HW3 - K-Means

107062130 陳奕君

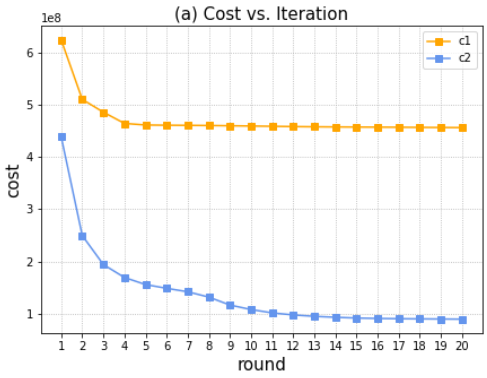
Method Explanation

1. Preprocess data
   - 參考 HW1 的 matrix multiplication，把 data, c1, c2 處理成 `(dim, (pid, value_dim)` ，方便之後依照 dimension 直接計算
   - 把 data 存成 `(pid, [v0, v1, ..., v57])` ，計算 centroid 用
2. Find centroid and closest_distance for each point (for loop 內)
   - 先用 `join` 把同個 dimension 的值都接在一起，再用 `map` 去計算 value 中 point 和 centroid 的 distance，最後 `reduceByKey` 把同一組 (point, centroid) 的 distance 相加，變成 `((pid, cid), distance)`
   - 為了後續比較各個 centroid 到底哪個最接近 point，用 `map` 把資料處理成 `(pid, (cid, distance))`
   - 接著用 `reduceByKey` 比較 distance，只回傳比較比較小的 distance 的 value，變成 `(pid, (cid, closest_distance))`
3. Match point and centroid (for loop 內)
   - 用 `map` 把 closest_distance 拿掉，變成 `(pid, cid)`
4. Recompute the centroid by points (for loop 內)
   - 先用 `join` 把 points_value 和前面得到的 (pid, cid) 合在一起，變成 `(pid, (cid, [v0, v1..., v57]))`
   - 用 `map` 把 pid 去掉，變成 `(cid, [v0, v1..., v57])` ，這些就是組成 centroid 的值
   - 利用 `reduceByKey` 把同個 centroid 的 value list 連在一起，再用 `map` 去將每個 dimension 的值相加再除以點數量，得到這輪新的 centroid 的值 `(cid, [v0, v1..., v57])`
   - 用跟 preprocess data 時一樣的方法把 centroid 變成 `(dim, (pid, value_dim)` 以利計算

Assignment Requirements

Problem (a). Euclidean Distance

Cost vs. Iteration + Percentage improvement values



a_cost_table

	C1	C2
Round 1	623660345.3064090	438747790.0279160
Round 2	509862908.2975460	249803933.62600300
Round 3	485480681.8720080	194494814.4063130
Round 4	463997011.685013	169804841.45154300
Round 5	460969266.5729970	156295748.80627600
Round 6	460537847.98276800	149094208.10896600
Round 7	460313099.65354600	142508531.61961500
Round 8	460003523.8894070	132303869.40653000
Round 9	459570539.3177350	117170969.83719100
Round 10	459021103.342291	108547377.17857000
Round 11	458490656.19198100	102237203.31799600
Round 12	457944232.5879740	98278015.74975670
Round 13	457558005.19867700	95630226.12177400
Round 14	457290136.3523020	93793314.05119300
Round 15	457050555.05956300	92377131.96821070
Round 16	456892235.6153550	91541606.25423890
Round 17	456703630.7370340	91045573.83042460
Round 18	456404203.01897500	90752240.10140800
Round 19	456177800.541993	90470170.18122730
Round 20	455986871.0273460	90216416.17563120

Percentage improvement values

```
c1:  26.885383292516984 %
c2:  79.43775029159896 %
```

Observation：

1. C1 has higher initial cost, and quickly converge.
2. C2's cost is continually decreasing, and C2 has more improvements too.
3. C2 has the better performance than C1.

We already knew:

- the Euclidean is more sensitive to outliers than the Manhattan.

Explanation for (a)

1. Maybe C1 are far from outliers in the beginning, so no matter which centroid those outliers belong to, they could only make a small move. Thus, the cost is high and stable.
2. Since C2 are as far apart as possible, they consider the outliers at the beginning. There may be some centroids that take care about outliers, the other take care about the most points. Therefore, they have lower cost and improve well.
3. Same reason with (2).

