دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

دانشکده ریاضی و علوم کامپیوتر

تمرین چهارم درس یادگیری ماشین

عنوان تمرین : Unsupervised Learning, Clustering

استاد درس :

دکتر اکبری

تاریخ انتشار : 26 آذر

تاریخ برگزاری کلاس رفع اشکال : 1 دی

مهلت تحویل : 8 دی

تاریخ برگزاری کارگاه : 16 دی

بخش اول : سؤالات 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[1] هر خوشه) نقاطA4 ، A7 و A8 باشند. الگوریتم k-means را تنها برای یک مرحله[2] 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 به‌دست‌آمده 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 به دست آورده و رسم کنید.

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

---

[1] centroid

[2] epoch

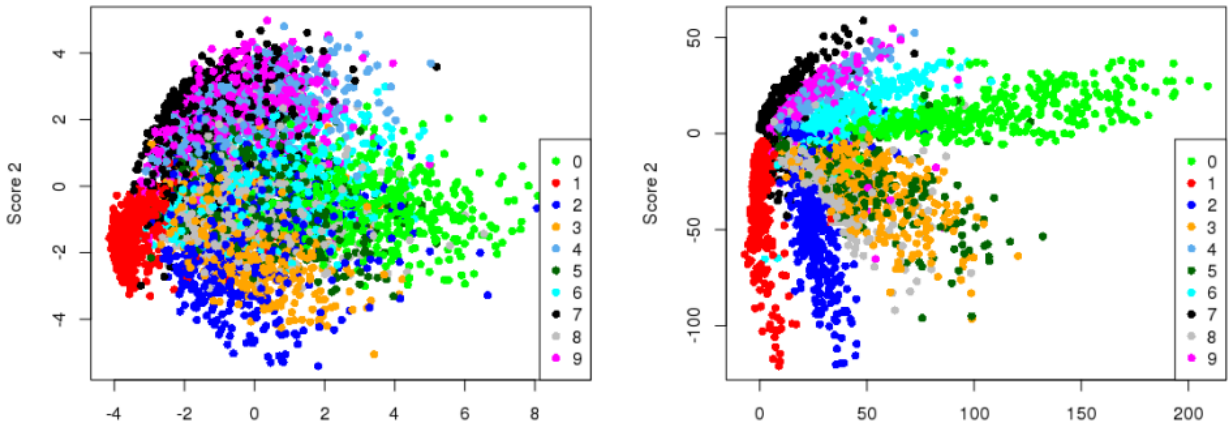
۲) از روش‌های تغییر نمایش[۳] یا کاهش بُعد داده‌ها می توان به PCA و Autoencoder اشاره کرد. با در نظر گرفتن این دو روش به سوالات زیر پاسخ دهید:

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

ب) اگر هدف ما استخراج ویژگی های مستقل یا ناهمبسته خطی[۴] 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 مجموعه‌داده ای با ارتباطات غیرخطی داریم و برای تغییر نمایش از دو الگوریتم PCA و AE استفاده کرده ایم که نتایج آن در دو نمودار زیر آمده است. کدام نتیجه متعلق به PCA و کدام یک متعلق به AEاست؟ دلیل خود را شرح دهید.



representation [3]

linearly uncorrelated [4]

بخش دوم : Reading Assignment

در این بخش باید درباره یکی از موضوعات زیر که در کلاس تدریس نشده، تحقیق و مطالعه کنید. هدف از این تمرین آشنایی با روند یادگیری مطالب جدید و همین‌طور ارائه این مطالب است. گزارش تحویلی شما باید بین دو الی سه صفحه باشد.

1. **Kernel PCA (link)**
2. **Variational Autoencoder (link)**
3. **Spectral clustering (link)**
4. **Linear Discriminant Analysis for Dimensionality Reduction (LDA) (link)**

- نحوه تخصیص موضوعات به شکل تصادفی بوده و از طریق فرمول (n % 4 + 1) به دست می‌آید که n دو رقم آخر شماره دانشجویی شما است.
- می‌توانید از هر منبع دیگری در کنار منبع کمکی استفاده کنید.

بخش سوم : پیاده‌سازی

دیتاست قسمت اول و دوم و سوم : Country-data

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 در اختیار گذاشته که شامل اطلاعات مرتبط با وضعیت اجتماعی، اقتصادی و فاکتورهای سلامتی می‌شود تا بتوان به‌طورکلی میزان توسعه هر کشور را بررسی کرد؛ بنابراین هدف مسئله این است که مجموعه کشورها را با توجه به اطلاعات موجود در داده‌ها و میزان توسعه آن‌ها به‌درستی خوشه‌بندی کنید. در انتها کشورهایی که بیشترین نیاز را به دریافت کمک دارند، به این سازمان پیشنهاد دهید.

