question | Like 658 views

Lab 4: 5d - error

Lab code has this problem:

```
---------------------------------------------------------------------------
Py4JJavaError                             Traceback (most recent call last)
<ipython-input-115-ea1370e4bfcf> in <module>()
      5     for regParam in regParams:
      6         model = (LogisticRegressionWithSGD
----> 7                  .train(hashTrainData, numIters, stepSize, regParam=regParam, regType=regType, intercept=includeIntercept)
)
      8         logLossVa = evaluateResults(model, hashValidationData)
      9         print ('\tstepSize = {0:.1f}, regParam = {1:.0e}: logloss = {2:.3f}'

/usr/local/bin/spark-1.3.1-bin-hadoop2.6/python/pyspark/mllib/classification.py in train(cls, data, iterations, step,
miniBatchFraction, initialWeights, regParam, regType, intercept)
    162                                   bool(intercept))
    163
--> 164         return _regression_train_wrapper(train, LogisticRegressionModel, data, initialWeights)
    165
    166

/usr/local/bin/spark-1.3.1-bin-hadoop2.6/python/pyspark/mllib/regression.py in _regression_train_wrapper(train_func, modelClass,
data, initial_weights)
    138     if initial_weights is None:
    139         initial_weights = [0.0] * len(data.first().features)
--> 140     weights, intercept = train_func(data, _convert_to_vector(initial_weights))
    141     return modelClass(weights, intercept)
    142

/usr/local/bin/spark-1.3.1-bin-hadoop2.6/python/pyspark/mllib/classification.py in train(rdd, i)
    160             return callMLlibFunc("trainLogisticRegressionModelWithSGD", rdd, int(iterations),
    161                                  float(step), float(miniBatchFraction), i, float(regParam), regType,
--> 162                                  bool(intercept))
    163
    164         return _regression_train_wrapper(train, LogisticRegressionModel, data, initialWeights)

/usr/local/bin/spark-1.3.1-bin-hadoop2.6/python/pyspark/mllib/common.py in callMLlibFunc(name, *args)
    118     sc = SparkContext._active_spark_context
    119     api = getattr(sc._jvm.PythonMLLibAPI(), name)
--> 120     return callJavaFunc(sc, api, *args)
    121
    122

/usr/local/bin/spark-1.3.1-bin-hadoop2.6/python/pyspark/mllib/common.py in callJavaFunc(sc, func, *args)
    111     """ Call Java Function """
    112     args = [_py2java(sc, a) for a in args]
--> 113     return _java2py(sc, func(*args))
    114
    115

/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 o1689.trainLogisticRegressionModelWithSGD.
: org.apache.spark.SparkException: Job aborted due to stage failure: Task 0 in stage 231.0 failed 1 times, most recent failure:
Lost task 0.0 in stage 231.0 (TID 341, localhost): java.lang.ArrayIndexOutOfBoundsException: 1305
        at org.apache.spark.mllib.linalg.BLAS$.dot(BLAS.scala:136)
        at org.apache.spark.mllib.linalg.BLAS$.dot(BLAS.scala:106)
        at org.apache.spark.mllib.optimization.LogisticGradient.compute(Gradient.scala:169)
        at org.apache.spark.mllib.optimization.GradientDescent$$anonfun$runMiniBatchSGD$1$$anonfun$1.apply(GradientDescen
t.scala:192)
        at org.apache.spark.mllib.optimization.GradientDescent$$anonfun$runMiniBatchSGD$1$$anonfun$1.apply(GradientDescen
t.scala:190)
        at scala.collection.TraversableOnce$$anonfun$foldLeft$1.apply(TraversableOnce.scala:144)
        at scala.collection.TraversableOnce$$anonfun$foldLeft$1.apply(TraversableOnce.scala:144)
        at scala.collection.Iterator$class.foreach(Iterator.scala:727)
        at org.apache.spark.InterruptibleIterator.foreach(InterruptibleIterator.scala:28)
        at scala.collection.TraversableOnce$class.foldLeft(TraversableOnce.scala:144)
```

```
        at org.apache.spark.InterruptibleIterator.foldLeft(InterruptibleIterator.scala:28)
        at scala.collection.TraversableOnce$class.aggregate(TraversableOnce.scala:201)
        at org.apache.spark.InterruptibleIterator.aggregate(InterruptibleIterator.scala:28)
        at org.apache.spark.rdd.RDD$$anonfun$28.apply(RDD.scala:988)
        at org.apache.spark.rdd.RDD$$anonfun$28.apply(RDD.scala:988)
        at org.apache.spark.rdd.RDD$$anonfun$29.apply(RDD.scala:989)
        at org.apache.spark.rdd.RDD$$anonfun$29.apply(RDD.scala:989)
        at org.apache.spark.rdd.RDD$$anonfun$14.apply(RDD.scala:634)
        at org.apache.spark.rdd.RDD$$anonfun$14.apply(RDD.scala:634)
        at org.apache.spark.rdd.MapPartitionsRDD.compute(MapPartitionsRDD.scala:35)
        at org.apache.spark.rdd.RDD.computeOrReadCheckpoint(RDD.scala:277)
        at org.apache.spark.rdd.RDD.iterator(RDD.scala:244)
        at org.apache.spark.scheduler.ResultTask.runTask(ResultTask.scala:61)
        at org.apache.spark.scheduler.Task.run(Task.scala:64)
        at org.apache.spark.executor.Executor$TaskRunner.run(Executor.scala:203)
        at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1145)
        at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:615)
        at java.lang.Thread.run(Thread.java:745)

Driver stacktrace:
        at org.apache.spark.scheduler.DAGScheduler.org$apache$spark$scheduler$DAGScheduler$$failJobAndIndependentStages(DAGSche
duler.scala:1204)
        at org.apache.spark.scheduler.DAGScheduler$$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)
```

