به نام خدا

تحلیل داده های حجیم

تمرین سری سوم

استاد: دکتر غلامپور

دانشجو:سجاد هاشم بیکی(98107077)

پاییز1401

سوال1:

الف) اگر همه سطرهای انتخاب شده مقدار صفر داشته باشند مقدار هش نامشخص خواهد بود.

احتمال اینکه همه سطرهای انتخاب شده صفر باشد برابر بدین صورت است:

$$P = \frac{\binom{n-k}{m}}{\binom{n}{m}} = \frac{\frac{(n-k)!}{m!\,(n-k-m)!}}{\frac{n!}{m!\,(n-m)!}} = \left(\frac{n-k}{n}\right)\left(\frac{n-k-1}{n-1}\right)\dots\left(\frac{n-k-m+1}{n-m+1}\right)$$

عبارات بالا همگی کوچک تر مساوی با $\left(\frac{n-k}{n}\right)$ میباشند بنابراین داریم:

$$P = \left(\frac{n-k}{n}\right)\left(\frac{n-k-1}{n-1}\right)\dots\left(\frac{n-k-m+1}{n-m+1}\right) \le \left(\frac{n-k}{n}\right)^m$$

ب)

$$P \le \left(\frac{n-k}{n}\right)^m = \left(1-\frac{k}{n}\right)^m = \left(\left(1-\frac{k}{n}\right)^{\frac{n}{k}}\right)^{\frac{mk}{n}} \approx \left(\frac{1}{e}\right)^{\frac{mk}{n}} = e^{-\frac{mk}{n}}$$

$$e^{-\frac{mk}{n}} \le e^{-10} \Longrightarrow \frac{mk}{n} \ge 10 \Longrightarrow k \ge \frac{10n}{m}$$

ج)

S1	S2
0	0
1	1
1	0

:

$$sim(S1, S2) = \frac{|S1 \cap S2|}{|S1 \cup S2|} = \frac{1}{2}$$

$$P[h(S1) = h(S2)] = \frac{1}{3}$$

با توجه به جایگشت تناوبی در یک سوم حالت ها مقدار هش یکسان میشود. که این مقدار مخالف مقدار similarity *میباشد*

سوال2:

الف:

با استفاده از نامساوی مارکوف داریم:

$$P\left[\sum_{i=1}^{L}|T\cap W_i|\geq 3L\right]\leq \frac{\mathbb{E}[\sum_{i=1}^{L}|T\cap W_i|]}{3L}=\frac{\sum_{i=1}^{L}\mathbb{E}[|T\cap W_i|]}{3L}=\frac{L\mathbb{E}[|T\cap W_i|]}{3L}$$

$$=\frac{\mathbb{E}[|T\cap W_1|]}{3}$$

به عنوان مثال W1 را انتخاب میکنیم.

توابع H به صورت $(c, c\lambda, p_1, p_2) - sensitive$ هستند بنابراین داریم:

$$g_1 = (h_{1i}, \dots, h_{1k})$$

$$if\ \ d(x,z) > c\lambda \ \longrightarrow P\left[h_{1i}(x) = h_{1i}(z)\right] \leq p_2$$

با فرض استقلال هش ها داریم:

$$if\ \ d(x,z) > c\lambda \ \longrightarrow P[g_1(x) = g_1(z)] \leq {p_2}^k = {p_2}^{\log_{\frac{1}{p_2}} n} = n^{-1} = \frac{1}{n}$$

اگر فاصله همه نقاط از z بیشتر از $c\lambda$ باشد،بنابراین n هستند که به احتمال $1/n$ به باکت یکسان میروند.نقاط آن باکت توزیع binomial(n,1/n) دارند. که امید ریاضی ان یک میباشد.بنابراین داریم:

$$\mathbb{E}[|T\cap W_1|] = 1 \longrightarrow P\left[\sum_{i=1}^{L}|T\cap W_i|\geq 3L\right]\leq \frac{\mathbb{E}[|T\cap W_1|]}{3}=\frac{1}{3}$$

ب:

