Lab 2: K-means in Spark

Youfei Zhang

Goal: Inplement the K-Means algorithm without using SparkML or SparkMLlib package.

Part 1: NBA Log File

For each player, we define the comfortable zone of shooting is as matrix of,

```
{SHOT DIST, CLOSE DEF DIST, SHOT CLOCK}
```

Develop an K-Means algorithm to classify each player's records into 4 comfortable zones. Considering the hit rate, which zone is the best for James Harden, Chris Paul, Stephen Curry and Lebron James.

Design

In this lab, I tried to implement kmeans with PySpark in two ways. One is an RDD based iteration, the other is based on Spark Dataframe. By comparision, the RDD based iteration is more efficient than the Spark Dataframe one.

1. RDD based Kmeans

1. Intialize spark session

```
spark = SparkSession.builder.appName("NBA-kmeans").getOrCreate()
```

2. Preprocessing: clean and filter into RDD Load the csv into a spark context as a Spark DataFrame, and filter based on player name and the matrix column names. Convert the DataFrame into RDD through map to prepare for the iteration.

```
df = spark.read.format("csv").load(sys.argv[1], header = "true", inferSchema = "true")
dataPts = df.filter(df.player_name == 'james harden').select('SHOT_DIST','CLOSE_DEF_DIST',
'SHOT_CLOCK').na.drop()
dataRDD = dataPts.rdd.map(lambda r: (r[0], r[1], r[2]))
```

3. K-Means Iteration

After preprocessing, I start to implement the Kmeans algorithm. First, randomly select 4 rows as intial centroid:

```
k = 4
int_centroid = dataRDD.takeSample(False, k)
```

Then, I run the K-Means algorithm iteratively. For each data point, we calculate their distances to the 4 initial centroids, and assign them to the cluster of their closest centroid. Next, for each cluster, we recalculate the new centroid by getting the mean of each column. By doing so, we have 4 new centriods, and we recalculate the distance between each data points to the new centroids. We iterate this process until the new centroids equal to the old centroids, or until the maximum times of iteration is reached.

- calculate each data's distance to the centroid
- asign each data to the closet cdntroid
- calculate the new centroid per cluster by finding their mean
- iteration: calculate each data's distance to the NEW centroid
- stop iteration when old_centroids = new_centroids or when the maximum number of iteration is reached

```
iters = 0
old_centroid = int_centroid

# set the maximum iteration as 40
for m in range(40):
    map1 = dataRDD.map(lambda x: closestCenter(x, old_centroid))
    reduce1 = map1.groupByKey()
    map2 = reduce1.map(lambda x: cal_centroid(x)).collect() # collect a list
    new_centroid = map2
    converge = 0
    for i in range(k):
        if new_centroid[i] == old_centroid[i]:
            converge += 1
        # check if the different is smaller than 0.03
        else:
            diff = 0.0009
            closeDiff = [round((a - b)**2, 6) for a, b in zip(new_centroid[i], old_centroid[i])]
            if all(v <= diff for v in closeDiff):
                converge += 1
    if converge >= 4:
        print("Converge at the %s iteration\n" %(iters))
        print("\nFinal Centroids: %s" %(new_centroid))
        break
    else:
        iters += 1
        print("Iteration - %s round" %(iters))
        old_centroid = new_centroid
        print('Update:',old_centroid,'\n')
```

The Result

Final cluster:

```
2019-04-23 15:27:59 INFO  DAGScheduler:54 - Job 11 finished: collect at /spark-examples/labs/
lab2/part1/./nba-kmeans-1.py:87, took 0.679742 s
Converge at the 7 iteration


Final Centroids: [[4.1054, 2.8148, 18.3343], [6.3922, 2.5587, 9.1581], [23.7822, 5.0147, 16.2
881], [21.3225, 4.2283, 5.7786]]
```

