به نام خدا

دانشگاه تهران

دانشکده مهندسی برق و کامپیوتر

درس داده کاوی پیشرفته

تمرین پنجم

نام و نام خانوادگی	عرفان شهابی
شماره دانشجویی	۸۱۰۱۰۳۱۶۶
تاریخ ارسال گزارش	۱۴۰۴.۰۳.۰۸

فهرست

نمودارها

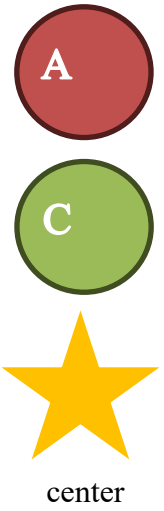
جدول‌ها

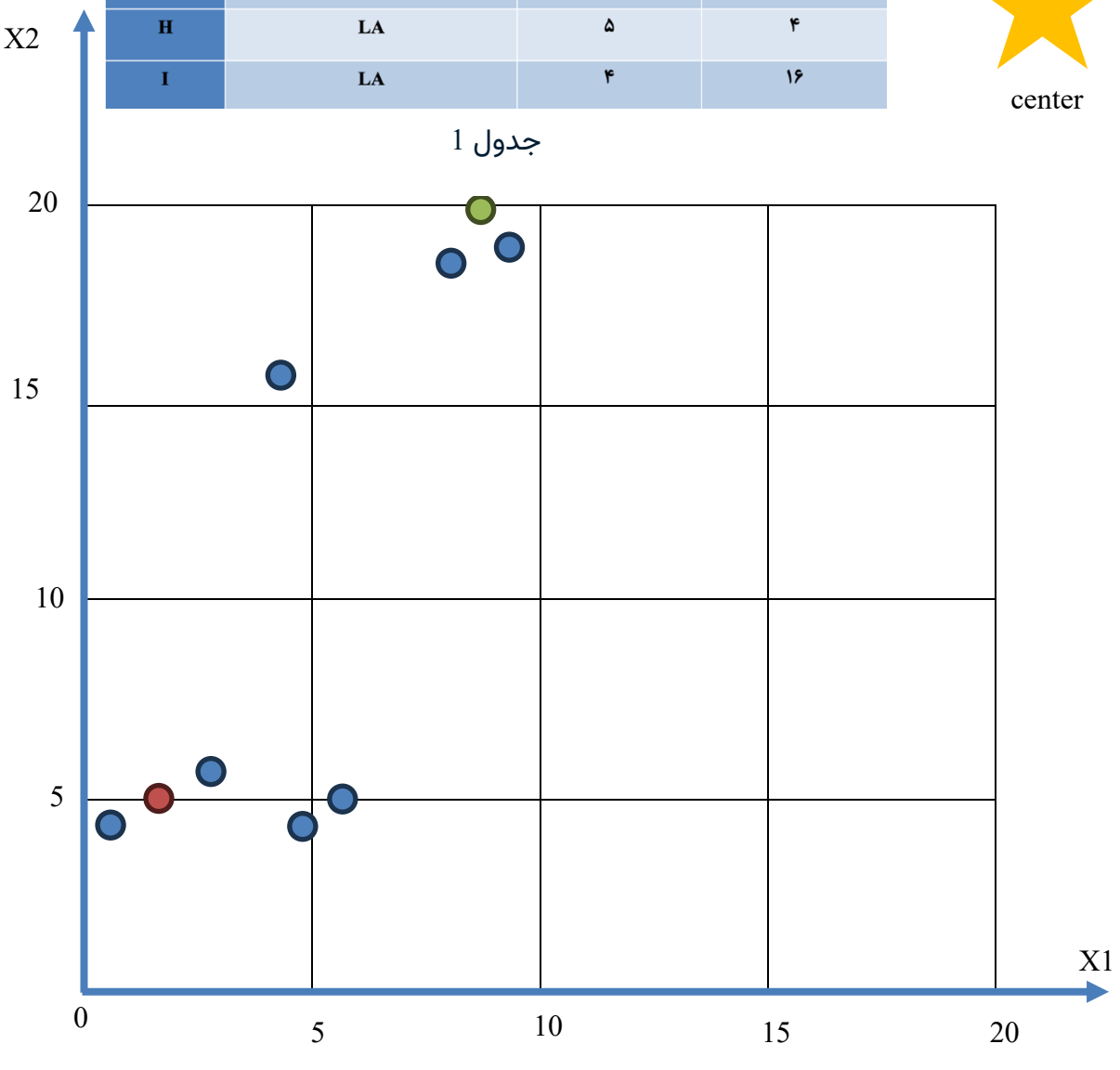
سوال ۱

- 1

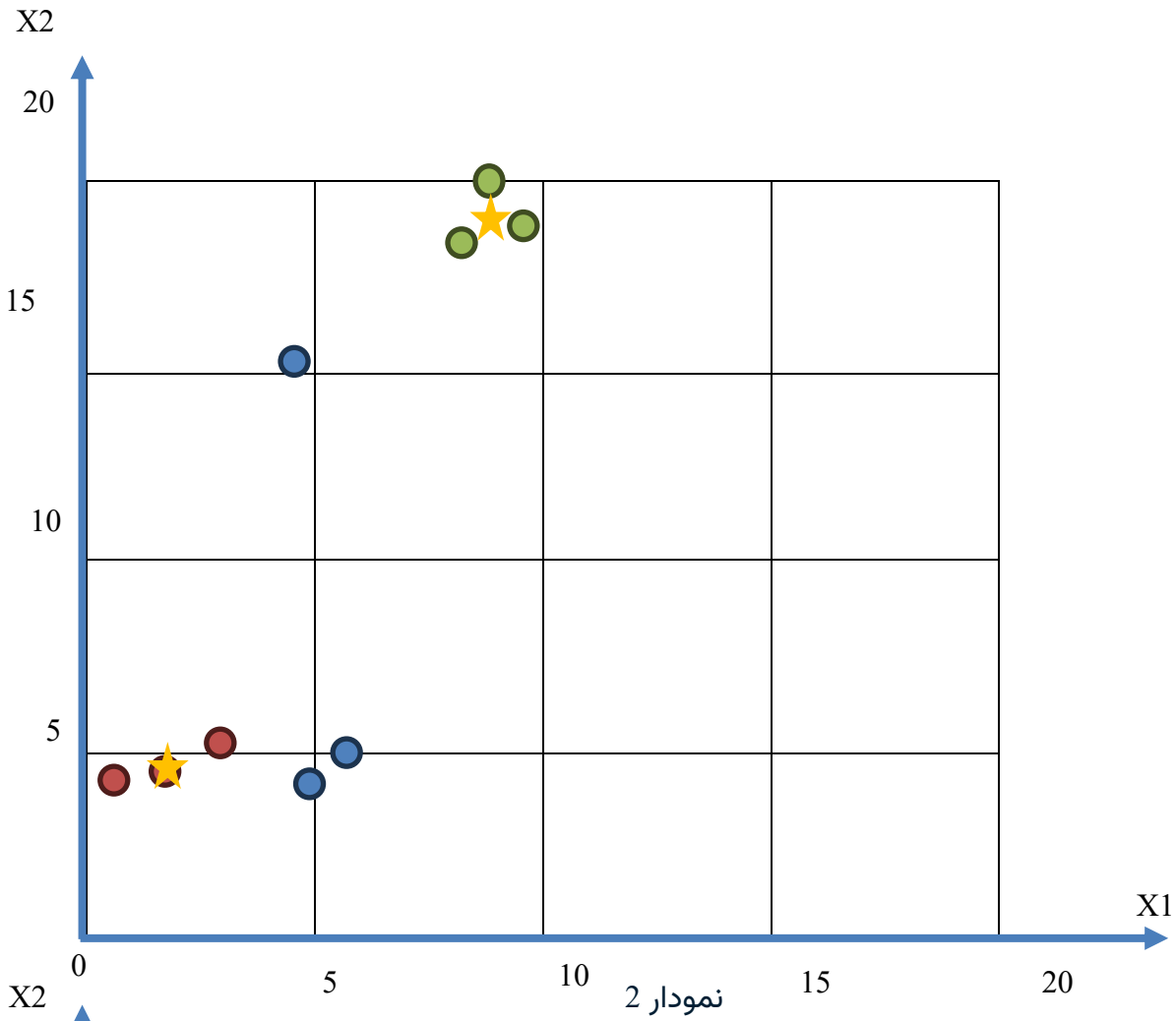
مشتری	Ground-truth	X1	X2
A	LA	۲	۵
B	LA	۳	۶
C	HA	۸	۲۰
D	HA	۷	۱۸
E	HA	۹	۱۹
F	LA	۱	۴
G	HA	۶	۵
H	LA	۵	۴
I	LA	۴	۱۶

