question | Like

417 views

Lab 4, 5d doesn't crash despite incorrect results

In lab 4, 5d I get the following results:

```
stepSize = 1.0, regParam = 1e-06: logloss = 0.470
        stepSize = 1.0, regParam = 1e-03: logloss = 0.470
        stepSize = 10.0, regParam = 1e-06: logloss = 0.448
        stepSize = 10.0, regParam = 1e-03: logloss = 0.450
Hashed Features Validation Logloss:
        Baseline = 0.528  (0.527602636661)
        LogReg = 0.448 (0.448112786335)
```

There are no errors until 5d.
LogReg is incorrect, since it should be 0.4481683608 according to the test.

In Lab 4, 5b I use the following approach for parseHashPoint:

1. I split the point using ','
2. the label is split[0]
3. the initial features are split[1:]
4. I create a list of enumerated features for each feature in features
5. I create a feature dictionary by calling hashFunction with the parameters numBuckets, enumerated features and False
6. I create a SparseVector with the parameters numBuckets, sorted(feature dictionary keys) and unsorted feature dictionary values, since I'm assuming they are all 1.0
7. The return is a LabeledPoint consisting of the label and the SparseVector

I sorted the indices by sorting the keys in the feature dictionary
I left the feature dictionary values unsorted since I'm assuming they are 1.0

When initially calling parseHashPoint to generate hashTrainData, hashValidationData etc. i use
1. rawTrainData, rawValidationData etc.
2. map(... parseHashPoint(with lambda, and numBucketsCTR as parameters))

Any suggestions?

lab4

Updated 7 days ago by Anonymous

the students' answer, *where students collectively construct a single answer*

Your assumption "I left the feature dictionary values unsorted since I'm assuming they are 1.0" is incorrect. hashed values will be incremented by the number of times that key appears after hash function. check the definition of function that return Hash dictionary

Updated 7 days ago by Anonymous and Sameer S

followup discussions *for lingering questions and comments*

Resolved Unresolved

Anonymous 7 days ago
How do I sort the feature dictionary keys and feature dictionary values, so that they are in sync?

Anonymous 7 days ago Should I use OrderedDict?

Anonymous 7 days ago I am not sure if this will amount to honor code violation but you need to first sort keys and then "reconstruct" the list using the sorted keys and iterating through their values..

Actually, there is an easier way to do it but that requires the latest version of python and the one we are using in virtual box is slightly dated..

Anonymous 7 days ago OrderedDict might not work for the Python version we have in the virtual box.. I didn't try but you could..

Anonymous 7 days ago Thanks. In modern organizations it's important to share and collaborate with colleagues. That's how evolution happens.

I was able to pass the test by adding the __init__ method to the SparseVector and by using dictionary as a parameter in SparseVector.
I changed line 6 and 7 from:

6. I create a SparseVector with the parameters numBuckets, sorted(feature dictionary keys) and unsorted feature dictionary values, since I'm assuming they are all 1.0
7. The return is a LabeledPoint consisting of the label and the SparseVector

To:

6. I create a SparseVector with the parameters numBuckets and the feature dictionary
7. I use the sparseVector __init__ metod
8. Return is same as in previous line 7

Ameet Talwalkar 6 days ago This is NOT a violation of the honor code. In fact, it's a great example of requesting feedback without explicitly posting code. Thank you both.

M Jansen 6 days ago Thanks, I needed this. Not sure about the init method. I created a SparseVector with the dictionary as argument, which works fine. Before I had tried to sort the dictionary by converting it to two lists. Something must have gone wrong there. All the tests passed up to 5d... with a logLoss ever so close, but not close enough.

But now it passes.
This post really helped!

Zhengchun Liu 5 days ago
this post helps me, get it finally ...

thanks a lot..

willhenry 4 days ago I am confused. I have always been able to just pass a dictionary as an input to SparseVector, but it sound like you had to modify the init method of SparseVector to do this? Why?

Arjuna Scagnetto 4 days ago they didn't change any line of code, except for the line you must change ;)

this post helped me thanks

willhenry 4 days ago what.. well strange. I have always been able to pass dictionaries. It worked for me, and I passed all tests on submission. I don't know why I could do this and others had to change the SparseVector code. Oh well.