The job process:

```
2019-04-23 15:27:58 INFO  BlockManagerInfo:54 - Added broadcast_20_piece0 in memory on 10.150
.0.7:43801 (size: 5.4 KB, free: 413.8 MB)
2019-04-23 15:27:58 INFO  ContextCleaner:54 - Cleaned accumulator 393
2019-04-23 15:27:58 INFO  ContextCleaner:54 - Cleaned accumulator 372
2019-04-23 15:27:58 INFO  ContextCleaner:54 - Cleaned accumulator 396
2019-04-23 15:27:58 INFO  ContextCleaner:54 - Cleaned accumulator 376
2019-04-23 15:27:58 INFO  ContextCleaner:54 - Cleaned accumulator 407
2019-04-23 15:27:58 INFO  ContextCleaner:54 - Cleaned accumulator 406
2019-04-23 15:27:58 INFO  ContextCleaner:54 - Cleaned accumulator 412
2019-04-23 15:27:58 INFO  ContextCleaner:54 - Cleaned accumulator 417
2019-04-23 15:27:58 INFO  ContextCleaner:54 - Cleaned accumulator 403
2019-04-23 15:27:58 INFO  ContextCleaner:54 - Cleaned accumulator 384
2019-04-23 15:27:58 INFO  ContextCleaner:54 - Cleaned accumulator 410
2019-04-23 15:27:58 INFO  ContextCleaner:54 - Cleaned accumulator 380
2019-04-23 15:27:58 INFO  ContextCleaner:54 - Cleaned accumulator 413
2019-04-23 15:27:58 INFO  ContextCleaner:54 - Cleaned accumulator 401
2019-04-23 15:27:58 INFO  ContextCleaner:54 - Cleaned accumulator 370
2019-04-23 15:27:58 INFO  ContextCleaner:54 - Cleaned accumulator 389
2019-04-23 15:27:58 INFO  ContextCleaner:54 - Cleaned accumulator 386
2019-04-23 15:27:58 INFO  ContextCleaner:54 - Cleaned accumulator 383
2019-04-23 15:27:58 INFO  MapOutputTrackerMasterEndpoint:54 - Asked to send map output locati
ons for shuffle 6 to 10.150.0.7:32950
```

2. Dataframe based Kmeans

1. Intialize spark session
2. Preprocessing: clean and filter

Load the csv into a spark context as a Spark DataFrame, and filter based on player name and the matrix column names.

```
df = spark.read.format("csv").load(sys.argv[1], header = "true", inferSchema = "true")
dataPts = df.filter(df.player_name == 'james harden').select('SHOT_DIST','CLOSE_DEF_DIST',
'SHOT_CLOCK').na.drop()
```

3. K-Means Iteration

First, randomly select 4 rows as intial centroid:

```
k = 4
int_centroid = dataPts.takeSample(False, k)
```

Then, run the iteration on Dataframe by adding a column 'Center' that idicate the datapoint's closet cluster from last round.

The function **closestCentroid** is wrapped into a udf to perform operations on the dataframe columns with:

```
def closestCentroid(col1, col2, col3):
    '''
    input: float value
    output: integer
    '''
    points = [col1, col2, col3]
    dist_list = []
    for c in newCenter:
        dist_list.append(euclDist(points, c))
    closest = float('inf') # an unbounded upper value for comparison
```

```
    index = -1
    for i, v in enumerate(dist_list):
        if v < closest:
            closest = v
            index = i
    return int(index)

minCenter = udf(closestCentroid, IntegerType())
```

The dataframe is first new column 'Center' that idicate the datapoint's closet cluster from the randomly chosen initial centroids. Then, for each cluster, new centers are computed based on Euclidean distance. Each centroid is compared to the prior centroids. If they are the same, or the absolute difference is within 0.03, then a convergence is achieved.

```
converge = 0
# calculate the new centroids for each cluster
for i in range(k):
    kCluster = rddCluster.filter(rddCluster.Center == i)
    n = kCluster.count()
    sumCol = [0] * 3
    sumCol[0] = calNewCentroid(kCluster, 'SHOT_DIST')
    sumCol[1] = calNewCentroid(kCluster, 'CLOSE_DEF_DIST')
    sumCol[2] = calNewCentroid(kCluster, 'SHOT_CLOCK')
    sumOfCols = [round(x / n, 4) for x in sumCol]
    newCenter[i] = sumOfCols
    print(newCenter)
    if newCenter[i] == oldCenter[i]:
        converge += 1
    elif:
        diff = 0.0009
        closeDiff = [round((a - b)**2, 6) for a, b in zip(newCenter[i], oldCenter[i])]
        if all(v <= diff for v in closeDiff):
            converge += 1
```

