هدف: آشنایی با مفاهیم پایه خوشه‌بندی و PCA

تذکر : ۱ ملاک اصلی انجام تمارین بخش پیاده سازی، گزارش است و ارسال کد بدون گزارش فاقد ارزش تهیه کنید و در آن برای هر سوال، تصاویر pdf است .لذا برای این بخش یک فایل گزارش مستقل در قالب ورودی، تصاویر خروجی و توضیحات مربوط به آن را ذکر کنید .سعی کنید توضیحات کامل و جامعی تهیه کنید .همچنین قابل توجه است که زبان مورد قبول برای بخش پیاده سازی، تنها پایتون میباشد.

تذکر : ۲ مطابق قوانین دانشگاه هر نوع کپی برداری و اشتراک کار دانشجویان غیر مجاز بوده و شدیدا برخورد خواهد شد .استفاده از کدها و توضیحات اینترنت به منظور یادگیری الزاما با ذکر منبع بلامانع است.

راهنمایی :در صورت نیاز میتوانید سواالت خود را در خصوص پروژه از تدریسیار درس، از طریق ایمیل زیر بپرسید.

E-mail: zahra.dehghanian97@gmail.com

ارسال : فایل های کد و گزارش خود را در قالب یک فایل فشرده با فرمتStudentID_DM04.zip ارسال نمایید. مهلت ارسال بخش نوشتاری تمرین تا ۱۴۰۰/۱۰/۲۳ است، مهلت ارسال بخش پیاده سازی تا ۱۴۰۰/۱۰/۳۰ می باشد.

*شایان ذکراست در مجموع برای تمارین ۷ روز تاخیر مجاز در نظر گرفته شده است و افزون بر آن هر روز تاخیر باعث کسر ۲۰ % نمره کل تمرین خواهد شد.

- **بخش نوشتاری**

۱) فرض کنید که میخواهید داده های زیر را به سه دسته خوشه بندی کنید. داده ها به صورت زیر می‌باشند

$$A_1(2,10), A_2(2,5), A_3(8,4), B_1(5,8), B_2(7,5), B_3(6,4), C_1(1,2), C_2(4,9).$$

فاصله را فاصله اقلیدسی در نظر بگیرید. فرض کنید که در ابتدا نقاط A_1 و B_1 و C_1 را مرکز هر خوشه در نظر بگیرید. با استفاده از الگوریتم K_Means :

الف) مرکز سه خوشه را بعد از یک بار اجرای الگوریتم به دست آورید.
ب) سه خوشه نهایی را به دست آورید.

۲) الف) شرایطی را توضیح دهید که در آن خوشه‌بندی مبتنی بر چگالی، مناسب‌تر از خوشه‌بندی مبتنی بر تقسیم‌بندی و خوشه‌بندی سلسله مراتبی است. ویژگی‌های هر رویکرد را توضیح دهید، مثال‌هایی از کاربردهای خوشه‌بندی ارائه کنید تا از استدلال خود پشتیبانی کنید.

ب) توضیح دهید که چه زمانی dbscan خوب کار نمی‌کند و چه زمانی خوب کار می‌کند؟ چرا؟

ج) کدام یک از موارد زیر در مورد الگوریتم خوشه‌بندی dbscan درست است؟ پاسخ خود را برای هر گزینه توضیح دهید.

۱. برای نقاط داده که باید در یک خوشه باشند، آن‌ها باید در یک آستانه فاصله تا یک نقطه مرکزی باشند.

۲ . این مدل دارای فرضیات قوی برای توزیع نقاط داده در فضای داده است.

۳ . پیچیدگی زمانی بسیار بالایی برای مرتبه $O(n^3)$ دارد.

۴ . نیازی به دانش قبلی درمورد تعداد خوشه‌های مد نظر نیست.

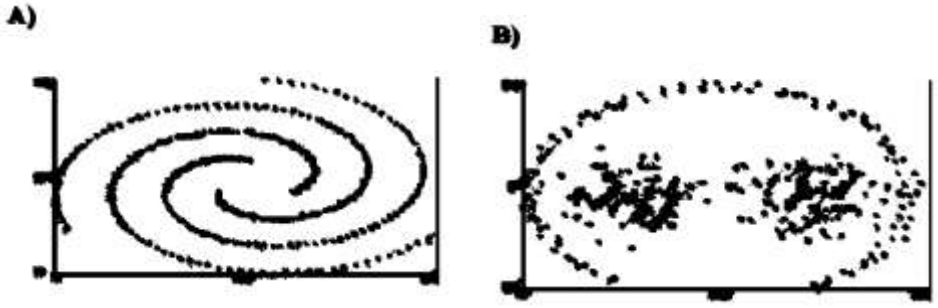
۵ . این روش نسبت به داده پرت مقاوم است.

