به نام خدا

دانشگاه صنعتی امیرکبیر

(پلی تکنیک تهران)

دانشکده مهندسی کامپیوتر

گزارش تمرین دوم درس کلان داده‌ها

استاد درس: دکتر مصطفی حقیر چهرقانی

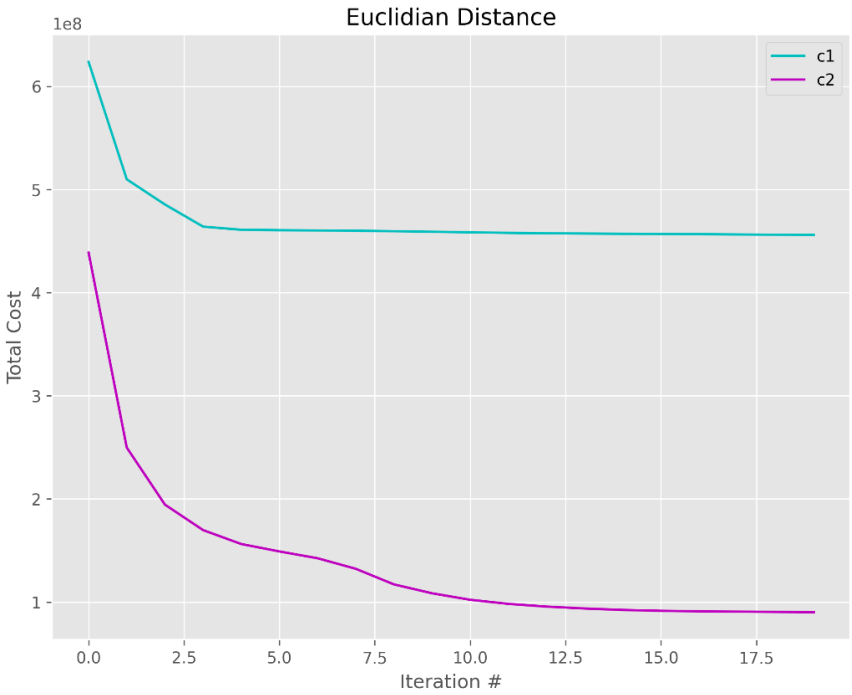
دانشجو: فاطمه غلام‌زاده

۹۹۱۳۱۰۰۳

نیم سال دوم ۱۳۹۹-۱۴۰۰

سوال ۱

الف) نمودار فاصله اقلیدسی برای تکرارهای ۱ تا ۲۰ و برای C1 و C2 رسم شده است:



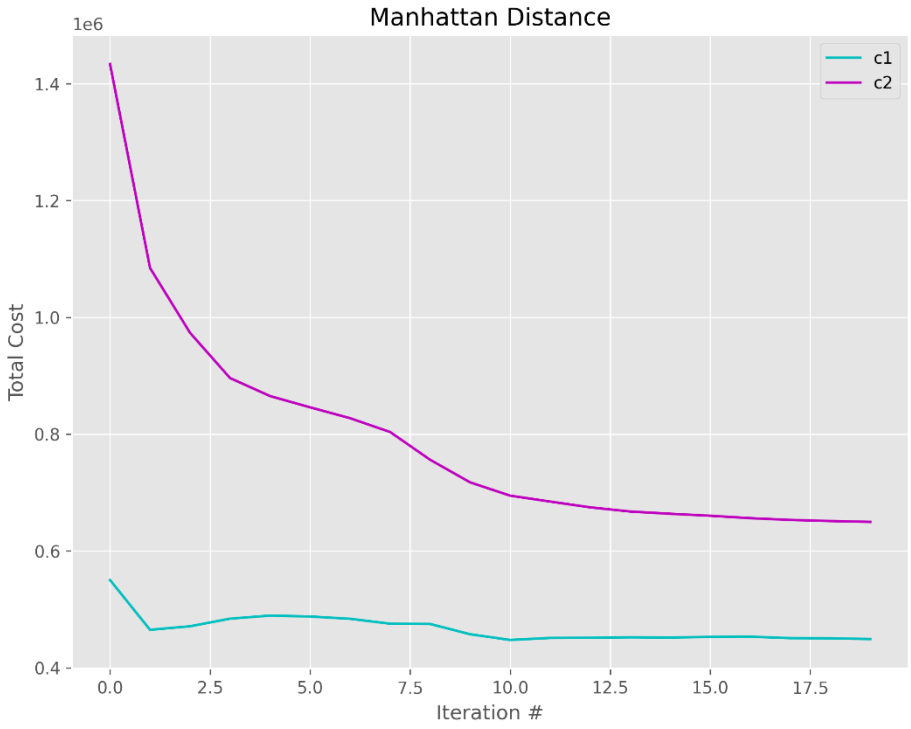
ب)

