Large Scale Data Analysis Exam

Lucian Tirca (hrn947@alumni.ku.dk)

June 15, 2017

1 Deep Learning

1.1 Variance

```
Variance for redshift is 0.010498
```

1.2 Deep Network for Regression

I have tested both the GradientDescentOptimizer and the AdamOptimizer, and saw that the AdamOptimizer is superior in performance. However, plots of both are available below.

Some of the lines I have changed were:

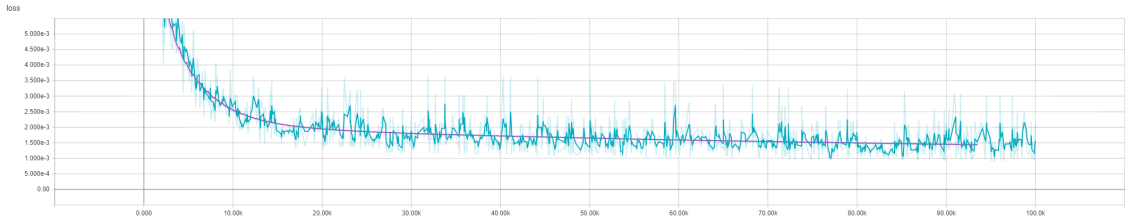
```
#change the activation to ReLU
y_1 = tf.matmul(x_data , W_1 ) + b_1
y_1 = tf.nn.relu(y_1)
```

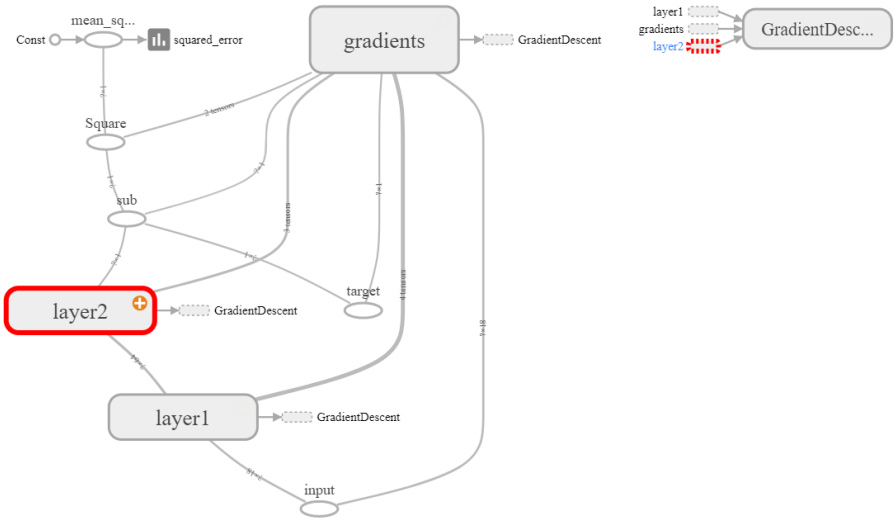
```
#change the loss function to MSE
loss = tf.reduce_mean(tf.square(model_output    y_target), name='
    ↪ mean_squared_error')
```

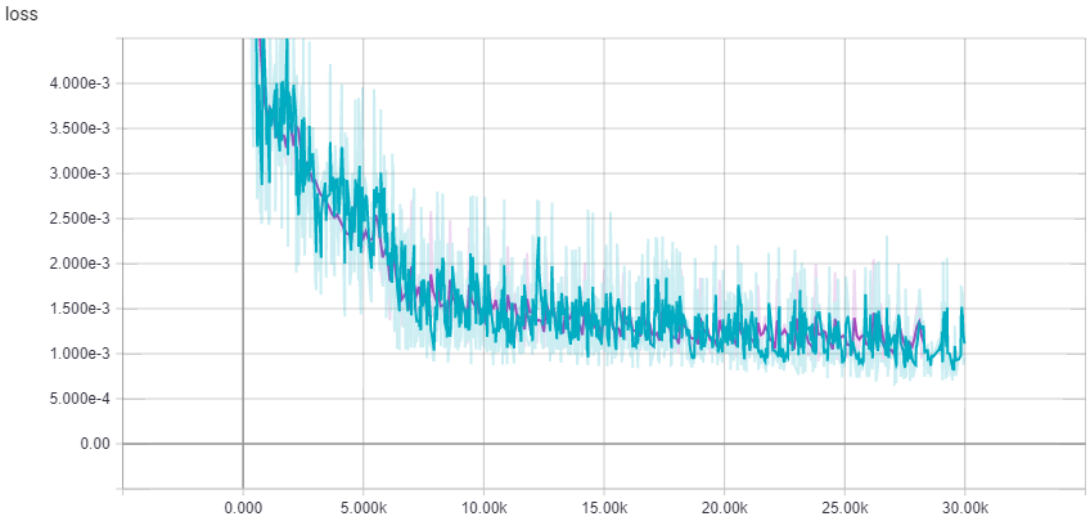
```
#change the optimizer to ADAM
# Declare optimizer
my_opt =  tf.train.AdamOptimizer(FLAGS.lr)
train_step = my_opt.minimize(loss)
```

