داده کاوی پیشرفته

تمرین اول

دکتر محبوبه ریاحی

نیم سال اول ۱۴۰۴-۱۴۰۵

سوال اول)

فرض کنید دو پایگاه‌داده با ویژگی‌های زیر در اختیار دارید:

- پایگاه‌داده A شامل تراکنش‌هایی با اقلام بسیار متنوع ولی توزیع ناهمگن فراوانی‌ها است (اقلام پرتکرار اندک‌اند ولی بسیاری از اقلام فقط یک یا دو بار ظاهر می‌شوند).
- پایگاه‌داده B شامل اقلام اندک ولی با فراوانی بالا و نسبتاً یکنواخت است.

اکنون قصد دارید با هر دو پایگاه‌داده، الگوهای پرتکرار (Frequent Itemsets) را استخراج کنید.

با توجه به ویژگی‌های ساختاری این دو پایگاه‌داده:

الف) تحلیل کنید که در هر مورد، استفاده از کدام الگوریتم Apriori ، FP-Growth ، یا ECLAT از نظر کارایی محاسباتی و حافظه‌ای مناسب‌تر است و چرا؟

ب) نشان دهید که چگونه ویژگی‌هایی نظیر Downward Closure Property و ساختار داده‌ای مورد استفاده (Hash-tree یا FP-tree یا Vertical Format) باعث می‌شود یک روش در یک نوع داده بهتر و در نوع دیگر ضعیف‌تر عمل کند.

ج) اگر بخواهید الگوریتم منتخب خود را برای داده A یا B بهبود دهید، چه تغییر یا ترکیب روشی (Hybridization) را پیشنهاد می‌کنید و منطق تحلیلی آن چیست؟

سوال دوم)

فرض کنید پایگاه‌داده تراکنشی زیر را در اختیار دارید و حداقل میزان حمایت (minsup) برابر ۵۰٪ است:

TID	اقلام خریداری‌شده
1	A, B, C, D
2	A, C, D
3	A, B, C

A, B, D	4
B, C, D	5

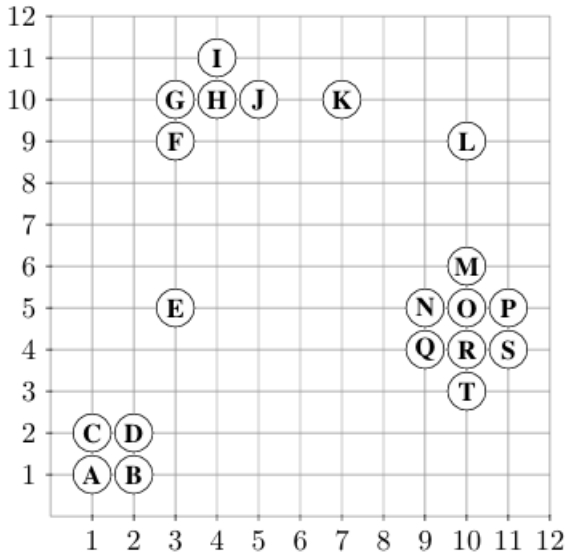
الف) با استناد به خاصیت بسته‌شدگی نزولی (Downward Closure)، کلیه الگوهای پرتکرار بسته (Closed Itemsets) را مشخص کنید.

ب) اکنون برای همین داده، پایه الگویی شرطی (Conditional Pattern Base) برای قلم «C» را بسازید و درخت شرطی آن را تا سطح دوم رسم کنید. سپس همه الگوهای پرتکرار مرتبط با «C» را از روی آن استنتاج کنید.

ج) اگر یکی از تراکنش‌ها (مثلاً T5) حذف شود، تحلیل کنید که این حذف چگونه ممکن است بر ساختار شاخه‌های FP-tree و مجموعه الگوهای بسته تأثیر بگذارد، بدون اینکه کل درخت از ابتدا ساخته شود.

سوال سوم)

دادگان زیر را در نظر بگیرید.



الف) با بکارگیری الگوریتم خوشه بندی DBSCAN با پارامترهای eps=2.1 ، minpts=4 و معیار فاصله منهتن روی دادگان بالا، برچسب نقاط داده(مرکزی، حاشیه و نویزی) و خوشه ها را تعیین کنید.

