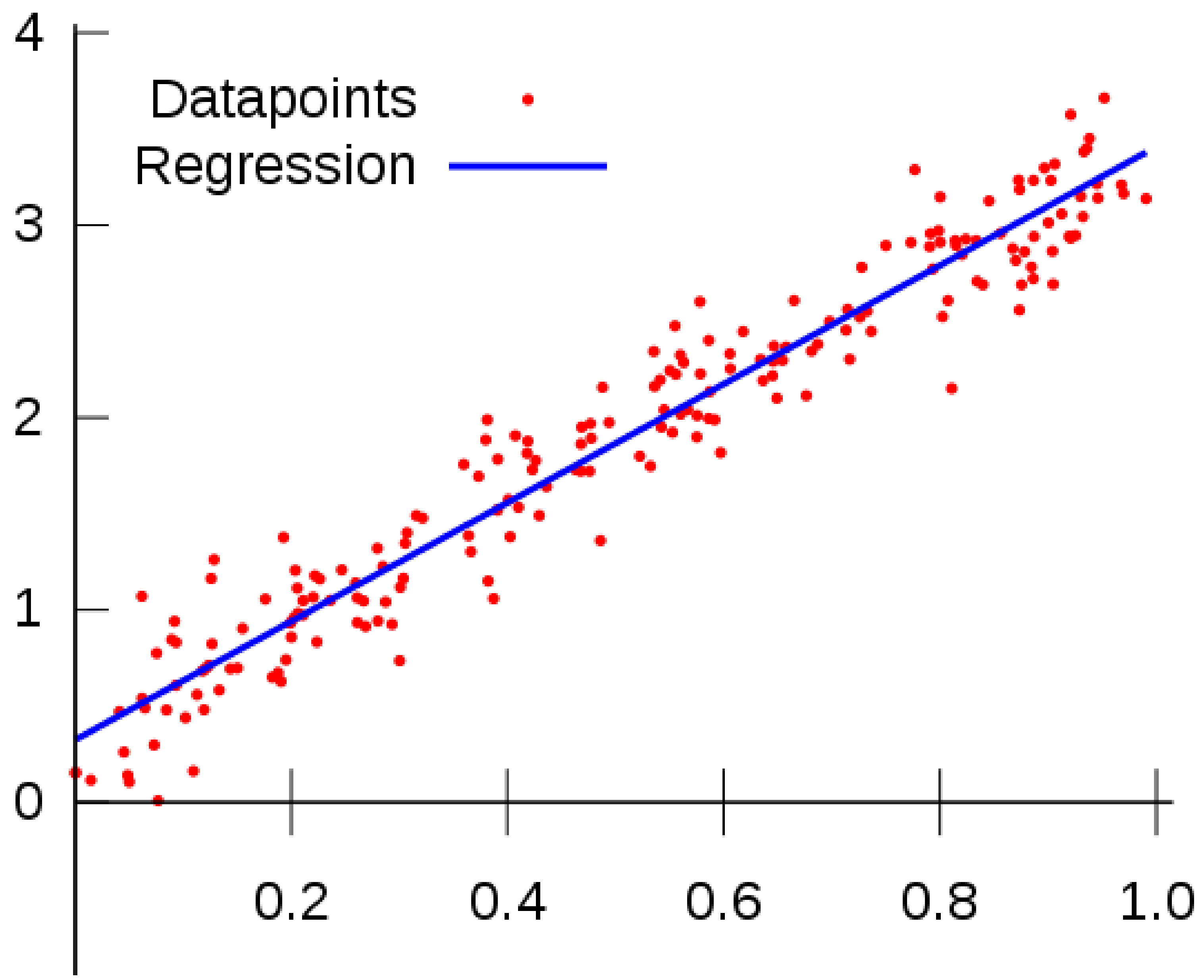
Practice 2

Regression

Problem

- ➢ Construct spark environment in your local computer and use three regression methods: Linear least squares, Lasso and Ridge regression
- ➢ Note that Ridge and Lasso regression have regularization term, so they may be able to avoid overfitting problem. But Least Square regression can't.



- Use predefined function in pyspark.**mllib.regression**

Dataset

- **Artificial dataset from pyspark tutorial**
  - This data is given from the reference link on the bottom
  - You can see whatever you want about pyspark mllib in this link.

- **Dataset format**
  - The first number is target
  - The remains are features

- **You can download the training and test dataset on i-campus**

https://github.com/apache/spark/blob/master/data/mllib/ridge-data/lpsa.data

Practice 2

1. **Use predefined classes in *pyspark.mllib.regression : LinearRegressionWIthSGD(), RidgeRegressionWithSGD(), LassoWithSGD().*** ***Please refer to hyperlinks below***

   Parameters for each method

   - LinearRegressionWIthSGD : iteration = 100, step = 0. 1
   - RidgeRegressionWithSGD : iteration = 100, step = 0.001, regParam = 0.01
   - LassoWithSGD : iteration = 100, step = 0.001, regParam = 0.01