lab4

Updated 11 days ago by Renat Bekbolatov and Anonymous

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

Click to start off the wiki answer

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

Renat, Thank you for sharing your fix with the group.

Updated 11 days ago by Ameet Talwalkar

followup discussions *for lingering questions and comments*

Resolved Unresolved

Anonymous 11 days ago
It probably has to do with indexing of features: 0-based vs 1-based
Lab requires 0-based (one of the tests uses *sum*).

Resolved Unresolved

Chester Parrott 11 days ago
I'm having the same issue; this should be a trivial answer (everything else in the lab is correct and this is just a simply python syntax add.) Not sure what's going on with this...

Resolved Unresolved

Anonymous 11 days ago
Nevermind, I figured it out - turns out I made a mistake at an earlier step.

Chester Parrott 11 days ago mind sharing the mistake? I'm having the same issue...

Renat Bekbolatov 11 days ago Sure, it was the place where I create SparseVector: I was passing length of the incoming data for each data point, but they vary in size, obvious when looking back.
So instead, I specify the correct vector size, which in our case is the number of buckets. I am pretty sure this is what you are doing as well, because of where this variable is set. Placing it earlier in the code will allow for its reuse.

Chester Parrott 11 days ago That was the exact issue; changing it as you suggested fixed my issue. Thank you very much!

Chester Parrott 11 days ago I also had to restart my notebook after this (completely unrelated, I'm sure.)

Alun 11 days ago Mine failed with the dimension error but I had correctly sized SparseVector. After pulling my hair out for a while that 'duh' moment - I had accidentally used '`model0`' (instead of '`model`') in '`evaluateResults()`' which doesn't show up as a problem until 5d :(

Chris Jones 10 days ago Just ran into this too. Thanks, Renat! This was annoying since the previous problem 5c mentioned how the len(sparsevector) != numBuckets and I ignored that.

YUANWEN (CHUCK) WANG 10 days ago Thanks for the information, that really helps. I got similar error, my mistake is when doing the feature hashing, I didn't set the sparse vector length as the number of buckets, instead, I set it to be the number of non-zeros elements. It didn't give me errors previously, but at the modeling part, the model couldn't run.

