به نام خدا

دانشگاه تهران

دانشکدگان فنی

دانشکده مهندسی برق و کامپیوتر

درس داده کاوی

تمرین سوم CA3

محمد ناصری

۸۱۰۱۰۰۴۸۶

خرداد ماه ۱۴۰۱

سوال اول

الف

GT Cluster	Entertainment	Financial	Foreign	Metro	National	Sport	Total
#1	1	1	0	11	4	676	**693**
#2	27	89	333	827	253	33	**1562**
#3	326	465	8	105	16	29	**949**
Total	**354**	**555**	**341**	**943**	**273**	**738**	**3204**

مقدار انتروپی کل خوشه‌بندی از فرمول زیر قابل محاسبه است:

$$H(\Omega) = \sum H(\omega) * \frac{N_\omega}{N}$$

که در آن داریم:

- $\Omega = \{w_1, w_2, \ldots.\}$ مجموعه کلاسترهاست
- مقدار $H(w)$ انتروپی یک کلاستر است
- N_w تعداد نقطه در کلاستر w است
- N تعداد کل نقطه هاست.

برای بدست آوردن انتروپی یک کلاستر نیز داریم:

$$H(w) = -\sum \underbrace{\frac{|w_c|}{n_w}}_{P(w\_c)} * \log_2 \frac{|w_c|}{n_w}$$

n_w تعداد نقاط در خوشه w است

$|w\_c|$ تعداد نقاط طبقه بندی شده به عنوان c در خوشه w است

با تعاریف ذکر شده در بالا بدست می‌آید:

$$\mathrm{H}(1) = \left(\left(\frac{1}{693}\right) * \log\left(\frac{1}{693}\right)\right) + \left(\left(\frac{1}{693}\right) * \log\left(\frac{1}{693}\right)\right) + \left(\left(\frac{11}{693}\right) * \log\left(\frac{11}{693}\right)\right) + \left(\left(\frac{4}{693}\right) * \log\left(\frac{4}{693}\right)\right) + \left(\left(\frac{676}{693}\right) * \log\left(\frac{676}{693}\right)\right)\Bigg) = 0.14$$

$$H(2) = \Bigg(\left(\left(\frac{27}{1562}\right) * \log\left(\frac{27}{1562}\right)\right) + \left(\left(\frac{89}{1562}\right) * \log\left(\frac{89}{1562}\right)\right) + \left(\left(\frac{333}{1562}\right) * \log\left(\frac{333}{1562}\right)\right) + \left(\left(\frac{827}{1562}\right) * \log\left(\frac{827}{1562}\right)\right) + \left(\left(\frac{253}{1562}\right) * \log\left(\frac{253}{1562}\right)\right) + \left(\left(\frac{33}{1562}\right) * \log\left(\frac{33}{1562}\right)\right)\Bigg) = 1.27$$

$$H(3) = \Bigg(\left(\left(\frac{326}{949}\right) * \log\left(\frac{326}{949}\right)\right) + \left(\left(\frac{465}{949}\right) * \log\left(\frac{465}{949}\right)\right) + \left(\left(\frac{8}{949}\right) * \log\left(\frac{8}{949}\right)\right) + \left(\left(\frac{105}{949}\right) * \log\left(\frac{105}{949}\right)\right) + \left(\left(\frac{16}{949}\right) * \log\left(\frac{16}{949}\right)\right) + \left(\left(\frac{29}{949}\right) * \log\left(\frac{29}{949}\right)\right)\Bigg) = 1.17$$

$$\mathrm{H}(\mathrm{T}) = \left(\left(\frac{693}{3204}\right) * 0.14\right) + \left(\left(\frac{1562}{3204}\right) * 1.27\right) + \left(\left(\frac{949}{3204}\right) * 1.17\right) = 0.99$$

ب)

مقدار purity برای یک خوشه بندی از فرمول زیر محاسبه میشود:

$$Purity = \frac{1}{N}\sum_{m\epsilon M}\max_{d\epsilon D}|m \cap d|$$

با توجه به فرمول بالا داریم:

$$Purity = \frac{676 + 827 + 465}{3204} = 0.614$$

ج)