Distances for all pairs of centroids

Euclidean - c1

a_c1_ec

	1	2	3	4	5	6	7	8	9	10
1	0	692.1578865536190	3490.2586403239800	205.75027883457200	346.71882253414800	512.6122467083620	444.7310005470830	566.2019922936280	1282.7708445154400	307.66912835221500
2	0	0.0	2798.8010531589400	897.6589863450830	1038.8268882911400	1204.0781989863100	1136.3273438149300	1257.4495275592400	669.8902282318560	412.076077167744
3	0	0.0	0.0	3695.114191079640	3836.9066381524600	4002.6890825730300	3934.8715588311600	4056.1355729719100	2294.57964158953	3195.9239010108900
4	0	0.0	0.0	0.0	142.43887392408600	309.50632446700000	241.7301145044180	363.2628951046580	1474.9454213642700	504.63411599571500
5	0	0.0	0.0	0.0	0.0	167.1498001315250	99.54554331498560	220.90178372040600	1615.852353440390	646.9305638786210
6	0	0.0	0.0	0.0	0.0	0.0	67.91186107588500	53.78989116172680	1782.2030486002300	814.0761501339800
7	0	0.0	0.0	0.0	0.0	0.0	0.0	121.63372043718300	1715.2531997144100	746.3355586141010
8	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1835.6396718448900	867.8230790917460
9	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	975.3204225971710
10	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Euclidean - c2

a_c2_ec

	1	2	3	4	5	6	7	8	9	10
1	0	15760.12247224590	14110.83439088330	9045.32023455239	5567.684524118410	1924.6240815733200	1100.8590503593300	402.89054961102900	2105.4425755929600	3169.003772849920
2	0	0.0	11524.505650179800	6743.88410019246	10192.525007384200	14455.11937212130	14682.450992891100	15362.417960805100	13674.707531226200	12597.039559531100
3	0	0.0	0.0	9545.879403387190	10883.382187801400	12233.959804503600	13208.002933714400	13786.484182516100	12508.95738096870	11938.376127029500
4	0	0.0	0.0	0.0	3494.2224155718100	7718.222009696620	7957.775949135470	8644.807041005520	6947.820636329120	5876.330199605870
5	0	0.0	0.0	0.0	0.0	4404.562590797220	4492.458214360270	5169.9372911256400	3488.1585187816200	2407.918794485800
6	0	0.0	0.0	0.0	0.0	0.0	1182.864189045910	1615.7882361392900	1313.3274934048700	2153.771471752400
7	0	0.0	0.0	0.0	0.0	0.0	0.0	698.4881359277910	1010.1976652126800	2085.4606764073000
8	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1702.7926583344800	2768.6077191659200
9	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1080.534943939510
10	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Manhattan - c1

a_c1_mh

	1	2	3	4	5	6	7	8	9	10
1	0	728.9243139995410	3797.8990780180700	212.18109038688200	374.8904224625680	577.4020758222030	499.1578939470970	645.7697774866790	1731.0643066429400	406.70122510554600
2	0	0.0	3072.8886904466500	935.8853379058190	1100.833091013830	1303.8957233218900	1225.351713180190	1372.0922054892200	1005.2930456989200	490.9280581546790
3	0	0.0	0.0	4001.0380519416300	4170.304532612550	4372.788718801670	4294.952834221330	4440.719767637060	2513.422660256410	3396.4200031055900
4	0	0.0	0.0	0.0	171.36515415561500	375.2479208943980	296.25472354226700	443.49844537759100	1934.086959807070	609.749321623295
5	0	0.0	0.0	0.0	0.0	204.52292364239300	125.59678617977000	272.93491284339400	2102.86492281106	779.3972267080750
6	0	0.0	0.0	0.0	0.0	0.0	79.40168444202220	69.58987631971820	2306.3802505970700	983.019680661424
7	0	0.0	0.0	0.0	0.0	0.0	0.0	147.86570909768600	2227.555856782110	904.3702500156850
8	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2374.545430478820	1050.9162214834800
9	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1327.5839795031100
10	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

