به نام خدا

دانشگاه تهران

دانشکده مهندسی برق و کامپیوتر

درس داده کاوی

تمرین پنجم

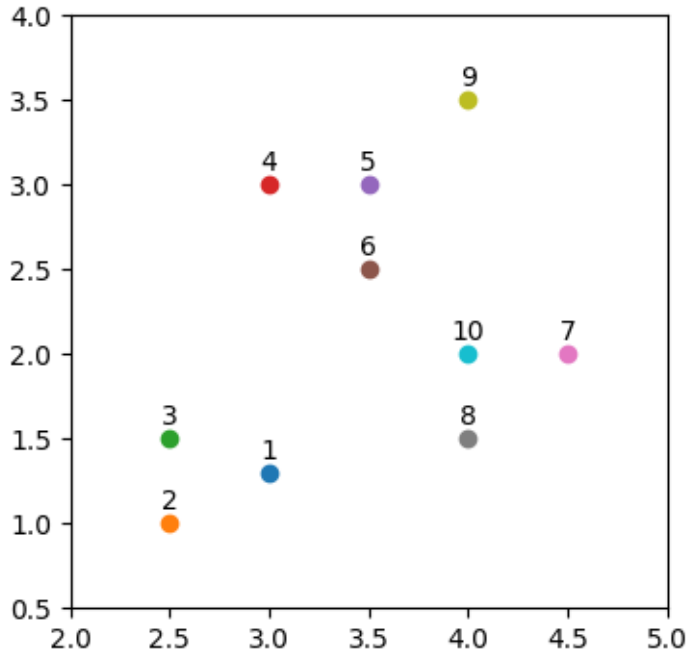
طراح	محمدرضا علائی mr.alaei@ut.ac.ir
تاریخ بارگذاری	۱۴۰۳/۰۳/۰۲
مهلت ارسال	۱۴۰۳/۰۳/۱۲

فهرست

بخش تشریحی

سوال اول

مجموعه‌ای از نقاط در مختصات دکارتی در نمودار ۱ آورده شده‌است.



نمودار ۱

فاصله‌ی منهتن میان هر جفت نقاط نیز در جدول ۱ آورده شده‌است. برای پاسخ به سوالات این بخش، این معیار را به عنوان فاصله‌ی بین نقاط در نظر بگیرید.

جدول ۱. فاصله‌ی منهتن میان نقاط

Points	1	2	3	4	5	6	7	8	9	10
1		0.8	0.7	1.7	2.2	1.7	2.2	1.2	3.2	1.7
2			0.5	2.5	3	2.5	3	2	4	2.5
3				2	2.5	2	2.5	1.5	3.5	2
4					0.5	1	2.5	2.5	1.5	2
5						0.5	2	2	1	1.5
6							1.5	1.5	1.5	1
7								1	2	0.5
8									2	0.5
9										1.5

الف- با استفاده از پارامترهای زیر، الگوریتم DBSCAN را بر روی این نقاط اجرا کنید و خوشه‌ها، نقاط مرزی[۱]، نقاط مرکزی[۲] و داده‌های پرت[۳] را بدست آورید.

- Eps: 0.6, MinPts: 3
- Eps: 0.9, MinPts: 3
- Eps: 1.2, MinPts: 3

ب- با توجه به نتایج خوشه‌بندی، اثر تنظیم پارامترها را بر روی خوشه‌های شناسایی‌شده بررسی نمایید.

ج- با توجه به خوشه‌بندی انجام‌شده با مقادیر پارامتر Eps: 0.9, MinPts: 3، می‌توانید نقاطی را بیابید که نسبت به یکدیگر Directly Density-reachable باشند ولی در یک خوشه قرار نگرفته باشند؟ همچنین می‌توانید سناریویی را ارائه دهید که دو نقطه از طریق یک نقطه‌ی سوم، Density-connected باشند ولی نسبت به یکدیگر Directly Density-reachable نباشند؟

د- حال با استفاده از الگوریتم KMeans، داده‌ها را خوشه‌بندی کنید. تعداد خوشه‌ها را ۳ و نقاط مرکزی[۴] ابتدایی را سه نقطه‌ی ۶، ۸ و ۱۰ در نظر بگیرید.

ه- خوشه‌های به‌دست‌آمده از دو روش فوق را مقایسه کنید. در کدام روش تاثیر بیشتری از داده‌های پرت در روند خوشه‌بندی مشاهده می‌شود؟ چه راه‌حلی را برای این مورد پیشنهاد می‌کنید؟

---

[۱] Border points
[۲] Core points
[۳] Outliers
[۴] Centroids

و- نمودارهای Dendrogram الگوریتم خوشه‌بندی Agglomerative را با دو روش single-link و complete-link رسم کنید.

ز- دو مورد از نقاط ضعف روش‌های Agglomerative را بیان کنید.