هرکدام از مقادیر precision و recall از فرمولهای زیر محاسبه میشود:

$$Precision_i = 1/n_i \max_j(n_{ij})$$

$$Precision(1) = \frac{676}{693} = 0.975$$
$$Precision(2) = \frac{827}{1562} = 0.529$$
$$Precision(3) = \frac{465}{949} = 0.489$$

$$Recall_i = \frac{n_{ij}}{|Tji|}$$

$$Recall(1) = \frac{676}{738} = 0.915$$
$$Recall(2) = \frac{827}{943} = 0.876$$
$$.Recall(3) = \frac{465}{555} = 0.837$$

برای بدست آوردن معیار F از فرمول زیر استفاده میشود:

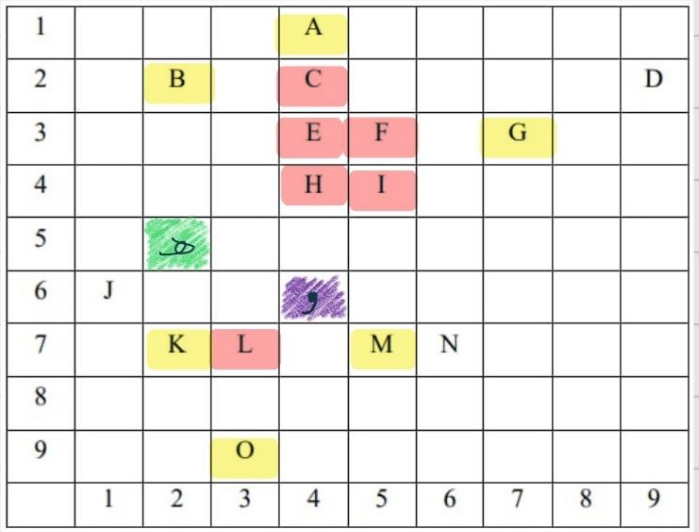
$$F(i,j) = \frac{2 * precision(i,j) * recall(i,j)}{precision(i,j) + recall(i,j)}$$

مانند قبل تنها موارد top را محاسبه میکنیم:

$$F(1) = \frac{(2 * 0.975 * 0.915)}{(0.975 + 0.915)} = 0.944$$
$$F(2) = \frac{(2 * 0.529 * 0.876)}{(0.529 + 0.876)} = 0.66$$
$$F(3) = \frac{(2 * 0.489 * 0.837)}{(0.489 + 0.837)} = 0.618$$

سوال دوم

$\begin{bmatrix} eps = 2,1 \\ MinPts = 4 \end{bmatrix}$



الف) با توجه به مقادیر eps و MinPts نقاط هسته نقاطی هستند که در فاصله منهتن (2.1) خود تعداد 4 همسایه حد اقل داشته باشند ← {C, E, H, F, I و L}

ب) نقاطی که در فاصله منهتن (2.1) از نقاط هسته باشند ولی کمتر از 4 همسایه داشته باشند.

↳ {A, B, K, M, G, O}

ج) نقاطی از I مستقیماً قابل دسترس هستند که در فاصله منهتن (2.1) باشند و نقطه هسته باشند

↳ {F, E, H}

د) برای مستقیماً قابل دسترس بودن باید نقطه مد نظر هسته باشد که M هسته نیست

ه) خانه (2,5) که با رنگ سبز مشخص شده است با H و H دو فاصله دارد و می تواند هسته باشد

و) خانه (4,6) که با رنگ بنفش مشخص شده است با مجاورت با H و H و H نقطه هسته شده و خوشه ها را ترکیب می کند

سوال سوم

بسیاری از تکنیک‌های خوشه‌بندی سلسله مراتبی انباشته‌ای بر روی یک رویکرد واحد هستند: با شروع با نقاط منفرد به عنوان خوشه، به طور متوالی دو نزدیکترین خوشه را ادغام کنید تا فقط یک خوشه باقی بماندالگوریتم مربوطه به صورت زیر است:

1. محاسبه ماتریس مجاورت
2. نزدیکترین دو خوشه را ادغام کنید.
3. ماتریس مجاورت را با فاصله بین خوشه جدید و خوشه های اصلی به روز کنید
4. اینکار را ادامه دهید تا زمانی که فقط یک خوشه باقی بماند.

