Homework #2

COEN 242 Big data

Please complete the given ParallelRegression.py file and follow the instructions in the notebook hw2-notebook.html to answer the following questions. Please attach your screenshots in this report template and export the report as a PDF file after finishing it.

Hint: All your need is to follow hw2-notebook.html.

Q1 Read Data (10 pts)

Please answer the following questions:

- What are the feature dimensions of datasets large and small?

  Feature dimensions(d) large: 65, small: 9

- What are the sample numbers of datasets large and small?

  large.train.count() = 1000 small.train.count() = 100

  large.test.count() = 111 small.test.count() = 11

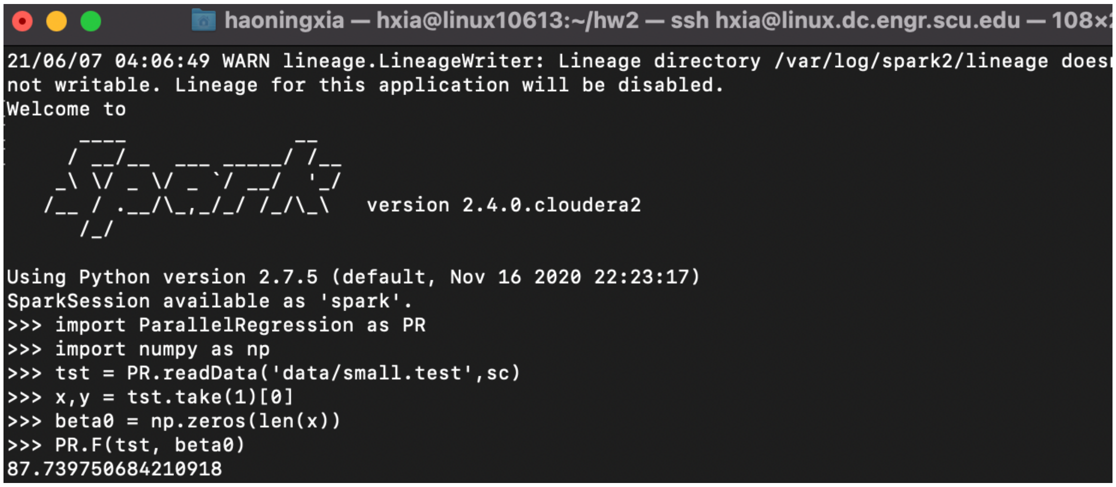
  sample numbers(n) large: 1111 small: 111

- readData will return a (key, value) paired RDD. What do the key and value represent in a regression context?

Key represent feature, value represent target value.

Q2 Implement Parallelized F (30 pts)

Please attach a screenshot of running the example given in the notebook.