جدول 1

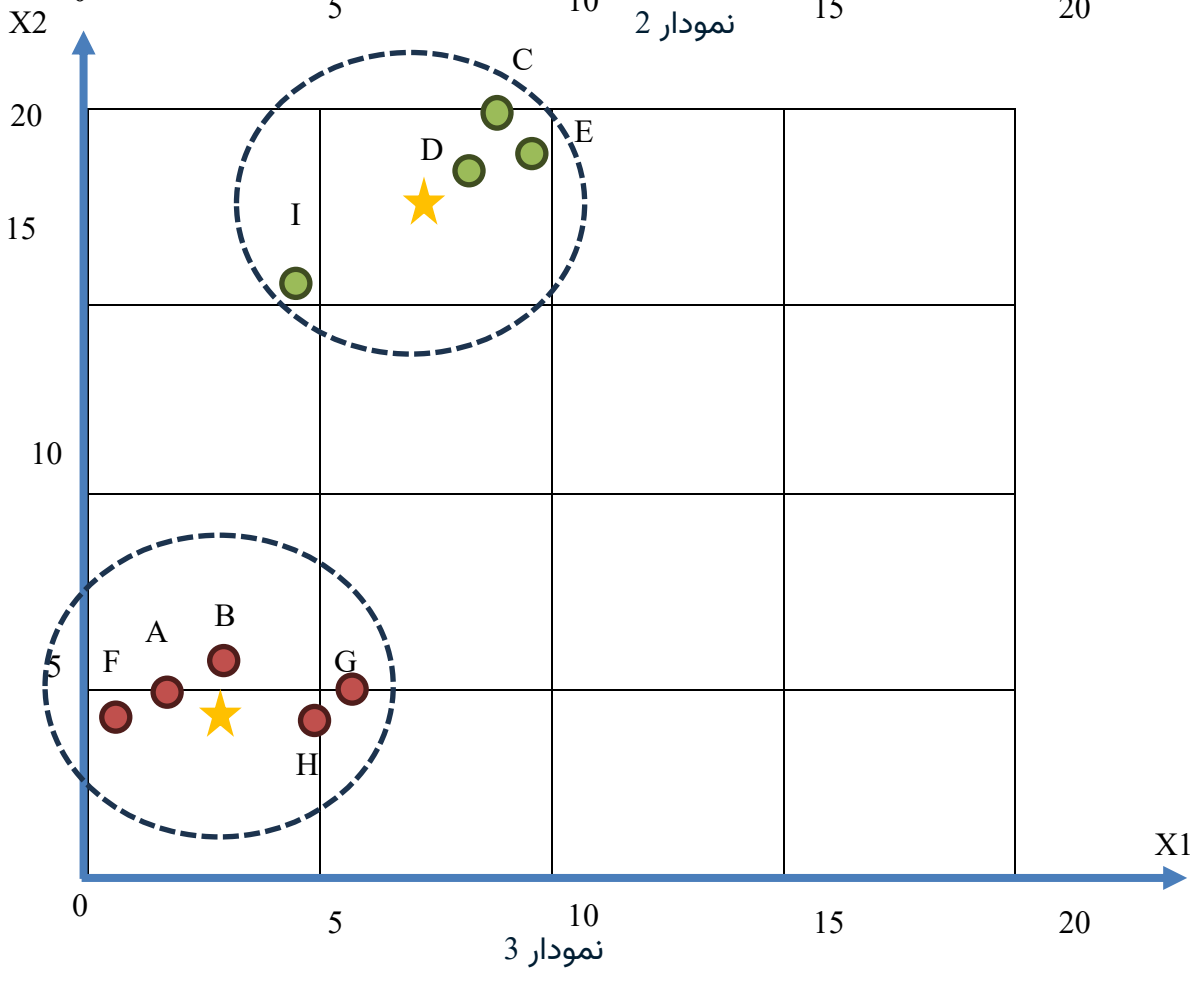




نمودار 1



نمودار 2



نمودار 3

LA: A, B, F, G, H

HA: C, D, E, I

LA (Ground Truth): A, B, F, H, I

HA (Ground Truth): C, D, E, G

Clustering Ground T	LA	HA
LA	A, B, F, H	I
HA	G	C, D, E

Clustering Ground T	C1	C2	Sum of Classes
G1	4	1	5
G2	1	3	4
Sum of Clusters	5	4	9

$$Purity = \frac{1}{9}(4+3) = \frac{7}{9}$$

$$I(C,G) = -\sum\sum P_{ij}\log\left(\frac{P_{ij}}{P_{Ci}P_{Gj}}\right)$$

$$NMI = \frac{I(C,G)}{\sqrt{(H(G)\,H(C))}}$$

$$p_{ij} = \frac{c_i \cap c_j}{n}, p_{ci} = \frac{C_i}{n}, p_{Gj} = \frac{G_j}{n}$$

$$P_{LA} = \frac{5}{9}, P_{HA} = \frac{4}{9}, P_{GLA} = \frac{5}{9}, P_{GHA} = \frac{4}{9}$$