When convergence are larger than 4 (all the four new centroids are similar to the prior centroids), we break out of the loop. Otherwise, we keep within the iteration. At the end of every round of iteration, the old centroid is updated to be the new centroid.

```
    iters += 1
    print("Iteration - %s round" %(iters))
    if converge >= 4:
        print("Converge at the %s iteration\n" %(iters))
        break
    else:
        iters += 1
        print("Iteration - %s round" %(iters))
        old_centroid = new_centroid
        print('Update:',old_centroid,'\n')
```

Part 2: Bonus 1

Based on the NY Parking Violation data, given a Black vehicle parking illegally at 34510, 10030, 34050 (streat codes). What is the probability that it will get an ticket? (very rough prediction).

Design

The K-Means algorithm is implemented with PySpark with the following steps:

1. Initialze spark session

2. Load in the dataset as DataFrame for preprocessing. Filter the dataframe color in black, and then selecting columns of Street Code. Transform the filter dataframe into rdd.

```
black_list = ['BK', 'BLK', 'BK/', 'BK.', 'BLK.', 'BLAC', 'Black', 'BCK', 'BC', 'B LAC']
black = df.filter(df['Vehicle Color'].isin(black_list))
dataPts = black.select(black['Street Code1'], black['Street Code2'], black['Street
Code3']).na.drop()
dataRDD = dataPts.rdd.map(lambda r: (r[0], r[1], r[2]))
```

3. K-Means

After preprocessing, I start to implement the Kmeans algorithm.

First, randomly select 4 rows as intial centroid:

```
ini_centroid = dataRDD.takeSample(False, 4)
```

Then, the K-Means algorithm was run iteratively. For each data point, we calculate their distances to the 4 initial centroids, and assign them to the cluster of their closest centroid. Next, for each cluster, we recalculate the new centroid by getting the mean of each column. By doing so, we have 4 new centriods, and we recalculate the distance between each data points to the new centroids. We iterate this process until the new centroids equal to the old centroids, or until the maximum times of iteration is reached.

- calculate each data's distance to the centroid
- asign each data to the closet cdntroid
- calculate the new centroid per cluster by finding their mean
- iteration: calculate each data's distance to the NEW centroid
- stop iteration when old_centroids = new_centroids or when the maximum number of iteration is reached

```
iters = 0
old_centroid = int_centroid

for m in range(40):
    map1 = dataRDD.map(lambda x: closestCenter(x, old_centroid))
    reduce1 = map1.groupByKey()
    map2 = reduce1.map(lambda x: cal_centroid(x)).collect() # collect a list
    new_centroid = map2
    converge = 0
    for i in range(k):
        if new_centroid[i] == old_centroid[i]:
            converge += 1
        else:
            diff = 0.0009
            closeDiff = [round((a - b)**2, 6) for a, b in zip(new_centroid[i], old_centroid[i])]
            if all(v <= diff for v in closeDiff):
                converge += 1
    if converge >= 4:
        print("Converge at the %s iteration\n" %(iters))
        print("\nFinal Centroids: %s" %(new_centroid))
        break
    else:
        iters += 1
        print("Iteration - %s round" %(iters))
        old_centroid = new_centroid
        print('Update:',old_centroid,'\n')
```

4. Probability of Getting Tickets

Given a car parked at street_code = [34510, 10030, 34050], the probability that a car will get tickets is defined by:

** % of gettiing tickets = (total tickets in the cluster)/((average illegal car in a street) * (the number of street codes in the cluster)) **

```
street_code = [34510, 10030, 34050]
closet = closestCenter(street_code, new_centroid)
map3 = dataRDD.filter(lambda x: closestCenter(x, new_centroid)[0] == closet[0]).collect()
count = len(map3)
token = dict(Counter(map3))
counter = len(token)
maxV = max(token.items(),key = itemgetter(1))[1]
probability = round(count/(maxV * counter), 6)
print ("Probability of getting ticets:\n")
print (probability)
```