با توجه به توابع H داریم:

$$if \;\; d(x,z) \leq \lambda \;\longrightarrow P[h_i(x) = h_i(z)] \geq p_1$$

با فرض استقلال هش ها داریم:

$$if \;\; d(x,z) \leq \lambda \;\longrightarrow P\big[g_j(x) = g_j(z)\big] = \prod_{i=1}^{k} P\big[h_{ji}(x) = h_{ji}(z)\big] \geq {p_1}^k$$

$$= p_1^{\log_{\frac{1}{p_2}} n} = n^{\log_{\frac{1}{p_2}} p1}$$

$$if \;\; d(x,z) \leq \lambda \;\longrightarrow P\big[g_j(x) \neq g_j(z)\big] \leq 1 - n^{\log_{\frac{1}{p_2}} p1}$$

با فرض استقلال g_j داریم:

$$P\left[\forall\, 1 \le j \le L, g_j(x) \ne g_j(z)\right] \le \left(1 - n^{\log_{\frac{1}{p_2}} p1}\right)^L = \left(1 - n^{\log_{\frac{1}{p_2}} p1}\right)^{n^\rho}$$
$$= \left(1 - n^{-\log_{p_2} p1}\right)^{n^\rho}$$

$$P\left[\forall\, 1 \le j \le L, g_j(x) \ne g_j(z)\right] \le \left(1 - \frac{1}{n^{\log_{p_2} p1}}\right)^{n^{\log_{p_2} p_1}}$$

اگر مقدار n را بزرگ در نظر بگیریم داریم:

$$P\left[\forall\, 1 \le j \le L, g_j(x) \ne g_j(z)\right] < \frac{1}{e}$$

سوال3:

الف:

با همان قالب گفته شده در سوال((key=(plate ,date),value = [Device Code List]) rdd ساخته شده بدین شکل میباشد:

```
rdd_a_1.take(10)
```

```
[(('9354665', '2022-01-08'), ['900156', '900101', '900212', '900264']),
 (('9803489', '2022-01-08'),
  ['631368',
   '631355',
   '900164',
   '230204',
   '900185',
   '145',
   '631829',
   '631765',
   '230101']),
 (('9824503', '2022-01-08'), ['900151']),
 (('9799819', '2022-01-08'), ['900247', '900104']),
 (('46805196', '2022-01-08'), ['22010122', '100701297']),
 (('93857155', '2022-01-08'), ['900236', '22010079', '900139', '100700812']),
 (('70465920', '2022-01-08'),
  ['22010120', '100701293', '22010110', '22010138']),
 (('68806444', '2022-01-08'), ['100701298', '22010047']),
 (('8257418', '2022-01-08'), ['900151', '22010031', '900117']),
 (('9401649', '2022-01-08'), ['900104'])]
```

ب:

دیتای sample:

یک مسیر فرضی انتخاب میکنیم.بدین صورت که چند دوربین را به صورت رندوم انتخاب کرده و آن را بردار مسیر فرضی در نظر میگیریم.

حال کسینوس زاویه میان دو بردار مسیر فرضی و مسیر خودروها را محاسبه میکنیم.

هر چه مقدار کسینوس به یک نزدیک تر باشد به معنای زاویه کمتر میان ان دو بردار و شباهت بیشتر آنهاست.

در اینجا یک مسیر فرضی شامل 4 دوربین را در نظر گرفتیم.

پنج مقدار بزرگتر برابر 0.5 و0.354 و0.289 و0.25 و 0.224 میباشد.

مولفه های هر زوج بدین صورت میباشد:

```
(key =cos(theta), value = ((plate,date),[Device Code List]))
```

بیشترین شباهت ها به ازای پنج مقدار اول:

```
(0.5, (('11064269', '2022-01-08'), ['22010058'])
(0.354, (('12094907', '2022-01-08'), ['22010057', '22010058']))
(0.289, (('10238525', '2022-01-08'), ['900155', '22010057', '22010058']))
(0.25, (('68635642', '2022-01-08'), ['22010058', '22010057', '100701150',
'100701145']))
(0.224, (('69018599', '2022-01-08'), ['631633', '900273', '100701252',
'22010058', '22010057']))
```