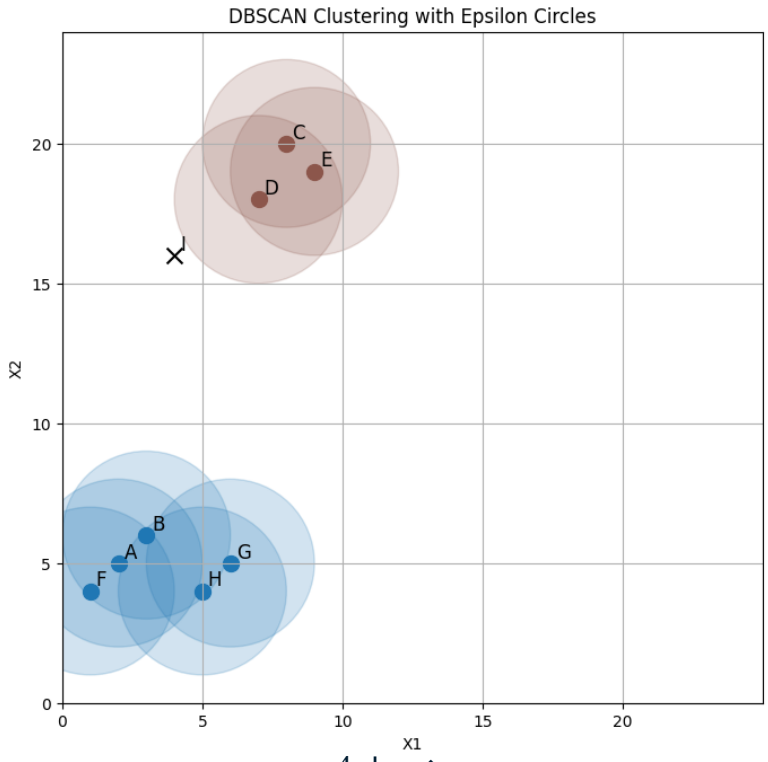
$$I = \frac{4}{9}\times\lg\left(\frac{\frac{4}{9}}{\frac{5}{9}\times\frac{5}{9}}\right) + \frac{1}{9}\times\lg\left(\frac{\frac{1}{9}}{\frac{4}{9}\times\frac{5}{9}}\right) + \frac{1}{9}\times\lg\left(\frac{\frac{1}{9}}{\frac{5}{9}\times\frac{4}{9}}\right) + \frac{3}{9}\times\lg\left(\frac{\frac{3}{9}}{\frac{4}{9}\times\frac{4}{9}}\right) = 0.23$$

$$for\ Classes: H(G) = -\left(\frac{5}{9}\lg\left(\frac{5}{9}\right) + \frac{4}{9}\lg\left(\frac{4}{9}\right)\right) = 0.991$$

$$for\ Clusters: H(C) = -\left(\frac{5}{9}\lg\left(\frac{5}{9}\right) + \frac{4}{9}\lg\left(\frac{4}{9}\right)\right) = 0.991$$

$$NMI = \frac{I}{\sqrt{(H(G), H(c))}} = \frac{0.23}{\sqrt{0.982}} = 0.232$$

- 3



نمودار 4

Clustering Ground T	LA	HA
LA	A, B, F, H	-
HA	G	C, D, E

Clustering Ground T	C1	C2	Sum of Classes
G1	4	0	4
G2	1	3	4
Sum of Clusters	5	3	8

LA: A, B, F, G, H

HA: C, D, E

LA (Ground Truth): A, B, F, H, I

HA (Ground Truth): C, D, E, G

$$Purity = \frac{1}{8}(4 + 3) = \frac{7}{8}$$

$$I(C, G) = -\sum\sum P_{ij} \log\left(\frac{P_{ij}}{P_{Ci} P_{Gj}}\right)$$

$$NMI = \frac{I(C,G)}{\sqrt{(H(G)\,H(C))}}$$

$$p_{ij} = \frac{c_i \cap c_j}{n}, p_{ci} = \frac{C_i}{n}, p_{Gj} = \frac{G_j}{n}$$

$$P_{LA} = \frac{5}{9}, P_{HA} = \frac{3}{9}, P_{GLA} = \frac{5}{9}, P_{GHA} = \frac{4}{9}$$

$$I = \frac{4}{9} \times \lg\left(\frac{\frac{4}{9}}{\frac{5}{9} \times \frac{5}{9}}\right) + \frac{0}{9} \times \lg\left(\frac{0}{\frac{3}{9} \times \frac{5}{9}}\right) + \frac{1}{9} \times \lg\left(\frac{\frac{1}{9}}{\frac{5}{9} \times \frac{4}{9}}\right) + \frac{3}{9} \times \lg\left(\frac{\frac{3}{9}}{\frac{3}{9} \times \frac{4}{9}}\right) = 0.49$$

$$for\ Classes: H(G) = -\left(\frac{5}{9}\lg\left(\frac{5}{9}\right) + \frac{4}{9}\lg\left(\frac{4}{9}\right)\right) = 0.991$$

$$for\ Clusters: H(C) = -\left(\frac{5}{8}\lg\left(\frac{5}{8}\right) + \frac{4}{8}\lg\left(\frac{4}{8}\right)\right) = 0.923$$

$$NMI = \frac{I}{\sqrt{(H(G), H(c))}} = \frac{0.49}{\sqrt{0.915}} = 0.52$$

نتایج purity و NMI برای این دو روش خوشه بندی به شرح زیر است:

KMeans:

Purity = 0.78, NMI = 0.232

LA: A, B, F, G, H

HA: C, D, E, I

DBSCAN:

Purity = 0.875, NMI = 0.52

LA: A, B, F, G, H

HA: C, D, E

LA (Ground Truth): A, B, F, H, I

HA (Ground Truth): C, D, E, G

همانطور که از نتایج مشخص است، روش DBSCAN بر اساس معیارهای ارزیابی ما عملکرد بهتری دارد. اگر به خوشه های ایجاد شده توجه کنیم، تنها تفاوت در نتایج این دو روش خوشه بندی، داده I است. داده های در خوشه LA قرار دارد در صورتی که روش KMeans به اشتباه آن را

