Project Report

UE19CS322 Big Data Project 2

Title: Machine Learning with Spark MLlib

Dataset: **Sentiment Analysis**

Team ID: BD_198_200_367_503

SRN	Name
PES1UG19CS198	JUSTIN JAMES
PES1UG19CS200	K MANISH GOWD
PES1UG19CS367	RAGHUTTAM G
PES1UG19CS503	SREESHA I N

- Design details

We created 3 files

stream.py: to stream the data in batches

extract.py: to train the models,

test.py: to test the models

visualization.py: to visualize data

- Surface level implementation details about each unit

Functions in **stream.py**

def connectTCP():

def streamDataset(tcp_connection, dataset_type):

def streamCSVFile(tcp_connection, input_file):

Functions in **extract.py**

def p_process(rdd): To do preprocessing

Removing punctuations in string using regex

Converting multiple white-spaces into single whitespace

Converting to Dataframe and training the model

Vectorizer : to vectorize the words in tweets

percetron_train_model

bernoulli_train_model

sgd_classifier_train_model

mini_batch_kmeans_cluster_train_model

The above are used to train the model

Functions in **test.py**

def p_process(rdd):

def loadData():

Tests the model

- Reason behind design decisions

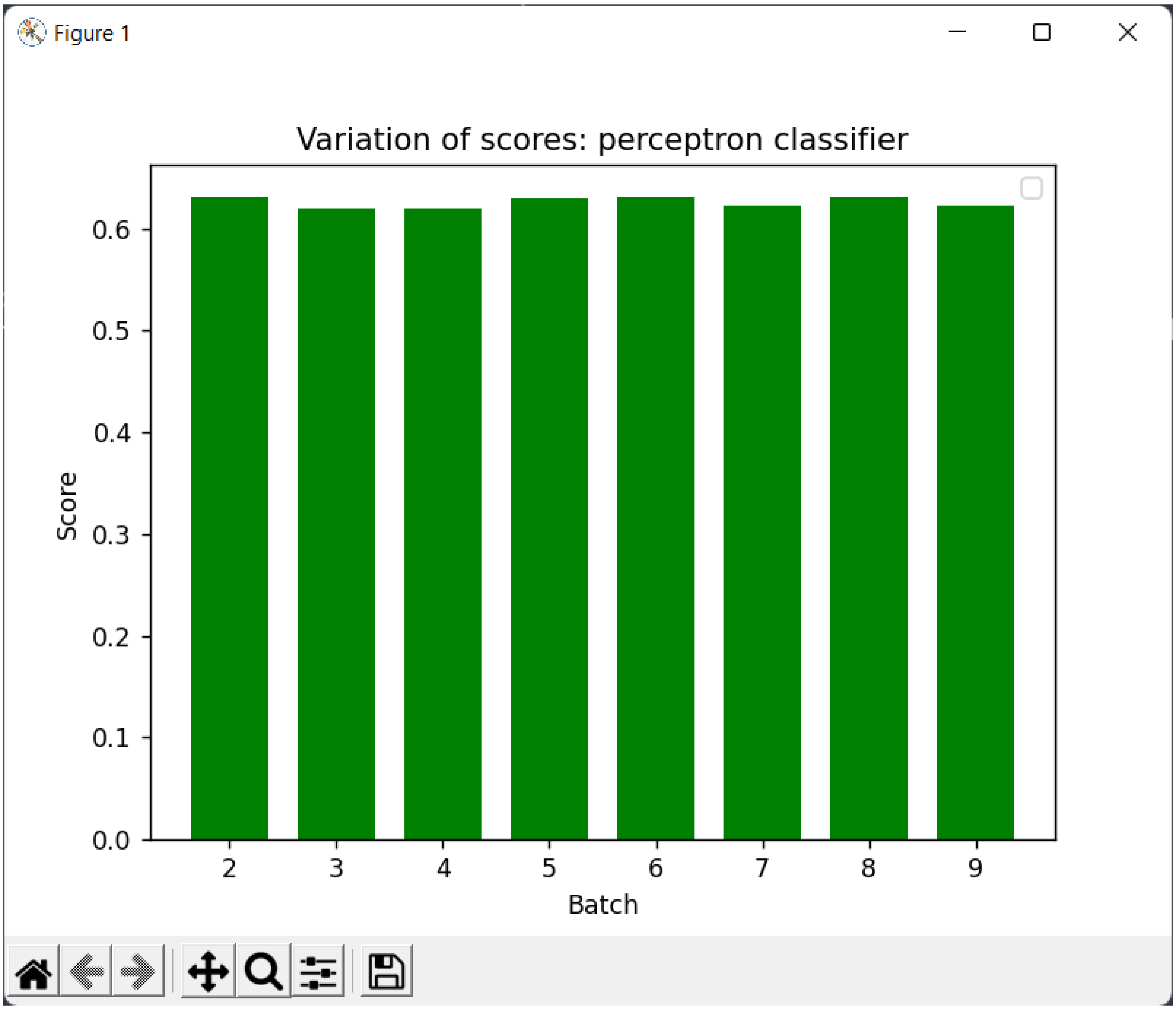
Because it is simple and easy to understand