Single-link: New distance equals Min of distances

	P1	P2	P3	P4	P5
P1	0.00	0.10	0.41	0.55	0.35
P2	0.10	0.00	0.64	0.47	0.98
P3	0.41	0.64	0.00	0.44	0.85
P4	0.55	0.47	0.44	0.00	0.76
P5	0.35	0.98	0.85	0.76	0.00

Min distance = p1 → p2

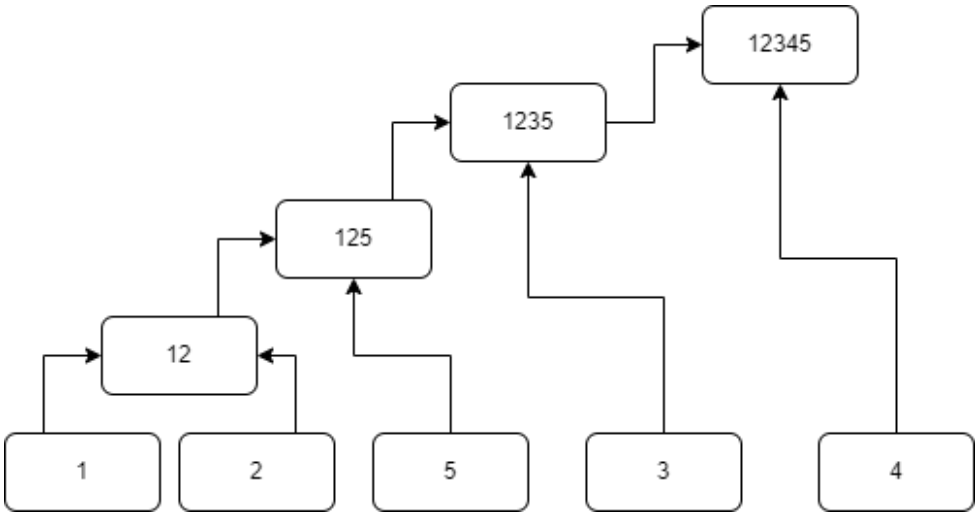
	P12	P3	P4	P5
P12	0.00	0.64	0.55	0.35
P3	0.41	0.00	0.44	0.85
P4	0.55	0.44	0.00	0.76
P5	0.35	0.85	0.76	0.00

Min = p12 → p5

	P125	P3	P4
P125	0.00	0.41	0.55
P3	0.41	0.00	0.44
P4	0.55	0.44	0.00

Min = p123 → p3

	P1235	P4
P1235	0.00	0.44
P4	0.44	0.00



Complete-link: New distance equals maximum of distances

	P1	P2	P3	P4	P5
P1	0.00	0.10	0.41	0.55	0.35
P2	0.10	0.00	0.64	0.47	0.98
P3	0.41	0.64	0.00	0.44	0.85
P4	0.55	0.47	0.44	0.00	0.76
P5	0.35	0.98	0.85	0.76	0.00