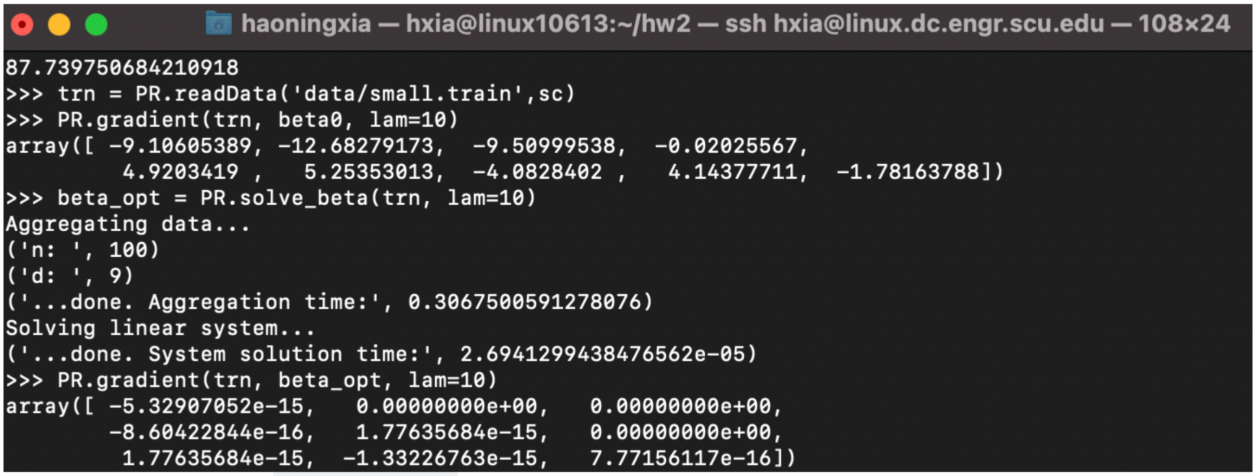

haoningxia — hxia@linux10613:~/hw2 — ssh hxia@linux.dc.engr.scu.edu — 108×

```
21/06/07 04:06:49 WARN lineage.LineageWriter: Lineage directory /var/log/spark2/lineage does
not writable. Lineage for this application will be disabled.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.4.0.cloudera2
      /_/

Using Python version 2.7.5 (default, Nov 16 2020 22:23:17)
SparkSession available as 'spark'.
>>> import ParallelRegression as PR
>>> import numpy as np
>>> tst = PR.readData('data/small.test',sc)
>>> x,y = tst.take(1)[0]
>>> beta0 = np.zeros(len(x))
>>> PR.F(tst, beta0)
87.739750684210918
```

Q3 Implement Parallelized Gradient (30 pts)

Please attach a screenshot of running the example given in the notebook and answer the questions below.

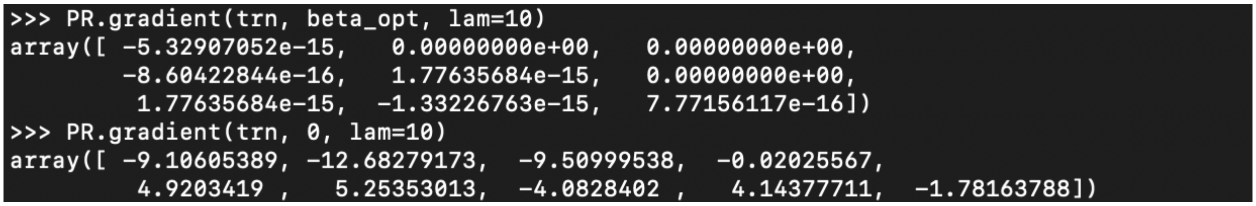

haoningxia — hxia@linux10613:~/hw2 — ssh hxia@linux.dc.engr.scu.edu — 108×24

```
87.739750684210918
>>> trn = PR.readData('data/small.train',sc)
>>> PR.gradient(trn, beta0, lam=10)
array([ -9.10605389, -12.68279173,  -9.50999538,  -0.02025567,
         4.9203419 ,   5.25353013,  -4.0828402 ,   4.14377711,  -1.78163788])
>>> beta_opt = PR.solve_beta(trn, lam=10)
Aggregating data...
('n: ', 100)
('d: ', 9)
('...done. Aggregation time:', 0.3067500591278076)
Solving linear system...
('...done. System solution time:', 2.6941299438476562e-05)
>>> PR.gradient(trn, beta_opt, lam=10)
array([ -5.32907052e-15,   0.00000000e+00,   0.00000000e+00,
        -8.60422844e-16,   1.77635684e-15,   0.00000000e+00,
         1.77635684e-15,  -1.33226763e-15,   7.77156117e-16])
```

- Why the return of `PR.gradient` is a vector?