توضیحات ستون‌های داده به‌صورت زیر است:

- country: Name of the country
- child_mort: Death of children under 5 years of age per 1000 live births
- exports: Exports of goods and services per capita. Given as %age of the GDP per capita
- health: Total health spending per capita. Given as %age of GDP per capita
- imports: Imports of goods and services per capita. Given as %age of the GDP per capita
- Income: Net income per person
- Inflation: The measurement of the annual growth rate of the Total GDP
- life_expec: The average number of years a newborn child would live if the current mortality patterns are to remain the same
- total_fer: The number of children that would be born to each woman if the current age-fertility rates remain the same
- gdpp: The GDP per capita. Calculated as the Total GDP divided by the total population

قسمت اول: خوشه‌بندی با استفاده از روش‌های K-means و GMM

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

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

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

K-means:

۳.۱) از روش Elbow برای به‌دست‌آوردن تعداد بهینهٔ خوشه‌ها در روش k-means استفاده کنید و نتیجه را ارائه دهید.

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

۳.۳) این بار با استفاده از معیار silhouette تعداد بهینه خوشه‌ها را در روش k-means به دست آورید و با روش Elbow مقایسه کنید. (جهت پیاده‌سازی می‌توانید از تابع silhouette_score در کتابخانه sklearn استفاده کنید.)

۳.۴) الگوریتم k-means را با تعداد بهینهٔ خوشه‌ها اجرا کنید و شمارهٔ خوشهٔ هر داده را ارائه دهید.

۳.۵) 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.

GMM (Gaussian Mixture Model) :

۴.۱) با استفاده از معیار silhouette تعداد بهینه خوشه‌ها را در روش GMM به دست آورید.

۴.۲) الگوریتم GMM را با تعداد بهینهٔ خوشه‌ها اجرا کنید و شمارهٔ خوشهٔ هر داده را ارائه دهید.

۴.۳) همان سه ویژگی که در مرحلهٔ قبل انتخاب کردید را با استفاده از نمودار scatter دو به دو رسم کنید. هر خوشه را به تفکیک رنگ در نمودار نمایش دهید و نتایج را تفسیر کنید.

۴.۴) نتایج را با مدل k-means مقایسه کنید.

نکته: برای پیاده‌سازی الگوریتم‌های k-means و GMM می‌توانید از کتابخانه sklearn استفاده کنید.

قسمت دوم: خوشه‌بندی با استفاده از روش Spectral Clustering

در این قسمت شما باید الگوریتم خوشه‌بندی Spectral را به‌صورت زیر پیاده‌سازی کنید و نتایج را با روش‌های قبل و همچنین مدل از پیش تعریف شده Sklearn مقایسه کنید.

۱.۱) برای به‌دست‌آوردن گراف نزدیک‌ترین همسایه از kneighbors_graph یا radius_neghbors_graph در کتابخانه Skleran استفاده کنید.

۱.۲) ماتریس Laplasian گراف را به دست آورده و مقادیر و بردارهای ویژه را به دست آورید.

۱.۳) یک الگوریتم خوشه‌بندی مانند الگوریتم K-means را روی بردارهای ویژه اجرا کنید.

۲) با استفاده از معیار silhouette تعداد بهینه خوشه‌ها را در الگوریتم Spectral Clustering که خودتان پیاده‌سازی کردید به دست آورید.

۳) این الگوریتم را با تعداد بهینهٔ خوشه‌ها اجرا کنید و شمارهٔ خوشهٔ هر داده را ارائه دهید.

۴) همان سه ویژگی که در مرحلهٔ قبل انتخاب کردید را با استفاده از نمودار scatter دو به دو رسم کنید. هر خوشه را به تفکیک رنگ در نمودار نمایش دهید و نتایج را تفسیر کنید.

۵) شماره‌های ۲ تا ۴ را برای مدل آماده الگوریتم Spectral Clustering در کتابخانهٔ Sklearn انجام دهید و نتایج را با الگوریتمی که خودتان پیاده‌سازی کردید مقایسه کنید.

۶) نتایج به‌دست‌آمده با این روش را با روش‌های k-means و GMM مقایسه کنید.

قسمت سوم: PCA

در این بخش هدف کاهش بعد توسط الگوریتم PCA و پیداکردن مؤلفه‌های اصلی[5] داده است .

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

۲) چه تعداد از مؤلفه‌های اساسی می‌توانند توزیع داده‌ها را به‌خوبی توضیح دهند؟ برای بیان نتایج از نمودارPercentage of Explained Variance بر حسب مؤلفه‌ها استفاده کنید و تحلیل این معیار تصمیم‌گیری را در گزارش ذکر کنید.

۳) با استفاده از نتایج به‌دست‌آمده، بعد داده را کاهش دهید. (مؤلفه‌های اساسی را نگه داشته و مابقی را حذف کنید.)

۴) بر روی داده‌های به‌دست‌آمده عملیات خوشه‌بندی را با روش‌های k-means، GMM و Spectral از شمارهٔ ۲ تا ۴ قسمت قبل تکرار کنید (یعنی تعداد بهینهٔ خوشه‌ها را به دست آورید، الگوریتم را اجرا کنید و نمودار scatter را رسم کنید.) و نتایج را با قسمت قبل مقایسه و تفسیر کنید.