ب) با در نظر گرفتن معیار فاصله منهتن، دو دندروگرام با بکارگیری single-link و complete-link روی دادگان بالا ترسیم کنید.

سوالات پیاده سازی:

سوال چهارم)

هدف این سوال، درک عمیق چگونگی کاهش هزینه‌ی محاسباتی Apriori با استفاده از دو تکنیک بهینه‌سازی کلاسیک در یک محیط تک‌سیستمی (In-Memory) است:

- DHP(Dynamic Hashing and Pruning) : هرس (Pruning) کاندیداهای C_2 با استفاده از تابع درهم سازی(Hash function).
- Sampling (نمونه‌برداری): محدود کردن دامنه‌ی جستجو و کاهش اسکن‌های پایگاه داده.

الف) الگوریتم Apriori را پیاده سازی نمایید بطوریکه ورودی های این الگوریتم شامل دیتاست تراکنش ها ، minsup_fraction (مثلاً ۰.۰۲) و خروجی ها شامل یک دیکشنری به صورت {itemset: support_count}، زمان اجرا و پیک حافظه باشد.

ب) الگوریتم Apriori را با بهینه‌سازی Direct Hashing and Pruning(DHP) پیاده‌سازی کنید بطوریکه خروجی یک دیکشنری به صورت {itemset: support_count}،زمان اجرا و پیک حافظه باشد.

در همان پاس اول که L_1 پیدا می‌شود، جدول هش دوآیتمی (Pair Hash Table) را برای تمامی زوج‌های موجود در تراکنش‌ها ایجاد کنید. تنها زوج‌هایی را به عنوان کاندیداهای C_2 در نظر بگیرید که سایز سطل هش مربوط به آن‌ها (Hashing Bucket Count) از آستانه‌ی پشتیبانی (minsup) بالاتر باشد. سپس الگوریتم Apriori را از L_2 به بعد طبق روال معمول در الگوریتم Apriori ادامه دهید(توجه: این روش تنها C_2 را بهینه می‌کند و برای $C_3, C_4, \ldots$ الگوریتم به روش استاندارد باز می‌گردد).

ج) الگوریتم Apriori را بهمراه تکنیک نمونه‌برداری پیاده‌سازی کنید(خروجی الگوریتم مشابه الگوریتم های قبلی).

د) آزمایش و مقایسه الگوریتم ها روی یک دیتاست واقعی: از فایل groceries.csv و retail_transactional.csv استفاده کنید. مقدار minsup را روی مقادیر [۰.۰۵, ۰.۰۳, ۰.۰۲, ۰.۰۱] تست کنید. برای هر مقدار minsup، سه الگوریتم standard_apriori ، dhp_apriori و sampling_apriori (با نرخ نمونه‌برداری ۱۰٪) را اجرا کنید. تعداد آیتم‌ست‌های مکرر، زمان اجرا و پیک حافظه را ثبت کنید. سپس، نمودارهای زمان اجرا و حافظه مصرفی بر حسب minsup را رسم کنید.

ه) همچنین در کد خود شمارنده‌هایی اضافه کنید تا در خروجی ، مقادیر زیر چاپ شوند:

- تعداد کاندیداهای C_2 در روش استاندارد ($|C_{2\_standard}|$).
- تعداد کاندیداهای C_2 پس از هرس ($|C_{2\_DHP}|$).
- تعداد آیتم‌ست‌های کاندیدای نهایی که در پاس دوم نمونه‌برداری نیاز به تأیید دارند ($|C_{sample\_confirm}|$).

نمودارها و جداول زیر را ترسیم کنید و تحلیل خود را ارایه دهید:

- چگونه کاهش minsup بر تعداد کاندیداها (در C_2) و زمان/حافظه در هر سه روش تأثیر می‌گذارد.