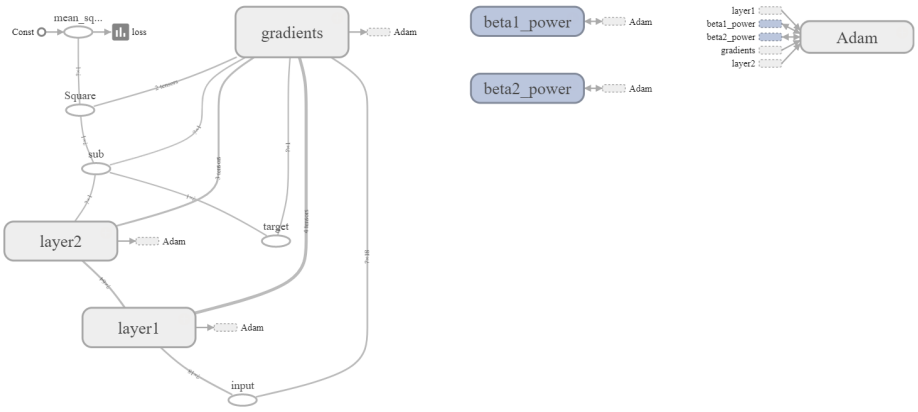
```
Iteration: 30000 / 30000
final training accuracy: 0.000976738
final test accuracy:  0.00126395
final validation accuracy:  0.0010332
INFO:tensorflow:Restoring parameters from tensor_logs/bestNetwork
best training accuracy: 0.000954459
best test accuracy:  0.00125532
best validation accuracy:  0.000915237
```

1.3 TensorBoard









2 Nearest Neighbors

2.1 Backward Selection

```
Fitting for columns [0, 1, 2, 8, 9] for validation...
CPU times: user 2.44 s, sys: 4 ms, total: 2.45 s
Wall time: 721 ms
Best MSE for Validation is 0.188903 ...
Fitting for columns [0, 1, 2, 8, 9] for test...
CPU times: user 26 s, sys: 64 ms, total: 26.1 s
Wall time: 6.87 s
Best MSE for test is 0.223736 ...
```

2.2 Computation of Nearest Neighbors

Considering the dimensionality of the dataset, the best way to select features using exact neighbors is (a) Brute-force . This is because we can exploit the fact that the number of points is small(1000), therefore the per-feature cost is $\binom{1000}{2}$ euclidean distance computations. A k-d tree would be prohibitive because it has to check 100000 dimensions for splits when fitting the model, and LSH only approximates the distance.

3 Locality Sensitive Hashing

3.1 Brute Force

Due to a omission in my code (forgetting to convert the sparse Jaccard similarity matrix to dense after computing the dot product $Q \cdot P^T$), the computation took an exceptionally long time (2.5 hours). However, after fixing it the computation was sped up. An important thing to add is that the indexes in the point set P are true indices , thus for the real indices one should add 100 to them.

```
#first run
('Bruteforce computation took ', 8975.425826072693, 'sec')
#second run after fixing Jaccard similarity matrix type
('Bruteforce computation took ', 283.36371898651123, 'sec')
('Average Jaccard similarity: ', 0.010597417134263177)
Similar value number 1 :J(0,733) = 1.000000
Similar value number 2 :J(1,173) = 0.910112
Similar value number 3 :J(3,269) = 0.952381
Similar value number 4 :J(9,140) = 1.000000
Similar value number 5 :J(9,512) = 1.000000
Similar value number 6 :J(10,765) = 0.952381
Similar value number 7 :J(11,87) = 1.000000
Similar value number 8 :J(12,67) = 1.000000
Similar value number 9 :J(19,761) = 1.000000
Similar value number 10 :J(23,199) = 1.000000
Similar value number 11 :J(27,216) = 1.000000
Similar value number 12 :J(29,138) = 1.000000
Similar value number 13 :J(33,666) = 1.000000
Similar value number 14 :J(36,224) = 0.857143
Similar value number 15 :J(39,23530) = 1.000000
Similar value number 16 :J(39,23715) = 0.889796
Similar value number 17 :J(42,694) = 1.000000
Similar value number 18 :J(43,221) = 1.000000
Similar value number 19 :J(43,248) = 1.000000
Similar value number 20 :J(48,211) = 1.000000
Similar value number 21 :J(53,697) = 1.000000
Similar value number 22 :J(59,624) = 1.000000
Similar value number 23 :J(61,211) = 1.000000
Similar value number 24 :J(69,27473) = 1.000000
Similar value number 25 :J(73,1196) = 0.894737
Similar value number 26 :J(76,216) = 1.000000
Similar value number 27 :J(79,55) = 1.000000
Similar value number 28 :J(84,316) = 0.857143
Similar value number 29 :J(85,145) = 0.977273
Similar value number 30 :J(85,8516) = 0.977273
Similar value number 31 :J(85,8888) = 0.977273
Similar value number 32 :J(85,10782) = 0.977273
Similar value number 33 :J(85,13027) = 0.977273
Similar value number 34 :J(85,13111) = 1.000000
Similar value number 35 :J(85,29155) = 0.977273
Similar value number 36 :J(85,32824) = 1.000000
Similar value number 37 :J(89,548) = 1.000000
Similar value number 38 :J(89,658) = 0.888889
Similar value number 39 :J(92,106) = 0.903846
Similar value number 40 :J(92,616) = 0.882353
Similar value number 41 :J(92,659) = 0.903846
Similar value number 42 :J(92,2302) = 0.882353
Similar value number 43 :J(92,32757) = 0.882353
```

```
Similar value number 44 :J(93,668) = 0.802817
Similar value number 45 :J(96,3) = 0.823529
Similar value number 46 :J(96,618) = 0.869565
```

3.2 LSH Framework

We want to find all pairs between P and Q such that documents with $J(x,q) \geq 0.8$ are reported with probability $\sigma_1 \geq 0.9$ and disimillar documents with $J(x,q) \leq 0.4$ are reported with probability $\sigma_2 \leq 0.01$.