- Takeaway from the project

Machine learning with streaming is a very complicated and dynamic problem which requires careful planning and execution

Visualization:

Batch size 10000



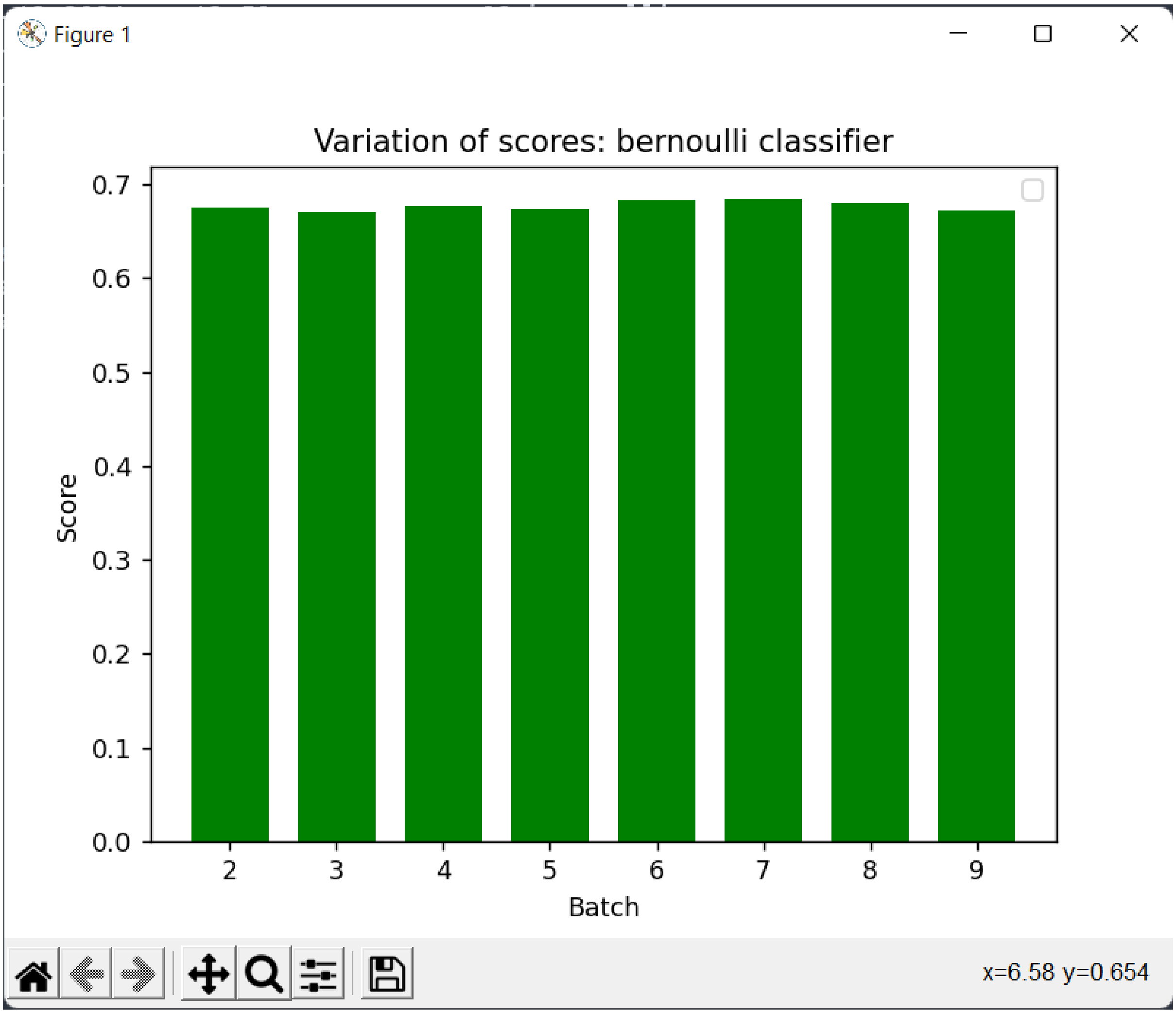


Figure 1
Variation of scores: bernoulli classifier
Score
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
2
3
4
5
6
7
8
9
Batch
x=6.58 y=0.654

