موضوع هفته:

Clustering (K-Means, DBSCAN, Hierarchical)و معیارهای ارزیابی خوشه‌بندی

سؤالات تحلیلی / ریاضی کوتاه

دو تمرین برنامه‌نویسی

بخش اول – سؤالات تحلیلی / مفهومی

سؤال ۱

تابع هدف الگوریتم K-Means معمولاً به صورت مجموع مربعات فاصله‌ی نقاط هر خوشه از مرکز آن خوشه (WCSS) تعریف می‌شود. فرمول محاسبه مرکز جدید را از روی این تابع هدف به دست بیاورید.

سؤال ۲

روش Elbow یکی از روش‌های رایج برای انتخاب تعداد خوشه‌ها در K-Means است. توضیح دهید:

- در این روش چه مقداری بر حسب K روی نمودار رسم می‌شود و شکل کلی نمودار چه انتظاری داریم باشد؟
- منظور از "نقطه‌ی آرنج" چیست و چگونه می‌توان آن را به صورت شهودی تشخیص داد؟
- در چه شرایطی ممکن است نمودار WCSS آرنج واضحی نداشته باشد و این موضوع چه چیزی درباره‌ی ساختار داده به ما می‌گوید؟

سؤال ۳

در خوشه‌بندی بدون ناظر، برچسب‌های خوشه (۰،۱،۲،...) هیچ معنای ذاتی خاصی ندارند و ممکن است بین اجرای الگوریتم یا نسبت به برچسب‌های واقعی کلاس‌ها جابه‌جا شوند. توضیح دهید:

- چرا برای مقایسه‌ی خوشه‌های K-Means با برچسب‌های واقعی کلاس‌ها، معمولاً از الگوریتم Hungarian (یا روش‌های مشابه) استفاده می‌شود؟
- اگر بدون استفاده ازHungarian ، خوشه‌ها را به صورت مستقیم با کلاس‌ها مقایسه کنیم، چه نوع خطای مفهومی یا عددی ممکن است رخ دهد؟
- نقش این الگوریتم در به‌دست‌آوردن "بهترین نگاشت ممکن بین خوشه‌ها و کلاس‌ها" چیست؟

سؤال ۴

دو معیار Adjusted Rand Index (ARI) و Normalized Mutual Information (NMI) از معیارهای مهم برای ارزیابی کیفیت خوشه‌بندی در مقایسه با برچسب‌های واقعی هستند. مقایسه کنید:

- هر کدام به طور شهودی چه چیزی را اندازه‌گیری می‌کنند؟
- هر دو معیار چگونه نسبت به جابه‌جایی برچسب خوشه‌ها (label permutation) ناپایا یا پایا هستند؟
- در چه شرایطی ممکن است مقدار ARI و NMI برای یک خوشه‌بندی اختلاف قابل توجهی داشته باشند و این اختلاف چگونه قابل تفسیر است؟

سؤال ۵

معیارهای ارزیابی مانند ARI و NMI با دقت (Accuracy) چه تفاوتی دارند؟ توضیح دهید:

- چرا استفاده‌ی مستقیم از دقت برای ارزیابی خروجی یک الگوریتم خوشه‌بندی (بدون ناظر) چندان مناسب نیست؟
- چه ویژگی‌هایی در ARI و NMI وجود دارد که آن‌ها را برای سنجش کیفیت خوشه‌بندی جذاب‌تر می‌کند (مثلاً تصحیح نسبت به شانس یا در نظرگرفتن اطلاعات مشترک)؟
- اگر تنها به دقت تکیه کنیم، چه نوع سوءبرداشت‌هایی درباره‌ی خوب بودن یا بد بودن خوشه‌بندی ممکن است به‌وجود آید؟

سؤال ۶

به صورت مفهومی تفاوت‌های اصلی بین سه روشK-Means ، DBSCANو Hierarchical Clustering را بیان کنید. در پاسخ خود حتماً به موارد زیر اشاره کنید:

- فرض هر الگوریتم درباره‌ی شکل خوشه‌ها (کروی، چگالتی، سلسله‌مراتبی و ...)
- نیاز یا عدم نیاز به دانستن تعداد خوشه‌ها قبل از اجرا
- حساسیت نسبت به مقیاس داده و پارامترها مثلاً epsilon و min_samples در DBSCAN یا نوع linkage در Hierarchical
- رفتار هر الگوریتم در مواجهه با نویز و نقاط دورافتاده(outlier)