درصد تغییر هزینه الگوریتم بین اجرای صفرم و دهم، با فاصله اقلیدسی:

initialization	*Percentage of change after 10 iterations*
C1	26.5%
C2	76.7%

نتیجه: همانطور که مشاهده می‌شود درصد تغییر هزینه الگوریتم بین اجرای صفرم و دهم با در نظر گرفتن C2 به عنوان مراکز اولیه، بیشتر از C1 شده است. بنابراین مقداردهی اولیه C2 بهتر بوده است چون هزینه را بیشتر کاهش داده است. می‌توان این نتیجه را گرفت که انتخاب مراکز اولیه خوشه بندی به صورتی که این مراکز تا حد امکان از هم فاصله داشته باشند انتخاب مناسبی است و بهتر از انتخاب رندوم مراکز اولیه عمل می‌کند. چون در این صورت خوشه ها از هم فاصله دارند و هم پوشانی کمتری ایجاد می‌شود.

ج) نمودار فاصله منهتن برای تکرارهای ۱ تا ۲۰ و برای C1 و C2 رسم شده است:



د)

درصد تغییر هزینه الگوریتم بین اجرای صفرم و دهم، با فاصله منهتن :

initialization	*Percentage of change after 10 iterations*
C1	18.7%
C2	51.6%

نتیجه: همانطور که مشاهده می‌شود باز هم درصد تغییر هزینه الگوریتم بین اجرای صفرم و دهم با در نظر گرفتن C2 به عنوان مراکز اولیه، بیشتر از C1 شده است. بنابراین مقداردهی اولیه C2 بهتر بوده است چون هزینه را بیشتر کاهش داده است. مقدار بهبود نسبت به همین بهبود برای حالتی که تابع هزینه اقلیدسی بود کمتر است. چون فایل C2 با در نظر گرفتن فاصله اقلیدسی مراکز خوشه ها را از هم فاصله دار انتخاب کرده است و ما اینجا از تابع هزینه منهتن استفاده کردیم. اما باز هم مشاهده می شود که بهبود داریم چون وقتی فاصله اقلیدسی زیاد می‌شود فاصله منهتن هم زیاد می‌شود.

سوال ۲

الف)

برای هر مقدار ویژه λ از MM^T داریم:

$$MM^T v = v\lambda$$

که در آن v بردار ویژه متناظر با مقدار ویژه λ است. $(\lambda \neq 0)$

هر دو طرف رابطه را از سمت چپ در M^T ضرب می‌کنیم:

$$M^T M M^T v = M^T v \lambda \xrightarrow[\text{شرکت‌پذیری}]{\text{خاصیت}} \boxed{M^T M (M^T v) = (M^T v)\lambda}$$

بنابراین طبق این رابطه، λ یک مقدار ویژه از $M^T M$ هم هست و بردار ویژه‌ی متناظر با این مقدار ویژه، برابر با $M^T v$ است. بنابراین بردارهای ویژه‌شان متعامد هم برابر نیستند مگر اینکه $M^T = I$ باشد که در اینصورت $M^T v = v$.