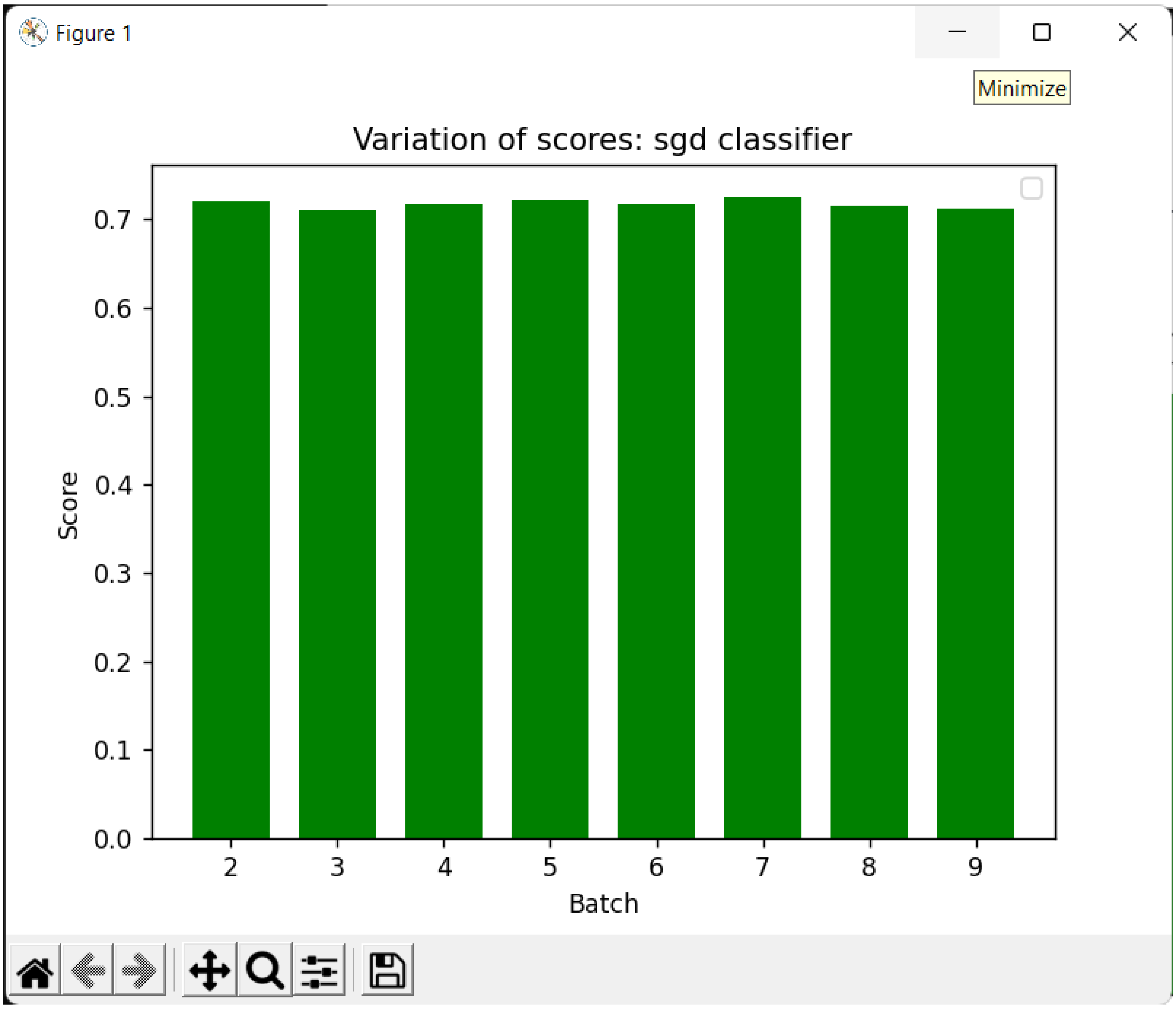
