به نام خدا

دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)
دانشکده مهندسی کامپیوتر و فناوری اطلاعات

تمرین سری دوم داده کاوی

توضیحات:

- پاسخ به تمرین‌ها باید به صورت انفرادی صورت گیرد و در صورت مشاهده هرگونه تقلب نمره صفر برای کل تمرین منظور خواهد شد.
- تمیزی و خوانایی گزارش تمرین از اهمیت بالایی برخوردار است.
- لطفا گزارش تمرین خود را در قالب یک فایل PDF با نام **«HW2_StudentNumber.pdf»** در سایت درس در مهلت معین بارگزاری نمایید.
- در صورت داشتن اشکال می‌توانید از طریق ایمیل **datamining.spring2021@gmail.com** با تدریس‌یار درس در ارتباط باشید.

نیم‌سال دوم ۰۰-۹۹

بخش اول: سوالات تئوری

۱) میخواهیم نقاط زیر را با الگوریتم‌های مختلف خوشه‌بندی در سه خوشه جای دهیم. تابع فاصله را اقلیدسی در نظر بگیرید:

	x	***y***
A_1	1	2
A_2	6	3
A_3	8	4
A_4	2	5
A_5	7	5
A_6	4	6
A_7	5	7
A_8	2	8

الف) با اجرای الگوریتم k-means خوشـــه‌ها و مراکز آنها را در پایان هر مرحله (تا ۳ مرحله) مشـــخص کنید. (فرض کنید در ابتدا A_3 و A_4 و A_8 در خوشــه‌ی اول قرار گرفته اند و A_2 و A_5 و A_7 در خوشــه‌ی دوم و A_1 و A_6 هم در خوشه سوم قرار دارند)

ب) با اجرای الگوریتم PAM در سه مرحله، خوشه‌ها و مراکز آنها را در پایان هر مرحله مشخص کنید. در هنگام نیاز به انتخاب نقاط تصادفی، آنها را از بالا به پایین به ترتیب جدول در نظر بگیرید، یعنی نقطه رندوم اول (که به عنوان medoid خوشه اول انتخاب می‌شود) A_1 و نقطه رندوم دوم A_2 و ... به همین ترتیب خواهد بود.

۲) یکی از روش‌های یافتن تعداد بهینه خوشه‌ها (k) را در هنگام استفاده از الگوریتم k-means توضیح دهید.

۳) برای هر یک از روش‌های خوشه‌بندی زیر، توضیح کوتاهی ارائه داده و مزایا و معایب آن‌ها را بیان کنید:

الف) k-medoids

ب) CLARA

ج) DBSCAN

د) OPTICS

ه) BIRCH

و) CHAMELEON

۴) کدام موارد زیر در مورد الگوریتم DBSCAN صحیح می‌باشند؟ پاسخ خود را برای هر گزینه توضیح دهید:

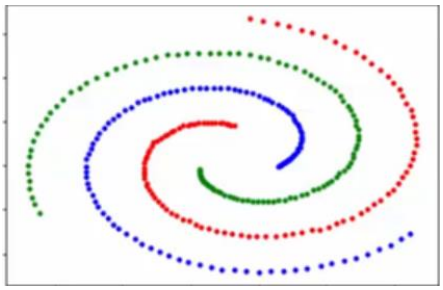
الف) برای اینکه نقاط داده در یک خوشه قرار گیرند، باید در فاصله‌ی آستانه‌ای از یک نقطه هسته (core point) باشند.

ب) این الگوریتم نسبت به داده‌های پرت (outliers) مقاوم است.

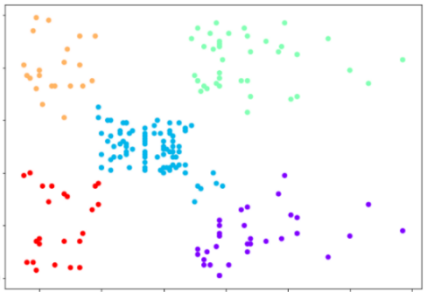
ج) پیچیدگی زمانی این الگوریتم از مرتبه $O(n^3)$ است.