- در دیتاست groceries و retail ، عملکرد نسخه‌ی DHP و Sampling را نسبت به نسخه‌ی استاندارد مقایسه کنید.(آیا DHP می‌تواند زمان اجرا را به‌دلیل کاهش $|C_2|$ بهبود بخشد؟ آیا Sampling با صرفه‌جویی در اسکن‌ها از DHP بهتر عمل می‌کند؟)

و) (امتیازی) رفتار الگوریتم‌ها را در دو دیتاستی که در اختیار شما قرار گرفته‌اند(groceries وretail)تحلیل کنید. تحلیل خود را مستقیماً به مفاهیم زیر مرتبط کنید:

- انفجار کاندیداها(Candidate Explosion): توضیح دهید چگونه الگوریتم DHP با استفاده از جدول هش، باعث کاهش تعداد کاندیداهای C_2 می‌شود و چرا این کاهش در دیتاست‌های متراکم‌تر(مثل retail)بیشتر اثر دارد.
- هزینه‌ی I/O : توضیح دهید که تکنیک Sampling چگونه با کاهش تعداد اسکن‌های پایگاه داده، هزینه‌ی I/O را کاهش می‌دهد، و چرا در دیتاست‌های بزرگ‌تر، اثر آن بیشتر است.
- ریسک Sampling(Risk) : تحلیل کنید که در چه شرایطی احتمال از دست رفتن الگوهای واقعی وجود دارد، و این ریسک بیشتر مربوط به چه نوع آیتم‌ست‌هایی است.

سوال پنجم)

در این سوال هدف درک تفاوت میان دو الگوریتم خوشه‌بندی مبتنی بر مرکز (K-Means) و دیگری مبتنی بر چگالی (DBSCAN) است. از فایل داده‌ی two_moons_X_only.csv شامل ۲۰۰۰ نقطه دوبعدی در شکل دو هلال در هم تنیده به عنوان داده ورودی استفاده نمایید.

الف) الگوریتم K-Means را پیاده‌سازی نمایید. سپس الگوریتم را با K=2 روی دادگان ورودی اجرا کنید. خروجی شامل برچسب خوشه‌ها و مقدار مجموع خطاهای مربعات (Total SSE) باشد.

ب) الگوریتم DBSCAN را پیاده‌سازی نمایید. سپس، پارامترهای ε و MinPts را طوری انتخاب کنید که دقیقاً دو خوشه (دو هلال) تشکیل شوند و کمترین تعداد نقاط نویز داشته باشید. تعداد خوشه‌ها و تعداد نقاط نویزی را گزارش کنید.

ج) ضریب Silhouette را برای هر دو الگوریتم گزارش کنید(توجه کنید که در هنگام محاسبه این ضریب برای الگوریتم DBSCAN، نقاط نویز DBSCAN را نادیده بگیرید).

د) نتایج SSE و Silhouette را برای K-Means تحلیل کنید. توضیح دهید چرا ممکن است K-Means با وجود SSE پایین، خوشه‌بندی نادرستی ارائه دهد(پاسخ خود را با اشاره به مفهوم مرکز ثقل (centroid) و فرض کروی بودن خوشه‌ها در K-Means بنویسید).

ه) مقدار Purity را برای K-Means و DBSCAN محاسبه و مقایسه کنید.

و) نتایج سه معیار SSE ، Silhouette و Purity را مقایسه و تحلیل کنید.

ز) توضیح دهید چرا در داده‌های غیرکروی، Purity و Silhouette معیارهای قابل‌اعتمادتری از SSE هستند.

برای ارسال تمرین به نکات زیر توجه کنید.

۱- ملاک اصلی انجام تمرین گزارش آن است و ارسال کد بدون گزارش فاقد ارزش است. برای این تمرین یک فایل گزارش در قالب pdf تهیه کنید و در آن برای هر سوال،سعی کنید توضیحات کامل و جامعی تهیه کنید.

۲- زبان برنامه نویسی برای انجام تمرین ها، پایتون در نظر گرفته شده است.

۳- تحویل تمرین از طریق سامانه LMS می‌باشد.

۴- تمرین را در قالب یک فایل فشرده شامل پاسخ تشریحی، نوتبوک و گزارش بارگذاری کنید و نام فایل به صورت زیر باشد.

$$HW1 - First\_name - Last\_name - Student\_Number$$

۵- پاسخ به سوالات تشریحی و پیاده‌سازی به صورت انفرادی بوده و در صورت مشاهده هر گونه تقلب، نمره ای تعلق نخواهد گرفت.

۶- پیاده‌سازی باید در نوتبوک و نتایج آن قابل مشاهده باشد.

موفق باشید.