مقادیر بالا به ازای مسیر فرضی

{['22010058'],['631702'],['22009911'],['109']} میباشد.

```
sorted(rdd_cosine_similarity.collect(),reverse=True)
```

```
[(0.5, (('9987866', '2022-01-08'), ['22010058'])),
 (0.5, (('99802064', '2022-01-08'), ['22010058'])),
 (0.5, (('99778379', '2022-01-08'), ['22010058'])),
 (0.5, (('99707047', '2022-01-08'), ['22010058'])),
 (0.5, (('99618654', '2022-01-08'), ['22010058'])),
 (0.5, (('99596045', '2022-01-08'), ['22010058'])),
 (0.5, (('99508099', '2022-01-08'), ['22010058'])),
 (0.5, (('99507827', '2022-01-08'), ['22010058'])),
 (0.5, (('99386250', '2022-01-08'), ['22010058'])),
 (0.5, (('9932501', '2022-01-08'), ['22010058'])),
 (0.5, (('99201637', '2022-01-08'), ['22010058'])),
 (0.5, (('9919129', '2022-01-08'), ['631702'])),
 (0.5, (('99053517', '2022-01-08'), ['22010058'])),
 (0.5, (('99053327', '2022-01-08'), ['22010058'])),
 (0.5, (('9895649', '2022-01-08'), ['22010058'])),
 (0.5, (('9874460', '2022-01-08'), ['22010058'])),
 (0.5, (('9873937', '2022-01-08'), ['22010058'])),
 (0.5, (('98459344', '2022-01-08'), ['109'])),
 (0.5, (('9835768', '2022-01-08'), ['22010058'])),
 (0.5, (('98329372', '2022-01-08'), ['22010058'])),
 (0.5, (('98293450', '2022-01-08'), ['22010058'])),
 (0.5, (('98245505', '2022-01-08'), ['22010058'])),
 (0.5, (('9822656', '2022-01-08'), ['22010058'])),
 (0.5, (('98215824', '2022-01-08'), ['631702'])),
 (0.5, (('98172210', '2022-01-08'), ['631702'])),
 (0.5, (('9792949', '2022-01-08'), ['22010058'])),
```

```
(0.354, (('99979911', '2022-01-08'), ['22010057', '22010058'])),
(0.354, (('9996760', '2022-01-08'), ['22010057', '22010058'])),
(0.354, (('9961678', '2022-01-08'), ['22010057', '22010058'])),
(0.354, (('9961375', '2022-01-08'), ['22010057', '22010058'])),
(0.354, (('9960007', '2022-01-08'), ['22010057', '22010058'])),
(0.354, (('99437514', '2022-01-08'), ['900174', '22010058'])),
(0.354, (('9936950', '2022-01-08'), ['22010057', '22010058'])),
(0.354, (('9916320', '2022-01-08'), ['22010057', '22010058'])),
(0.354, (('9911000', '2022-01-08'), ['22010057', '22010058'])),
(0.354, (('9904018', '2022-01-08'), ['22010057', '22010058'])),
(0.354, (('9879607', '2022-01-08'), ['22010057', '22010058'])),
(0.354, (('9876084', '2022-01-08'), ['22010057', '22010058'])),
(0.354, (('9874884', '2022-01-08'), ['22010057', '22010058'])),
(0.354, (('9850869', '2022-01-08'), ['22010057', '22010058'])),
(0.354, (('9832100', '2022-01-08'), ['22010057', '22010058'])),
(0.354, (('9829391', '2022-01-08'), ['22010057', '22010058'])),
(0.354, (('9825266', '2022-01-08'), ['22010057', '22010058'])),
(0.354, (('9818380', '2022-01-08'), ['22010057', '22010058'])),
(0.354, (('9815821', '2022-01-08'), ['22010057', '22010058'])),
(0.354, (('9814894', '2022-01-08'), ['22010057', '22010058'])),
(0.354, (('9812799', '2022-01-08'), ['22010057', '22010058'])),
(0.354, (('9794777', '2022-01-08'), ['900162', '22010058'])),
(0.354, (('9789789', '2022-01-08'), ['22010057', '22010058'])),
(0.354, (('9787508', '2022-01-08'), ['22010057', '22010058'])),
```

```
[(0.289, (('9901330', '2022-01-08'), ['100701297', '22010057', '22010058'])),
 (0.289, (('9795827', '2022-01-08'), ['900113', '900104', '22010058'])),
 (0.289, (('9751846', '2022-01-08'), ['22010047', '22010057', '22010058'])),
 (0.289, (('9745726', '2022-01-08'), ['22010039', '22010058', '22010059'])),
 (0.289, (('97220619', '2022-01-08'), ['100701119', '22010057', '22010058'])),
 (0.289, (('96755449', '2022-01-08'), ['22010057', '22010058', '22010059'])),
 (0.289, (('9667739', '2022-01-08'), ['22010047', '22010057', '22010058'])),
 (0.289, (('9624530', '2022-01-08'), ['900179', '22010057', '22010058'])),
 (0.289, (('96232708', '2022-01-08'), ['900164', '22010058', '100700880'])),
 (0.289, (('9552931', '2022-01-08'), ['22010057', '22010058', '900193'])),
 (0.289, (('9542820', '2022-01-08'), ['22010057', '631634', '22010058'])),
 (0.289, (('9330932', '2022-01-08'), ['100701252', '22010057', '22010058'])),
 (0.289, (('9309629', '2022-01-08'), ['900158', '22010057', '22010058'])),
 (0.289, (('92887143', '2022-01-08'), ['109', '100700841', '900259'])),
 (0.289, (('9248768', '2022-01-08'), ['900226', '22010057', '22010058'])),
 (0.289, (('92239319', '2022-01-08'), ['100700804', '100700834', '22010058'])),
 (0.289, (('9212339', '2022-01-08'), ['22010058', '631829', '631363'])),
 (0.289, (('9189515', '2022-01-08'), ['22010057', '22010058', '22010085'])),
 (0.289, (('91760473', '2022-01-08'), ['635713', '22010057', '22010058'])),
 (0.289, (('91458551', '2022-01-08'), ['900155', '100700841', '22010058'])),
 (0.289, (('9088770', '2022-01-08'), ['631757', '22010058', '22010083'])),
```