د) این الگوریتم نیازی به دانستن تعداد خوشه‌ها پیش از انجام خوشه‌بندی ندارد.

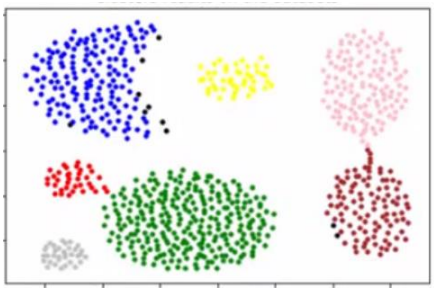
۵) در صورتی که توزیع داده‌های ما مشابه شکل‌های زیر باشد، در هر مورد، بهتر است از کدام روش برای خوشه‌بندی آنها استفاده کنیم؟ k-means (با فرض اطلاع از تعداد خوشه‌ها) یا DBSCAN یا هر دو؟ چرا؟



الف



ب



ج

د

۶) در جدول زیر فواصل پنج داده از هم ذکر شده است:

	A	B	C	D	E
A		1.23	2.44	0.85	2.04
B	1.23		0.74	1.2	0.98
C	2.44	0.74		1.34	1.4
D	0.85	1.2	1.34		0.87
E	2.04	0.98	1.4	0.87	

این داده‌ها را با استفاده از روش خوشه‌بندی سلسله‌مراتبی، یک بار با single-link و یک بار با complete-link خوشه‌بندی کرده و درخت dendrogram آن‌ها را رسم کنید. (دقت داشته باشید درخت‌های رسم شده باید دقیق باشند و ترتیب طبقات مشخص باشد)

بخش دوم: سوالات عملی

در این قسمت می‌خواهیم الگوریتم‌های k-means و DBSCAN را برای خوشه‌بندی دو مجموعه‌داده‌ی Iris و Worms به کار بگیریم. این دو مجموعه‌داده در اختیار شما قرار داده شده است. مجموعه‌ی اول، یک مجموعه‌داده‌ی معروف در مسائل خوشه‌بندی است که اطلاعات مربوط به اندازه‌گیری‌های انجام شده بر روی ابعاد سه گونه گل، به همراه برچسب هر کدام را دارا می‌باشد. مجموعه‌داده‌ی دوم نقاط ثبت شده از جسم تعدادی موجود میکروسکوپی است که البته به صورت مصنوعی تولید شده است.

همانطور که می‌دانید الگوریتم k-means یک الگوریتم هوشمند نیست، به این معنی که ما باید تعداد خوشه‌ها را برایش تعیین کنیم. در این الگوریتم ابتدا تعدادی مرکز[1] به صورت تصادفی ایجاد می‌شود. سپس فاصله‌ی همه‌ی نقاط از این مراکز خوشه‌ها محاسبه شده و هر نقطه به خوشه با نزدیک‌ترین مرکز تعلق می‌یابد. در مرحله‌ی بعد، میانگین نقاط هر خوشه محاسبه شده و مرکز را برابر آن قرار می‌دهیم. مراحل بالا را چندین بار تکرار می‌کنیم تا زمانی که هیچ کدام از مراکز تغییری نکند یا تعلق هیچ نقطه‌ای تغییر نکند یا به حد تعیین شده برای تعداد iteration های الگوریتم برسیم (یا ...).