Let us first assume we have columns C_1 and C_2 such that $J(C_1, C_2) = 0.8$.

Since $J(C_1, C_2) \geq 0.8$ we want (C_1, C_2) to be a candidate pair: We want them to hash to at least 1 common bucket (at least one band is identical)

$\mathbb{P}[(C_1, C_2)$ identical in one particular band$] = 0.8^r$

$\mathbb{P}[(C_1, C_2)$ not similar in all of the b bands$] = (1 - 0.8^r)^b$

Therefore, about $(1-0.8^r)^b$ of the 80%-similar columns are false negative(we will miss them).

We will find $1 - (1 - 0.8^r)^b$ pairs of truly similar documents.

Now we assume we have columns C_1 and C_2 such that $J(C_1, C_2) = 0.4$.

Since $J(C_1, C_2) \leq 0.8$ we want (C_1, C_2) to hash to no common bucket (all bands should be different)

$\mathbb{P}[(C_1, C_2)$ identical in one particular band$] = 0.4^r$

$\mathbb{P}[(C_1, C_2)$ identical in at least one of the b bands$] = 1 - (1 - 0.4^r)^b$

In other words, approximately $1 - (1 - 0.4^r)^b$ pairs of documents with similarity 40% end up becoming candidate pairs.

They are false positives since we will have to examine them (they are candidate pairs) but then it will turn out their similarity is below threshold 0.8.

We now need to solve the system of inequalities: $\begin{cases} (1 - 0.8^r)^b < 1 - \sigma_1 \\ 1 - (1 - 0.4^r)^b < \sigma_2 \end{cases} \implies$

$$\begin{cases} (1 - 0.8^r)^b < 0.1 \\ 1 - (1 - 0.4^r)^b < 0.01 \end{cases}$$

For the purpose of this implementation, I have chosen to use $r = 8$ and $b = 13$ after verifying that they solve the system using a "brute-force" approach written in Python.

3.3 Verification

```
('Generating MinHash signatures took ', 303.0851049423218, 'sec')
('Splitting into LSH buckets took ', 0.8080439567565918, 'sec')
('Candidate size: ', 1931)
('Number of true pair in the candidate set: ', 43)
('Number of true pairs in bruteforce: ', 46)
('False negatives: ', 0.9782608695652174)
('False positives: ', 0.9777317452097359)
('Probability that a far away pair is in the candidate set: ',
    ↪ 0.0008969085829825304)
```

We can see that these results correspond to the theoretical limits. The difference in terms of execution speed is huge (partly because of the problem with my initial computation), but it can also be proven from the number of operations made . The brute-force approach compares the

100 documents in the query set with another 39761 points, each with 28102 words. Generating MinHash signatures is a lot faster due to its $O(N_{documents} \cdot k)$ runtime(for every document we compute k hashes). The LSH part is almost instant, since splitting into bands greatly reduces the dimensionality of dataset. We can see that we have less than 10% error rate (43 in the candidate set vs 46 in the true set), and the probability that a far away pair is in the candidate set is much less than 0.01.

3.4 Optimization

In order to optimize, we have to impose the additional constraint $k = r \cdot b$ is minimum where k is the size of the MinHash signature(number of universal hash functions generated). This ensures that the space we need to store the LSH tables/Signature matrix is minimized, and the runtime is smaller. We can formulate this as a constrained optimization problem:

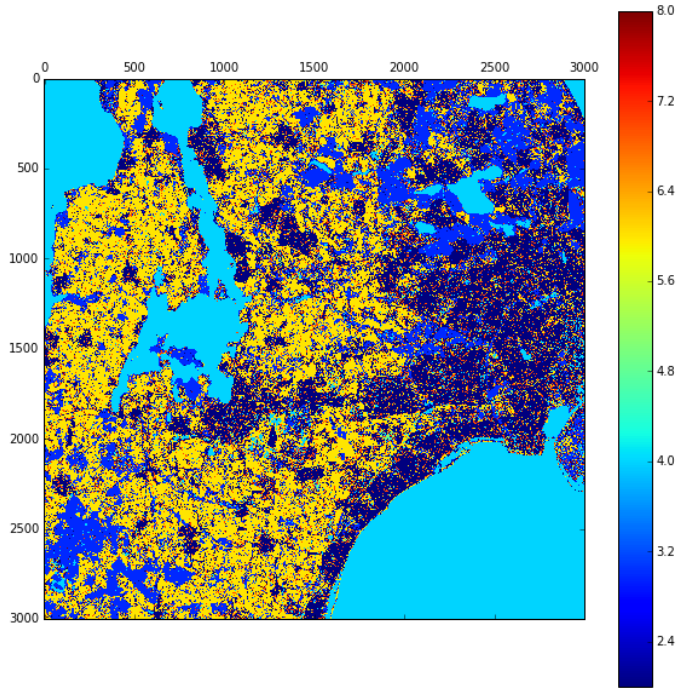
$$\begin{cases} \text{argmin } f(r,b) = r \cdot b \text{ s.t.} \\ (1-0.8^r)^b - 0.1 < 0 \\ 1-(1-0.4^r)^b - 0.01 < 0 \end{cases}$$

This would allow us to use numerical optimization techniques such as the KKT conditions to find a minimum.This can be further parametrized in terms of the Jaccard similarity desired for the similar and dissimilar document thresholds on order to provide a more generic solution. However, that is out of the scope of this course.

