گزارش تکلیف سوم درس الگوریتم های علوم داده

نام و نام خانوادگی: فاطمه ترودی

شماره دانشجویی: ۴۰۳۴۲۲۰۴۸

نام استاد: دکتر سعیدرضا خردپیشه

نیمسال دوم ۰۴-۱۴۰۳

فهرست مطالب

۱. سوالات تئوری

۱/۱. سوال ۱

الف) مدل ترکیبی گاوسی (Gaussian Mixture Model – GMM)

- GMM یک روش خوشه‌بندی نرم است که فرض می‌کند داده‌ها از ترکیبی از توزیع‌های گاوسی تولید شده‌اند. هر خوشه با یک توزیع گاوسی (میانگین و کوواریانس) تعریف می‌شود. GMM از الگوریتم انتظار-بیشینه‌سازی (EM) برای یادگیری پارامترهای این توزیع‌ها استفاده می‌کند. در مرحله انتظار، احتمال تعلق هر نقطه داده به هر خوشه محاسبه می‌شود، و در مرحله بیشینه‌سازی، پارامترهای توزیع‌ها به‌روزرسانی می‌شوند.
- سناریوهای مناسب: GMM برای داده‌هایی که خوشه‌ها شکل‌های غیرکروی (مانند بیضی) دارند یا همپوشانی بین خوشه‌ها وجود دارد، مناسب است. به عنوان مثال، در تحلیل داده‌های زیستی یا تصاویر که خوشه‌ها ممکن است توزیع‌های پیچیده داشته باشند.
- مزایا: انعطاف‌پذیری در مدل‌سازی اشکال مختلف خوشه‌ها و ارائه احتمالات تعلق.
- معایب: حساس به مقداردهی اولیه و محاسبات سنگین برای داده‌های بزرگ.

ب) K-Means++

- K-means++ نسخه بهبودیافته الگوریتم K-means است که مراکز اولیه خوشه‌ها را به صورت هوشمند انتخاب می‌کند. ابتدا یک مرکز به صورت تصادفی انتخاب می‌شود، سپس مراکز بعدی با احتمالی متناسب با فاصله از مراکز قبلی انتخاب می‌شوند. سپس، مانند K-means، هر نقطه به نزدیک‌ترین مرکز تخصیص می‌یابد و مراکز خوشه‌ها به‌روزرسانی می‌شوند تا زمانی که همگرایی رخ دهد.
- سناریوهای مناسب: K-means++ برای داده‌هایی با خوشه‌های کروی و تفکیک‌پذیر مناسب است، مانند داده‌های مشتری بر اساس ویژگی‌های عددی (مثلاً درآمد و سن). همچنین برای داده‌های بزرگ که نیاز به محاسبات سریع دارند، مناسب است.
- مزایا: سرعت بالا و انتخاب بهتر مراکز اولیه نسبت به K-means معمولی.
- معایب: فرض خوشه‌های کروی و حساسیت به نویز و نقاط پرت.

ج) خوشه‌بندی طیفی (Spectral Clustering)

- خوشه‌بندی طیفی داده‌ها را به صورت یک گراف نشان می‌دهد، جایی که نقاط داده گره‌ها و شباهت بین آن‌ها یال‌ها هستند. ماتریس مجاورت بر اساس شباهت (مانند فاصله گاوسی) ساخته می‌شود. سپس، مقادیر ویژه و بردارهای ویژه ماتریس لاپلاسین گراف محاسبه می‌شوند. داده‌ها به فضای کاهش‌یافته (با استفاده از بردارهای ویژه) نگاشت می‌شوند و در این فضا از الگوریتمی مانندK-means برای خوشه‌بندی استفاده می‌شود.
- سناریوهای مناسب :این روش برای داده‌هایی با خوشه‌های غیرخطی یا پیچیده (مانند حلقه‌ها یا اشکال مارپیچی) مناسب است. به عنوان مثال، در تحلیل شبکه‌های اجتماعی یا خوشه‌بندی تصاویر.
- مزایا: توانایی مدل‌سازی خوشه‌های غیرکروی و پیچیده.
- معایب: نیاز به محاسبات سنگین برای داده‌های بزرگ و حساسیت به انتخاب پارامتر شباهت.