سوال دوم

زمانی که برچسب خوشه‌ی هر داده مشخص باشد، می‌توانیم از معیارهای خارجی[۱] برای سنجش کیفیت خوشه‌بندی استفاده کنیم. این معیارها با در نظر گرفتن دسته‌های واقعی داده‌ها، به ارزیابی نتایج یک خوشه‌بندی می‌پردازند.

فرض کنید بر روی مجموعه‌ای از داده‌های مشتریان، خوشه‌بندی انجام گرفته و الگوهای رفتاری هر کدام از مشتریان در خرید را نیز به عنوان برچسب‌های gold، در اختیار داریم. مشتریان به طور کلی در ۴ گروه دسته‌بندی شده‌اند. در جدول ۲، نتایج خوشه‌بندی و توزیع مشتریان هر دسته آورده شده‌است.

جدول ۲. نتایج خوشه‌بندی

	Impulse Buyers	Discount Seekers	Loyal Customers	Infrequent Shoppers
Cluster 1	30	10	5	5
Cluster 2	10	20	15	5
Cluster 3	5	15	25	5

دسته‌بندی G را Ground truth و خوشه‌بندی انجام‌شده را **خوشه‌بندی C** در نظر بگیرید. سپس به سوالات زیر پاسخ دهید.

الف- انتروپی شرطی[۲] G به شرط هر کدام از خوشه‌ها را محاسبه کنید. سپس با توجه به معیار محاسبه‌شده، خوشه‌ها را مقایسه نمایید.

ب- انتروپی شرطی G با داشتن خوشه‌بندی C را محاسبه و گزارش نمایید.

ج- معیار Mutual Information میان دسته‌ها و خوشه‌ها، بیانگر چه مفهومی می‌تواند باشد؟

د- معیار خلوص[۳] را برای این خوشه‌بندی محاسبه کنید.

---

[۱] Extrinsic measures

[۲] Conditional Entropy

[۳] Purity

بخش عملی

شرح دادگان

مجموعه داده‌ی مربوط به کارت‌های اعتباری مشتریان در فایل تمرین قرار گرفته‌است. این مجموعه‌داده شامل ۱۸ ستون است و اطلاعات مربوط به حساب‌های مشتریان در آن آورده شده‌است. شرح هر کدام از ستون‌های این مجموعه‌داده را می‌توانید در جدول ۳ مشاهده کنید.

جدول ۳. شرح ستون‌های مجموعه‌داده

نام ستون	شرح
CLIENT_ID	شناسه‌ی مشتری
ACCOUNT_BALANCE	موجودی قابل برداشت
BALANCE_UPDATE_FREQUENCY	بسامد تراکنش‌های مشتری*
TOTAL_PURCHASES	مجموع مبلغ خریدهای مشتری
SINGLE_PURCHASE_AMOUNT	بیشترین مبلغ خرید یک‌باره
INSTALLMENT_PURCHASES_AMOUNT	مجموع مبلغ قسط‌های مشتری
ADVANCE_CASH_AMOUNT	مجموع مبلغ پیش‌پرداخت‌های مشتری
PURCHASES_UPDATE_FREQUENCY	بسامد خریدهای مشتری*
SINGLE_PURCHASE_FREQUENCY	بسامد خریدهای یک‌باره*
INSTALLMENT_PURCHASES_FREQUENCY	بسامد خریدهای قسطی*
CASH_ADVANCE_FREQUENCY	بسامد پیش‌پرداخت‌های انجام‌شده
CASH_ADVANCE_TRANSACTIONS	تعداد پیش‌پرداخت‌های انجام‌شده
PURCHASES_TRANSACTION_COUNT	تعداد خریدهای انجام‌شده
CREDIT_MAXIMUM	مبلغ اعتبار کارت
AMOUNT_PAID	کل مبلغ پرداخت‌شده توسط مشتری
MINIMUM_PAYMENT_AMOUNT	حداقل مبلغ پرداختی اقساط
FULL_PAYMENT_PERCENTAGE	درصد پرداخت کامل انجام‌شده توسط مشتری
CREDIT_CARD_TENURE	مدت اعتبار کارت برای مشتری

*بسامدها در مقیاسی از ۰ تا ۱ هستند. به صورتی که مقیاس ۱ به معنای بسیار پرتکرار بودن و مقیاس ۰ به معنای به ندرت رخدادن پیشامد مورد نظر خواهد بود.

آشنایی با داده‌ها

در ابتدا یک بررسی اولیه بر روی داده‌ها انجام دهید. مجموعه‌داده را بارگذاری کرده و به موارد زیر پاسخ دهید.