4 Tree Ensembles for Huge Data

4.1 Big Trees and Random Subsets

```
Frequency of appearance: [      0       0 2212058 1353200 2094475
    ↪ 22530 2690463  536110   91164]
Most frequent : 6
```



4.2 Runtime

As stated on SKlearn's documentation(http://scikit-learn.org/stable/modules/tree.html), the construction time of a DecisionTree is $O(N_{samples}N_{features}\log(N_{samples}))$ and the query time is $O(\log(N_{samples}))$. We train 100 of them , therefore the time would be: $O(100 \cdot 10000 \cdot N_{features}\log(10000) = O(13.3 \cdot 10^6 N_{features})$

4.3 Detecting Rare Instances

The decision tree is a perfect choice for detecting rare instances in this specific case. Because all the other labels are 0, the gini index with the other parameters used in this model will attempt to find the best splits, and will create them where the space can be partitioned into 2 different labels. However, the problem here comes from the way we sample.

We sample 10000 out of $10^9 = 1/10^5$. We do this 100 times, so the class frequency in the sample is $1/10^3$. Therefore, there should be at least 1000 class 1 samples so that there would be on average 1 per tree. We only have 19, so the answer is we cannot detect the anomaly.

5 Hadoop

5.1 Airline Statistics

a) Shortest Flight Distance

The minimal distance for airline ID 20436 is 332.

```
19393    137.0
```

```
19690    84.0
19790    94.0
19805    83.0
19930    31.0
19977    108.0
20304    30.0
20366    49.0
20409    68.0
20416    105.0
20436    332.0
21171    236.0
```

```
17/06/10 22:06:44 WARN streaming.StreamJob:  file option is
    ↪ deprecated, please use generic option  files instead.
packageJobJar: [mapper.py, reducer.py, /tmp/hadoop
    ↪ unjar6830261437728606957/] [] /tmp/streamjob8038751277777585943
    ↪ .jar tmpDir=null
17/06/10 22:06:45 INFO client.RMProxy: Connecting to ResourceManager
    ↪ at localhost/127.0.0.1:8050
17/06/10 22:06:45 INFO client.RMProxy: Connecting to ResourceManager
    ↪ at localhost/127.0.0.1:8050
17/06/10 22:06:45 INFO mapred.FileInputFormat: Total input paths to
    ↪ process : 6
17/06/10 22:06:45 INFO mapreduce.JobSubmitter: number of splits:6
17/06/10 22:06:45 INFO mapreduce.JobSubmitter: Submitting tokens for
    ↪ job: job_1497110946339_0010
17/06/10 22:06:45 INFO impl.YarnClientImpl: Submitted application
    ↪ application_1497110946339_0010
17/06/10 22:06:45 INFO mapreduce.Job: The url to track the job: http
    ↪ ://lsdabox:8088/proxy/application_1497110946339_0010/
17/06/10 22:06:45 INFO mapreduce.Job: Running job:
    ↪ job_1497110946339_0010
17/06/10 22:06:51 INFO mapreduce.Job: Job job_1497110946339_0010
    ↪ running in uber mode : false
17/06/10 22:06:51 INFO mapreduce.Job:  map 0% reduce 0%
17/06/10 22:07:04 INFO mapreduce.Job:  map 78% reduce 0%
17/06/10 22:07:06 INFO mapreduce.Job:  map 100% reduce 0%
17/06/10 22:07:15 INFO mapreduce.Job:  map 100% reduce 100%
17/06/10 22:07:15 INFO mapreduce.Job: Job job_1497110946339_0010
    ↪ completed successfully
17/06/10 22:07:16 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=39698443
                FILE: Number of bytes written=80249909
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=211806569
                HDFS: Number of bytes written=137
```

```
26              HDFS: Number of read operations=21
27              HDFS: Number of large read operations=0
28              HDFS: Number of write operations=2
29      Job Counters
30              Launched map tasks=6
31              Launched reduce tasks=1
32              Data local map tasks=6
33              Total time spent by all maps in occupied slots (ms)
                    ↪ =76319
34              Total time spent by all reduces in occupied slots (ms
                    ↪ )=6568
35              Total time spent by all map tasks (ms)=76319
36              Total time spent by all reduce tasks (ms)=6568
37              Total vcore milliseconds taken by all map tasks=76319
38              Total vcore milliseconds taken by all reduce tasks
                    ↪ =6568
39              Total megabyte milliseconds taken by all map tasks
                    ↪ =78150656
40              Total megabyte milliseconds taken by all reduce tasks
                    ↪ =6725632
41      Map Reduce Framework
42              Map input records=2777469
43              Map output records=2777463
44              Map output bytes=34143511
45              Map output materialized bytes=39698473
46              Input split bytes=672
47              Combine input records=0
48              Combine output records=0
49              Reduce input groups=12
50              Reduce shuffle bytes=39698473
51              Reduce input records=2777463
52              Reduce output records=12
53              Spilled Records=5554926
54              Shuffled Maps =6
55              Failed Shuffles=0
56              Merged Map outputs=6
57              GC time elapsed (ms)=1336
58              CPU time spent (ms)=17750
59              Physical memory (bytes) snapshot=1659801600
60              Virtual memory (bytes) snapshot=13414629376
61              Total committed heap usage (bytes)=1251475456
62      Shuffle Errors
63              BAD_ID=0
64              CONNECTION=0
65              IO_ERROR=0
66              WRONG_LENGTH=0
67              WRONG_MAP=0
68              WRONG_REDUCE=0
69      File Input Format Counters
```

```
                Bytes Read=211805897
        File Output Format Counters
                Bytes Written=137
17/06/10 22:07:16 INFO streaming.StreamJob: Output directory: out/1
    ↪ a_shortest_flight_distance
```