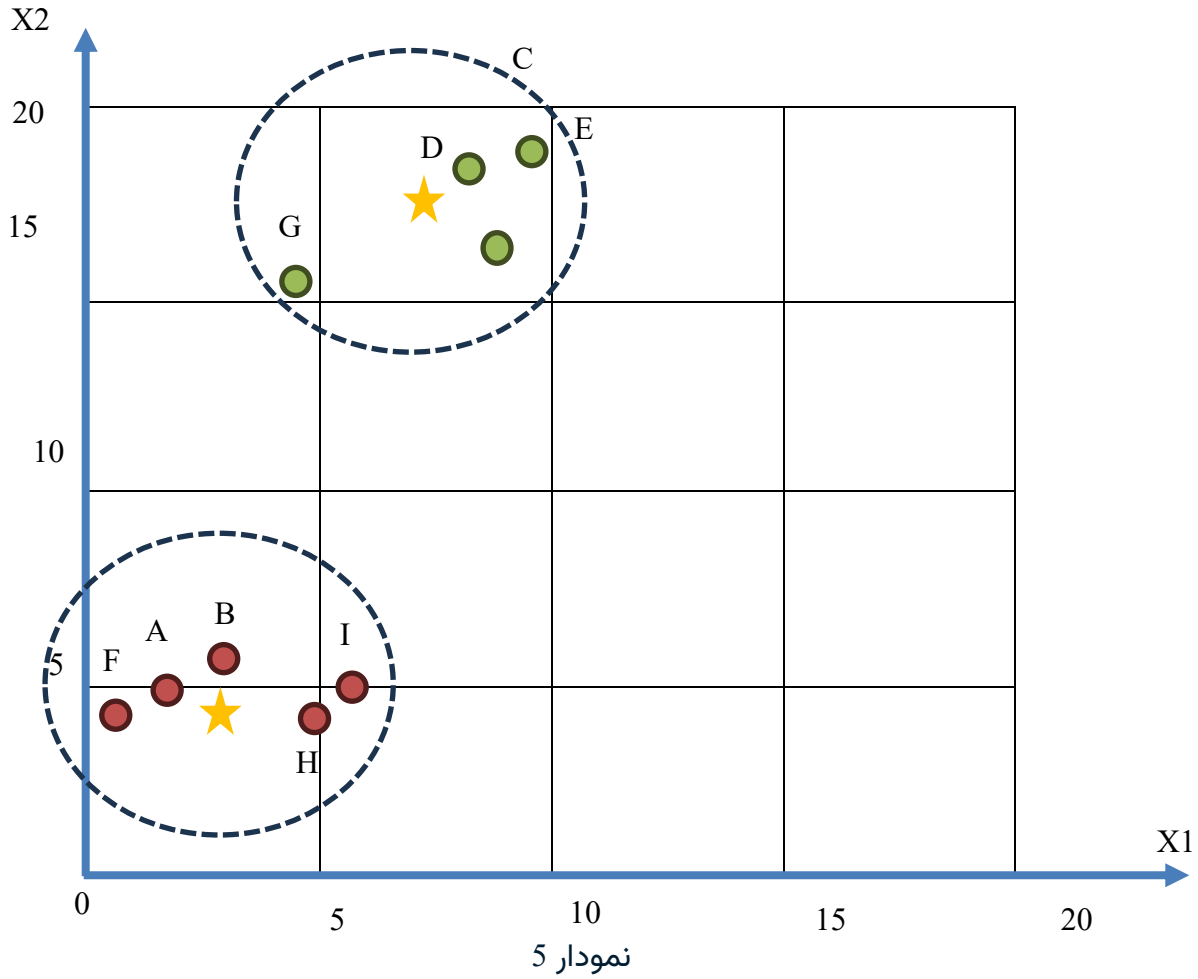
در خوشه HA دسته بندی کردی. اما روش DBSCAN این داده را در هیچ کدام از دسته ها قرار نداده به همین دلیل در این روش خوشه بندی ما خلوص بیشتر و مقدار NMI نزدیک تری به ۱ داریم.

- 4

با توجه به داده های I و G، داده ها دارای انسجام مناسب نیستند و داده های مرزی هستند. به همین دلیل با توجه به این موضوع که الگوریتم KMeans به داده های مرزی و outlier ها حساس است، الگوریتم خوبی برای این خوشه ها نمی باشد. همانطور که از نتایج مشخص است و در بالا گفته شد، بر خلاف الگوریتم KMeans، الگوریتم DBSCAN داده I را خوشه بندی نکرده و در خوشه اشتباه قرار نداده است. به همین دلیل خوشه ها در این الگوریتم از خلوص بالا تر و مقدار NMI بیشتری برخوردارند.

- 5

همانطور که گفته شد، داده های I و G در خوشه های اشتباه پراکنده هستند و با جابجایی این دو نقطه، الگوریتم KMeans می تواند به خوشه بندی مشابه با Ground Truth دست یابد:



نمودار 5

برای داده های جدید، ماتریس آشفتگی به شرح زیر است:

LA: A, B, F, H, I

HA: C, D, E, G

LA (Ground Truth): A, B, F, H, I

HA (Ground Truth): C, D, E, G

Clustering Ground T	LA	HA
LA	A, B, F, H, I	-
HA	-	C, D, E, G

Clustering Ground T	C1	C2	Sum of Classes
G1	5	0	5
G2	0	4	4
Sum of Clusters	5	4	9

$$Purity = \frac{1}{9}(5+4) = \frac{9}{9} = 1$$

$$I(C,G) = -\sum\sum P_{ij}\log\left(\frac{P_{ij}}{P_{Ci}P_{Gj}}\right)$$

$$NMI = \frac{I(C,G)}{\sqrt{(H(G)\,H(C))}}$$

$$p_{ij} = \frac{c_i \cap c_j}{n}, p_{ci} = \frac{C_i}{n}, p_{Gj} = \frac{G_j}{n}$$

$$P_{LA} = \frac{5}{9}, P_{HA} = \frac{4}{9}, P_{GLA} = \frac{5}{9}, P_{GHA} = \frac{4}{9}$$

$$I = \frac{5}{9}\times\lg\left(\frac{\frac{5}{9}}{\frac{5}{9}\times\frac{5}{9}}\right) + \frac{0}{9}\times\lg\left(\frac{\frac{0}{9}}{\frac{4}{9}\times\frac{5}{9}}\right) + \frac{0}{9}\times\lg\left(\frac{\frac{0}{9}}{\frac{5}{9}\times\frac{4}{9}}\right) + \frac{4}{9}\times\lg\left(\frac{\frac{4}{9}}{\frac{4}{9}\times\frac{4}{9}}\right) = 0.991$$

$$for\ Classes: H(G) = -\left(\frac{5}{9}\lg\left(\frac{5}{9}\right) + \frac{4}{9}\lg\left(\frac{4}{9}\right)\right) = 0.991$$

$$for\ Clusters: H(C) = -\left(\frac{5}{9}\lg\left(\frac{5}{9}\right) + \frac{4}{9}\lg\left(\frac{4}{9}\right)\right) = 0.991$$

$$NMI = \frac{I}{\sqrt{(H(G), H(c))}} = \frac{0.991}{\sqrt{0.982}} = 1$$

که همانطور که از نتایج نشان میدهد، با تغییر این دو داده، ما میتوانیم به خلوص و NMI یک دست پیدا کنیم که نشان دهنده این است که خوشه بندی کاملا درست انجام گرفته است.