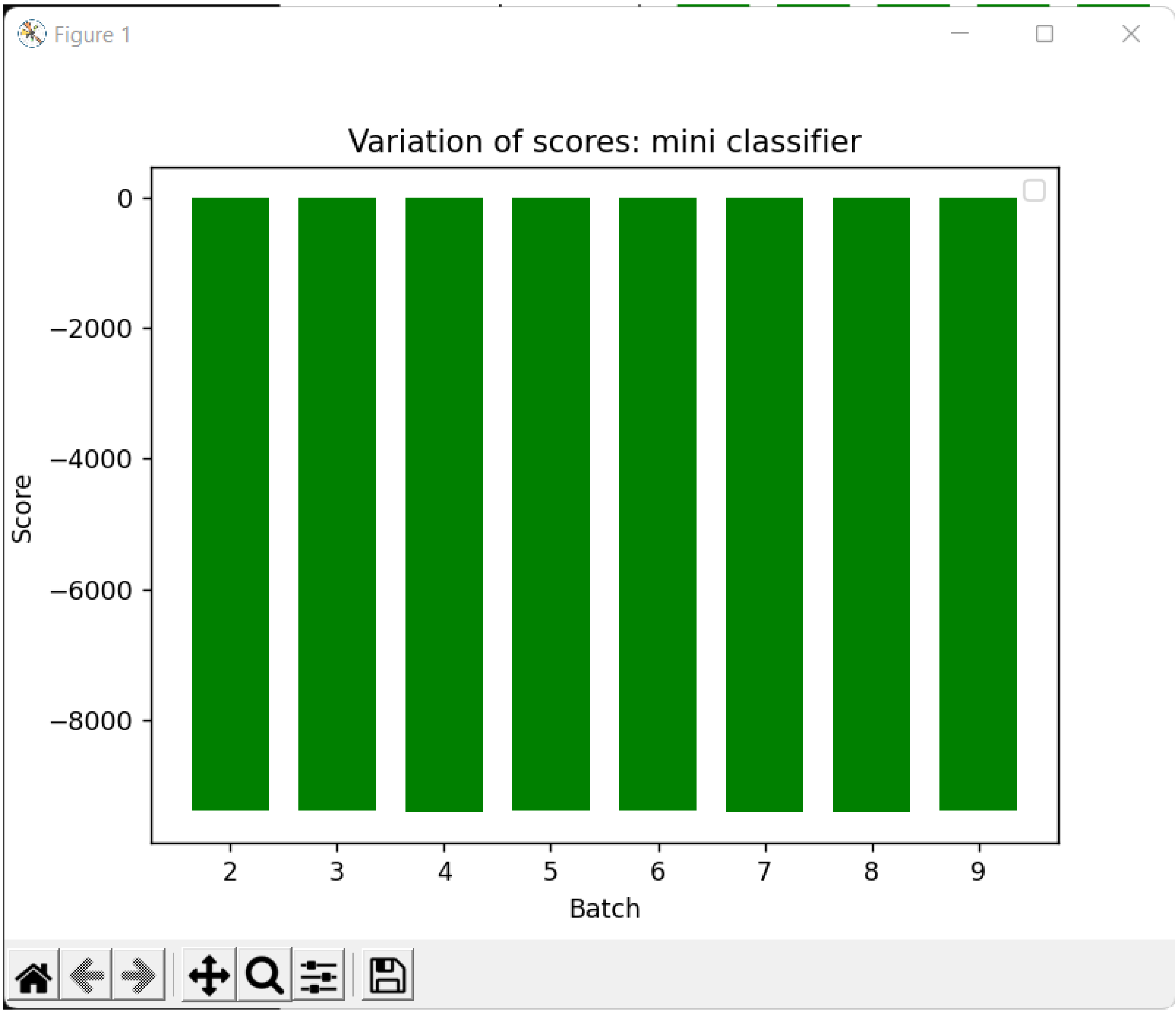