Zhengchun Liu 4 days ago Hi,

This sentence helped me to pass "Your assumption "I left the feature dictionary values unsorted since I'm assuming they are 1.0" is incorrect. **hashed values will be incremented by the number of times that key appears after hash function. check the definition of function that return Hash dictionary**"

Anonymous 1 day ago Hi, so these are the results that I am getting. I have not applied sorting or changed any of the SparseVector. Is the result that I am getting correct or not, I am not really sure. (this is passing the test, but looking at the discussion above wanted to confirm)

```
stepSize = 1.0, regParam = 1e-06: logloss = 0.470
        stepSize = 1.0, regParam = 1e-03: logloss = 0.470
        stepSize = 10.0, regParam = 1e-06: logloss = 0.448
        stepSize = 10.0, regParam = 1e-03: logloss = 0.450
Hashed Features Validation Logloss:
        Baseline = 0.528(0.5276026)
        LogReg = 0.448(0.4481684)
```

Francis Kim 1 day ago Thanks M Jansen! passing dictionary as argument worked!! Didn't know you could do this, 8-)

Jagrut Sharma 8 hours ago Beautiful. Subtle point - finally got 5d and 5e to be exactly what they should be.

Resolved Unresolved

Peter Szabo 6 days ago
Hey!

I'm still sucked.

```
<code>def parseHashPoint(point, numBuckets):
</code>l=split, get 0th
f_temp=split, get a list of th elese elements
f=create a list of tuples from f_temp by (i, f_temp[i])
hash_res=hashed f with numBuckets
feat = create a sparse vector from hash_res
return a labeledPoint from l and feat
```

Output:

```
[LabeledPoint(0.0, (32768,[1305,2883,3807,4814,4866,4913,6952,7117,9985,10316,11512,11722,12365,13893,14735,15816,16198,17761,19274,21604,22256,22563,22785,24855,25202,25533,25721,26487,26656,27668,28211,29152,29402,29873,30039,31484,32493,32708],[1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0]))]
```

Every value in my return is 1.0. What did I wrong? How can I make it pass 5d? Plz help, I'm a python starter.
My submission ID is :

```
2087768-bcada4e70111a42722855c2c1c94e328:ip-172-31-25-214
```

Peter Szabo 6 days ago For hashTrainData.take(1) i got:

```
[LabeledPoint(0.0, (32768,[1305,2883,3807,4814,4866,4913,6952,7117,9985,10316,11512,11722,12365,13893,14735,15816,16198,17761,19274,21604,22256,22563,22785,24855,25202,25533,25721,26487,26656,27668,28211,29152,29402,29873,30039,31484,32493,32708],[1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0]))]
```