شبه‌کد مربوط به این الگوریتم به این صورت است:

Algorithm 1 K-Means Clustering

```
1: while True do
2:     for i = 1 to m... do
3:         c^(i) := index (1 to k) of cluster centroid closest to x^(i)
4:     end for
5:     for k = 1 to K... do
6:         μ_(k) := average (mean) of points assigned to cluster k.
7:     end for
8: end while
```

۱) ابتدا تابعی بنویسید که با گرفتن یک مجموعه‌داده‌ی ورودی، k (تعداد خوشه‌ها) و max_iteration (حداکثر تعداد تکرار دستورات حلقه)، الگوریتم k-means را روی داده‌ها اجرا کند و به عنوان خروجی، لیست برچسب‌های نقاط را برگرداند. دقت کنید که پیاده‌سازی این قسمت باید کاملاً توسط شما (بدون استفاده از کتابخانه‌های آماده برای این کار) انجام شود و همچنین تابع باید مستقل از ابعاد مجموعه ورودی، درست عمل کند. (یعنی با فرض n-بعدی بودن مجموعه‌داده باشد)

۲) تابعی بنویسید که با گرفتن مجموعه‌داده و برچسب داده‌ها، **میانگین فاصله‌ی نقاط تا مرکز خوشه‌ی متناظر با هر کدام** را محاسبه کند. اگر فاصله‌ی نقاط از مرکز خوشه را به عنوان خطا در نظر بگیریم، میانگین

[1] centroid

این مقادیر، که با عنوان Mean Absolute Error (MAE) شـــناخته می‌شـــود، یک معیار برای ارزیابی میزان دقت الگوریتم خواهد بود.

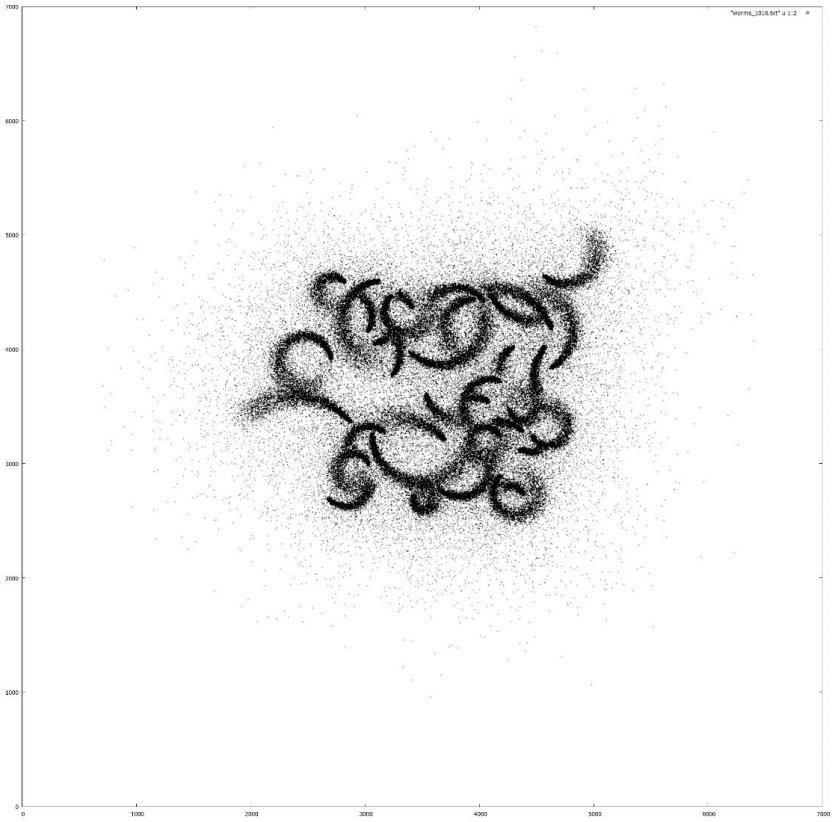
۳) حال تابع ســـوال اول را برای مجموعه‌داده‌ی Iris، به ازای k های ۱، ۲، ۳، ۴ و ۵ اجرا کنید. برای هر k، میانگین فاصله‌ی داده‌ها را تا مرکز خوشه‌های متناظرشان (خروجی تابع سوال دو) بیابید و نمودار آن را به ازای k های مختلف رســـم کنید. (دقت کنید که پیش از اجرا باید ســـتون آخر را که برچســـب هر رکورد اســـت از مجموعه‌داده حذف کنید)

۴) با استفاده از روش elbow توضیح دهید که کدام k مناسب‌ترین گزینه برای خوشه‌بندی است.