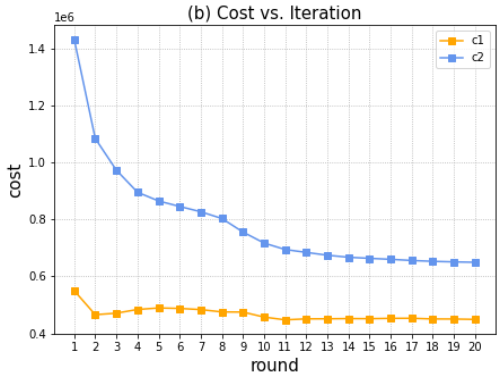
Manhattan - c2

a_c2_mh

	1	2	3	4	5	6	7	8	9	10
1	0	15772.614899885300	20215.645980206500	9533.17084939759	5604.200489099260	3088.0543184882400	1311.039156985080	471.26571999995200	2369.4121590413600	3349.657086008630
2	0	0.0	16003.499	7219.196666666670	10221.031	16105.3475	14909.169510714300	15434.460040787600	13950.575945454500	12776.883065217400
3	0	0.0	0.0	10690.484333333300	14613.552	17509.90275	18912.605410714300	19748.93569338960	17851.806836363600	16873.243673913000
4	0	0.0	0.0	0.0	3935.2926666666700	8896.389208333330	8228.355075	9065.404333333330	7168.732963636360	6190.679311594200
5	0	0.0	0.0	0.0	0.0	5893.070125	4696.975382142860	5221.252805907170	3737.707	2564.1705434782600
6	0	0.0	0.0	0.0	0.0	0.0	1781.8226714285700	2619.8113862517600	2162.8021454545500	3337.7462608695700
7	0	0.0	0.0	0.0	0.0	0.0	0.0	840.7225236939920	1068.9399724026000	2137.7882566770200
8	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1901.208756322720	2883.734536812820
9	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1176.4504256917000
10	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Problem (b). Manhattan Distance

Cost vs. Iteration + Percentage improvement values



Percentage improvement values

```
c1:  18.378954327236944 %
c2:  54.68569434813399 %
```

Observation：

1. C2 has higher initial cost.
2. C2's cost is continually decreasing, and has more improvements.
3. C1 seems to converge at the beginning.
4. C1 has the better performance than C2.

We already knew:

- the Euclidean is more sensitive to outliers than the Manhattan.

b_cost_table

	C1	C2
Round 1	550117.1420000000	1433739.3100000000
Round 2	465617.3485236670	1084488.77696488
Round 3	470902.29650127500	973431.714662042
Round 4	483932.3931689980	895934.5925630710
Round 5	489207.57581724900	865128.3352940820
Round 6	487597.67386148900	845846.6470313490
Round 7	483507.3944688320	827219.5827561260
Round 8	475258.8666639480	803590.3456011110
Round 9	474863.62392036900	756039.5172761210
Round 10	457210.606910489	717332.9025432300
Round 11	447522.40673446000	694587.9252526880
Round 12	450872.48677093400	684444.5019967910
Round 13	451157.7926116740	674574.7475478560
Round 14	451922.28116581000	667409.4699160270
Round 15	451572.9468517020	663556.6278215030
Round 16	452744.14279976800	660162.7772287570
Round 17	453082.73028718400	656041.3222947130
Round 18	450583.67086029900	653036.7540731600
Round 19	450368.74931674200	651112.4262522730
Round 20	449011.363725519	649689.0131843530

Explanation for (b)

1. Since C2 has taken outliers into consideration, the cost will be higher when using Manhattan distance.
2. Since Manhattan distance is less sensitive to outliers, it focus on dealing with concetrated data. So it could make C2 move quickly to the centroid of most data points.
3. Maybe C1 are close to most concentrated data points in the beginning, so the centroid won't have large difference by using Manhattan distance.
4. Combining (1)(2)(3), the main reason is considering outliers or not. We could conclude that Manhattan distance is suitable for concentrated data and randomized choosing centroid, and Euclidean distance is suitable for the data with outliers.

Distances for all pairs of centroids

Euclidean - c1

b_c1_ec

	1	2	3	4	5	6	7	8	9	10
1	0	2219.1772770509900	9948.044077639340	528.6997575475610	413.3650612071160	827.7188856579820	681.0349895443140	917.127382961168	832.1474343052390	729.0563485517980
2	0	0.0	7767.945602579570	2734.0498544640400	2628.4908097297600	3044.4778721252700	2898.7128939326600	3133.460130173690	1812.4545744763500	1491.3573457511600
3	0	0.0	0.0	10433.061351319800	10361.367486044600	10773.530838116900	10626.48859681840	10862.965776276400	9340.275232170260	9236.840021697920
4	0	0.0	0.0	0.0	221.37279398782300	375.1561884837040	249.3791882654920	457.25965255434800	1156.5833757939900	1251.1583460671200
5	0	0.0	0.0	0.0	0.0	415.9899852468610	270.74879157775900	505.07106661337900	1171.964205896140	1137.135265913950
6	0	0.0	0.0	0.0	0.0	0.0	147.04697388149500	89.4909165524104	1529.4640115204100	1553.1238066646700
7	0	0.0	0.0	0.0	0.0	0.0	0.0	236.51462239245400	1391.550421146710	1407.404400267170
8	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1613.555789411290	1642.1286873773300
9	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	709.4077855501390
10	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Euclidean - c2