الف- تعداد سطرهای مجموعه‌داده را گزارش نمایید. در هر کدام از ستون‌ها چه تعدادی از سطرها مقدار ندارند؟ استراتژی پیشنهادی شما برای این سطرها چیست؟

ب- چولگی داده‌ها و تعداد داده‌های پرت در هر کدام از ستون‌های عددی را گزارش نمایید. می‌توانید از مقایسه‌ی مقادیر میانگین و میانه برای بررسی چولگی داده‌ها استفاده کنید. استراتژی پیشنهادی شما برای این موارد چیست؟

ج- تعداد مقادیر یکتا در هر کدام از ستون‌ها را گزارش نمایید. آیا در این مجموعه‌داده، سطر تکراری وجود دارد؟

د- نمودار مستطیلی[1] هر کدام از ستون‌های عددی را رسم نموده و تحلیل خود را از بررسی نمودارها را بیان کنید. به عنوان مثال، مشتریان خریدهای یک‌باره را بیشتر ترجیح داده‌اند یا خریدهای قسطی؟ حداقل ۵ مورد از برداشت‌های خود را از نمودارهای رسم‌شده ذکر کنید. تحلیل‌های بیشتر شامل نمره **امتیازی** خواهد بود.

ه- میزان همبستگی مقدارها را در ستون‌های عددی بررسی نمایید. در یک نمودار Heatmap مقادیر مطلق همبستگی بیشتر از ۰٫۴ را نمایش دهید. با توجه به نموداری که رسم نموده‌اید، آیا می‌توان در مورد نگه‌داشتن یا حذف بعضی از ستون‌ها اظهار نظر کرد؟ تحلیل خود را از این نمودار بیان کنید.

[1] Histogram

آماده‌سازی داده‌ها

در قسمت قبل، با شاخصه‌های آماری و ویژگی‌های ستون‌های مجموعه‌داده آشنا شدید. حال به آماده‌سازی داده‌ها برای اجرای الگوریتم‌های خوشه‌بندی می‌پردازیم.

الف- استراتژی‌های پیشنهادی خود را برای مقادیر Null، چولگی داده‌ها و داده‌های پرت، بر روی داده‌ها اجرا نمایید.

ب- بررسی کنید که مشکلات مورد نظر در مجموعه‌داده برطرف شده‌اند یا خیر.

الگوریتم‌های خوشه‌بندی

در این مرحله به اجرای الگوریتم‌های خوشه‌بندی بر روی داده‌ها می‌پردازیم.

الف- تعداد خوشه‌های بهینه را بیابید. روش کار خود و معیار سنجش را بیان کنید.

ب- از روش KMeans برای خوشه‌بندی داده‌ها استفاده کنید.

ج- نتیجه‌ی خوشه‌بندی را در یک نمودار دو بعدی رسم نمایید.

(امتیازی) این کار را با دو روش PCA و t-SNE اجرا نموده و نتایج را مقایسه کنید. به نظر شما، کدام یک از این روش‌ها خروجی بهتری را در بازنمایی خوشه‌ها ارائه داده‌است؟ با مطالعه‌ی ویژگی‌های هر کدام از این دو روش، این تفاوت در خروجی‌ها را چگونه می‌توان توجیه کرد؟

د- در هر خوشه، میانگین مجموعه‌داده را برای هر یک از ویژگی‌های عددی محاسبه و گزارش کنید. بدیهی است این مرحله باید بر روی داده‌های خام و تغییر نیافته‌ی مجموعه‌داده اعمال شود.

ه- **(امتیازی)** سعی کنید توصیفی از اعضای هر یک از خوشه‌ها ارائه دهید.

ملاحظات

- تمامی نتایج شما باید در یک فایل فشرده با عنوان DM_CA5_StudentID تحویل داده 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.
- بخش اصلی نمره به گزارش شما تعلق می‌گیرد و دستیاران الزامی برای اجرای تمام کدهای شما در صورتی که در گزارش به آن‌ها اشاره‌ای نکرده باشید ندارند. لطفاً تمام موارد مورد نیاز را در گزارش ذکر کنید.
- کدهای نوشته شده برای هر بخش را با نام مناسب مشخص کرده و به همراه گزارش تکلیف ارسال کنید. همه‌ی کدهای پیوست گزارش بایستی قابلیت اجرای مجدد داشته باشند. در صورتی که برای اجرا مجدد آنها نیاز به تنظیمات خاصی می‌باشد بایستی تنظیمات مورد نیاز را نیز در گزارش خود ذکر کنید.
- برای تحویل تمرین‌ها از چارچوب قرار داده شده در سامانه و کانال تلگرام استفاده کنید.
- در صورت قصد ارسال تمرین به صورت دیگر (انگلیسی، latex و ...)، لطفاً پیش از ارسال با دستیار مسئول تمرین هماهنگ کنید.
- توجه کنید این تمرین باید به صورت تک نفره انجام شود و پاسخ‌های ارائه شده باید نتیجه فعالیت فرد نویسنده باشد (هم‌فکری خارج از چارچوب و به اتفاق هم نوشتن تمرین نیز ممنوع است). در صورت مشاهده تقلب برای همه‌ی افراد مشارکت کننده، نمره تمرین، صفر در نظر گرفته خواهد شد.
- در صورت بروز هرگونه مشکل با ایمیل زیر در ارتباط باشید:

mr.alaei@ut.ac.ir

مهلت تحویل: ۱۲ خرداد ۱۴۰۳

مهلت تحویل با تاخیر: ۱۹ خرداد ۱۴۰۳