در قسمت بعد، برای اجرای الگوریتم DBSCAN می‌توانید از کتابخانه SKlearn استفاده کنید:

https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html

۵) مجموعه‌داده‌ی Worms را خوانده و نمودار آن را رســـم کنید. دقت کنید که به دلیل زیاد بودن نقاط این مجموعه‌داده شـــاید نیاز باشـــد تا اندازه‌ی brush را برای نقاط کاهش دهید. نمودار این مجموعه‌داده باید مشـــابه شکل زیر باشد:



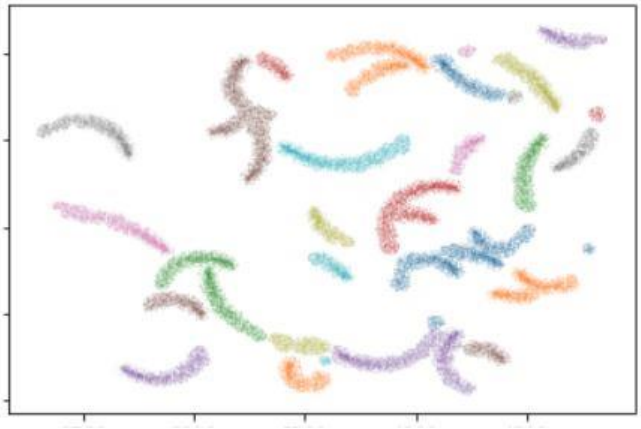
۶) یک بار این مجموعه‌داده را با تابع k-means سوال اول (که خودتان نوشتید) با یک k دلخواه خوشه‌بندی کرده و نتیجه را رسم کنید. (بهتر است تابعی بنویسید که با گرفتن یک مجموعه‌داده و برچسب‌های آن، نمودار آن‌ها را با رنگ‌های مختلف رسم کند تا قابل فهم باشد)

همانطور که از خروجی قسمت قبل مشخص است، k-means نمی‌تولند این نوع مجموعه‌داده را به خوبی خوشه‌بندی کند. پس سراغ الگوریتم مناسب‌تری برای خوشه‌بندی این مجموعه‌داده می‌رویم.

۷) در این بخش سعی می‌کنیم با استفاده از الگوریتم DBSCAN این مجموعه‌داده را خوشه‌بندی کنیم، بطوری که بر اساس خروجی آن، یک حدس نسبتاً دقیق از تعداد این موجودات در مجموعه‌داده به دست آوریم.

از آنجایی که الگوریتم DBSCAN نیازی به تعیین تعداد خوشه‌ها ندارد، وظیفه شما این است که دو پارامتر اصلی آن را یعنی **اپسیلون و حداقل تعداد نقاط هر خوشه** به نحوی تعیین کنید که خروجی دقیق‌تر شود. پس با آزمودن مقادیر مختلف این دو پارامتر و رسم نمودار این نقاط متناسب با خوشه‌های به دست آمده، بهترین حالتی که می‌توانید را ثبت کرده و آن را رسم کنید و به همراه تعداد خوشه های به دست آمده (تعداد برچسب‌های خروجی) در گزارش خود ذکر کنید. دقت داشته باشید که در این قسمت به دنبال یک پاسخ مشخص نیستیم، اما انتظار می‌رود پاسخ شما خیلی دور از واقعیت نباشد. (راهنمایی: تعداد موجودات کمتر از ۱۵ و بیشتر از ۵۰ نباشد)

یک نمونه خروجی خوشه‌بندی شده مجموعه‌داده Worms با الگوریتم DBSCAN بصورت شکل زیر است:



موفق و سربلند باشید