  The gradient of a function is a vector field. It is obtained by applying the vector operator V to the scalar function f(x, y)

- Please show the comparison result between $\|\nabla\beta|_{\beta=0}\|$ and $\|\nabla\beta|_{\beta=\beta_{*ridge}}\|$

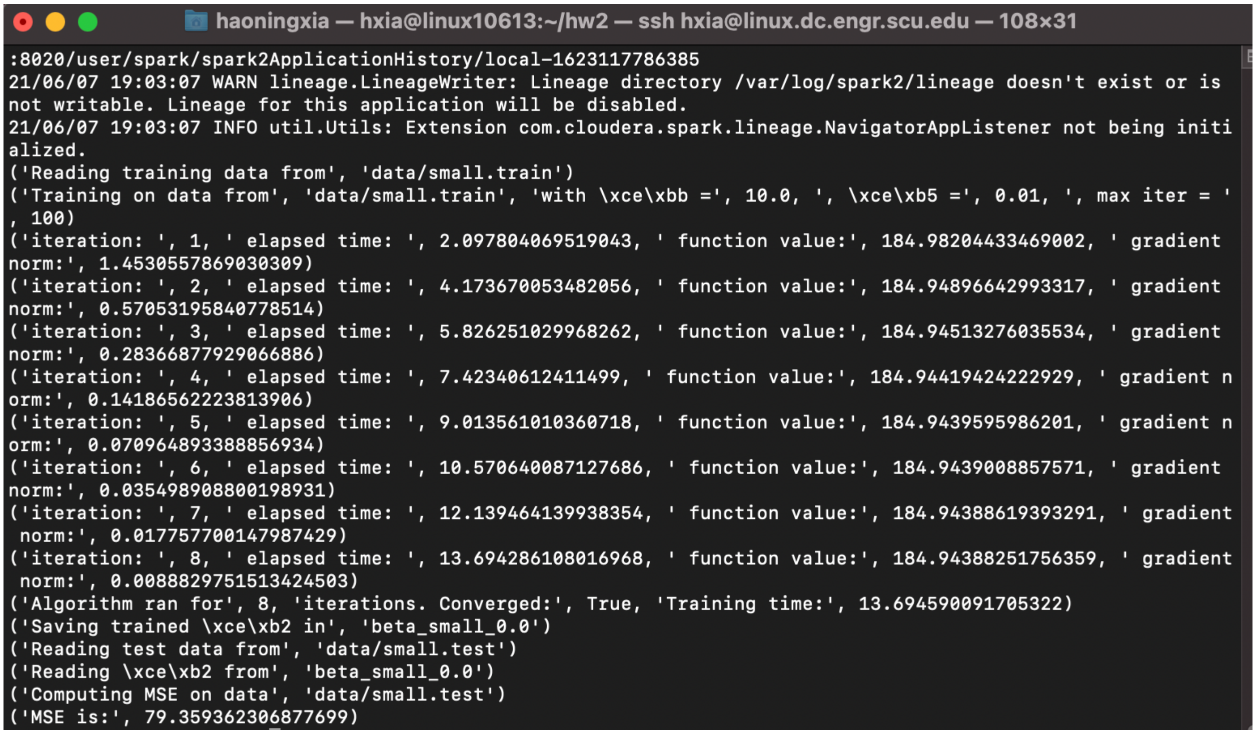


```
>>> PR.gradient(trn, beta_opt, lam=10)
array([ -5.32907052e-15,   0.00000000e+00,   0.00000000e+00,
        -8.60422844e-16,   1.77635684e-15,   0.00000000e+00,
         1.77635684e-15,  -1.33226763e-15,   7.77156117e-16])
>>> PR.gradient(trn, 0, lam=10)
array([ -9.10605389, -12.68279173,  -9.50999538,  -0.02025567,
         4.9203419 ,   5.25353013,  -4.0828402 ,   4.14377711,  -1.78163788])
```

Q4 Implement Parallelized Gradient Descent (30 pts)