کل دیتا:

نتایج این قسمت برای **کل دیتا** و مسیر فرضی (['22009806'], ['1001001'], [['900195'], ['100701298']]) بدین شکل میباشد:

```
[(0.5, (('91978972', '2022-01-08'), ['100701298'])),
 (0.5, (('74609983', '2022-01-08'), ['100701298'])),
 (0.5, (('40201285', '2022-01-08'), ['100701298'])),
 (0.5, (('20250397', '2022-01-08'), ['100701298'])),
 (0.354, (('9423700', '2022-01-08'), ['22010121', '100701298'])),
 (0.354, (('88467288', '2022-01-08'), ['22010112', '100701298'])),
 (0.354, (('8738634', '2022-01-08'), ['100701298', '22010128'])),
 (0.354, (('60482918', '2022-01-08'), ['100701298', '22010125'])),
 (0.354, (('26108324', '2022-01-08'), ['100701298', '22010127'])),
 (0.354, (('20211489', '2022-01-08'), ['100701298', '22010059'])),
 (0.289,
  (('60123951', '2022-01-08'), ['100701298', '100701119', '100701297'])),
 (0.289, (('42083983', '2022-01-08'), ['22010122', '100701298', '22010054'])),
 (0.25,
  (('26173221', '2022-01-08'),
   ['100701298', '900202', '900155', '100701157'])),
 (0.224,
  (('19745155', '2022-01-08'),
   ['205202', '205201', '169', '22009972', '100701298'])),
 (0.204,
  (('104633602', '2022-01-08'),
   ['100700820', '22010110', '900216', '100701298', '631765', '900158']))]
```

```
[(0.5, (('99912490', '2022-01-08'), ['100701298'])),
 (0.5, (('99756221', '2022-01-08'), ['100701298'])),
 (0.5, (('99690837', '2022-01-08'), ['100701298'])),
 (0.5, (('99581728', '2022-01-08'), ['100701298'])),
 (0.5, (('99499590', '2022-01-08'), ['100701298'])),
 (0.5, (('9947174', '2022-01-08'), ['100701298'])),
 (0.5, (('9946603', '2022-01-08'), ['100701298'])),
 (0.5, (('99459802', '2022-01-08'), ['100701298'])),
 (0.5, (('99360929', '2022-01-08'), ['100701298'])),
 (0.5, (('99341435', '2022-01-08'), ['100701298'])),
 (0.5, (('99065953', '2022-01-08'), ['100701298'])),
 (0.5, (('98989470', '2022-01-08'), ['100701298'])),
 (0.5, (('98987107', '2022-01-08'), ['100701298'])),
 (0.5, (('98967124', '2022-01-08'), ['100701298'])),
 (0.5, (('98902509', '2022-01-08'), ['100701298'])),
 (0.5, (('98850992', '2022-01-08'), ['100701298'])),
 (0.5, (('9868734', '2022-01-08'), ['100701298'])),
 (0.5, (('98678237', '2022-01-08'), ['100701298'])),
 (0.5, (('9864580', '2022-01-08'), ['100701298'])),
 (0.5, (('98615835', '2022-01-08'), ['100701298'])),
```

```
[(0.35355, (('99940317', '2022-01-08'), ['100701298', '22010128'])),
 (0.35355, (('9990496', '2022-01-08'), ['100701298', '100701156'])),