Min distance = p1 → p2

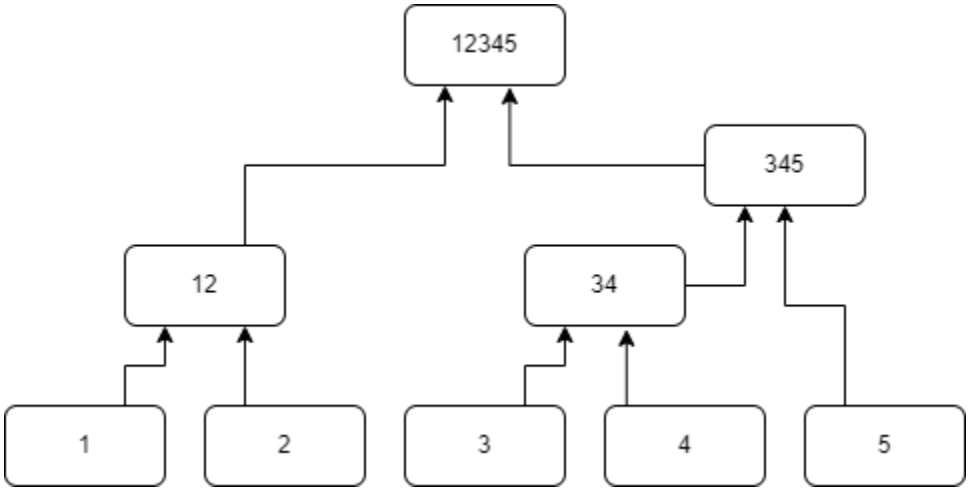
	P12	P3	P4	P5
P12	0.00	0.64	0.98	0.98
P3	0.64	0.00	0.44	0.85
P4	0.98	0.44	0.00	0.76
P5	0.98	0.85	0.76	0.00

Min distance = p3→p4

	P12	P34	P5
P12	0.00	0.98	0.98
P34	0.98	0.00	0.85
P5	0.98	0.85	0.00

Min distance = p34→p5

	P12	P34
P12	0.00	0.98
P34	0.98	0.00



سوال چهارم

در قدم اول پس از مشاهده و بررسی دیتاست و مطالعه توضیحات، ستون‌های admission_type_idو discharge_disposition_idو admission_source_id را با مقادیر categorical خود جایگزین میکنیم تا بتوانیم مقادیر null آنها را تشخیص داده و جایگزین کنیم. در این مساله نیاز نداریم تا از One-hot برای دسته‌بندی ها استفاده کنیم چون مساله unsupervised و کلاسترینگ است و فاصله نقاط با یکدیگر مدنظر است.

همچنین برای مقادیر string موجود در ۳ ستون diag نیز null قرار میدهیم.

سپس به جای همه علامت سوال‌ها و عبارت‌های NONE مقدار null قرار میدهیم تا بعدا آنهارا مدیریت کنیم.

پس از این مراحل، با ساخت جدولی میزان مقادیر null هر ستون را محاسبه و نمایش میدهیم. با مشاهده جدول تعدادی ستون دارای تعداد بسیار بالایی Null هستند که آنهارا از دیتاست حذف میکنیم.

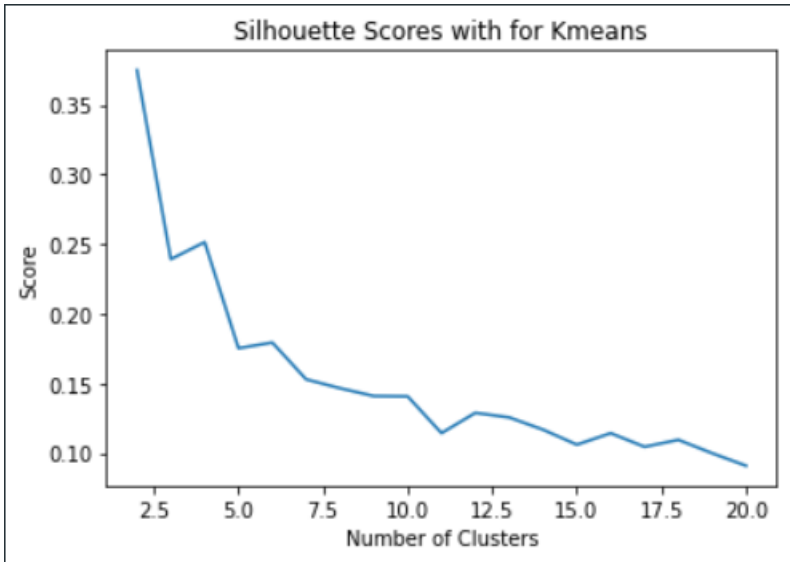
پس از آن برای پر کردن Null ها برای داده‌های categorical از مد آنها و برای داده‌های عددی از میانگین استفاده میکنیم.