Figure 1
Minimize
Variation of scores: sgd classifier
Score
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
2
3
4
5
6
7
8
9
Batch



For batch size 1000

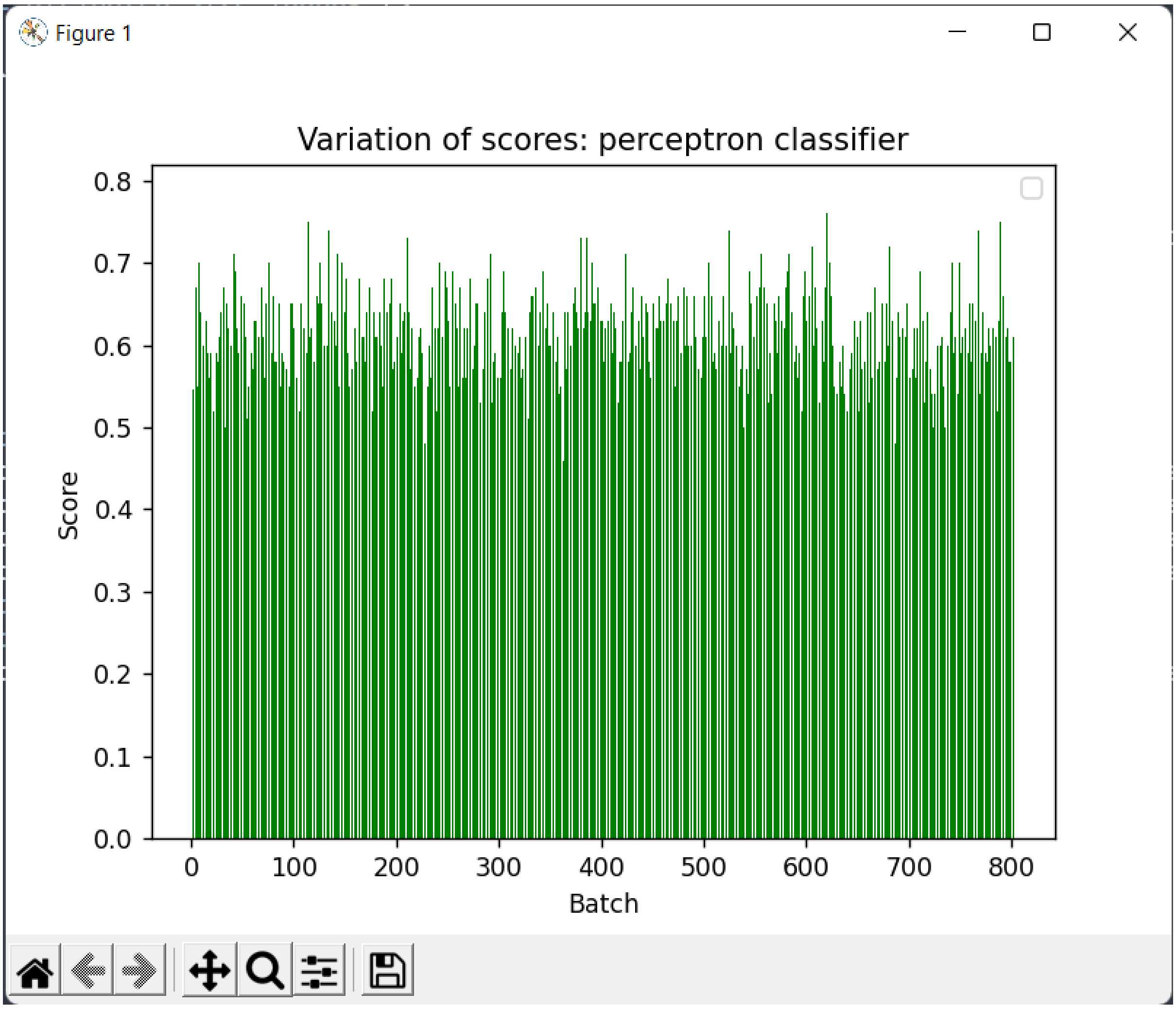


Figure 1
Variation of scores: perceptron classifier
Score
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Batch
0
100
200
300
400
500
600
700
800