 (0.35355, (('9984587', '2022-01-08'), ['100701298', '22010125'])),
 (0.35355, (('99754917', '2022-01-08'), ['100701298', '22010059'])),
 (0.35355, (('99701280', '2022-01-08'), ['100701298', '22010109'])),
 (0.35355, (('99666385', '2022-01-08'), ['100701298', '100701156'])),
 (0.35355, (('99648337', '2022-01-08'), ['22010122', '100701298'])),
 (0.35355, (('99622156', '2022-01-08'), ['100701298', '100701119'])),
 (0.35355, (('99555580', '2022-01-08'), ['100701298', '22010110'])),
 (0.35355, (('9948758', '2022-01-08'), ['100701298', '22010125'])),
 (0.35355, (('9947130', '2022-01-08'), ['100701298', '22010127'])),
 (0.35355, (('9944299', '2022-01-08'), ['100701298', '22010122'])),
 (0.35355, (('99357744', '2022-01-08'), ['100701298', '22010047'])),
 (0.35355, (('9933056', '2022-01-08'), ['900195', '100700857'])),
 (0.35355, (('99163653', '2022-01-08'), ['100701298', '22010110'])),
 (0.35355, (('9903907', '2022-01-08'), ['100701298', '22010110'])),
```

```
(0.224,
 (('11288858', '2022-01-08'),
  ['900167', '100700857', '900195', '100700894', '100700979'])),
(0.224,
 (('11046089', '2022-01-08'),
  ['22010109', '631347', '22010139', '22010138', '100701298'])),
(0.224,
 (('11029416', '2022-01-08'),
  ['100701297', '22010122', '100701298', '22010111', '100701156'])),
(0.224,
 (('10954671', '2022-01-08'),
  ['100700881', '631758', '900214', '22010122', '100701298'])),
(0.224,
 (('10907087', '2022-01-08'),
  ['100701297', '22010112', '100701298', '22010125', '100701156'])),
(0.224,
 (('10729005', '2022-01-08'),
  ['22010134', '22010113', '22010139', '22010112', '100701298'])),
(0.224,
 (('10643018', '2022-01-08'),
  ['100701297', '100701293', '100701298', '22010120', '22010127'])),
```

```
[(0.204,
  (('98617349', '2022-01-08'),
   ['900167', '900235', '900195', '100700979', '900222', '900273'])),
 (0.204,
  (('9830777', '2022-01-08'),
   ['22010087', '22010096', '22010099', '22010122', '100701298', '22010095'])),
 (0.204,
  (('9752615', '2022-01-08'),
   ['900167', '100700804', '100700834', '100700978', '900195', '900164'])),
 (0.204,
  (('9480696', '2022-01-08'),
   ['22010126',
    '100701297',
    '22010122',
    '100701298',
    '22010125',
    '22010054'])),
```

```
(0.189,
 (('9992889', '2022-01-08'),
  ['900256',
   '900215',
   '900253',
   '100701059',
   '169',
   '100701298',
   '100701156'])),
(0.189,
 (('9853369', '2022-01-08'),
  ['22010060',
   '100701297',
   '100701293',
   '22010110',
   '22010138',
   '100701298',
   '22010137'])),
```