ب)

$$M = U \Sigma V^T$$

$$M^T = (U \Sigma V^T)^T = (V^T)^T \Sigma^T U^T = V \Sigma^T U^T$$

Σ یک ماتریس قطری است پس داریم: $\Sigma^T = \Sigma$

$$\Rightarrow M^T = V \Sigma U^T$$

$$M^T M = V \Sigma \underbrace{U^T U} \Sigma V^T = \boxed{V \Sigma^2 V^T}$$

ماتریس U، orthonormal است

پس: $U^T U = I$

ج)

```
U
[[-0.27854301  0.5       ]
 [-0.27854301 -0.5       ]
 [-0.64993368  0.5       ]
 [-0.64993368 -0.5       ]]

 Sigma
 [7.61577311 1.41421356]

Vh
[[-0.70710678 -0.70710678]
 [-0.70710678  0.70710678]]
```

مقادیر ویژه $M^T M$ به صورت مرتب شده :

```
eigenValues
[58.  2.]
```

لیست بردارهای ویژه $M^T M$:

```
eigenVectors
[[ 0.70710678 -0.70710678]
 [ 0.70710678  0.70710678]]
```

نتیجه اول: ستون‌های ماتریس eigenvectors که در واقع همان بردارهای ویژه متناظر با مقادیر ویژه ماتریس $M^T M$ هستند همان ستون‌های V هستند که در 1- و 1 ضرب شده اند.

نتیجه دوم: هر singular value از ماتریس M ریشه‌ی دوم یک مقدارویژه از $M^T M$ است.

سوال ۳

الف)

Tii: نشان دهنده تعداد اقلامی است که user i می‌پسندد که در گراف دوبخشی کاربر-اقلام نشان دهنده درجه خروجی راس مربوط به user i است.

Tij: تعداد اقلامی که user i و user j هر دو می‌پسندند که در گراف دوبخشی کاربر-اقلام تعداد همسایه‌های مشترک user i و user j این را نشان می‌دهد.

ب)

اگر یک ماتریس قطری از سمت راست در یک ماتریس ضرب شود باعث scale شدن ستون‌های آن ماتریس می‌شود بنابراین برای نرمال کردن ماتریس اقلام، R را در $Q^{-1/2}$ ضرب می‌کنیم: $RQ^{-1/2}$

مشابه قسمت الف، ماتریس شباهت اقلام را تعریف می‌کنیم:

$$S_I = (RQ^{-1/2})^T RQ^{-1/2} = \boxed{Q^{-1/2} R^T R Q^{-1/2}}$$

هم‌چنین مشابه S_I، S_U را داریم:

$$S_U = (R^T P^{-1/2})^T R^T P^{-1/2} = \boxed{P^{-1/2} R R^T P^{-1/2}}$$

ج)

روش کاربر-کاربر:

$$\Gamma_{us} = \sum_{x \in users} Cos\text{-}Sim(x,u) * R_{xs}$$

$$\Gamma = S_U R = P^{-1/2} R R^T P^{-1/2} R$$

روش اقلام-اقلام:

$$\Gamma_{us} = \sum_{x \in items} R_{ux} * Cos\text{-}Sim(x,s)$$

$$\Gamma = R S_I = R Q^{-1/2} R^T R Q^{-1/2}$$

د)

۱-

ماتریس P :

```
[[35.  0.  0. ...  0.  0.  0.]
 [ 0. 26.  0. ...  0.  0.  0.]
 [ 0.  0. 44. ...  0.  0.  0.]
 ...
 [ 0.  0.  0. ...  5.  0.  0.]
 [ 0.  0.  0. ...  0. 30.  0.]
 [ 0.  0.  0. ...  0.  0. 19.]]
```

ماتریس Q:

```
[[1089.    0.    0. ...    0.    0.    0.]
 [   0. 3350.    0. ...    0.    0.    0.]
 [   0.    0. 3187. ...    0.    0.    0.]
 ...
 [   0.    0.    0. ...  358.    0.    0.]
 [   0.    0.    0. ...    0.  294.    0.]
 [   0.    0.    0. ...    0.    0.    6.]]
```

۲-

```
[(908.480053476128, 96, '"FOX 28 News at 10pm"'),
(861.1759992873301, 74, '"Family Guy"'),
(827.6012954743582, 45, '"2009 NCAA Basketball Tournament"'),
(784.7819589039739, 60, '"NBC 4 at Eleven"'),
(757.6011181024228, 9, '"Two and a Half Men"')]
```

۳-

```
[(31.364701678342396, 96, '"FOX 28 News at 10pm"'),
(30.001141798877764, 74, '"Family Guy"'),
(29.396797773402554, 60, '"NBC 4 at Eleven"'),
(29.22700156150048, 45, '"2009 NCAA Basketball Tournament"'),
(28.97127767405556, 82, '"Access Hollywood"')]
```