b) Late Arrival Counts

The percentage of flight delays for airline ID 20436 is 0.295.

```
19393   (643370, 234624, 0.3646797332794504)
19690   (37887, 12158, 0.3209016285269354)
19790   (454372, 117601, 0.25882096608065636)
19805   (459324, 165059, 0.3593520042497235)
19930   (88162, 26422, 0.29969828270683513)
19977   (258574, 69094, 0.26721170728688887)
20304   (299512, 97456, 0.32538262239910254)
20366   (251522, 77122, 0.30662128958898227)
20409   (139891, 55437, 0.3962871092493441)
20416   (67525, 31417, 0.4652647167715661)
20436   (43954, 12978, 0.29526322974018293)
21171   (33370, 14084, 0.42205573868744384)
```

```
17/06/11 14:49:18 WARN streaming.StreamJob:  file option is
    ↪ deprecated, please use generic option  files instead.
packageJobJar: [mapper.py, reducer.py, /tmp/hadoop
    ↪ unjar908628992761082978/] [] /tmp/streamjob3468781984105564986.
    ↪ jar tmpDir=null
17/06/11 14:49:18 INFO client.RMProxy: Connecting to ResourceManager
    ↪ at localhost/127.0.0.1:8050
17/06/11 14:49:19 INFO client.RMProxy: Connecting to ResourceManager
    ↪ at localhost/127.0.0.1:8050
17/06/11 14:49:19 INFO mapred.FileInputFormat: Total input paths to
    ↪ process : 6
17/06/11 14:49:19 INFO mapreduce.JobSubmitter: number of splits:6
17/06/11 14:49:19 INFO mapreduce.JobSubmitter: Submitting tokens for
    ↪ job: job_1497183961540_0002
17/06/11 14:49:19 INFO impl.YarnClientImpl: Submitted application
    ↪ application_1497183961540_0002
17/06/11 14:49:19 INFO mapreduce.Job: The url to track the job: http
    ↪ ://lsdabox:8088/proxy/application_1497183961540_0002/
17/06/11 14:49:19 INFO mapreduce.Job: Running job:
    ↪ job_1497183961540_0002
17/06/11 14:49:24 INFO mapreduce.Job: Job job_1497183961540_0002
    ↪ running in uber mode : false
17/06/11 14:49:24 INFO mapreduce.Job:  map 0% reduce 0%
17/06/11 14:49:38 INFO mapreduce.Job:  map 49% reduce 0%
17/06/11 14:49:40 INFO mapreduce.Job:  map 100% reduce 0%
17/06/11 14:49:49 INFO mapreduce.Job:  map 100% reduce 100%
```

```
17/06/11 14:49:50 INFO mapreduce.Job: Job job_1497183961540_0002 completed successfully
17/06/11 14:49:50 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=27774636
                FILE: Number of bytes written=56402197
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=211806569
                HDFS: Number of bytes written=509
                HDFS: Number of read operations=21
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=6
                Launched reduce tasks=1
                Data local map tasks=6
                Total time spent by all maps in occupied slots (ms)=80362
                Total time spent by all reduces in occupied slots (ms)=6571
                Total time spent by all map tasks (ms)=80362
                Total time spent by all reduce tasks (ms)=6571
                Total vcore milliseconds taken by all map tasks=80362
                Total vcore milliseconds taken by all reduce tasks=6571
                Total megabyte milliseconds taken by all map tasks=82290688
                Total megabyte milliseconds taken by all reduce tasks=6728704
        Map Reduce Framework
                Map input records=2777469
                Map output records=2777463
                Map output bytes=22219704
                Map output materialized bytes=27774666
                Input split bytes=672
                Combine input records=0
                Combine output records=0
                Reduce input groups=12
                Reduce shuffle bytes=27774666
                Reduce input records=2777463
                Reduce output records=12
                Spilled Records=5554926
                Shuffled Maps =6
                Failed Shuffles=0
                Merged Map outputs=6
                GC time elapsed (ms)=1066
                CPU time spent (ms)=20010
```

```
                Physical memory (bytes) snapshot=1559564288
                Virtual memory (bytes) snapshot=13405638656
                Total committed heap usage (bytes)=1257766912
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=211805897
        File Output Format Counters
                Bytes Written=509
17/06/11 14:49:50 INFO streaming.StreamJob: Output directory: out/1
    ↪ b_late_arrival_counts
```