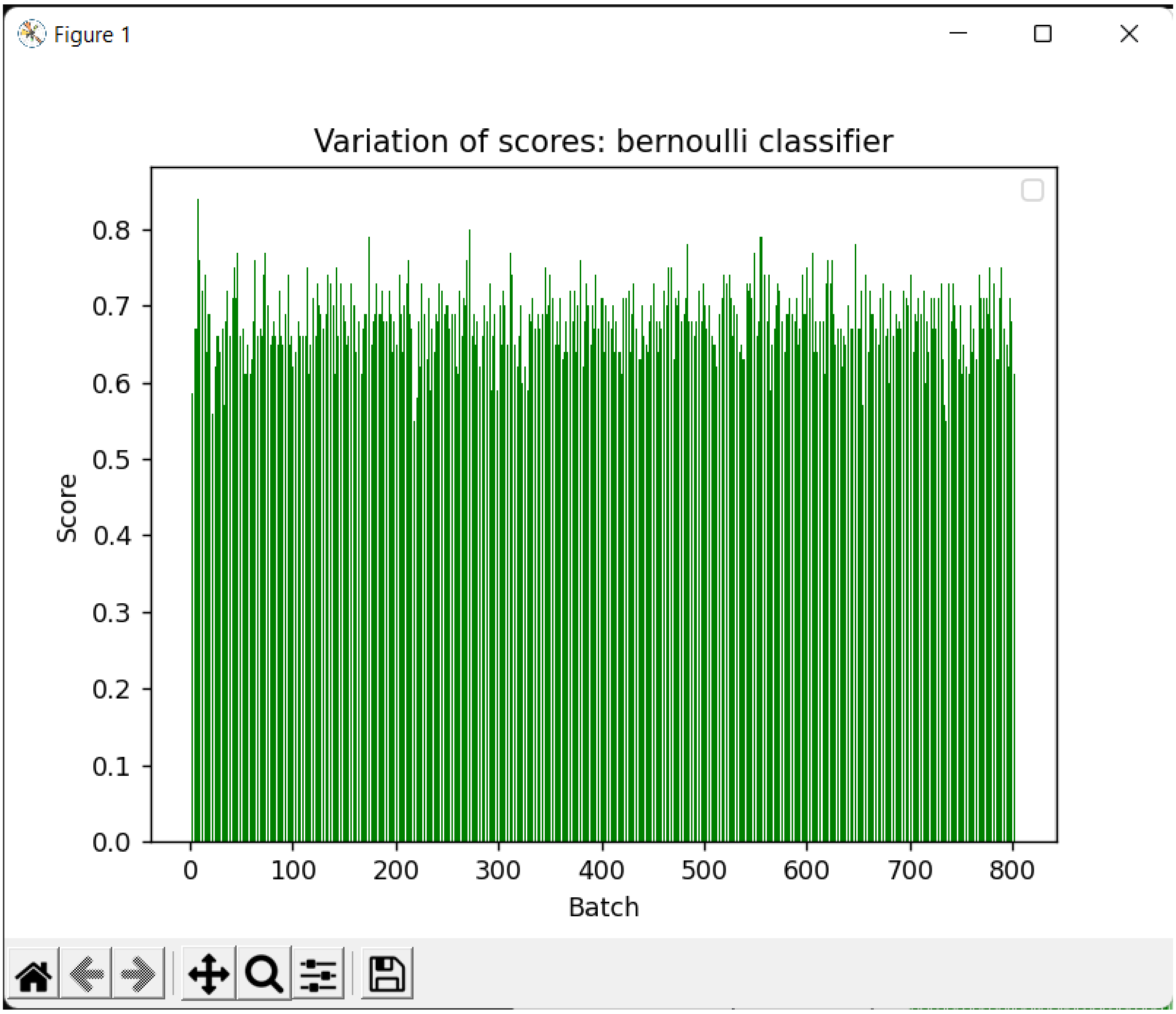


Figure 1
Variation of scores: bernoulli classifier
Score
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0
100
200
300
400
500
600
700
800
Batch

