سوالات تشریحی

۱- نقاط A1 تا A8 را در نظر بگیرید. با از استفاده الگوریتم k-means و فاصله اقلیدسی، نقاط داده شده را در ۳ خوشه، خوشه‌بندی کنید. ماتریس فاصله این نقاط بر اساس متر اقلیدسی به شکل زیر است:

	x	y
A1	2	12
A2	3	5
A3	8	4
A4	6	13
A5	13	5
A6	10	6
A7	2	2
A8	4	13

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	7.0711	10	4.123	13.038	10	10	2.236
A2		0	5.099	8.544	10	7.071	3.162	8.062
A3			0	9.22	5.099	2.828	6.325	9.849
A4				0	10.63	8.062	11.7	2
A5					0	3.162	11.4	12.04
A6						0	8.944	9.22
A7							0	11.18
A8								0

فرض کنید نقاط اولیه (مرکز[۱] هر خوشه) نقاط A4، A7 و A8 باشند. الگوریتم k-means را تنها برای یک مرحله[۲] اجرا کنید. در انتهای این مرحله به سوالات زیر پاسخ دهید:

الف) هر نقطه متعلق به کدام خوشه است؟

ب) مرکز خوشه‌های جدید را مشخص کنید.

ج) در یک صفحه مختصات تمام ۸ نقطه را کشیده و خوشه‌های بدست آمده بعد از مرحله اول را به همراه مرکز های جدید آن‌ها رسم کنید.

د) چه تعداد تکرار دیگر از الگوریتم برای همگرایی آن نیاز است؟ نتایج هر مرحله (نقاط متعلق به هر خوشه و مرکز آن) را بدست آورده و رسم کنید.

ه) با توجه به فاصله اقلیدسی نقاط، آیا خوشه‌بندی بدست آمده بهترین خوشه‌بندی ممکن است؟ استدلال خود را شرح دهید و اگر خوشه بندی به دست آمده بهترین نیست، راهکاری برای حل آن ارائه کرده و توضیح دهید.

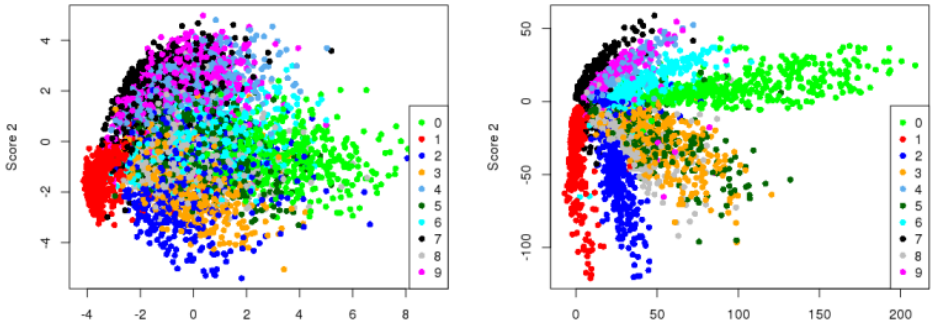
۲- از روش های تغییر نمایش[3] یا کاهش بُعد داده‌ها می توان به PCA و Autoencoder اشاره کرد. با در نظر گرفتن این دو روش به سوالات زیر پاسخ دهید:

الف) مزایا و معایب این دو روش را بیان کنید و آن ها را با یکدیگر مقایسه کنید. (۴ مورد)

ب) اگر هدف ما استخراج ویژگی‌های مستقل یا ناهمبسته خطی[4] باشد کدام روش را توصیه میکنید؟ علت را شرح دهید.

ج) آیا ممکن است خروجی این دو روش یکسان شود؟ در صورت مثبت بودن پاسخ، چگونگی این اتفاق را توضیح دهید.

د) فرض کنید مجموعه داده‌ای با ارتباطات غیرخطی داریم و برای تغییر نمایش از دو الگوریتم PCA و AE استفاده کرده‌ایم که نتایج آن در دو نمودار زیر آمده است. کدام نتیجه متعلق به PCA و کدام یک متعلق به AE است؟ دلیل خود را شرح دهید.



پیاده‌سازی

دیتاست

سازمانی مردم‌نهاد به نام HELP قصد دارد کمک‌های بشردوستانه خود را به کشورهایی که بیشترین نیاز به آن را دارند اختصاص دهد. به این منظور دیتاستی به نام Country data در اختیار گذاشته که شامل اطلاعات مرتبط با وضعیت اجتماعی، اقتصادی و فاکتورهای سلامتی می‌شود تا بتوان به طور کلی میزان توسعه هر کشور را بررسی کرد. بنابراین هدف مسئله این است که مجموعه کشورها را با توجه به اطلاعات موجود در داده ها و میزان توسعه آن‌ها به درستی خوشه‌بندی کنید. در انتها کشورهایی که بیشترین نیاز را به دریافت کمک دارند، به این سازمان پیشنهاد دهید.