سؤال ۷

تابع هدف K-Means یک تابع غیرمحدب (non-convex) نسبت به تمام مراکز خوشه به صورت هم‌زمان است. توضیح دهید:

- این ویژگی چه تأثیری روی فرایند بهینه‌سازی و جواب نهایی الگوریتم دارد؟
- چرا اجرا کردن الگوریتم با نقاط اولیه‌ی مختلف initialization های متفاوت می‌تواند به جواب‌های متفاوت منجر شود؟
- در عمل چه راهکارهایی برای کاهش اثر مینیمم‌های محلی به کار می‌رود ؟

بخش دوم – تمرین‌های برنامه‌نویسی(Python / sklearn)

نکته‌ی مهم: در هر دو تمرین این بخش، لازم نیست الگوریتم‌ها را از صفر پیاده‌سازی کنید؛ می‌توانید از توابع موجود در کتابخانه‌ی sklearnو سایر کتابخانه‌های مرسوم استفاده کنید. تمرکز باید بر طراحی آزمایش، تحلیل نتایج، انتخاب پارامترها و استفاده از معیارهای ارزیابی باشد، نه صرفاً نوشتن چند خط کد.

تمرین ۱ – خوشه‌بندی داده‌ی تصاویر ارقام و ارزیابی با معیارهای پیشرفته

در این تمرین از دیتاست ارقام دست‌نویس (digits) موجود در کتابخانه‌ی sklearn استفاده کنید. این دیتاست شامل تصاویر ۸×۸ از ارقام ۰ تا ۹ است و به‌طور طبیعی ۱۰ کلاس دارد.

الف) داده را بارگذاری کنید، شکل کلی آن (تعداد نمونه‌ها، تعداد ویژگی‌ها، تعداد کلاس‌ها) را گزارش دهید و در صورت نیاز مقادیر ویژگی‌ها را نرمال‌سازی یا استاندارد کنید. چند نمونه‌ی تصویری را به صورت شبکه‌ای نمایش دهید تا شهودی از داده به‌دست آورید.

ب) الگوریتم K-Means را برای چند مقدار متفاوت از K (برای مثال در بازه‌ای حول ۱۰، مثل ۶ تا ۱۴) روی این داده اجرا کنید. برای هر مقدارK ، مقادیر زیر را محاسبه و ذخیره کنید:

- مقدار WCSS یا همان inertia گزارش‌شده توسطK-Means
- یک معیار داخلی مانند silhouette score برای خوشه‌بندی بدون استفاده از برچسب‌های واقعی
- نمودار تغییرات این مقادیر را بر حسب K رسم کنید و از دید خودتان تحلیل کنید که آیا می‌توان "نقطه‌ی مناسب" برای Kرا از روی این نمودارها حدس زد یا خیر.

ج) یکی از مقدارهای K را که منطقی‌تر به نظر می‌رسد (مثلاً نزدیک به ۱۰ و با رفتار مناسب در نمودارهای بخش قبل) انتخاب کنید و K-Means را با این K چند بار و باinitialization های متفاوت اجرا کنید. نشان دهید آیا مقدار تابع هدف و ساختار خوشه‌ها در اجراهای مختلف کاملاً یکسان است یا خیر و این رفتار را با بحث مینیمم‌های محلی مرتبط کنید.

د) با استفاده از برچسب‌های واقعی ارقام، برای خروجی خوشه‌بندیK-Means برای همان K انتخابی معیارهای ارزیابی زیر را محاسبه کنید:

- Adjusted Rand Index (ARI)
- Normalized Mutual Information (NMI)

برای این کار ابتدا باید بین خوشه‌ها و برچسب‌های واقعی یک نگاشت مناسب برقرار کنید می‌توانید از الگوریتم Hungarian یا پیاده‌سازی آماده‌ی آن استفاده کنید .گزارش کنید پس از اعمال این نگاشت، دقت ظاهری (Accuracy) چقدر می‌شود و آن را در کنار ARI و NMI تفسیر کنید؛ توضیح دهید هر کدام از این اعداد چه زاویه‌ای از کیفیت خوشه‌بندی را نشان می‌دهند و آیا بالابودن یکی از آن‌ها لزوماً به معنای بالابودن دیگری است یا خیر.