۱/۲. سوال ۲

برای خوشه‌بندی داده‌هایی که شامل متغیرهای عددی (مانند درآمد) و دسته‌ای (مانند جنسیت) هستند، استراتژی‌های زیر استفاده می‌شوند:

- **رمزگذاری متغیرهای رسته‌ای (Categorical):**
  - **One-Hot Encoding**: هر دسته به یک بردار باینری تبدیل می‌شود. این روش برای متغیرهای با تعداد دسته‌های کم مناسب است.
  - **رمزگذاری مبتنی بر شباهت:** از معیارهایی مانند فاصله گاور (Gower Distance) استفاده می‌شود که می‌تواند متغیرهای عددی و دسته‌ای را به طور همزمان مدیریت کند.
- **معیارهای فاصله ترکیبی:** فاصله گاور یا معیارهای مشابه برای محاسبه شباهت بین نقاط داده با متغیرهای مختلط استفاده می‌شود. این معیار فاصله‌های عددی را نرمال‌سازی می‌کند و برای متغیرهای دسته‌ای از تطابق یا عدم تطابق استفاده می‌کند.
- **الگوریتم‌های خاص:** الگوریتم‌هایی مانند K-Prototypes (ترکیبی از K-means برای متغیرهای عددی و K-modes برای متغیرهای رسته‌ای) برای این نوع داده‌ها طراحی شده‌اند.
- **تبدیل متغیرها:** متغیرهای عددی را می‌توان به دسته‌های گسسته تبدیل کرد (مثلاً تبدیل درآمد به بازه‌های کم، متوسط، زیاد) و سپس از روش‌های خوشه‌بندی دسته‌ای استفاده کرد.
- **وزن‌دهی به متغیرها:** برای متعادل کردن تأثیر متغیرهای عددی و دسته‌ای، می‌توان وزن‌های متفاوتی به هر نوع متغیر اختصاص داد.

۱/۳. سوال ۳

مقایسه خوشه‌بندی نرم و سخت:

- **خوشه‌بندی سخت (Hard Clustering)**:
  - هر نقطه داده دقیقاً به یک خوشه تعلق دارد (مانند K-means).
  - مزایا: محاسبات ساده‌تر و سریع‌تر، نتایج قابل تفسیر
  - معایب: عدم انعطاف‌پذیری در داده‌هایی با همپوشانی خوشه‌ها یا نقاط مرزی.
- **خوشه‌بندی نرم (Soft Clustering)**:
  - هر نقطه داده می‌تواند به چندین خوشه با احتمالات مختلف تعلق داشته باشد (مانند GMM).
  - مزایا: مدل‌سازی بهتر داده‌های پیچیده با همپوشانی یا عدم قطعیت.
  - معایب: پیچیدگی محاسباتی بالاتر و نیاز به تفسیر احتمالات.

سناریوهای مناسب برای خوشه‌بندی نرم:

- **داده‌های با همپوشانی:** مانند تحلیل داده‌های زیستی که خوشه‌ها ممکن است مرزهای مشخصی نداشته باشند.
- **داده‌های مبهم:** در مواردی که نقاط داده ممکن است به چندین گروه تعلق داشته باشند (مثلاً مشتریان با رفتارهای خرید متنوع).
- **نیاز به احتمالات تعلق:** در کاربردهایی مانند بازاریابی که دانستن احتمال تعلق مشتری به یک بخش خاص مفید است.
- **داده‌های غیرکروی:** GMM می‌تواند خوشه‌هایی با اشکال مختلف را مدل کند، برخلاف K-means که فرض خوشه‌های کروی دارد.

۱/۴. سوال ۴

بله، خوشه‌بندی می‌تواند برای تشخیص ناهنجاری (Anomaly Detection) استفاده شود، زیرا نقاطی که به هیچ خوشه‌ای تعلق ندارند یا از خوشه‌ها فاصله زیادی دارند، می‌توانند به عنوان ناهنجاری شناسایی شوند.