The result is as following:

```
2019-04-23 17:44:37 INFO  DAGScheduler:54 - ResultStage 21 (collect at /spark-examples/labs/l
ab2/part2/./parking-kmeans.py:90) finished in 0.110 s
2019-04-23 17:44:37 INFO  DAGScheduler:54 - Job 12 finished: collect at /spark-examples/labs/
lab2/part2/./parking-kmeans.py:90, took 0.304983 s
Converge at the 8 iteration


Final Centroids: [[20571.907, 29129.7116, 30848.4605], [49073.6102, 13849.8898, 17240.4661],
[54765.4206, 59625.0561, 52164.3832], [5006.5079, 2299.828, 1649.1402]]
2019-04-23 17:44:37 INFO  SparkContext:54 - Starting job: collect at /spark-examples/labs/lab
2/part2/./parking-kmeans.py:115
```

Part 3: Different Parallelism

Please try different levels of parallelism: 2, 3, 4, 5

Design

To try out different levels of parallelism, the configuration for spark-submit was set by:

```
--conf spark.default.parallelism = 2
```

It can be observed that with higher level of parallelism (-> 5), a convergence is achieved. While when parallelism is lower (2 or 3), no convergence was achieved until the maximum iteration was reached.

In addition, with a parallelism of 5 rather than 4, the convergence is achieved faster. (0.529500s vs 0.549806s)

Parallelism = 2:

```
2019-04-23 15:58:54 INFO  DAGScheduler:54 - Job 43 finished: collect at /spark-examples/labs/
lab2/part3/./nba-kmeans-3.py:86, took 0.429484 s
Iteration - 40 round
Update: [[23.876, 5.0318, 16.3141], [4.0666, 2.786, 17.9298], [7.6774, 2.6829, 8.4829], [21.7
673, 4.2738, 5.7392]]
```

Parallelism = 3:

```
2019-04-23 16:07:54 INFO  DAGScheduler:54 - Job 43 finished: collect at /spark-examples/labs/
lab2/part3/./nba-kmeans-3.py:86, took 0.507689 s
Iteration - 40 round
Update: [[22.1216, 4.2449, 5.478], [23.8189, 5.0402, 16.2664], [8.3097, 2.836, 8.5697], [4.01
73, 2.768, 17.891]]
```

Parallelism = 4:

```
2019-04-23 16:14:15 INFO  DAGScheduler:54 - ResultStage 35 (collect at /spark-examples/labs/l
ab2/part3/./nba-kmeans-3.py:86) finished in 0.194 s
2019-04-23 16:14:15 INFO  DAGScheduler:54 - Job 19 finished: collect at /spark-examples/labs/
lab2/part3/./nba-kmeans-3.py:86, took 0.549806 s
Converge at the 15 iteration


Final Centroids: [[6.4024, 2.5723, 9.1307], [23.7756, 4.9935, 16.1945], [21.2841, 4.2365, 5.6
852], [4.1076, 2.8058, 18.3176]]
```

Parallelism = 5:

```
2019-04-23 16:16:52 INFO  TaskSchedulerImpl:54 - Removed TaskSet 51.0, whose tasks have all c
ompleted, from pool
2019-04-23 16:16:52 INFO  DAGScheduler:54 - ResultStage 51 (collect at /spark-examples/labs/l
ab2/part3/./nba-kmeans-3.py:86) finished in 0.200 s
2019-04-23 16:16:52 INFO  DAGScheduler:54 - Job 27 finished: collect at /spark-examples/labs/
lab2/part3/./nba-kmeans-3.py:86, took 0.529500 s
Converge at the 23 iteration


Final Centroids: [[8.3097, 2.836, 8.5697], [23.8189, 5.0402, 16.2664], [22.1216, 4.2449, 5.47
8], [4.0173, 2.768, 17.891]]
2019-04-23 16:16:52 INFO  AbstractConnector:318 - Stopped Spark@380f39d6{HTTP/1.1,[http/1.1]}
{0.0.0.0:4040}
```