۳) از ماتریس شباهت در جدول زیر برای انجام خوشه‌بندی سلسله مراتبی با لینک تک و کامل (min , max) استفاده کنید. نتایج خود را با کشیدن یک دندروگرام نشان دهید. در رسم باید به روشنی ترتیب ادغام نقاط نشان داده شود. (معیار اندازه‌گیری شباهت را به عنوان کمترین فاصله دو نقطه در خوشه‌های مختلف در نظر بگیرید)

	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

۴) در تشخیص کیفیت روش‌های خوشه‌بندی برای داده‌های دو بعدی، چشم انسان به سرعت و موثر عمل می‌کند. آیا می‌توانید یک روش طراحی کنید که به انسان‌ها کمک کند تا خوشه‌های داده های سه بعدی را شناسایی و کیفیت خوشه‌بندی را ارزیابی کند؟ برای داده‌های با ابعاد بالاتر چطور؟

۵) تصویر زیر را در نظر بگیرید . کدام یک از الگوریتم‌های خوشه‌بندی در خوشه‌بندی دقیق داده‌های داده‌شده به خوبی عمل خواهند کرد؟ k-means، dbscan، یا هر دو؟ توضیح دهید که چرا فکر می‌کنید که یک الگوریتم خوب عمل می‌کند یا خوب عمل نمی‌کند.



۶) بسیاری از الگوریتم‌های خوشه‌بندی مبتنی بر تقسیم‌بندی که به طور خودکار تعداد خوشه‌ها را تعیین می‌کنند، ادعا می‌کنند که این یک مزیت است. دو موقعیت را نام ببرید که در آن‌ها این مورد صدق نمی‌کند.

۷) داده های زیر را با الگوریتم PCA به یک بعد تقلیل دهید.

(۲,۱) , (۳, ۵), (۴, ۳), (۵, ۶), (۶, ۷), (۷, ۸)

- **بخش پیاده سازی:**

در این بخش شما از الگوریتم خوشه‌بندی K-means استفاده خواهید کرد و یاد می‌گیرید که مجموعه داده ۲ بعدی ساده را خوشه‌بندی کنید.

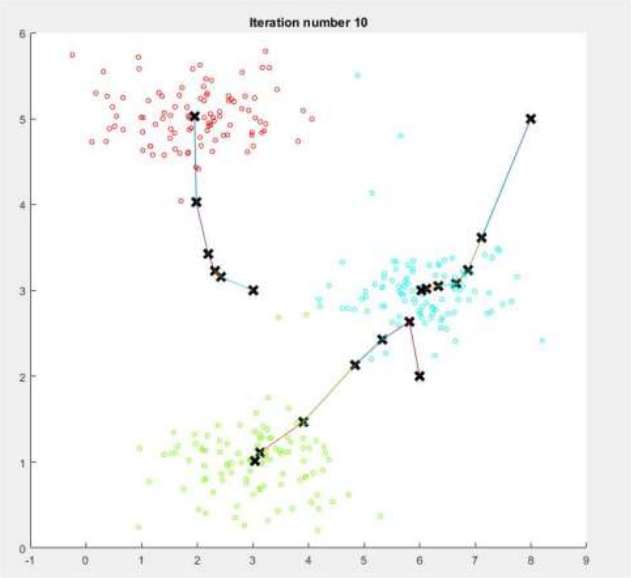
۱) همانطور که در کلاس درس گفته شد، ایده اصلی پشت K-means یک فرآیند تکراری است که با حدس زدن مراکز خوشه اولیه شروع می‌شود و سپس این حدس را با تخصیص مکرر نقاط داده به نزدیک‌ترین مرکز خوشه خود بهبود می‌بخشد. شبه کد K-means به شرح زیر است :

```
// X is the matrix of input data, each row is a data point and each column is a data feature
// K is the number of desired clusters
// n is the number of iteration that we want the k-means to run on the data
// we first randomly Initialize centers
centers = initialize_centers(X, K);
// centers is a matrix with K rows , each row is one center and each column is a feature
for i = 1 to n
        // cluster assignment step: Assign each data point to the
        // closest center.idx is vector and  idx(i) is  the index
        // of the center assigned to example data point i
        idx = findClosestCenters(X, centroids);

        // move centers step: recompute each center based on the mean of the data
        // assigned to that center
        centers = computeMeans(X, idx, K);
end for
```

توجه داشته باشید که برای اجرای K-means شما باید توابع findClosestCenters، computeMeans و initializecenters را کامل کنید. همچنین از فاصله اقلیدسی برای اندازه‌گیری فاصله بین نقاط داده استفاده کنید.

بعد از تکمیل الگوریتم، آن را روی Dataset1 اجرا کنید. تعداد تکرار را حداقل ۱۵ در نظر بگیرید و K-means را با ۲، ۳ و ۴ دسته اجرا کنید. بعد از هر بار اجرا، نقاط داده هر خوشه را با رنگ متفاوتی مثل شکل زیر رسم کنید.



۲) همانطور که می‌دانید، الگوریتم PCA برای کاهش ابعاد دیتاست استفاده می‌شود.
در این تمرین ابتدا الگوریتم PCA را پیاده سازی کنید، سپس Dataset2 داده شده را به وسیله تابع StandardScaler در کتابخانه sklearn برای n_component = 2 به دست آورید و بر روی دیتاست داده شده fit کنید.
حال به وسیله تابع train_test_split در کتابخانه sklearn ، با پارامتر۰,۲ به دو دسته train و test تقسیم کنید .
سپس به وسیله الگوریتم SVM و به کمک کتابخانه sklearn ، متریک accuracy را به دست آورید و مقدار آن را گزارش دهید.

توجه : برای پیاده سازی الگوریتم PCA **نمی‌توانید** از کتابخانه sklearn استفاده کنید.