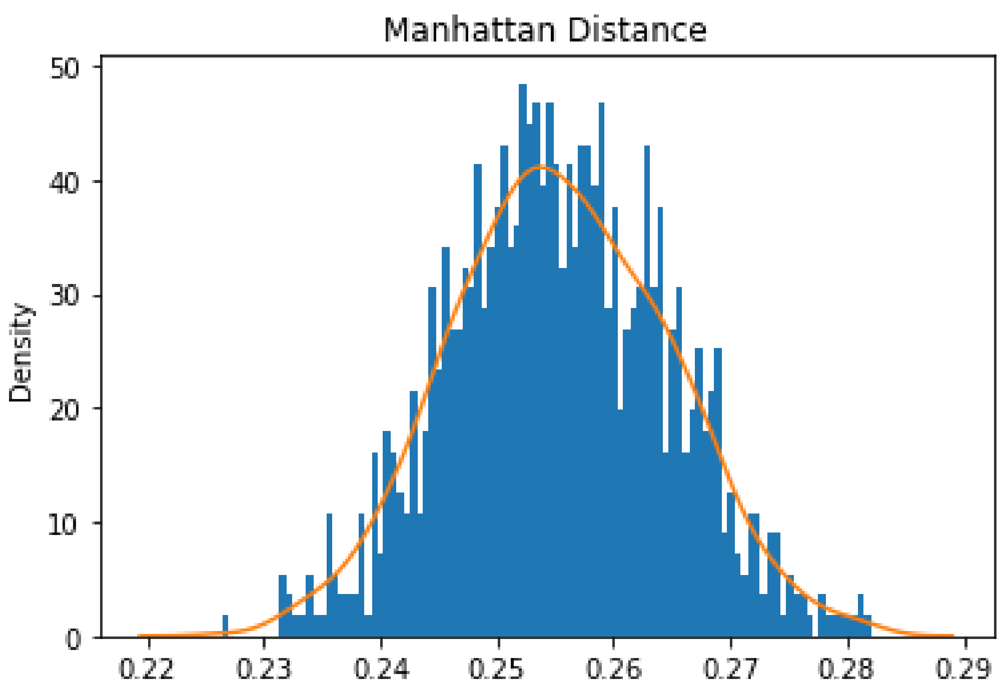




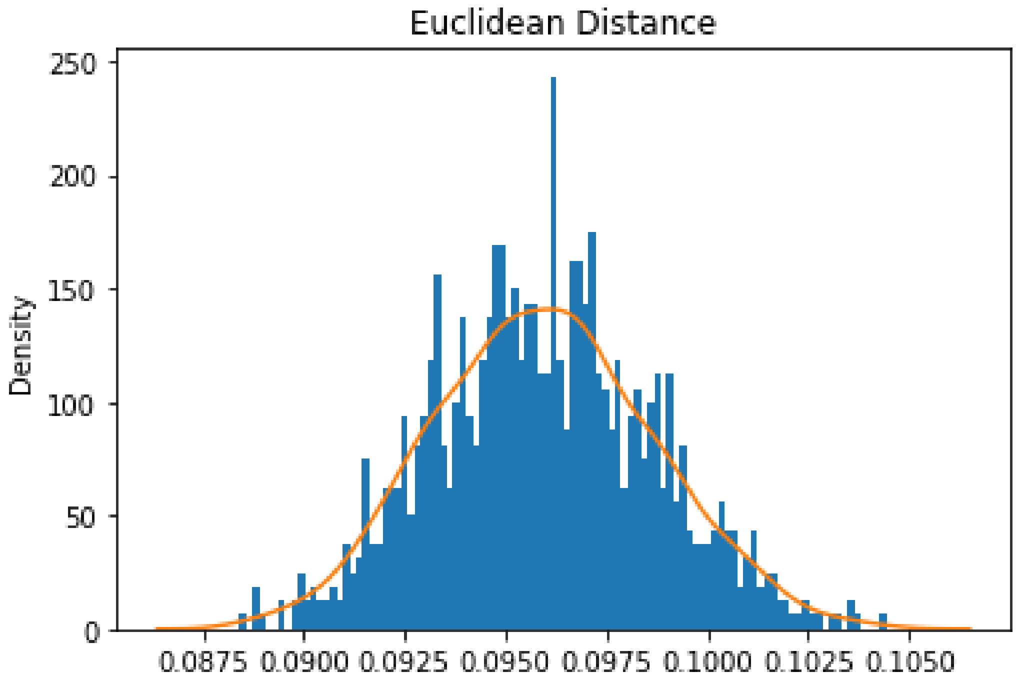

```
Mean Square Error: 0.0009210423904651979
```




```
Mean Absolute Error: 0.025533758898824453
Mean Manhattan Distance: 0.25533758898824455
```



```
Mean Euclidean Distance: 0.09593250489392745
```




```
[ ]:
```