سوال ۲

- 1

کاربر	Feature 1	Feature 2	Feature 3	Feature 4	Label
U1	۰.۹	۰.۲	۰.۱	۰.۵	A
U2	۰.۸۸	۰.۱۵	۰.۲	۰.۴۵	A
U3	۰.۹۲	۰.۲۲	۰.۰۵	۰.۵۵	B
U4	۰.۳	۰.۷	۰.۶	۰.۸	B
U5	۰.۳۲	۰.۶۸	۰.۶۵	۰.۷۵	B
U6	۰.۴	۰.۶۶	۰.۷	۰.۷۸	A
U7	۰.۶	۰.۴	۰.۴	۰.۶	C
U8	۰.۵۵	۰.۴۵	۰.۳۵	۰.۶۵	B
U9	۰.۵۸	۰.۴۲	۰.۳۷	۰.۶۲	B
U10	۰.۵۶	۰.۴۴	۰.۳۶	۰.۶۳	B

جدول 2

Class Cluster	G1	G2	G3	Sum
C1	2	1	0	3
C2	1	2	0	3
C3	0	3	1	4
Sum	3	6	1	

Ground truth:

G1: U1, U2, U6| G2: U3, U4, U5, U8, U9, U10| G3: U7

Clustering:

C1: U1, U2, U3| C2: U4, U5, U6| C3: U7, U8, U9, U10

$$Purity = \frac{1}{10}(2+2+1) = \frac{5}{10}$$

$$I(C,G) = -\sum\sum P_{ij}\log\left(\frac{P_{ij}}{P_{Ci}P_{Gj}}\right)$$

$$NMI = \frac{I(C,G)}{\sqrt{(H(G)\,H(C))}}$$

$$p_{ij} = \frac{c_i \cap c_j}{n}, p_{ci} = \frac{C_i}{n}, p_{Gj} = \frac{G_j}{n}$$

$$P_{C1} = \frac{3}{10}, P_{C2} = \frac{3}{10}, P_{C3} = \frac{4}{10}, P_{G1} = \frac{3}{10}, P_{G2} = \frac{6}{10}, P_{G3} = \frac{1}{10}$$

$$I = \frac{2}{10}\times\lg\left(\frac{\frac{2}{10}}{\frac{3}{10}\times\frac{3}{10}}\right) + \frac{1}{10}\times\lg\left(\frac{\frac{1}{10}}{\frac{3}{10}\times\frac{6}{10}}\right) + \frac{1}{10}\times\lg\left(\frac{\frac{1}{10}}{\frac{3}{10}\times\frac{3}{10}}\right)$$
$$+\frac{2}{10}\times\lg\left(\frac{\frac{2}{10}}{\frac{3}{10}\times\frac{6}{10}}\right) + \frac{3}{10}\times\lg\left(\frac{\frac{3}{10}}{\frac{4}{10}\times\frac{6}{10}}\right) + \frac{1}{10}\times\lg\left(\frac{\frac{1}{10}}{\frac{4}{10}\times\frac{1}{10}}\right) = 0.42$$

$$for\ Clusters: H(C) = -\left(\frac{3}{10}\lg\left(\frac{3}{10}\right) + \frac{3}{10}\lg\left(\frac{3}{10}\right) + \frac{4}{10}\lg\left(\frac{4}{10}\right)\right) = 1.57$$

$$for\ Classes: H(G) = -\left(\frac{3}{10}\lg\left(\frac{3}{10}\right) + \frac{6}{10}\lg\left(\frac{6}{10}\right) + \frac{1}{10}\lg\left(\frac{1}{10}\right)\right) = 1.29$$

$$NMI = \frac{I}{\sqrt{H(G), H(c)}} = \frac{0.42}{\sqrt{2.0253}} = 0.207$$

- 2

نتایج به دست آمده در قسمت اول نشان می دهد که خلوص در خوشه ها مقدار قابل قبولی دارد. اما دلیل اینکه NMI در این سناریو پایین است، به تعریف این دو معیار بر میگردد. Purity همانطور که از نامش مشخص است خلوص را نشان می دهد. یعنی نشان میدهد چند درصد از داده های موجود در آن خوشه واقعا به آن خوشه تعلق دارد و طبیعتا داده های غالب روی این معیار تاثیر می گذارند. اما NMI اطلاعات مشترک بین لیبل های واقعی و خوشه ها که با آنتروپی نرمال شده را نشان می دهد و با این کار بین دسته بندی اصلی و تعداد خوشه ها تعادل ایجاد می کند. برای این سناریو می توان به دلایل زیر برای پایین بودن NMI نسبت به Purity اشاره کرد:

- همه خوشه ها مخلوط هستند و هر خوشه حداقل دو کلاس را شامل می شود.
- هیچ خوشه ای به طور خاص با یک کلاس Ground Truth هم راستا نیست.
- توزیع برچسب های Ground Truth در خوشه ها به صورت پراکنده و ناهمگون است.

بنابر این Mutual Information بین خوشه بندی و Ground Truth پایین است چون دانستن اطلاعات یک خوشه اطلاعات کمی در مورد کلاس واقعی می دهد.

- 3

بر فرض اگر خوشه بندی به شکل زیر باشد:

خوشه ۱: U1, U2, U3

خوشه ۲: U4, U5, U6, U8, U9, U10

خوشه ۳: U7

Ground truth:

G1: U1, U2, U6| G2: U3, U4, U5, U8, U9, U10| G3: U7

Purity: $\frac{1}{10}(2+5+1)=\frac{8}{10}$

$$P_{C1}=\frac{3}{10}, P_{C2}=\frac{6}{10}, P_{C3}=\frac{1}{10}, P_{G1}=\frac{3}{10}, P_{G2}=\frac{6}{10}, P_{G3}=\frac{1}{10}$$

$$I=\frac{2}{10}\times\lg\left(\frac{\frac{2}{10}}{\frac{3}{10}\times\frac{3}{10}}\right)+\frac{1}{10}\times\lg\left(\frac{\frac{1}{10}}{\frac{3}{10}\times\frac{6}{10}}\right)+\frac{1}{10}\times\lg\left(\frac{\frac{1}{10}}{\frac{6}{10}\times\frac{3}{10}}\right)$$
$$+\frac{5}{10}\times\lg\left(\frac{\frac{5}{10}}{\frac{6}{10}\times\frac{6}{10}}\right)+\frac{1}{10}\times\lg\left(\frac{\frac{1}{10}}{\frac{1}{10}\times\frac{1}{10}}\right)=0.63$$

$$for\ Clusters: H(C) = -\left(\frac{3}{10}\lg\left(\frac{3}{10}\right) + \frac{6}{10}\lg\left(\frac{6}{10}\right) + \frac{1}{10}\lg\left(\frac{1}{10}\right)\right) = 1.29$$

$$for\ Classes: H(G) = -\left(\frac{3}{10}\lg\left(\frac{3}{10}\right) + \frac{6}{10}\lg\left(\frac{6}{10}\right) + \frac{1}{10}\lg\left(\frac{1}{10}\right)\right) = 1.29$$

$$NMI = \frac{I}{\sqrt{H(G), H(c)}} = \frac{0.64}{1.29} = 0.488$$

همانطور که مشاهده می شود، با خوشه بندی پیشنهادی جدید، purity نه تنها افت نداشته بلکه به ۸۰٪ رسیده است و مقدار NMI نیز تقریبا دو برابر شده و به مقدار 0.488 رسیده است.

- 4

۴.۱- با افزایش ابعاد، فاصله بین داده ها به صورت غیر واقعی افزایش پیدا میکند. همچنین تفاوت بین نزدیک ترین و دور ترین نقاظ کاهش می یابد و همه داده ها تقریبا به یک اندازه از هم دور به نظر میرسند. در نتیجه در این سناریو مفهوم نزدیک در فضا بی معنی می شود و خوشه بندی بر پایه فاصله (مانند KMeans) به شدت آسیب می بیند.