b_c2_ec

	1	2	3	4	5	6	7	8	9	10
1	0	15747.234225995100	14100.144687804200	9032.333022974990	5554.786693382850	2006.7026683699500	1338.161125522520	514.627037539365	1571.2434198099000	3022.6608840606800
2	0	0.0	11524.505650179800	6743.88410019246	10192.525007384200	14474.554115568200	14412.056615494400	15239.87707116480	14328.226191823100	12731.397634823900
3	0	0.0	0.0	9545.879403387190	10883.382187801400	12167.79387138700	13125.351004065800	13684.606757319800	12643.985638343200	12006.39461776130
4	0	0.0	0.0	0.0	3494.2224155718100	7742.6281172713400	7694.276701483220	8521.197863130880	7588.404540222490	6009.820222838370
5	0	0.0	0.0	0.0	0.0	4452.971684507900	4219.760574098210	5047.516256062290	4167.636533016610	2542.5693542004600
6	0	0.0	0.0	0.0	0.0	0.0	1405.1090803330800	1637.7294382123900	910.9943878357350	2124.2633623517300
7	0	0.0	0.0	0.0	0.0	0.0	0.0	827.840658014979	566.5510174100830	1684.516011855050
8	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1081.3793348240200	2511.458858671230
9	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1649.3891719444800
10	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Manhattan - c1

b_c1_mh

	1	2	3	4	5	6	7	8	9	10
1	0	2341.017218605990	11929.300151785700	651.187488095238	496.33152119047600	947.7432364650630	770.7373834915090	1056.7995006298200	1260.5105607142900	737.7135730519480
2	0	0.0	9597.441187096770	2778.9457620967700	2830.1445281720400	3280.3591681796100	3104.2857711482100	3388.9826482565900	2380.460958064520	1605.2701287064200
3	0	0.0	0.0	12323.287569444400	12421.263080000000	12871.483428978200	12695.554202097900	12979.13318018670	10775.939185714300	11196.78698181820
4	0	0.0	0.0	0.0	335.951213	558.4692581658290	382.4633301282050	667.5332295988940	1653.8258869047600	1379.165172979800
5	0	0.0	0.0	0.0	0.0	452.86133068676700	276.3264914965030	561.8492485408020	1755.105532761910	1226.6603546666700
6	0	0.0	0.0	0.0	0.0	0.0	177.59316237364000	110.21762404606000	2205.3073830102900	1677.666863534850
7	0	0.0	0.0	0.0	0.0	0.0	0.0	287.42970773365600	2028.901615784220	1500.9934102564100
8	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2314.667454756970	1786.8113163428900
9	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1006.3678264069300
10	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Manhattan - c2

b_c2_mh

	1	2	3	4	5	6	7	8	9	10
1	0	15757.691264458300	20200.259435786000	9517.668232509730	5588.853634537060	3281.488246764410	1430.2086778758800	602.9548488263030	2102.553977579330	3211.4557560606600
2	0	0.0	16003.499	7219.196666666670	10221.031	16325.2705	14506.48588961040	15335.957402597400	14980.056095890400	12922.931357142900
3	0	0.0	0.0	10690.484333333300	14613.552	17521.517666666700	18775.12146103900	19602.262814935100	18111.885424657500	16995.133535714300
4	0	0.0	0.0	0.0	3935.2926666666700	9116.0245	8090.510188311690	8918.813116883120	7771.22207762557	6312.5300119047600
5	0	0.0	0.0	0.0	0.0	6110.8325	4293.5019025974000	5123.066808441560	4768.923	2710.0565000000000
6	0	0.0	0.0	0.0	0.0	0.0	1855.5799090909100	2682.5692337662300	1358.795894977170	3413.0361785714300
7	0	0.0	0.0	0.0	0.0	0.0	0.0	833.4302824675320	674.8275699163850	1784.5120454545500
8	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1500.824883695070	2613.9973051948100
9	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2062.251068003910
10	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0