الف) DBSCAN برای تشخیص ناهنجاری

- DBSCAN خوشه‌ها را بر اساس چگالی نقاط داده تشکیل می‌دهد. نقاطی که در مناطق کم‌چگالی قرار دارند و به هیچ خوشه‌ای تعلق نمی‌گیرند، به عنوان ناهنجاری برچسب‌گذاری می‌شوند.
- نحوه تشخیص ناهنجاری: نقاطی که به عنوان نویز (Noise) شناسایی می‌شوند، ناهنجاری‌ها هستند. این روش برای داده‌هایی با خوشه‌های غیرکروی و چگالی متفاوت مناسب است.
- مزایا: نیازی به مشخص کردن تعداد خوشه‌ها ندارد و می‌تواند ناهنجاری‌ها را به طور خودکار شناسایی کند.
- معایب: حساس به انتخاب پارامترهای چگالی (مانند شعاع و حداقل تعداد نقاط).

ب) GMM برای تشخیص ناهنجاری

- GMM احتمال تعلق هر نقطه به خوشه‌ها را محاسبه می‌کند. نقاطی که احتمال تعلق آن‌ها به همه خوشه‌ها بسیار کم است (یعنی در مناطق کم‌احتمال توزیع‌های گاوسی قرار دارند) به عنوان ناهنجاری شناسایی می‌شوند.
- نحوه تشخیص ناهنجاری: یک آستانه برای احتمال تعلق تعریف می‌شود (مثلاً نقاط با احتمال کمتر از ۰.۰۱). این نقاط به عنوان ناهنجاری برچسب‌گذاری می‌شوند.
- مزایا: توانایی مدل‌سازی خوشه‌های پیچیده و ارائه احتمالات برای تصمیم‌گیری.
- معایب: نیاز به تنظیم تعداد خوشه‌ها و حساسیت به مقداردهی اولیه.

۱/۵. سوال ۵

چالش‌ها

- **تسلط خوشه‌های بزرگ:** در داده‌های نامتعادل، خوشه‌های بزرگ (با تعداد نقاط زیاد) می‌توانند الگوریتم را تحت تأثیر قرار دهند و خوشه‌های کوچک نادیده گرفته شوند.
- **تشخیص نادرست خوشه‌ها:** الگوریتم‌هایی مانند K-means ممکن است نقاط خوشه‌های کوچک را به خوشه‌های بزرگ‌تر تخصیص دهند.
- **حساسیت به نویز:** نقاط پرت در خوشه‌های کوچک ممکن است به اشتباه به عنوان خوشه‌های مجزا شناسایی شوند.
- **انتخاب تعداد خوشه‌ها:** در داده‌های نامتعادل، روش‌هایی مانند نمودار Elbow Plot ممکن است خوشه‌های کوچک را نادیده بگیرند.

راهکارها

- **نمونه‌برداری متعادل:** از روش‌هایی مانند Oversampling (برای خوشه‌های کوچک) یا Undersampling (برای خوشه‌های بزرگ) برای متعادل کردن داده‌ها استفاده کنید.
- **وزن‌دهی به نقاط:** در الگوریتم‌هایی مانند GMM، می‌توان به نقاط خوشه‌های کوچک وزن بیشتری داد تا تأثیر آن‌ها افزایش یابد.
- **استفاده از الگوریتم‌های مبتنی بر چگالی:** الگوریتم‌هایی مانند DBSCAN یا HDBSCAN می‌توانند خوشه‌های کوچک را در داده‌های نامتعادل بهتر شناسایی کنند.
- **پیش‌پردازش داده‌ها:** نرمال‌سازی یا استانداردسازی داده‌ها برای کاهش تأثیر مقیاس‌های مختلف و استفاده از معیارهای فاصله مناسب (مانند فاصله گاور).
- **روش‌های ترکیبی:** از خوشه‌بندی سلسله‌مراتبی برای شناسایی خوشه‌های بزرگ و سپس خوشه‌بندی مبتنی بر چگالی برای خوشه‌های کوچک استفاده کنید.
- **اعتبارسنجی خوشه‌ها:** از معیارهای ارزیابی مانند شاخص Silhouette Score یا شاخص Davies-Bouldin برای اطمینان از کیفیت خوشه‌بندی استفاده کنید.