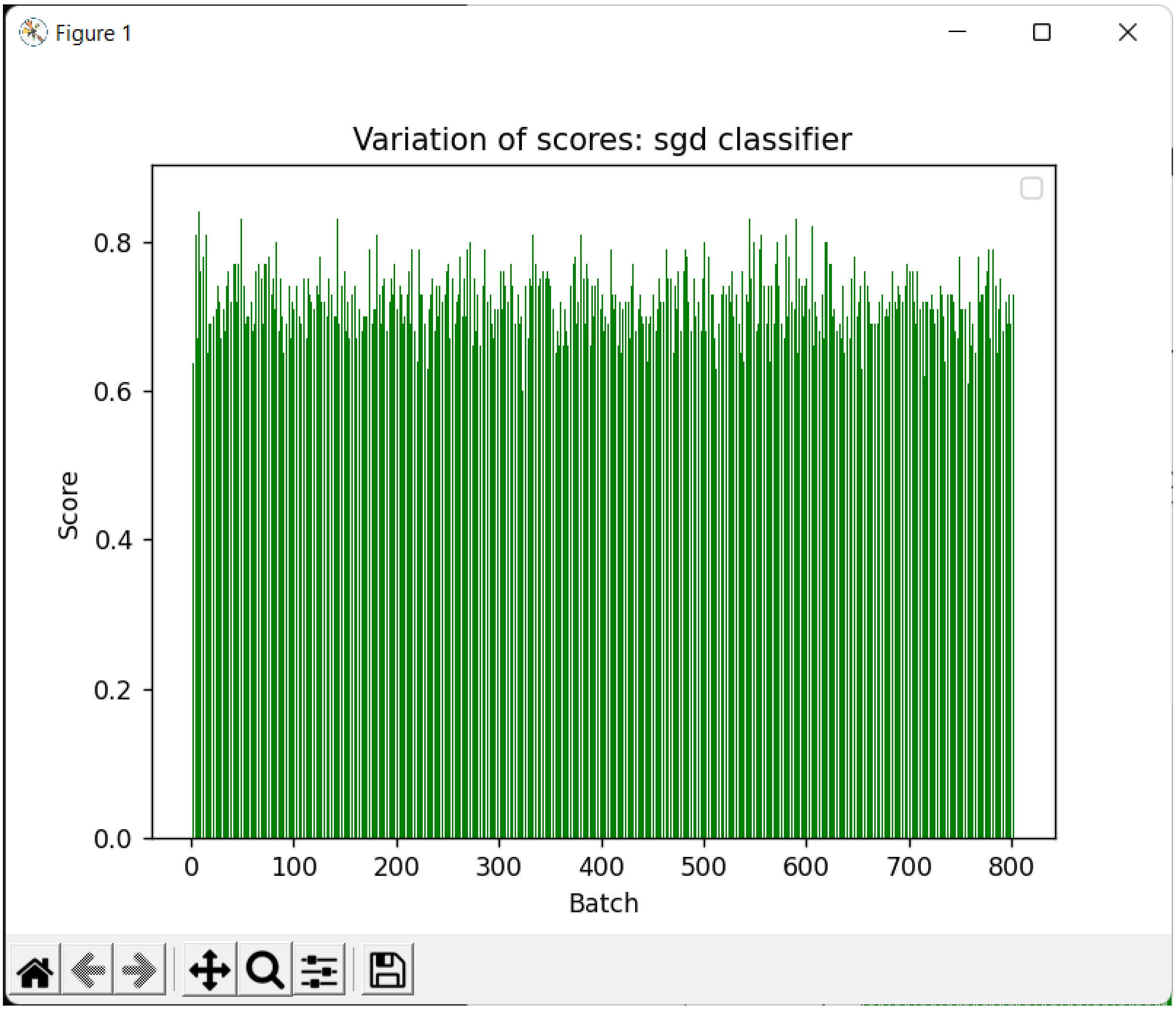


Figure 1
Variation of scores: sgd classifier
Score
0.0
0.2
0.4
0.6
0.8
0
100
200
300
400
500
600
700
800
Batch