هـ) در یک جمع‌بندی کوتاه، تحلیل کنید که K-Means روی این دیتاست چه نقاط قوت و چه محدودیت‌هایی دارد؛ به‌خصوص به این نکته اشاره کنید که بعضی ارقام از نظر شکل و شدت پیکسل‌ها شبیه هم هستند و این موضوع چگونه می‌تواند باعث تداخل خوشه‌ها شود، حتی اگر مدل به‌صورت عددی عملکرد نسبتاً خوبی داشته باشد.

تمرین ۲ – مقایسه‌ی DBSCAN و Hierarchical Clustering روی داده‌ی غیرکروی با نویز

در این تمرین می‌خواهید تفاوت رفتار دو الگوریتم DBSCAN و Hierarchical Clustering را روی داده‌هایی با شکل‌های غیرکروی و وجود نویز بررسی کنید. برای این کار می‌توانید از یکی از توابع تولید داده در sklearn مانند داده‌های دو هلالی معروف که توسط تابعی مثل make_moons تولید می‌شوند استفاده کنید و در صورت تمایل مقداری نویز تصادفی به داده اضافه نمایید.

الف) یک دیتاست دوبعدی با حداقل دو خوشه‌ی غیرکروی (مثلاً دو هلال در کنار هم) و تعداد مناسبی نقاط نویزی تولید کنید. داده را رسم کنید و توضیح دهید ساختار شهودی خوشه‌ها و نویزها چگونه است. در صورت نیاز داده را مقیاس‌بندی (scaling) کنید و دلیل این انتخاب را بیان کنید.

ب) الگوریتم DBSCAN را با چند مجموعه پارامتر متفاوت مقادیر مختلف epsilon و min_samples روی این داده اجرا کنید. برای هر تنظیم، تعداد خوشه‌های شناسایی‌شده، تعداد نقاط نویزی و شکل ظاهری خوشه‌ها را گزارش و روی نمودار نمایش دهید. نتایج را مقایسه کنید و توضیح دهید چگونه تغییر epsilon و min_samples می‌تواند به "بیش‌خوشه‌بندی"، "کم‌خوشه‌بندی" یا علامت‌گذاری بیش‌ازحد نقاط به‌عنوان نویز منجر شود.

ج) روی همان داده، الگوریتم Hierarchical Clustering از نوع Agglomerative را با حداقل دو نوع linkage متفاوت مثلاً single و complete یا average و ward اجرا کنید. برای هر حالت، نمودار سلسله‌مراتبی (dendrogram) را رسم کنید و نشان دهید اگر در ارتفاع‌های مختلف آن را "برش" دهید، چه تعداد خوشه و با چه ساختاری به دست می‌آید. توضیح دهید آیا این روش بدون دانستن تعداد خوشه‌ها هم می‌تواند به شما در انتخاب تعداد خوشه‌ی مناسب کمک کند یا خیر.

د) نتایج DBSCAN و Hierarchical Clustering را از دو جنبه‌ی زیر با هم مقایسه کنید:

- توانایی در شناسایی شکل‌های غیرکروی و جدا کردن خوشه‌های هلالی از یکدیگر
- توانایی در شناسایی و برچسب‌زدن نقاط نویزی
- در تحلیل خود توضیح دهید هر الگوریتم بر چه فرض‌هایی تکیه می‌کند و این فرض‌ها چرا در این نوع داده می‌توانند به تفاوت محسوس در خروجی‌ها منجر شوند.

هـ) در پایان، یک سناریوی واقعی (مثلاً در حوزه‌ی تشخیص الگو، سامانه‌های فضایی، داده‌های سنسوری یا شبکه‌های اجتماعی) پیشنهاد کنید که در آن ترجیح دهید از DBSCAN استفاده کنید و یک سناریوی دیگر که در آن Hierarchical Clustering مناسب‌تر است. برای هر سناریو کوتاه توضیح دهید ساختار داده و نیاز مسئله چگونه با نقاط قوت آن الگوریتم هم‌خوانی دارد.