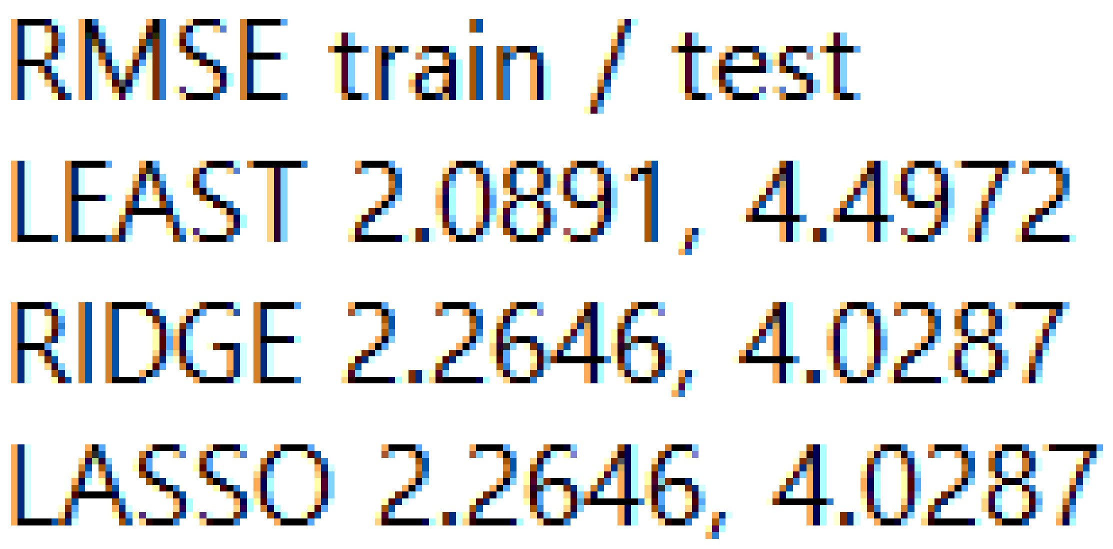
2. **After training the models, calculate the root mean square error(RMSE) using all data points for each algorithm.**
3. **Write a simple report with RMSE of each algorithm.**

https://spark.apache.org/docs/latest/mllib-linear-methods.html#linear-least-squares-lasso-and-ridge-regression

https://spark.apache.org/docs/latest/api/python/pyspark.mllib.html#pyspark.mllib.regression

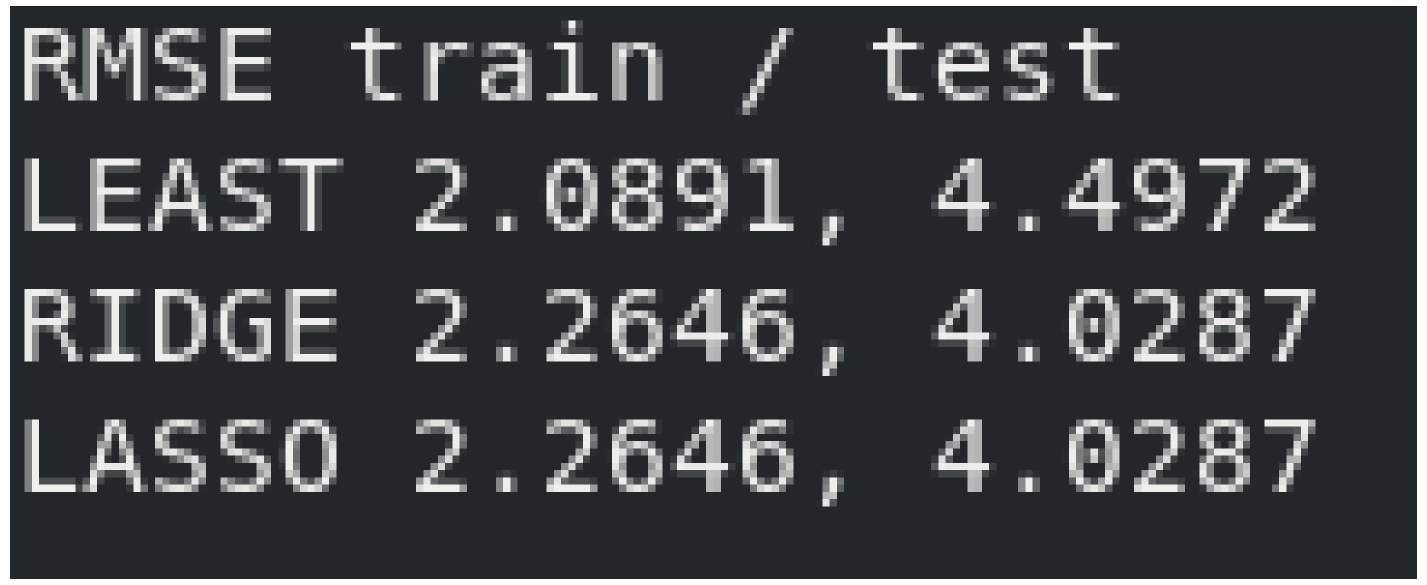
Submission

1. Submit “result.txt” file which includes Root Mean Squared Error(RMSE) of Least Square, Ridge and Lasso regression.
2. You must write the result of applying your trained model to training data points and test data points.
3. Your results.txt file must be like following.



```
RMSE train / test
LEAST 2.0891, 4.4972
RIDGE 2.2646, 4.0287
LASSO 2.2646, 4.0287
```

<Windows>



```
RMSE train / test
LEAST 2.0891, 4.4972
RIDGE 2.2646, 4.0287
LASSO 2.2646, 4.0287
```

<Linux>