Resolved Unresolved

Hamster 6 days ago
Can't figure out what is wrong, please help:

```
[u'0,1,1,5,0,1382,4,15,2,181,1,2,,2,68fd1e64,80e26c9b,fb936136,7b4723c4,25c83c98,7e0ccccf,de7995b8,1f89b562,a73ee510,a8cd5504,b2cb9c98,37c9c164,2824a5f6,1adce6ef,8ba8b39a,891b62e7,e5ba7672,f54016b9,21ddcdc9,b1252a9d,07b5194c,,3a171ecb,c5c50484,e8b83407,9727dd16']
0
-----
[(0, u'1'), (1, u'1'), (2, u'5'), (3, u'0'), (4, u'1382'), (5, u'4'), (6, u'15'), (7, u'2'), (8, u'181'), (9, u'1'), (10, u'2'), (11, u''), (12, u'2'), (13, u'68fd1e64'), (14, u'80e26c9b'), (15, u'fb936136'), (16, u'7b4723c4'), (17, u'25c83c98'), (18, u'7e0ccccf'), (19, u'de7995b8'), (20, u'1f89b562'), (21, u'a73ee510'), (22, u'a8cd5504'), (23, u'b2cb9c98'), (24, u'37c9c164'), (25, u'2824a5f6'), (26, u'1adce6ef'), (27, u'8ba8b39a'), (28, u'891b62e7'), (29, u'e5ba7672'), (30, u'f54016b9'), (31, u'21ddcdc9'), (32, u'b1252a9d'), (33, u'07b5194c'), (34, u''), (35, u'3a171ecb'), (36, u'c5c50484'), (37, u'e8b83407'), (38, u'9727dd16')]
-----
```

```
---------------------------------------------------------------------------
Py4JJavaError                             Traceback (most recent call last)
<ipython-input-151-e47ab883dff7> in <module>()
     31 hashTestData.cache()
     32 
---> 33 print hashTrainData.take(1)

/usr/local/bin/spark-1.3.1-bin-hadoop2.6/python/pyspark/rdd.py in take(self, num)
   1222 
   1223             p = range(partsScanned, min(partsScanned + numPartsToTry, totalParts))
-> 1224             res = self.context.runJob(self, takeUpToNumLeft, p, True)
   1225 
   1226             items += res

/usr/local/bin/spark-1.3.1-bin-hadoop2.6/python/pyspark/context.py in runJob(self, rdd, partitionFunc, partitions, allowLocal)
    840         mappedRDD = rdd.mapPartitions(partitionFunc)
    841         port = self._jvm.PythonRDD.runJob(self._jsc.sc(), mappedRDD._jrdd, javaPartitions,
--> 842                                           allowLocal)
    843         return list(_load_from_socket(port, mappedRDD._jrdd_deserializer))
    844 

/usr/local/bin/spark-1.3.1-bin-hadoop2.6/python/lib/py4j-0.8.2.1-src.zip/py4j/java_gateway.py in __call__(self, *args)
    536         answer = self.gateway_client.send_command(command)
    537         return_value = get_return_value(answer, self.gateway_client,
--> 538                 self.target_id, self.name)
    539 
    540         for temp_arg in temp_args:

/usr/local/bin/spark-1.3.1-bin-hadoop2.6/python/lib/py4j-0.8.2.1-src.zip/py4j/protocol.py in get_return_value(answer, gateway_client, target_id, name)
    298                 raise Py4JJavaError(
    299                     'An error occurred while calling {0}{1}{2}.\n'.
--> 300                     format(target_id, '.', name), value)
    301             else:
    302                 raise Py4JError(

Py4JJavaError: An error occurred while calling z:org.apache.spark.api.python.PythonRDD.runJob.
: org.apache.spark.SparkException: Job aborted due to stage failure: Task 0 in stage 308.0 failed 1 times, most recent failure: Lost task 0.0 in stage 308.0 (TID 444, localhost): org.apache.spark.api.python.PythonException: Traceback (most
```

```
recent call last):
  File "/usr/local/bin/spark-1.3.1-bin-hadoop2.6/python/pyspark/worker.py", line 101, in main
    process()
  File "/usr/local/bin/spark-1.3.1-bin-hadoop2.6/python/pyspark/worker.py", line 96, in process
    serializer.dump_stream(func(split_index, iterator), outfile)
  File "/usr/local/bin/spark-1.3.1-bin-hadoop2.6/python/pyspark/serializers.py", line 236, in dump_stream
    vs = list(itertools.islice(iterator, batch))
  File "<ipython-input-151-e47ab883dff7>", line 26, in <lambda>
  File "<ipython-input-151-e47ab883dff7>", line 14, in parseHashPoint
  File "<ipython-input-135-eebb81d621f6>", line 23, in hashFunction
TypeError: 'int' object is not iterable

        at org.apache.spark.api.python.PythonRDD$$anon$1.read(PythonRDD.scala:135)
        at org.apache.spark.api.python.PythonRDD$$anon$1.<init>(PythonRDD.scala:176)
        at org.apache.spark.api.python.PythonRDD.compute(PythonRDD.scala:94)
        at org.apache.spark.rdd.RDD.computeOrReadCheckpoint(RDD.scala:277)
        at org.apache.spark.CacheManager.getOrCompute(CacheManager.scala:70)
        at org.apache.spark.rdd.RDD.iterator(RDD.scala:242)
        at org.apache.spark.api.python.PythonRDD$WriterThread$$anonfun$run$1.apply(PythonRDD.scala:243)
        at org.apache.spark.util.Utils$.logUncaughtExceptions(Utils.scala:1618)
        at org.apache.spark.api.python.PythonRDD$WriterThread.run(PythonRDD.scala:205)

Driver stacktrace:
        at org.apache.spark.scheduler.DAGScheduler.org$apache$spark$scheduler$DAGScheduler$$failJobAndIndependentStages(DA
GScheduler.scala:1204)
        at org.apache.spark.scheduler.DAGScheduler$$anonfun$abortStage$1.apply(DAGScheduler.scala:1193)
        at org.apache.spark.scheduler.DAGScheduler$$anonfun$abortStage$1.apply(DAGScheduler.scala:1192)
        at scala.collection.mutable.ResizableArray$class.foreach(ResizableArray.scala:59)
        at scala.collection.mutable.ArrayBuffer.foreach(ArrayBuffer.scala:47)
        at org.apache.spark.scheduler.DAGScheduler.abortStage(DAGScheduler.scala:1192)
        at org.apache.spark.scheduler.DAGScheduler$$anonfun$handleTaskSetFailed$1.apply(DAGScheduler.scala:693)
        at org.apache.spark.scheduler.DAGScheduler$$anonfun$handleTaskSetFailed$1.apply(DAGScheduler.scala:693)
        at scala.Option.foreach(Option.scala:236)
        at org.apache.spark.scheduler.DAGScheduler.handleTaskSetFailed(DAGScheduler.scala:693)
        at org.apache.spark.scheduler.DAGSchedulerEventProcessLoop.onReceive(DAGScheduler.scala:1393)
        at org.apache.spark.scheduler.DAGSchedulerEventProcessLoop.onReceive(DAGScheduler.scala:1354)
        at org.apache.spark.util.EventLoop$$anon$1.run(EventLoop.scala:48)


</span></span></span></span></span></span></span></span></span></span>
```

Hamster 5 days ago Never mind, mixed the order of parameters, _o_0_/