سوال ۳: یکی از ایده های اصلی در کلاسترینگ استفاده از فاصله بین نقاط است. آیا این روش همیشه جواب می‌دهد؟ در چه شرایطی این روش می‌تواند نتیجه منفی بدهد.

روش‌های مبتنی بر فاصله در کلاسترینگ (مانند K-Means یا Hierarchical Clustering) از فاصله بین نقاط برای تعیین شباهت یا نزدیکی استفاده می‌کنند. اگرچه این روش‌ها در بسیاری از سناریوها مفید هستند، اما در شرایط خاصی ممکن است به نتایج نامطلوب منجر شوند. برخی از این شرایط:

۱. داده‌های غیرمحدب یا پیچیده: روش‌های مبتنی بر فاصله برای کلاسترهایی که به صورت محدب هستند (مانند دایره یا بیضی) عملکرد خوبی دارند. اما اگر داده‌ها شکل‌های پیچیده‌ای داشته باشند (مثل کلاسترهای دایره‌ای داخل یکدیگر یا داده‌های خطی پیچیده)، این روش‌ها ممکن است نتوانند کلاسترها را به درستی تشخیص دهند.

۲. مقیاس‌گذاری نادرست ویژگی‌ها: اگر ویژگی‌های داده‌ها در مقیاس‌های مختلف باشند و بدون نرمال‌سازی یا استانداردسازی استفاده شوند، فاصله‌ها به‌طور نادرست محاسبه می‌شوند. این می‌تواند باعث شود که ویژگی‌هایی با مقیاس بزرگ‌تر، اهمیت بیشتری پیدا کنند.

۳. داده‌های نویزی یا با نویز بالا: داده‌هایی که دارای نویز یا نقاط پرت (outliers) زیادی هستند می‌توانند محاسبات فاصله را مختل کنند، زیرا این نقاط فاصله‌های غیرمعمولی ایجاد می‌کنند که تأثیر منفی روی الگوریتم دارند.

۴. ابعاد بالا (مشکل Curse of Dimensionality): در داده‌های با ابعاد بالا، فاصله بین نقاط تمایل دارد یکنواخت شود (یعنی تفاوت فاصله بین نزدیک‌ترین و دورترین نقاط کاهش می‌یابد). این باعث می‌شود که الگوریتم‌های مبتنی بر فاصله قدرت خود را از دست بدهند و نتوانند کلاسترها را به خوبی تشخیص دهند.

۵. فرضیات نادرست درباره تعداد کلاسترها: بسیاری از الگوریتم‌های مبتنی بر فاصله (مانند K-Means) نیاز به تعداد کلاسترها به‌عنوان ورودی دارند. اگر تعداد کلاسترها به‌درستی مشخص نشود، نتایج کلاسترینگ ممکن است بی‌معنی باشد.

۶. همگن نبودن چگالی کلاسترها: اگر کلاسترها چگالی‌های متفاوتی داشته باشند (مثلاً برخی کلاسترها دارای نقاط متراکم و برخی دارای نقاط پراکنده باشند)، روش‌های مبتنی بر فاصله ممکن است نتوانند این کلاسترها را به‌درستی شناسایی کنند.

الگوریتم DBSCAN را توضیح دهید و همچنین توضیح دهید در کدام دسته از الگوریتم‌های کلاسترینگ قرار می‌گیرد. در ادامه تفاوت آن را با الگوریتم OPTICS توضیح دهید.

DBSCAN یکی از الگوریتم‌های کلاسترینگ مبتنی بر چگالی است که برای پیدا کردن کلاسترها در داده‌هایی که ساختار نامنظم دارند یا شامل نویز هستند، مناسب است.

پارامترها:

- ε: شعاعی که همسایگی یک نقطه را مشخص می‌کند.
- $minPts$: حداقل تعداد نقاط مورد نیاز برای تشکیل یک کلاستر در یک منطقه ε-همسایگی.

تعریف نقاط:
- نقطه مرکزی (Core Point): نقطه‌ای که حداقل $minPts$ نقطه (شامل خودش) در شعاع ε آن وجود داشته باشد.
- نقطه مرزی (Border Point): نقطه‌ای که در شعاع ε یک نقطه مرکزی قرار دارد اما خودش نقطه مرکزی نیست.
- نقطه نویز (Noise Point): نقطه‌ای که نه مرکزی است و نه مرزی.

الگوریتم از یک نقطه مرکزی شروع می‌کند و با گسترش منطقه ε، تمام نقاط مرتبط را به کلاستر اضافه می‌کند. این فرآیند ادامه می‌یابد تا تمام نقاط مرکزی و نقاط مرتبط پوشش داده شوند.

نیازی به تعیین تعداد کلاسترها از قبل ندارد. می‌تواند کلاسترهایی با شکل‌های نامنظم را شناسایی کند. در برابر نویز مقاوم است.

انتخاب مناسب پارامترهای ε و $minPts$ دشوار است. در داده‌های با ابعاد بالا، عملکرد آن ممکن است کاهش یابد.

DBSCAN در دسته الگوریتم‌های کلاسترینگ مبتنی بر چگالی قرار می‌گیرد. این دسته از الگوریتم‌ها کلاسترها را به‌عنوان مناطقی با چگالی بالا (نسبت به نقاط مجاور) تعریف می‌کنند و برای داده‌هایی که کلاسترهایی با شکل‌های پیچیده دارند، مناسب هستند.

تفاوت DBSCAN و OPTICS

الف) DBSCAN بر اساس پارامترهای ثابت ε و $minPts$ عمل می‌کند. این محدودیت می‌تواند باعث شود که در داده‌هایی با چگالی‌های متغیر، برخی کلاسترها شناسایی نشوند. OPTICS نسخه پیشرفته‌تر DBSCAN است که به جای استفاده از یک مقدار ثابت ε، از یک رویکرد تطبیقی برای شناسایی کلاسترها استفاده می‌کند. این الگوریتم می‌تواند چگالی‌های متغیر را مدیریت کند.

ب) DBSCAN تنها یک مجموعه از کلاسترها را ایجاد می‌کند و نقاط نویز را مشخص می‌سازد. OPTICS به جای تولید مستقیم کلاسترها، یک ساختار مرتب‌شده از داده‌ها بر اساس چگالی تولید می‌کند. این ساختار می‌تواند برای استخراج کلاسترهای مختلف با تنظیمات متفاوت استفاده شود.

ج) DBSCAN اگر داده‌ها دارای کلاسترهایی با چگالی‌های متفاوت باشند، ممکن است برخی کلاسترها را از دست بدهد. OPTICS به دلیل استفاده از یک رویکرد انعطاف‌پذیر برای تعیین شعاع ε، می‌تواند کلاسترهایی با چگالی‌های متغیر را به خوبی شناسایی کند.

د) DBSCAN کلاسترهای مشخصی تولید می‌کند. OPTICS یک ترتیب (ordering) از نقاط تولید می‌کند که برای استخراج کلاسترهای مختلف استفاده می‌شود.