Fuad Kamal 9 days ago Thanks Alun! I made exactly the same mistake.

Per Jonasson 9 days ago Thanks Alun! I had the same issue.

Anonymous 7 days ago Thanks~! After wasting many hours, I read this thread and found my mistake (model0) and fixed.

Luis Felipe Jimenez 6 days ago Thanks Renat, I have the same problem

Shahab Yunus 5 days ago Thanks. This was one pesky issue and hard to debug as the error message is not easy to interpret.

Andrew Lavers 5 days ago Thanks. Helped find both issuea.

Pasquale Grippo 3 days ago Great, thanks, had the same issue. This post saved me about 5 hours fixing !

gush 23 hours ago Thanks for the model <-> model0 hint! Save me so much time!

Resolved Unresolved

Zach Izham 8 days ago
Where is the fix???? Seriously wasted too long on this and am livid

Chester Parrott 8 days ago Look at how you've defined the SparseVector in 3d/e and 5b; this is where your woes lie, you need to use the number of buckets when declaring the length of your sparse vector, not the number of features in the point you are using to build the SV. If you use numBuckets instead, you will correct this error.

Prateek Jain 4 days ago Danke! This helped

Resolved Unresolved

Anonymous 5 days ago
Hi,
got stuck from 5b - 5e back and forth for almost 2 whole days now. please help !
i got exact same problem and i went back and fix the sparseVector's length my result for 5b is as below
[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]))]

but when i fix this problem, the original problem of 5d still is wrong and still i receive dimension mismatch problem -- see below
now now only is my 5d is wrong, my 5c also fail 1 test. so i guess i still have problem with 5b then ?
but how can i debug when i pass also the extra 2 tests provided when i read the post about 5b 5d and 5e ?
the two extra test is as below :
Test.assertEquals(hashValidationData.take(1)[0].features.indices[32],28405,"incorrect")
Test.assertEquals(hashValidationData.take(1)[0].features.values[32],2.0,"incorrect")

error message --
--- Py4JJavaError Traceback (most recent call last) <ipython-input-410-2b5061edcaea> in <module>() 7 .train(hashTrainData, numIters, stepSize, regParam=regParam, regType=regType, 8 intercept=includeIntercept)) ----> 9 logLossVa = evaluateResults(model, hashValidationData) 10 print ('\tstepSize = {0:.1f}, regParam = {1:.0e}: logloss = {2:.3f}' 11 .format(stepSize, regParam, logLossVa)) <ipython-input-382-2dafc1b66918> in evaluateResults(model, data) 11 """ 12 ---> 13 result=data.map(lambda x: computeLogLoss(getP(x.features,model0.weights,model0.intercept), x.label)).mean() 14 return result 15 /home/ubuntu/databricks/spark/python/pyspark/rdd.pyc in mean(self) 1076 2.0 1077 """ -> 1078 return self.stats().mean() 1079 1080 def variance(self):

/home/ubuntu/databricks/spark/python/pyspark/rdd.pyc in stats(self) 940 return left_counter.mergeStats(right_counter) 941 --> 942 return self.mapPartitions(lambda i: [StatCounter(i)]).reduce(redFunc) 943 944 def histogram(self, buckets): /home/ubuntu/databricks/spark/python/pyspark/rdd.pyc in reduce(self, f) 737 yield reduce(f, iterator, initial) 738 --> 739 vals = self.mapPartitions(func).collect() 740 if vals: 741 return reduce(f, vals) /home/ubuntu/databricks/spark/python/pyspark/rdd.pyc in collect(self) 711 """ 712 with SCCallSiteSync(self.context) as css: --> 713 port = self.ctx._jvm.PythonRDD.collectAndServe(self._jrdd.rdd()) 714 return list(_load_from_socket(port, self._jrdd_deserializer)) 715 /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:

Anonymous 5 days ago edit above *
the long error message is for 5d,
the 2 extra tests code is for 5b,
the 5c error is on averageSparsityHash but i pass averageSparsityOHE

Roger Meli 5 days ago My data is the same as yours and I still get the dimension mismatch as well.