۴.۲- با افزایش ابعاد به ۴۰۰ بعد، به نظر من کیفیت خوشه بندی کاهش پیدا میکند، الگوریتم دچار overfitting شده و به نویزها شده و یا ویژگی ها بی معنی می شوند. همچنین معیار هایی مانند NMI کاهش می یابند چرا که خوشه ها کمتر با برچسب های واقعی مطابقت دارند.

۴.۳- راهکار های پیشنهادی جهت حل این مشکلات:

- کاهش ابعاد: استفاده از روش هایی مانند t-SNE، PCA و یا UMAP برای حذف ویژگی های کم ارزش و فشرده سازی داده ها.

- Feature Selection: حذف ویژگی های نامرتبط با استفاده از اطلاعات آماری یا روش های یادگیری نظارت شده.

- استاندارد سازی و normalization: برای جلوگیری از تسلط ویژگی ها با مقیاس بزرگتر میتوان از normalization استفاده کرد.

- 5

برای استفاده از معیار Modularity برای داده هایی مانند این مثال، پاسخ کلی بله است ولی با یک پیش فرض اساسی:

به نظرم باید برای این کار داده ها را به گراف تبدیل کنیم. یعنی بین نمونه ها یال هایی تعریف شود مثلا بر اساس شباهت های معنایی در این صورت نود ها همان داده ها هستند و یال ها هم شباهت بین داده ها. در غیر این صورت به نظرم اگر نتوان یک گراف معنا دار از داده ها ساخت، یعنی رابطه بین داده ها گرافی نباشد یا اینکه داده ها خیلی نویزی باشند، آنگاه modularity کاربرد ندارد چرا که این معیار به ساختار گرافی وابسته است و بدون آن قابل تفسیر نیست.

سوال ۳ – بخش عملی

بخش ۱

1 -در این بخش ابتدا طبق خواسته سوال، دو دیتا ست label و titles را بارگزاری کردم و طبق گفته سوال داده ها را به صورت جداگانه نمایش دادم که نمایش این داده ها به شرح زیرند:

	label
0	18
1	18
2	3
3	3
4	7
...	...
19997	20
19998	20
19999	20

	title
0	How do I fill a DataSet or a DataTable from a ...
1	How do you page a collection with LINQ?
2	Best Subversion clients for Windows Vista (64bit)
3	Best Practice: Collaborative Environment, Bin ...
4	Visual Studio Setup Project - Per User Registr...
...	...
19995	Magento Custom Options VS Attributes
19996	How to solve 404 not found problem in Magento
19997	Want to add custom option from the frontend of...
19998	installing magento plugins without using magen...
19999	Magento : Call to a member function count() on...

همانطور که از نمایش داده شده مشخص است، داده ها شامل عنوان هایی از اسناد هستند که برای هر عنوان لیبل مشخصی وجود دارد. همچنین در ادامه بررسی کردم که تعداد داده های دو دیتاست عنوان ها و لیبل ها یکسان باشند که این موضوع برقرار بود و تعداد داده ها در هر دو دیتاست یکسان بودند.

- 2

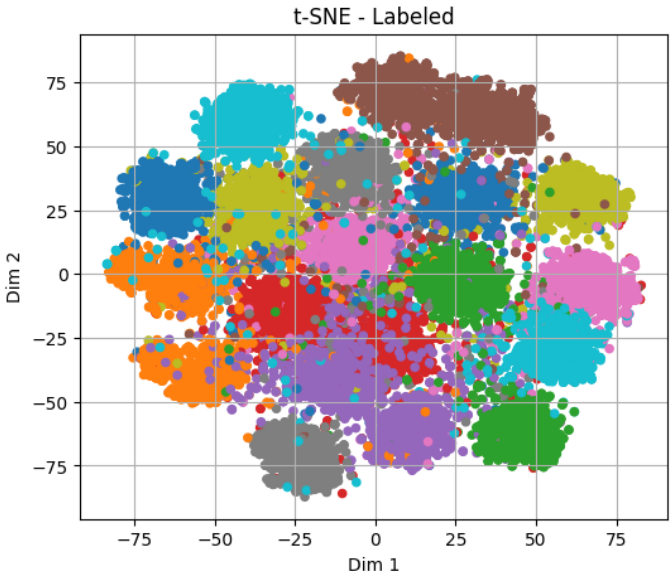
در این قسمت از تمرین نیز با توجه به خواسته سوال و قطعه کد درج شده در تمرین، مدل GTE را بارگزاری کردم و سپس برای دیتاست titels این مدل را اعمال کردم و خروجی بردارهای امبدینگ برای عنوان ها در این دیتا ست بود که نمونه ای از این بردارهای امبدینگ در زیر نمایش داده شده است:

```
array([[-3.1008607e-02,  4.6575769e-05,  1.1385731e-02, ...,
        -5.2819330e-02,  5.7307415e-04,  2.7543655e-02],
       [-6.4695075e-02,  6.4545602e-04,  1.3240832e-02, ...,
        -4.3488305e-02, -2.0164149e-02,  3.0185059e-02],
       [-7.8939125e-02, -3.6999058e-02, -1.3121576e-02, ...,
         2.4249829e-02,  4.3211434e-02,  1.8765777e-02],
       ...,
       [-9.0950847e-02, -3.9918222e-02,  6.2028289e-02, ...,
         4.9867928e-02, -2.9881403e-03,  6.4335816e-02],
       [-7.6423898e-02, -1.2492814e-02,  2.8748997e-02, ...,
         4.2816419e-02,  4.5070123e-02,  4.4149836e-03],
       [-5.4269336e-02, -2.3541616e-02,  3.9925877e-02, ...,
         4.3662619e-03,  2.1669243e-02,  1.3829525e-02]],
dtype=float32)
```