c) Mean and Standard deviation for Arrival Delay

The mean and standard deviation for airline ID 20436 are 1.717 and 46.18. The mean and Standard deviation have been computed using Wilbur's method : the mapper streams the data, the combiner uses a streaming version of the algorithm, while the reducers use the parallel execution version of the algorithm.

```
19393    3.17728375622    30.0668179309
19690    0.890991631985   23.1002169299
19790    1.40829752039    39.7652955697
19805    4.40986312904    44.274274348
19930    3.09850048934    26.7074300525
19977    0.189141271091   42.5316254377
20304    3.83094496718    43.7087638153
20366    3.12154796263    44.0636245694
20409    8.10612536937    44.3026657876
20416    11.9151427341    41.9729684622
20436    1.71770491562    46.1810372781
21171    7.13658958085    37.2946361144
```

```
17/06/11 19:21:58 WARN streaming.StreamJob:  file option is
    ↪ deprecated, please use generic option  files instead.
packageJobJar: [mapper.py, combiner.py, reducer.py, /tmp/hadoop
    ↪ unjar6775979134498773199/] [] /tmp/streamjob4651826839231852246
    ↪ .jar tmpDir=null
17/06/11 19:21:59 INFO client.RMProxy: Connecting to ResourceManager
    ↪ at localhost/127.0.0.1:8050
17/06/11 19:21:59 INFO client.RMProxy: Connecting to ResourceManager
    ↪ at localhost/127.0.0.1:8050
17/06/11 19:22:00 INFO mapred.FileInputFormat: Total input paths to
    ↪ process : 6
17/06/11 19:22:01 INFO mapreduce.JobSubmitter: number of splits:6
```

```
17/06/11 19:22:01 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1497183961540_0003
17/06/11 19:22:01 INFO impl.YarnClientImpl: Submitted application application_1497183961540_0003
17/06/11 19:22:01 INFO mapreduce.Job: The url to track the job: http://lsdabox:8088/proxy/application_1497183961540_0003/
17/06/11 19:22:01 INFO mapreduce.Job: Running job: job_1497183961540_0003
17/06/11 19:22:06 INFO mapreduce.Job: Job job_1497183961540_0003 running in uber mode : false
17/06/11 19:22:06 INFO mapreduce.Job:  map 0% reduce 0%
17/06/11 19:22:22 INFO mapreduce.Job:  map 37% reduce 0%
17/06/11 19:22:23 INFO mapreduce.Job:  map 54% reduce 0%
17/06/11 19:22:24 INFO mapreduce.Job:  map 100% reduce 0%
17/06/11 19:22:30 INFO mapreduce.Job:  map 100% reduce 100%
17/06/11 19:22:31 INFO mapreduce.Job: Job job_1497183961540_0003 completed successfully
17/06/11 19:22:31 INFO mapreduce.Job: Counters: 50
        File System Counters
                FILE: Number of bytes read=2528
                FILE: Number of bytes written=862580
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=211806569
                HDFS: Number of bytes written=413
                HDFS: Number of read operations=21
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Killed map tasks=1
                Launched map tasks=6
                Launched reduce tasks=1
                Data local map tasks=6
                Total time spent by all maps in occupied slots (ms)=93472
                Total time spent by all reduces in occupied slots (ms)=2460
                Total time spent by all map tasks (ms)=93472
                Total time spent by all reduce tasks (ms)=2460
                Total vcore milliseconds taken by all map tasks=93472
                Total vcore milliseconds taken by all reduce tasks=2460
                Total megabyte milliseconds taken by all map tasks=95715328
                Total megabyte milliseconds taken by all reduce tasks=2519040
        Map Reduce Framework
                Map input records=2777469
```

```
                Map output records=2777463
                Map output bytes=31262785
                Map output materialized bytes=2558
                Input split bytes=672
                Combine input records=2777463
                Combine output records=72
                Reduce input groups=12
                Reduce shuffle bytes=2558
                Reduce input records=72
                Reduce output records=12
                Spilled Records=144
                Shuffled Maps =6
                Failed Shuffles=0
                Merged Map outputs=6
                GC time elapsed (ms)=1294
                CPU time spent (ms)=17960
                Physical memory (bytes) snapshot=1592520704
                Virtual memory (bytes) snapshot=13418299392
                Total committed heap usage (bytes)=1264058368
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=211805897
        File Output Format Counters
                Bytes Written=413
17/06/11 19:22:31 INFO streaming.StreamJob: Output directory: out/1
    ↪ c_mean_std_arrival_delay
```