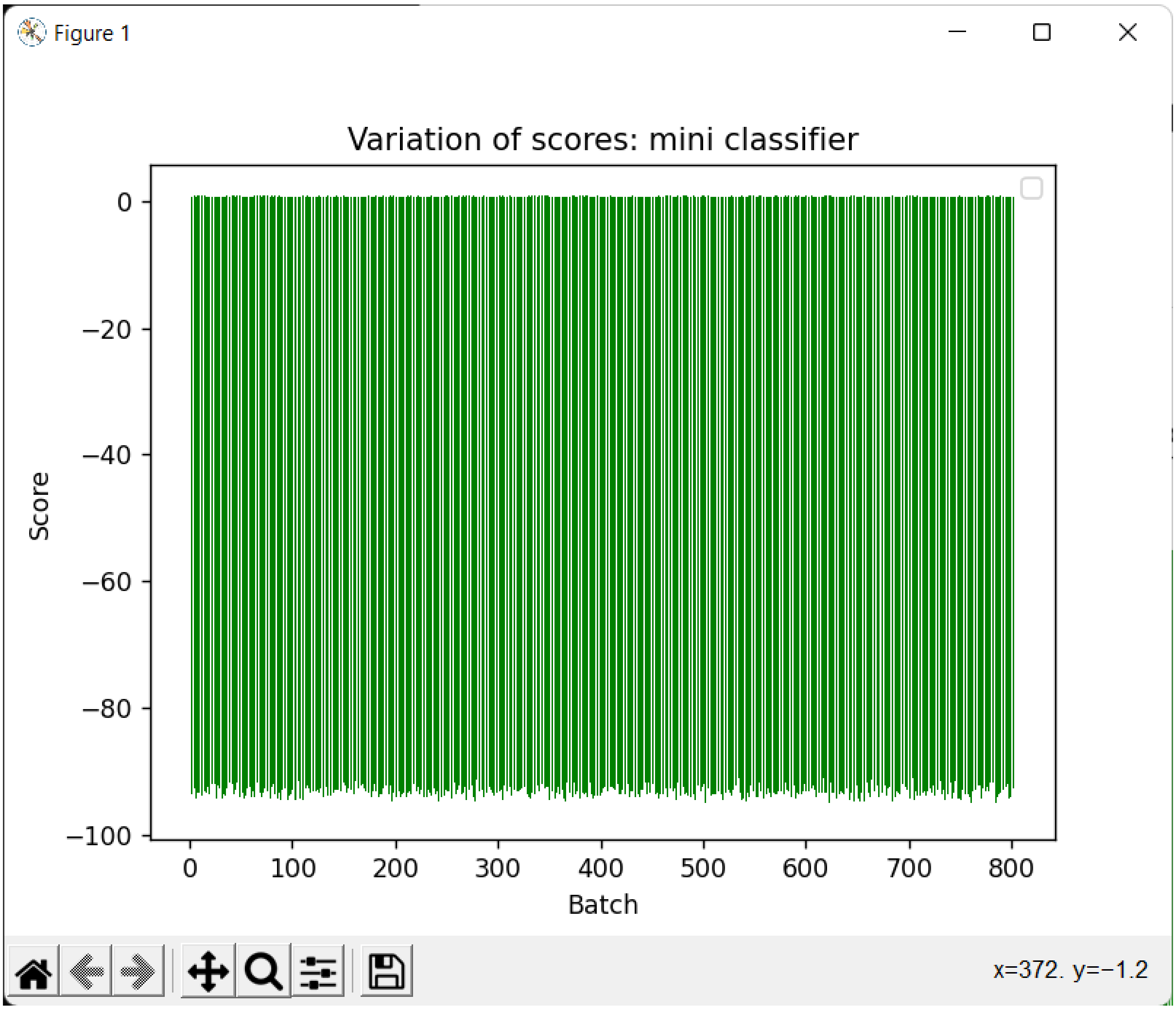


Figure 1
Variation of scores: mini classifier
Score
0
−20
−40
−60
−80
−100
0
100
200
300
400
500
600
700
800
Batch
x=372. y=−1.2