با توجه به دیتاست بدست امده که خالی از Null است، بر روی ستون‌های دیتاست برای دسته‌بندیها عمل encoding و برای عددی ها عمل نرمالسازی را انجام میدهیم.

تا اینجا مراحل پاکسازی انجام شده و آماده اجرای الگوریتم ها هستیم.

قبل از هرکاری به دلیل حجم بالای دیتاست، از آن نمونه گرفته و الگوریتم ها را بر روی نمونه اجرا میکنیم. (با توجه به تست انجام شده روی kmeans مقادیر بدست امده بر روی نمونه تقریبا با کل دیتاست برابری میکند)

در قدم اول الگوریتم kmean را به ازای تعداد کلاستر ۲ تا ۲۰ بر روی دیتاست اعمال کرده و نتایج معیار silhouette را برای این مقادیر Plot میکنیم.



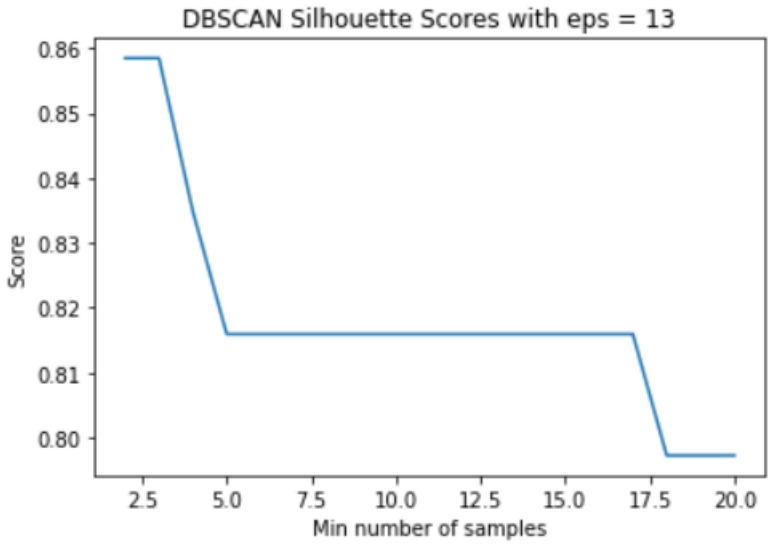
مشاهده میشود که بهترین تعداد کلاستر با توجه به این معیار برابر ۲ میباشد.

امتیاز Silhouette برای ارزیابی کیفیت خوشه‌های ایجاد شده با استفاده از الگوریتم‌های خوشه‌بندی مانند K-Means از نظر میزان خوشه‌بندی نمونه‌ها با نمونه‌های مشابه دیگر استفاده می‌شود. امتیاز Silhouette برای هر نمونه از خوشه های مختلف محاسبه می شود.

بهترین مقدار 1 و بدترین مقدار -1 است. مقادیر نزدیک به 0 نشان دهنده همپوشانی خوشه ها است. مقادیر منفی به طور کلی نشان می دهد که یک نمونه به خوشه اشتباهی اختصاص داده شده است، زیرا یک خوشه متفاوت شبیه تر است.

در مرحله بعد برای الگوریتم DBSCAN و با مقادیر مشخص شده تست کرده و نمودار میکشیم.

بهترین نمودار این الگوریتم به صورت زیر است:



این نشان‌دهنده این است که الگوریتم dbscan خوشه‌های مجزاتر و بهتری را نسبت به kmean ارائه داده است و بهتر از الگوریتم پیشین عمل کرده. ولی از مشکلات این لگوریتم فضای محاسباتی بالای آن است به طوریکه بنده در این آزمایش موفق به اجرای این الگوریتم بر روی تمام دیتاست نشدیم و به نتایج نمونه برداری اکتفا میکنیم.

در کتابخانه sklearn پیشنهاد شده به جای dbscan از optic که الگوریتمی مشابه dbscan دارد استفاده شود (که البته جوابهای متفاوتی خواهد داشت) اما با توجه به منابع کم موفق به اجرای این الگوریتم نیز نشدیم.