d) Top-10 Arrival Delays

The 10 most delayed flights for airline ID 20436 are 1285, 1298, 291, 1124, 720, 1135, 249, 1248, 733, 756. The mappers extract the relevant data, the combiners aggregate it into dictionaries and send it to the reducers.

```
19393   [('275', 7067.0), ('40', 7039.0), ('902', 6893.0), ('46',
    ↪ 6751.0), ('43', 6594.0), ('32', 6281.0), ('751', 6280.0),
    ↪ ('42', 6237.0), ('47', 6064.0), ('2678', 5999.0)]
19690   [('466', 2197.0), ('5', 1617.0), ('15', 1520.0), ('3',
    ↪ 1493.0), ('7', 1486.0), ('30', 1343.0), ('11', 1259.0), ('279',
    ↪  1168.0), ('43', 1050.0), ('19', 1028.0)]
19790   [('2780', 6272.0), ('2774', 4907.0), ('2779', 4708.0),
    ↪ ('2777', 4609.0), ('2776', 4178.0), ('2758', 4021.0), ('2781',
    ↪ 3960.0), ('2778', 3814.0), ('2687', 3770.0), ('2770', 3734.0)]
19805   [('238', 8792.0), ('2185', 8561.0), ('2285', 7879.0),
```

```
    ↪ ('2319', 7270.0), ('2447', 7180.0), ('1037', 7084.0), ('189',
    ↪ 6752.0), ('2167', 6729.0), ('1261', 6716.0), ('342', 6265.0)]
19930   [('337', 8235.0), ('306', 5982.0), ('304', 4285.0), ('381',
    ↪ 2997.0), ('307', 2686.0), ('331', 2572.0), ('382', 2475.0),
    ↪ ('223', 2314.0), ('386', 1977.0), ('318', 1867.0)]
19977   [('263', 5630.0), ('692', 5603.0), ('1469', 5083.0), ('322',
    ↪ 4746.0), ('385', 4739.0), ('2014', 4728.0), ('262', 4622.0),
    ↪ ('509', 4586.0), ('1431', 4490.0), ('300', 4392.0)]
20304   [('5223', 7518.0), ('7404', 5776.0), ('2922', 5474.0),
    ↪ ('5055', 5066.0), ('5723', 5025.0), ('5970', 5000.0), ('5454',
    ↪ 4954.0), ('2929', 4751.0), ('5056', 4465.0), ('4628', 4430.0)]
20366   [('2867', 6810.0), ('2865', 5897.0), ('5221', 5447.0),
    ↪ ('5107', 4856.0), ('2851', 4538.0), ('5526', 4353.0), ('5496',
    ↪ 4140.0), ('5348', 3993.0), ('2817', 3901.0), ('5116', 3848.0)]
20409   [('1262', 8795.0), ('499', 8591.0), ('698', 8157.0), ('432',
    ↪ 8088.0), ('2689', 7285.0), ('672', 7182.0), ('1371', 6971.0),
    ↪ ('1116', 6838.0), ('428', 6597.0), ('1198', 6480.0)]
20416   [('711', 12849.0), ('251', 10950.0), ('619', 10397.0),
    ↪ ('906', 10192.0), ('719', 9504.0), ('866', 8762.0), ('474',
    ↪ 8062.0), ('630', 7949.0), ('600', 7051.0), ('805', 6924.0)]
20436   [('1285', 3385.0), ('1298', 3082.0), ('291', 2917.0),
    ↪ ('1124', 2827.0), ('720', 2827.0), ('1135', 2823.0), ('249',
    ↪ 2712.0), ('1248', 2675.0), ('733', 2507.0), ('756', 2357.0)]
21171   [('330', 5208.0), ('941', 4971.0), ('1178', 4712.0), ('942',
    ↪ 4444.0), ('1935', 4208.0), ('938', 3933.0), ('927', 3923.0),
    ↪ ('948', 3772.0), ('945', 3670.0), ('593', 3623.0)]
```

```
17/06/12 06:10:53 WARN streaming.StreamJob:  file option is
    ↪ deprecated, please use generic option  files instead.
packageJarJar: [mapper.py, combiner.py, reducer.py, /tmp/hadoop
    ↪ unjar1435737619979480049/] [] /tmp/streamjob1769067731645710421
    ↪ .jar tmpDir=null
17/06/12 06:10:53 INFO client.RMProxy: Connecting to ResourceManager
    ↪ at localhost/127.0.0.1:8050
17/06/12 06:10:54 INFO client.RMProxy: Connecting to ResourceManager
    ↪ at localhost/127.0.0.1:8050
17/06/12 06:10:54 INFO mapred.FileInputFormat: Total input paths to
    ↪ process : 6
17/06/12 06:10:54 INFO mapreduce.JobSubmitter: number of splits:6
17/06/12 06:10:54 INFO mapreduce.JobSubmitter: Submitting tokens for
    ↪ job: job_1497239888790_0003
17/06/12 06:10:54 INFO impl.YarnClientImpl: Submitted application
    ↪ application_1497239888790_0003
17/06/12 06:10:54 INFO mapreduce.Job: The url to track the job: http
    ↪ ://lsdabox:8088/proxy/application_1497239888790_0003/
17/06/12 06:10:54 INFO mapreduce.Job: Running job:
    ↪ job_1497239888790_0003
17/06/12 06:10:59 INFO mapreduce.Job: Job job_1497239888790_0003
    ↪ running in uber mode : false
```

```
17/06/12 06:10:59 INFO mapreduce.Job:  map 0% reduce 0%
17/06/12 06:11:13 INFO mapreduce.Job:  map 48% reduce 0%
17/06/12 06:11:14 INFO mapreduce.Job:  map 59% reduce 0%
17/06/12 06:11:16 INFO mapreduce.Job:  map 67% reduce 0%
17/06/12 06:11:27 INFO mapreduce.Job:  map 83% reduce 0%
17/06/12 06:11:29 INFO mapreduce.Job:  map 100% reduce 0%
17/06/12 06:11:37 INFO mapreduce.Job:  map 100% reduce 100%
17/06/12 06:11:37 INFO mapreduce.Job: Job job_1497239888790_0003
    ↪ completed successfully
17/06/12 06:11:38 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=1271279
                FILE: Number of bytes written=3400019
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=211806569
                HDFS: Number of bytes written=2164
                HDFS: Number of read operations=21
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=6
                Launched reduce tasks=1
                Data local map tasks=6
                Total time spent by all maps in occupied slots (ms)
                    ↪ =159837
                Total time spent by all reduces in occupied slots (ms
                    ↪ )=6333
                Total time spent by all map tasks (ms)=159837
                Total time spent by all reduce tasks (ms)=6333
                Total vcore milliseconds taken by all map tasks
                    ↪ =159837
                Total vcore milliseconds taken by all reduce tasks
                    ↪ =6333
                Total megabyte milliseconds taken by all map tasks
                    ↪ =163673088
                Total megabyte milliseconds taken by all reduce tasks
                    ↪ =6484992
        Map Reduce Framework
                Map input records=2777469
                Map output records=2777463
                Map output bytes=44138581
                Map output materialized bytes=1271309
                Input split bytes=672
                Combine input records=2777463
                Combine output records=72
                Reduce input groups=12
                Reduce shuffle bytes=1271309
```

```
                Reduce input records=72
                Reduce output records=12
                Spilled Records=144
                Shuffled Maps =6
                Failed Shuffles=0
                Merged Map outputs=6
                GC time elapsed (ms)=1642
                CPU time spent (ms)=66850
                Physical memory (bytes) snapshot=1610190848
                Virtual memory (bytes) snapshot=13405368320
                Total committed heap usage (bytes)=1270874112
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=211805897
        File Output Format Counters
                Bytes Written=2164
17/06/12 06:11:38 INFO streaming.StreamJob: Output directory: out/1
    ↪ d_top10_arrival_delays
```