Anonymous 5 days ago find

in the function

```
evaluateResults(model, data)
```

i was using model0 to compute the return...

Roger Meli 5 days ago I checked my code and I was using model0 as well, Changed it to (model,data). Same dimension mismatch. I will keep looking for the mistake.

Anonymous 5 days ago Thanks it helps a lot :)

Roger Meli 5 days ago

```
---------------------------------------------------------------------------
Py4JJavaError                             Traceback (most recent call last)
<ipython-input-144-70008db9b4ce> in <module>()
      7                 .train(hashTrainData, numIters, stepSize, regParam=regParam, regType=regType,
      8                        intercept=includeIntercept))
----> 9         logLossVa = evaluateResults(model, hashValidationData)
     10         print ('\tstepSize = {0:.1f}, regParam = {1:.0e}: logloss = {2:.3f}'
     11                .format(stepSize, regParam, logLossVa))

<ipython-input-119-ae42c2fbb93f> in evaluateResults(model, data)
     10         float: Log loss for the data.
     11     """
---> 12     return data.map(lambda x: computeLogLoss(getP(x.features, model0.weights, model0.intercept), x.label)).su
m() / data.count()
     13
     14 logLossTrLR0 = evaluateResults(model0, OHETrainData)

/usr/local/bin/spark-1.3.1-bin-hadoop2.6/python/pyspark/rdd.py in sum(self)
    921         6.0
    922         """
--> 923         return self.mapPartitions(lambda x: [sum(x)]).reduce(operator.add)
    924
    925     def count(self):

/usr/local/bin/spark-1.3.1-bin-hadoop2.6/python/pyspark/rdd.py in reduce(self, f)
    737             yield reduce(f, iterator, initial)
    738
--> 739         vals = self.mapPartitions(func).collect()
    740         if vals:
    741             return reduce(f, vals)

/usr/local/bin/spark-1.3.1-bin-hadoop2.6/python/pyspark/rdd.py in collect(self)
    711         """
    712         with SCCallSiteSync(self.context) as css:
--> 713             port = self.ctx._jvm.PythonRDD.collectAndServe(self._jrdd.rdd())
    714         return list(_load_from_socket(port, self._jrdd_deserializer))
    715

/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.collectAndServe.
: org.apache.spark.SparkException: Job aborted due to stage failure: Task 0 in stage 1777.0 failed 1 times, most rece
nt failure: Lost task 0.0 in stage 1777.0 (TID 3452, localhost): org.apache.spark.api.python.PythonException: Traceba
ck (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/rdd.py", line 2252, in pipeline_func
    return func(split, prev_func(split, iterator))
  File "/usr/local/bin/spark-1.3.1-bin-hadoop2.6/python/pyspark/rdd.py", line 2252, in pipeline_func
    return func(split, prev_func(split, iterator))
  File "/usr/local/bin/spark-1.3.1-bin-hadoop2.6/python/pyspark/rdd.py", line 282, in func
    return f(iterator)
  File "/usr/local/bin/spark-1.3.1-bin-hadoop2.6/python/pyspark/rdd.py", line 923, in <lambda>
    return self.mapPartitions(lambda x: [sum(x)]).reduce(operator.add)
  File "<ipython-input-119-ae42c2fbb93f>", line 12, in <lambda>
  File "<ipython-input-1
```

Anonymous 5 days ago You need to reload evaluateresults() to make the effective

SMV 4 days ago I had 'dimension mismatch' error too.
The problem was traced to using model0 instead of model as mentioned above.
Thanks for the tip. Saved hours of debugging.

Roger Meli 4 days ago Not sure why it was not working for me but it is now. Thanks for your help.

Till Haenisch 4 days ago Thank you very much for your help, I had the same problem.