به نام خدا

دانشگاه تهران

دانشکدگان فنی

دانشکده مهندسی برق و کامپیوتر

درس داده‌کاوی

تمرین عملی ۳

خرداد ماه ۱۴۰۱

* فهرست

تمرینات تشریحی

۱- آیتم هایی از شش بخش متفاوت (Sport، National، Metro، Foreign، Financial، Entertainment) در سه خوشه قرار داده شده اند. به عبارت دیگر، Sport، National، Metro، Foreign، Financial و Entertainment، categoryهای ground truth را نشان می‌دهند و آیتم‌ها در سه خوشه‌ی #1، #2 و #3 خوشه‌بندی شده‌اند. جدول زیر تعداد آیتم های موجود در هر خوشه به تفکیک بخش آن‌ها را گزارش می‌کند. با استفاده از این جدول، به سوالات زیر پاسخ دهید:

ground truth / partition / 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**

الف- مقدار entropy را برای این خوشه بندی محاسبه کنید.

ب- مقدار purity کل را برای خوشه‌بندی محاسبه کنید.

ج- مقدار precision، recall، F-measure را برای هر خوشه محاسبه کنید.

۲- شکل زیر نحوه نشستن دانش آموزان A، ....، O در سالن امتحان نمایش داده شده است. مسئول برگزاری امتحان، با استفاده از الگوریتم DBSCAN اقدام به خوشه بندی دانش آموزان می کند. او از معیار فاصله منهتن (Manhattan)، مقدار eps برابر با 2.1 و MinPts برابر با 4 برای این منظور استفاده می کند. (مقدار minPts شامل آیتم مورد نظر نیز می شود). در پایان، مسئول برگزاری امتحان، دانش آموزان را به دو خوشه تقسیم می کند.

1				A					
2		B		C					D
3				E	F		G		
4				H	I				
5									
6	J								
7		K	L		M	N			
8									
9			O						
	1	2	3	4	5	6	7	8	9

الف- کدام دانش آموزان در نقاط هسته ای قرار گرفته اند؟ (Core Point)

ب- کدام دانش آموزان در نقاط مرزی قرار گرفته اند؟ (Border Point)

ج- کدام دانش آموزان به طور مستقیم از دانش آموز I قابل دسترسی هستند؟ (Directly Density Reachable)

د- کدام دانش آموزان به طور مستقیم از دانش آموز M قابل دسترسی هستند؟ (Directly Density Reachable)

ه- در صورت ورود دانش آموز P، محل نشستن او را به گونه ای مشخص کنید که به هر دو خوشه متصل (connected) باشد اما باعث ادغام دو خوشه نشود.

و- محل نشستن دانش آموز P را به گونه ای مشخص کنید که هر دو خوشه با هم ادغام شوند.

::: پاسخ به سوالات همراه با دلایل مشروح ذکر شود :::

۳- با استفاده از جدول شباهت زیر، Dendrogram های الگوریتم خوشه بندی Agglomerative را با دو روش single-link و complete-link رسم کنید.

	p1	p2	p3	p4	p5
p1	1.00	0.10	0.41	0.55	0.35
p2	0.10	1.00	0.64	0.47	0.98
p3	0.41	0.64	1.00	0.44	0.85
p4	0.55	0.47	0.44	1.00	0.76
p5	0.35	0.98	0.85	0.76	1.00

تمرین عملی

در سال ۲۰۱۴، گروهی از محققین در مقاله ای تحت عنوان Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records[۱] اقدام به جمع آوری اطلاعات بیماران دیابتی متعلق به ده ها بیمارستان و کلینیک درمانی در آمریکا کردند. بخشی از این اطلاعات به صورت ناشناس سازی شده (Anonymized) در اختیار عموم قرار گرفته است که شامل صد هزار آیتم با پنجاه ویژگی می باشد[۲]. با استفاده از این دیتاست (ضمیمه شده در تمرین) موارد خواسته شده زیر را به صورت یکجا (Jupyter Notebook) پیاده سازی کنید. همچنین خروجی های خواسته شده را در فایل گزارش خود ارائه دهید.

::: استفاده از کتابخانه های آماده مجاز می باشد :::

الف- دیتاهای موجود دارای مقادیر متفاوتی از انواع عددی و متنی می باشد. همچنین مقادیر پرت و null نیز در این دیتاست وجود دارد که باعث اختلال و کاهش دقت خوشه بندی می شود. دیتاهای موجود را در حد امکان نرمال‌سازی کنید و دلیل هر اقدام را به صورت مشروح در فایل گزارش بیان کنید.

ب- با استفاده از معیار سیلوئت، بهترین تعداد خوشه ها را در روش K-means و بهترین پارامتر های ورودی (eps, minPnt) در روش DBSCAN پیدا کنید و با توجه به مقادیر به دست آمده، بهترین نتیجه هر روش را در یک فایل CSV ذخیره نمایید. (تنها شامل ستون های encounter_id، kmean_label و dbscan_label)

ارائه نمودار ضریب سیلوئت نسبت به پارامتر های ورودی برای هر روش خوشه بندی در فایل گزارش الزامی می باشد.

[۱] https://www.hindawi.com/journals/bmri/2014/781670/

[۲] https://archive-beta.ics.uci.edu/ml/datasets/diabetes+130+us+hospitals+for+years+1999+2008

ملاحظات (حتما مطالعه شود)

تمامی نتایج شما باید در یک فایل فشرده با عنوان DM_CA3_StudentID تحویل داده شود.

- این فایل فشرده، بایستی حاوی یک فایل با فرمت PDF (گزارش تایپ شده) و یک فایل Jupyter Notebook باشد که کدهای نوشته شده را شامل شود.
- خوانایی و دقت بررسی‌ها در گزارش نهایی از اهمیت ویژه‌ای برخوردار است. به تمرین‌هایی که به صورت کاغذی تحویل داده شوند یا به صورت عکس در سایت بارگذاری شوند، ترتیب اثری داده نخواهد شد.
- گزارش به صورت تایپ شده در قالبPDF شامل شرح آزمایش‌های انجام شده، پارامترهای آزمایش، نتایج و تحلیل‌ها باشد. دقت داشته باشید که در تمامی تمرین‌ها، نمره‌ی اصلی به تفسیر و تحلیل شما تعلق می‌گیرد.
- مهلت تحویل تمرین به هیچ عنوان تمدید نخواهد شد و این تمرین، امکان تحویل با تاخیر را ندارد.
- توجه کنید این تمرین باید به صورت تک نفره انجام شود و پاسخ‌های ارئه شده باید نتیجه فعالیت فرد نویسنده باشد (همفکری و به اتفاق هم نوشتن تمرین نیز ممنوع است). در صورت مشاهده تقلب به همه افراد مشارکت کننده، نمره تمرین صفر و به استاد نیز گزارش می‌گردد.
- در صورت بروز هرگونه مشکل، با آقای همایونی از طریق ایمیل زیر در ارتباط باشید:

mailto:alihomayouni@ut.ac.ir

مهلت تحویل بدون جریمه: ۲۸ خرداد ۱۴۰۱