5.2 Landsat Statistics

The mapper fills up chunks of the data into a larger numpy array. Once the array is filled, they are used to generate predictions all at the same time, therefore speeding up computation considerably.

```
2.0     2055357
3.0     1354525
4.0     2187821
5.0     41202
6.0     2407063
7.0     771499
8.0     182533
```

```
17/06/14 20:48:21 WARN streaming.StreamJob:  file option is
    ↪ deprecated, please use generic option  files instead.
packageJobJar: [mapper.py, combiner.py, reducer.py, /tmp/hadoop
    ↪ unjar6829331547760204245/] [] /tmp/streamjob1376860636054957049
    ↪ .jar tmpDir=null
17/06/14 20:48:21 INFO client.RMProxy: Connecting to ResourceManager
    ↪ at localhost/127.0.0.1:8050
17/06/14 20:48:22 INFO client.RMProxy: Connecting to ResourceManager
    ↪ at localhost/127.0.0.1:8050
```

```
17/06/14 20:48:22 INFO mapred.FileInputFormat: Total input paths to process : 1
17/06/14 20:48:22 INFO net.NetworkTopology: Adding a new node: /default rack/127.0.0.1:50010
17/06/14 20:48:22 INFO mapreduce.JobSubmitter: number of splits:3
17/06/14 20:48:22 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1497462483237_0004
17/06/14 20:48:22 INFO impl.YarnClientImpl: Submitted application application_1497462483237_0004
17/06/14 20:48:22 INFO mapreduce.Job: The url to track the job: http://lsdabox:8088/proxy/application_1497462483237_0004/
17/06/14 20:48:22 INFO mapreduce.Job: Running job: job_1497462483237_0004
17/06/14 20:48:27 INFO mapreduce.Job: Job job_1497462483237_0004 running in uber mode : false
17/06/14 20:48:27 INFO mapreduce.Job:  map 0% reduce 0%
17/06/14 20:48:37 INFO mapreduce.Job:  map 6% reduce 0%
17/06/14 20:48:39 INFO mapreduce.Job:  map 10% reduce 0%
17/06/14 20:48:40 INFO mapreduce.Job:  map 15% reduce 0%
17/06/14 20:48:43 INFO mapreduce.Job:  map 17% reduce 0%
17/06/14 20:48:44 INFO mapreduce.Job:  map 22% reduce 0%
17/06/14 20:48:46 INFO mapreduce.Job:  map 25% reduce 0%
17/06/14 20:48:47 INFO mapreduce.Job:  map 30% reduce 0%
17/06/14 20:48:49 INFO mapreduce.Job:  map 38% reduce 0%
17/06/14 20:48:52 INFO mapreduce.Job:  map 44% reduce 0%
17/06/14 20:48:53 INFO mapreduce.Job:  map 46% reduce 0%
17/06/14 20:48:55 INFO mapreduce.Job:  map 55% reduce 0%
17/06/14 20:48:58 INFO mapreduce.Job:  map 63% reduce 0%
17/06/14 20:49:01 INFO mapreduce.Job:  map 67% reduce 0%
17/06/14 20:49:02 INFO mapreduce.Job:  map 78% reduce 0%
17/06/14 20:49:04 INFO mapreduce.Job:  map 89% reduce 0%
17/06/14 20:49:05 INFO mapreduce.Job:  map 100% reduce 0%
17/06/14 20:49:08 INFO mapreduce.Job:  map 100% reduce 100%
17/06/14 20:49:09 INFO mapreduce.Job: Job job_1497462483237_0004 completed successfully
17/06/14 20:49:09 INFO mapreduce.Job: Counters: 50
        File System Counters
                FILE: Number of bytes read=274
                FILE: Number of bytes written=490345
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=403699996
                HDFS: Number of bytes written=80
                HDFS: Number of read operations=12
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=3
```

```
46                Launched reduce tasks=1
47                Data local map tasks=2
48                Rack local map tasks=1
49                Total time spent by all maps in occupied slots (ms)
                      ↪ =102693
50                Total time spent by all reduces in occupied slots (ms
                      ↪ )=2689
51                Total time spent by all map tasks (ms)=102693
52                Total time spent by all reduce tasks (ms)=2689
53                Total vcore milliseconds taken by all map tasks
                      ↪ =102693
54                Total vcore milliseconds taken by all reduce tasks
                      ↪ =2689
55                Total megabyte milliseconds taken by all map tasks
                      ↪ =105157632
56                Total megabyte milliseconds taken by all reduce tasks
                      ↪ =2753536
57        Map Reduce Framework
58                Map input records=9000000
59                Map output records=9000000
60                Map output bytes=54000000
61                Map output materialized bytes=286
62                Input split bytes=339
63                Combine input records=9000000
64                Combine output records=21
65                Reduce input groups=7
66                Reduce shuffle bytes=286
67                Reduce input records=21
68                Reduce output records=7
69                Spilled Records=42
70                Shuffled Maps =3
71                Failed Shuffles=0
72                Merged Map outputs=3
73                GC time elapsed (ms)=454
74                CPU time spent (ms)=89150
75                Physical memory (bytes) snapshot=952541184
76                Virtual memory (bytes) snapshot=7662616576
77                Total committed heap usage (bytes)=680525824
78        Shuffle Errors
79                BAD_ID=0
80                CONNECTION=0
81                IO_ERROR=0
82                WRONG_LENGTH=0
83                WRONG_MAP=0
84                WRONG_REDUCE=0
85        File Input Format Counters
86                Bytes Read=403699657
87        File Output Format Counters
88                Bytes Written=80
```

```
89 17/06/14 20:49:09 INFO streaming.StreamJob: Output directory: out/2
   ↪ _landsat_stats
```