۱/۶. سوال ۶

الف) نحوه ساخت سلسله مراتب خوشه‌ها در خوشه‌بندی سلسله مراتبی:

- **مراحل الگوریتم:**
  ۱. هر نقطه داده به عنوان یک خوشه مستقل در نظر گرفته می‌شود.
  ۲. در هر مرحله، دو خوشه‌ای که کمترین افزایش را در مجموع واریانس درون خوشه‌ای (با استفاده از روش Ward) ایجاد می‌کنند، ادغام می‌شوند. روش Ward به دنبال حداقل کردن واریانس درون خوشه‌ها پس از ادغام است.
  ۳. این فرآیند تا زمانی ادامه می‌یابد که تمام نقاط در یک خوشه واحد ادغام شوند یا تعداد خوشه‌های مورد نظر به دست آید.
  ٤. نتیجه یک سلسله‌مراتب (دندروگرام) است که نشان می‌دهد خوشه‌ها چگونه ادغام شده‌اند.
- **روش Ward**: این روش از معیار حداقل واریانس استفاده می‌کند و خوشه‌هایی را ادغام می‌کند که کمترین افزایش را در مجموع مربعات فاصله‌های درون‌گروهی ایجاد کنند. این باعث می‌شود خوشه‌ها فشرده و کروی باشند.

ب) تعیین تعداد بهینه خوشه‌ها از Dendrogram:

- **Dendrogram:** دندروگرام یک نمودار درختی است که محور عمودی آن فاصله ادغام (یا افزایش واریانس) و محور افقی آن خوشه‌ها را نشان می‌دهد.
- **روش تعیین تعداد خوشه‌ها:**

۱. **برش در ارتفاع زیاد:** در دندروگرام، به دنبال نقاطی بگردید که فاصله عمودی بین ادغام‌ها (ارتفاع) به طور قابل توجهی افزایش می‌یابد. این نشان‌دهنده ادغام خوشه‌های غیرمشابه است. تعداد خوشه‌ها در این نقطه می‌تواند بهینه باشد.

۲. **قاعده Elbow:** نمودار فاصله ادغام در مقابل تعداد خوشه‌ها را رسم کنید و نقطه‌ای که کاهش فاصله کند می‌شود (مشابه روش Elbow در K-Means) را انتخاب کنید.

۳. **معیارهای ارزیابی:** از معیارهایی مانند شاخص Silhouette Score یا شاخص Davies-Bouldin برای اعتبارسنجی تعداد خوشه‌ها استفاده کنید.

٤. **دانش حوزه:** تعداد خوشه‌ها را بر اساس نیازهای کسب‌وکار (مثلاً تقسیم‌بندی مشتریان به گروه‌های معنی‌دار) انتخاب کنید.

ج) تفسیر خوشه‌ها با برش Dendrogram در ۳ خوشه:

با فرض برش دندروگرام در ۳ خوشه و ویژگی‌های داده (درآمد سالانه، سن، مجموع هزینه در سال گذشته)، می‌توان خوشه‌ها را به صورت زیر تفسیر کرد:

- **خوشه ۱: مشتریان جوان با درآمد و هزینه کم**
  - **ویژگی‌ها:** سن پایین، درآمد سالانه کم، هزینه سالانه کم.
  - **تفسیر رفتاری:** این مشتریان احتمالاً دانشجویان یا افراد تازه‌کار هستند که بودجه محدودی دارند و خریدهای ضروری انجام می‌دهند. استراتژی بازاریابی: ارائه تخفیفات یا محصولات اقتصادی.
- **خوشه ۲: مشتریان میانسال با درآمد و هزینه متوسط**
  - **ویژگی‌ها:** سن متوسط، درآمد متوسط، هزینه متوسط.
  - **تفسیر رفتاری:** این گروه ممکن است شامل خانواده‌ها یا افراد شاغل با درآمد ثابت باشد که خریدهای منظم و متعادل انجام می‌دهند. استراتژی بازاریابی: برنامه‌های وفاداری یا پیشنهادات خانوادگی.
- **خوشه ۳: مشتریان مسن‌تر با درآمد و هزینه بالا**

- **ویژگی‌ها:** سن بالاتر، درآمد بالا، هزینه بالا.
- **تفسیر رفتاری:** این مشتریان احتمالاً افراد حرفه‌ای یا بازنشستگان با توان مالی بالا هستند که خریدهای لوکس یا پرهزینه انجام می‌دهند. استراتژی بازاریابی: محصولات premium یا خدمات اختصاصی.