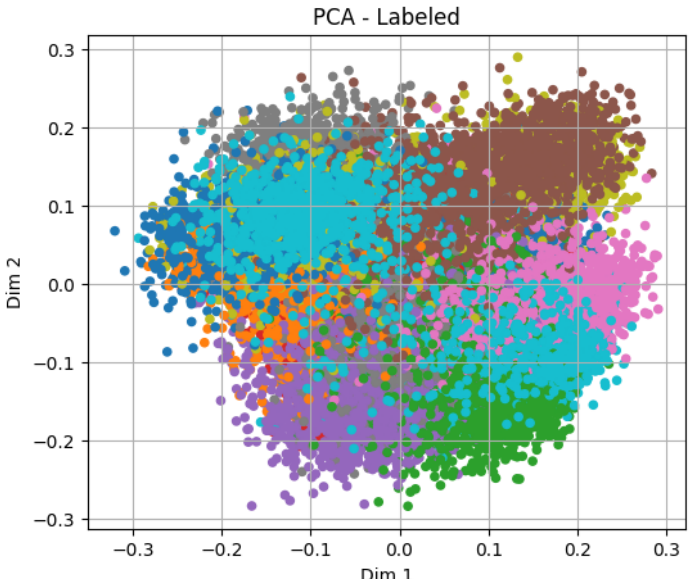
بخش ۲

- 1

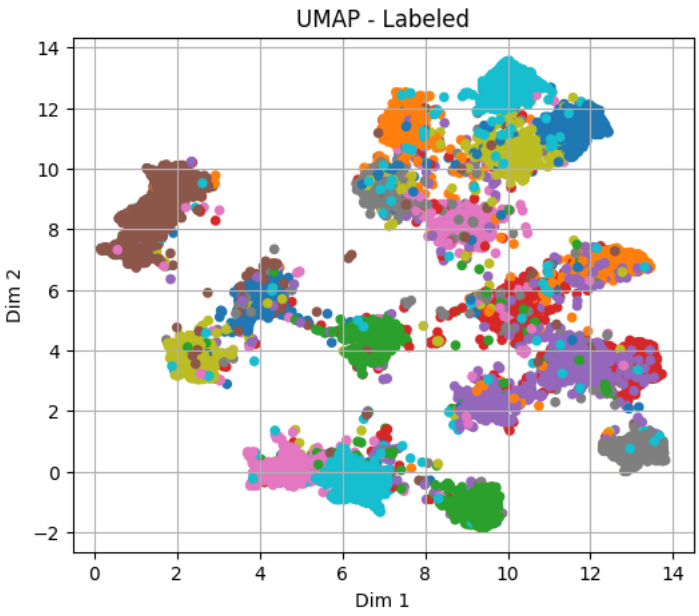
حالا در این مرحله با توجه به خواسته سوال، خوشه ها را با سه روش PCA، t-SNE و UMAP به فضای دو بعدی نگاشت کردم و نمودار حاصل از هر روش را هم رسم کردم. برای این کار از متد های PCA و TSNE از کتابخانه Sikitlearn و کتابخانه umap استفاده کردم که نمایش دو بعدی خوشه ها به شرح زیر است:



نمودار 6



نمودار 7



نمودار 8

آیا خوشه ها در فضا جدا هستند؟

در PCA خوشه ها به خوبی جدا نیستند. نقاط به شدت روی هم افتاده اند و همپوشانی زیادی بین کلاس ها دیده می شود.

در t-SNE خوشه ها به صورت مجزا و متراکم در فضا ظاهر شده اند و هر خوشه در ناحیه ای متمرکز شده است.

در روش UMAP مشابه t-SNE خوشه ها به صورت متراکم و مجزا در فضا ظاهر شده اند ولی با ساختار پیوسته تر. در این مدل برخی اتصال های بین خوشه ای دیده میشود.

آیا روش ها نتایج متفاوتی ارائه می دهند؟

روش PCA خطی و مبتنی بر واریانس است. این روش فقط واریانس های بزرگ را نگه میدارد و ساختارهای nonlinear را نمی فهمد. در نتیجه در این روش خوشه ها به خوبی جدا نیستند.

روش t-SNE نیز غیر خطی و با خقط همسایگی موضعی کار میکند و خوشه ها را در فضاهای مستقل جمع میکند ولی فاصله بین خوشه ها معنی ندارد. این روش برای visualization عالی است ولی برای خود clustering لزوما خوب عمل نمیکند.

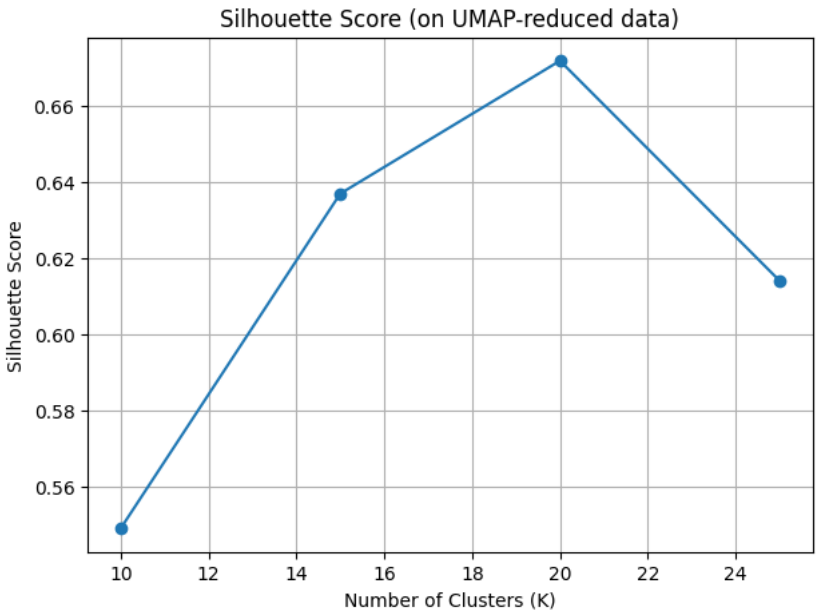
روش UMAP نیز مانند روش t-SNE به صورت غیر خطی عمل میکند ولی سعی بر حفظ ساختار درون خوشه ای و بین خوشه ای دارد. این روش نسبت به روش t-SNE پایدار تر است و ساختار global محلی را بهتر حفظ می کند به همین دلیل در الگوریتم های خوشه بندی قابل استفاده تر است.

انتخاب یکی از روش ها برای اجرای الگوریتم خوشه بندی:

با توجه به موارد گفته شده در بالا من روش UMAP را انتخاب کردم. این روش نسبت به t-SNE پایدار تر است و کمتر به پارامتر ها وابسته است. بر خلاف t-SNE، فاصله ها بین خوشه ها با معناتر هستند و ساختار global و local را باهم حقظ می کند و خروجی این روش برای الگوریتم هایی مانند KMeans بسیار مناسب است.

بخش ۳

در این بخش با توجه به خواسته سوال و با توجه به بخش قبل که روش UMAP را برای کاهش ابعاد انتخاب کردم،الگوریتم KMeans را با مقادیر مختلف K اجرا کردم و نتیجه هر مقدار را با توجه به معیار silhouette رسم کردم:



نمودار 9

معیار silhouette بین ۱- تا ۱+ است و هرچه به ۱ نزدیک تر باشد نشاندهنده این است که خوشه بندی بهتر است. عدد نزدیک تر به ۱ نشان می دهد که داده ها به خوشه خود نزدیک اند و از خوشه های دیگر دور است و مقدار نزدیک به ۰ نشان میدهد که داده ها روی مرز بین خوشه ها هستند و مقدار منفی نشان دهنده این است که برخی داده ها به خوشه اشتباهی نسبت داده شده اند. مقادیر به دست آمده برای این معیار برای مقادیر مختلف K به شرح زیر است:

```
K=10, Silhouette Score=0.5490
K=15, Silhouette Score=0.6370
K=20, Silhouette Score=0.6719
K=25, Silhouette Score=0.6141
```

طبق مقادیر به دست آمده بهترین مقدار برای تعداد خوشه ها برابر با ۲۰ است که با این تعداد خوشه و اجرای الگوریتم KMeans مقدار معیار Silhouette برابر با 0.6719 می شود که نشان میدهد با این مقدار K بیشترین جدایی بین خوشه ها اتفاق می افتد.