---

[5] Principal Component

دیتاست قسمت چهارم: shuttle

داده‌ها شامل ۹ ویژگی و ۲ برچسب هستند. (برچسب 1 : outlier و برچسب 0: inlier)

قسمت چهارم : شناسایی داده‌های outlier با استفاده از خوشه‌بندی

در این بخش هدف پیداکردن outlier ها در مجموعه‌داده گفته شده با روش‌های مختلف خوشه‌بندی است.

۱) الگوریتم k-means را با تعداد خوشه‌های ۲ اجرا کنید.

۲) خوشهٔ با تعداد داده‌های کمتر را خوشهٔ داده‌های outlier و خوشه با تعداد داده‌های بیشتر را خوشه داده‌های inlier در نظر بگیرید و برچسب‌هایی که نسبت داده‌اید را با برچسب واقعی مقایسه کنید و دقت مدل (accuracy و f1-score) را محاسبه کنید.

۳) این کار را با الگوریتم‌های GMM و Spectral نیز انجام دهید و نتایج را مقایسه کنید.

قسمت پنجم (امتیازی) : پیاده‌سازی الگوریتم Kmeans و GMM

برای پیاده‌سازی این بخش به موارد زیر توجه کنید.

- بعد از پیاده‌سازی الگوریتم‌های خوشه‌بندی Kmeans و GMM یک مجموعه‌داده با حداقل 150 هزار نمونه انتخاب کنید.
- سپس به کمک مدل‌هایی که پیاده‌سازی کردید خوشه‌بندی را روی مجموعه‌داده انتخابی و با تعداد خوشه‌های 5 انجام دهید. سپس همین روند را به کمک مدل‌های آماده کتابخانه Sklearn انجام دهید.
- در نهایت عملکرد مدل‌هایی که پیاده‌سازی کردید را از نظر زمان اجرا و همین‌طور شاخص‌های ارزیابی خوشه‌بندی مثل silhouette و ... ارزیابی و نتایج را گزارش دهید.

معیار ارزیابی شما

بخش اول سؤالات تشریحی (۴۰ نمره) : سؤال اول (۳۰ نمره) / سؤال دوم (۱۰ نمره)

بخش دوم Reading Assignment (۲۰ نمره)

بخش سوم : پیاده‌سازی (۲۴۰ نمره) :

- قسمت اول (۷۰ نمره)
- قسمت دوم (۷۰ نمره)
- قسمت سوم (۶۰ نمره)

- قسمت چهارم (۴۰ نمره)
- قسمت پنجم (امتیازی) : Kmeans (۴۰ نمره) و GMM (۶۰ نمره)
- این تمرین حدود ۳ نمره از ۲۰ نمره نهایی شما را شامل می‌شود.

نکات تکمیلی

✓ جهت انجام بخش پیاده سازی، از Jupyter استفاده کنید و فایل نهایی را با پسوند ipynb آپلود کنید.

✓ انجام این تمرین بسته به تسلط شما به مطالب درس و زبان پایتون حداقل بین ۶ الی ۸ روز از وقت مفید شما را خواهد گرفت. به همین علت انجام این تمرین را به‌روزهای پایانی موکول نکنید. همین‌طور باتوجه به برنامه فشرده کلاس و حجم زیاد مطالب، مهلت تحویل این تمرین تمدید نخواهد شد.

✓ **ارسال گزارش اجباری است.** نکته مهم در گزارش‌نویسی و سؤال تشریحی روشن بودن پاسخ است نه حجم زیاد، اگر فرضی برای حل سؤال استفاده می‌کنید حتماً آن را ذکر کنید، و پاسخ نهایی را به‌صورت واضح بیان کنید. گزارش کد و پاسخ سؤال تشریحی به‌صورت فایل pdf باشد.

✓ هرگونه شباهت در گزارش و پاسخ تشریحی به منزله تقلب است و کل نمره تمرین را نخواهید گرفت.

✓ فایل pdf مربوط به بخش اول و دوم تمرین و همین‌طور گزارش مربوط به بخش پیاده‌سازی را به همراه کدها را به‌صورت یکجا در قالب یک فایل zip در سامانه کورسز آپلود کنید (نام فایل = شماره دانشجویی)

✓ در صورت هرگونه ابهام درباره این تمرین می‌توانید در کلاس رفع اشکال سؤالات خودتان رو بپرسید و یا از طریق ایمیل‌های زیر با ما در ارتباط باشید.

ملیکا سپیدبند : melikasepidband@aut.ac.ir

امیر گودرزی : amirgoudarzi023@gmail.com

محمدعلی سفیدی اصفهانی : mohammadali.esfahani@aut.ac.ir

با آرزوی سلامتی و موفقیت برای شما عزیزان