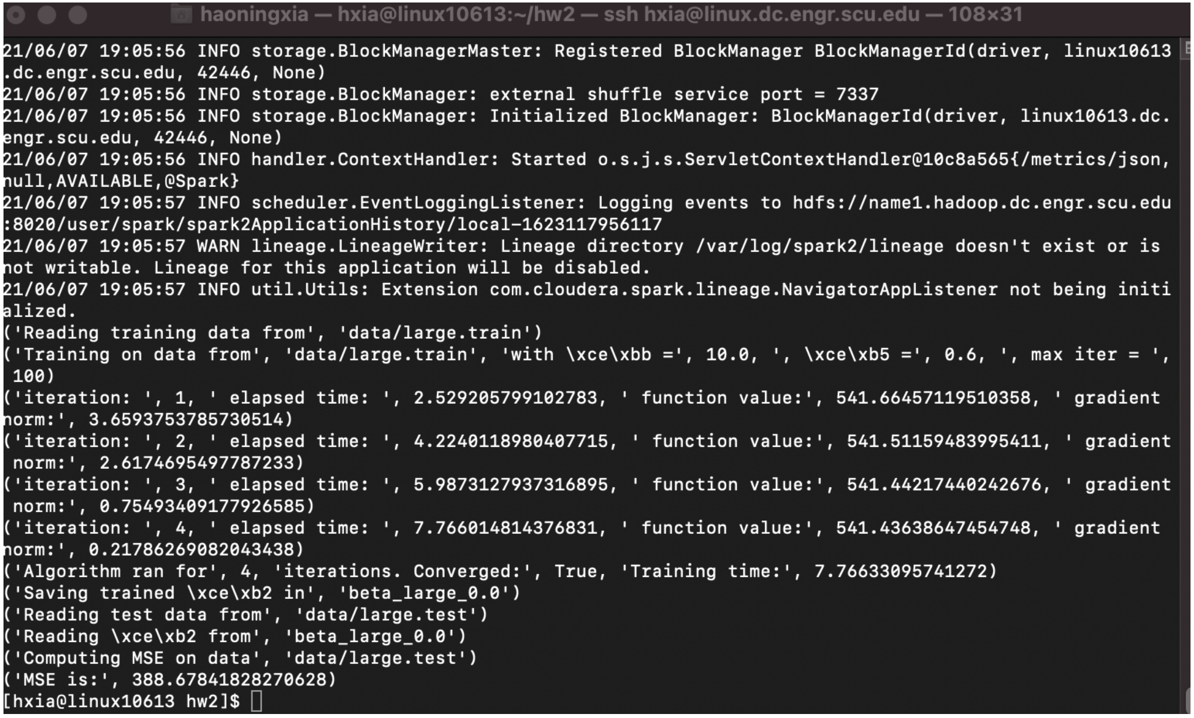
Please spark-submit your codes to train a Ridge regression model on large.train and test the model on large.test, by setting lambda=10 and epsilon=0.6. Please attach a screenshot of the running result below.

spark-submit ParallelRegression.py --train data/small.train --test data/small.test --beta beta_small_0.0 --lam 10.0 --eps 0.01