بخش ۴

1 - در این بخش طبق گفته سوال توزیع برچسب ها در خوشه بندی انجام شده بررسی شد که به شرح زیر است:

- Cluster 1 → فقط label 1 (مورد 983)
- Cluster 2 → فقط label 2 (مورد 1004)
- Cluster 4 → فقط label 14
- Cluster 7 → فقط label 6
- Cluster 9 → فقط label 10
- Cluster 11 → فقط label 13
- Cluster 12 → فقط label 4
- Cluster 13 → فقط label 8
- Cluster 15 → تقریباً فقط label 3
- Cluster 17 → تقریباً فقط label 7

این خوشه‌ها به احتمال زیاد شباهت معنایی دارند. الگوریتم به‌درستی توانسته داده‌های دارای موضوع مشابه را در یک خوشه قرار دهد.

خوشه هایی با اختلاط جزئی

- Cluster 0: شامل دو برچسب است ((628) label 24 و (436) label 0)
- Cluster 3: برچسب غالب ۱۷ است اما ۵ مورد از لیبل ۹ هم وجود دارد
- Cluster 5: برچسب غالب ۱۲ است اما ۱۲ مورد از لیبل ۱۶ هم وجود دارد
- Cluster 10: برچسب غالب با تعداد ۱۰۰۷ لیبل ۱۱ است اما چند مورد از برچسب های دیگر نیز وجود دارد
- Cluster 14: برچسب ۱۶ با ۶۰۲ داده و برچسب ۲۰ با ۳۸۲ داده وجود دارد
- Cluster 16: برچسب ۲۳ با ۵۵۱ داده و برچسب ۱۸ با ۴۵۲ داده وجود دارد
- Cluster 19: برچسب غالب ۱۹ است ولی ۵ مورد از لیبل ۰ هم وجود دارد

برخی خوشه‌ها احتمالاً موضوعات نزدیکی دارند یا embedding مدل در تمایز بین آن‌ها دقت کافی نداشته.

مثلاً شاید label 20 و 16 از نظر معنایی به هم نزدیک‌اند و در یک embedding خوشه‌بندی شده‌اند.

چه حالتی در توزیع برچسب ها می تواند نشانگر نیاز به خوشه بندی سلسله مراتبی باشد؟

اگر خوشه ای شامل چند زیر موضوع متمایز باشد یعنی در یک خوشه چند برچسب پر تکرار اما متفاوت وجود داشته باشد و یا خوشه های بزرگ شامل زیرخوشه های معنادار باشند، نشان دهنده این است که خوشه بندی flat کافی نیست و شاید بهتر باشد از خوشه بندی سلسله مراتبی استفاده شود.

- 2

تعداد خوشه ها تقریبا نزدیک به تعداد برچسب های حقیق است. تعداد برچسب های حقیقی برابر با ۲۵ عدد بود که تعداد خوشه هایی که ما در خوشه بندی خود داشتیم برابر با ۲۰ خوشه بود. اگر تعداد خوشه ها بسیار بیشتر و یا بسیار کمتر از تعداد کلاس های واقعی باشد، ممکن است over clustering و یا under vlustering رخ دهد به این معنا که اگر k زیاد باشد کلاس های واقعی در چند خوشه پخش می شوند و یا اگر k کم باشد خوشه ها شامل ترکیبی از چند کلاس خواهند بود.

- 3

معیار های Purity و NMI طبق خواسته سوال برای خوشه بندی انجام شده محاسبه شد که مقادیر این معیار ها به شرح زیر است:

Purity: 0.8632

NMI: 0.9606

که هر دو این مقادیر نشان دهنده این است که خوشه بندی به خوبی انجام گرفته است. purity نشان می دهد که به طور میانگین حدود ۸۶٪ از اعضای هر خوشه به برچسب غالب همان خوشه تعلق دارند. این مقدار بالاست و نشان میدهد که خوشه ها تا حد زیادی خالص هستند. باید دقت داشته باشیم که این معیار حساس به تعداد خوشه هاست یعنی اگر تعداد خوشه ها زیاد باشد، Purity ممکن است مصنوعی بالا باشد اما در این سناریو با توجه به این که تعداد خوشه ها از تعداد کلاس های حقیقی کمتر و نزدیک به تعداد کلاس های حقیقی است می توان اطمینان پیدا کرد که مقدار Purity در این بخش به خوبی نشان دهنده خلوص خوشه ها هست.

مقدار NMI هم برابر با ۹۶٪ بوده که به این معناست بین برچسب هاب واقعی و برچسب های خوشه بندی، اطلاعات مشترک بسیار زیادی وجود دارد. الگوریتم خوشه بندی توانسته ساختار برچسب های واقعی را تقریبا به طور کامل باز سازی کند. نکته ای که وجود دارد این است که NMI نرمالسازی شده است و در برابر تغییر تعداد خوشه ها مقاوم تر است و این مقدار بسیار بالا و مقدار Purity به دست آمده هم نشان می دهد دقت و هم راستایی قوی بین خوشه ها و کلاس های واقعی بر قرار است.

بخش ۵

- 1

در این بخش ۵ نمونه داده نزدیک به مرکز هر خوشه انتخاب شده و نمایش داده شد.

نتایج نشان می دهد که بیشتر خوشه ها موضوعات متمرکز و معنا داری دارند. در موارد محدودی مثل خوشه ۱۹، خوشه بندی چندان دقیق نبوده و علت آن ممکن است ترکیب موضوعات مختلف، نویز و یا مدل امبدینگ باشد. در مجموع به نظر من با توجه به Purity برابر با ۸۰٪ و نتایج چاپ شده از خوشه ها، اعضای هر خوشه غالبا به یک حوزه مشخص تغلق دارند و خوشه بندی به خوبی انجام گرفته است.

- 2

برای این قسمت من ابتدا سوال جدید را با استفاده از مدل امبدینگ، انکود کردم و با استفاده از روش UMAP ابعاد آن را به ۲ بعد کاهش دادم سپس با خوشه بندی که از قبل داشتیم، خوشه سوال مورد نظر رو با استفاده از kmeans.predict پیشبینی کردم که خوشه پیش بینی شده، خوشه ۲ بود.

با بررسی خوشه ۲ متوجه می شویم که داده های این خوشه بیشتر درباره پایگاه داده های oracle صحبت میکنند و میتوان گفت تخصیص سوال جدید به این خوشه با توجه به این که این سوال درباره پایگاه داده و sql است، به درستی انجام گرفته.

اظهارنامه استفاده از هوش مصنوعی

تأیید می‌کنم که از ابزارهای هوش مصنوعی مطابق با دستورالعمل‌های بارگذاری شده در سامانه Elearn درس به طور مسئولانه استفاده کرده‌ام. تمام اجزای کار خود را درک می‌کنم و آماده بحث شفاهی درباره آنها هستم.