جهت دانلود مجموعه داده و بررسی اطلاعات بیشتر می‌توانید به این لینک مراجعه کنید.

[3] representation
[4] linearly uncorrelated

https://drive.google.com/file/d/1oWeE67I4yyx4E2RUU-zJzvEO3yjb08zR/view

بخش اول: خوشه‌بندی

در این بخش هدف خوشه‌بندی داده‌ها، ارزیابی و تفسیر نتایج آن است.

عملیات

الف) پس از بررسی اولیه داده، با استفاده از ماتریس آشفتگی میزان همبستگی ویژگی‌ها را بررسی کنید. آیا می‌توان برخی از ویژگی‌ها را حذف کرد؟ علت آن را بیان کنید.

ب) آیا نیازی به نرمال کردن داده وجود دارد؟ علت پاسخ خود را بیان کرده و عملیات مورد نیاز را بر روی داده انجام دهید.

ج) با استفاده از الگوریتم k-means عملیات خوشه‌بندی را انجام دهید. برای تعیین تعداد بهینه خوشه‌ها، از روش Elbow کمک بگیرید و نتیجه را ارائه دهید.

د) معیار ارزیابی silhouette یکی از معیارهای ارزیابی کیفیت خوشه‌بندی است. درباره چگونگی ارزیابی این معیار تحقیق کرده و به صورت مختصر توضیح دهید.

ه) کیفیت خوشه ها را توسط معیار silhouette ارزیابی کرده و نتیجه را تفسیر کنید. (جهت پیاده‌سازی می‌توانید از تابع silhouette_score در کتابخانه sklearn استفاده کنید.)

و) به دلخواه خود سه ویژگی را انتخاب کرده و با استفاده از نمودار scatter آن‌ها را دو به دو رسم کنید. هر خوشه را به تفکیک رنگ در نمودار نمایش دهید و نتایج را تفسیر کنید.

بخش دوم: تغییر نمایش داده

در این بخش هدف کاهش بعد توسط الگوریتم PCA و پیدا کردن مولفه‌های اصلی داده است.

عملیات

الف) الگوریتم PCA را بر روی داده‌های نرمال شده اجرا کنید. (جهت پیاده‌سازی می‌توانید از تابع PCA در کتابخانه sklearn استفاده کنید.)

ب) چه تعداد از مولفه‌های اساسی[5] می‌توانند توزیع داده‌ها را به خوبی توضیح دهند؟ برای بیان نتایج از نمودار Percentage of Explained Variance بر حسب مولفه‌ها استفاده کنید و تحلیل این معیار تصمیم گیری را در گزارش ذکر کنید.

ج) با استفاده از نتایج بدست آمده بُعد داده را کاهش دهید. (مولفه‌های اساسی را نگه داشته و مابقی را حذف کنید.)

د) نمونه ها را پس از اعمال PCA بر اساس دو مولفه اصلی آن ها در مختصات دوبعدی رسم کنید و تفکیک پذیری بصری آن ها را با شکل رسم شده در بخش (و) قسمت قبل مقایسه کنید.

ه) عملیات خوشه‌بندی را بر روی داده‌های بدست آمده از قسمت (ج) تکرار کنید و نتایج آن را با بخش اول مقایسه و تفسیر کنید.

نکات مهم

- جهت انجام بخش پیاده سازی، از Jupyter استفاده کنید و فایل نهایی را با پسوند. ipynb آپلود کنید. همچنین شماره دانشجوی خود را به عنوان نام فایل در نظر بگیرید.
- نکته مهم در گزارش نویسی و سوال تشریحی روشن بودن پاسخ می‌باشد نه حجم زیاد، اگر فرضی برای حل سوال استفاده می‌کنید حتما آن را ذکر کنید، و پاسخ نهایی را به صورت واضح بیان کنید.
- هرگونه شباهت در گزارش و پاسخ تشریحی به منزله تقلب می‌باشد و کل نمره تمرین صفر در نظرگرفته می شود. **(می‌توانید از اینترنت به عنوان منبع کمکی هم در سوالات تشریحی و هم در سوالات پیاده سازی استفاده کنید، اما کپی برداری ممنوع می باشد و نمره‌ی صفر تعلق می‌گیرد)**
- گزارش و تحلیل های خواسته شده در کد و پاسخ سوالات تشریحی باید در یک فایل pdf باشد.
- **توجه شود: در این تمرین، پاسخ به سوالات تشریحی که تنها نیاز به توضیح و تشریح مسئله دارد ، باید به صورت تایپ شده باشد. اما در مسائل حل کردنی که نیاز به رسم یا استفاده از فرمول‌های ریاضی وجود دارد، تایپ ضرورتی ندارد.**
- فایل pdf و کد را بصورت یکجا در قالب یک فایل zip در سامانه کورسز آپلود کنید (نام فایل = شماره دانشجویی).