haoningxia — hxia@linux10613:~/hw2 — ssh hxia@linux.dc.engr.scu.edu — 108×31

```
:8020/user/spark/spark2ApplicationHistory/local-1623117786385
21/06/07 19:03:07 WARN lineage.LineageWriter: Lineage directory /var/log/spark2/lineage doesn't exist or is
not writable. Lineage for this application will be disabled.
21/06/07 19:03:07 INFO util.Utils: Extension com.cloudera.spark.lineage.NavigatorAppListener not being initi
alized.
('Reading training data from', 'data/small.train')
('Training on data from', 'data/small.train', 'with \xce\xbb =', 10.0, ', \xce\xb5 =', 0.01, ', max iter = '
, 100)
('iteration: ', 1, ' elapsed time: ', 2.097804069519043, ' function value:', 184.98204433469002, ' gradient
norm:', 1.4530557869030309)
('iteration: ', 2, ' elapsed time: ', 4.173670053482056, ' function value:', 184.94896642993317, ' gradient
norm:', 0.57053195840778514)
('iteration: ', 3, ' elapsed time: ', 5.826251029968262, ' function value:', 184.94513276035534, ' gradient
norm:', 0.28366877929066886)
('iteration: ', 4, ' elapsed time: ', 7.42340612411499, ' function value:', 184.94419424222929, ' gradient n
orm:', 0.14186562223813906)
('iteration: ', 5, ' elapsed time: ', 9.013561010360718, ' function value:', 184.9439595986201, ' gradient n
orm:', 0.070964893388856934)
('iteration: ', 6, ' elapsed time: ', 10.570640087127686, ' function value:', 184.9439008857571, ' gradient
norm:', 0.035498908800198931)
('iteration: ', 7, ' elapsed time: ', 12.139464139938354, ' function value:', 184.94388619393291, ' gradient
 norm:', 0.017757700147987429)
('iteration: ', 8, ' elapsed time: ', 13.694286108016968, ' function value:', 184.94388251756359, ' gradient
 norm:', 0.0088829751513424503)
('Algorithm ran for', 8, 'iterations. Converged:', True, 'Training time:', 13.694590091705322)
('Saving trained \xce\xb2 in', 'beta_small_0.0')
('Reading test data from', 'data/small.test')
('Reading \xce\xb2 from', 'beta_small_0.0')
('Computing MSE on data', 'data/small.test')
('MSE is:', 79.359362306877699)
```

```
spark-submit ParallelRegression.py --train data/large.train --test data/large.test --beta beta_large_0.0 --lam 10.0 --eps 0.6
```


haoningxia — hxia@linux10613:~/hw2 — ssh hxia@linux.dc.engr.scu.edu — 108×31

```
21/06/07 19:05:56 INFO storage.BlockManagerMaster: Registered BlockManager BlockManagerId(driver, linux10613
.dc.engr.scu.edu, 42446, None)
21/06/07 19:05:56 INFO storage.BlockManager: external shuffle service port = 7337
21/06/07 19:05:56 INFO storage.BlockManager: Initialized BlockManager: BlockManagerId(driver, linux10613.dc.
engr.scu.edu, 42446, None)
21/06/07 19:05:56 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@10c8a565{/metrics/json,
null,AVAILABLE,@Spark}
21/06/07 19:05:57 INFO scheduler.EventLoggingListener: Logging events to hdfs://name1.hadoop.dc.engr.scu.edu
:8020/user/spark/spark2ApplicationHistory/local-1623117956117
21/06/07 19:05:57 WARN lineage.LineageWriter: Lineage directory /var/log/spark2/lineage doesn't exist or is
not writable. Lineage for this application will be disabled.
21/06/07 19:05:57 INFO util.Utils: Extension com.cloudera.spark.lineage.NavigatorAppListener not being initi
alized.
('Reading training data from', 'data/large.train')
('Training on data from', 'data/large.train', 'with \xce\xbb =', 10.0, ', \xce\xb5 =', 0.6, ', max iter = ',
 100)
('iteration: ', 1, ' elapsed time: ', 2.529205799102783, ' function value:', 541.66457119510358, ' gradient
norm:', 3.6593753785730514)
('iteration: ', 2, ' elapsed time: ', 4.2240118980407715, ' function value:', 541.51159483995411, ' gradient
 norm:', 2.6174695497787233)
('iteration: ', 3, ' elapsed time: ', 5.9873127937316895, ' function value:', 541.44217440242676, ' gradient
 norm:', 0.75493409177926585)
('iteration: ', 4, ' elapsed time: ', 7.766014814376831, ' function value:', 541.43638647454748, ' gradient
norm:', 0.21786269082043438)
('Algorithm ran for', 4, 'iterations. Converged:', True, 'Training time:', 7.76633095741272)
('Saving trained \xce\xb2 in', 'beta_large_0.0')
('Reading test data from', 'data/large.test')
('Reading \xce\xb2 from', 'beta_large_0.0')
('Computing MSE on data', 'data/large.test')
('MSE is:', 388.67841828270628)
[hxia@linux10613 hw2]$ 
```