به نام خدا

دانشگاه تهران

دانشکده مهندسی برق و کامپیوتر

# درس داده کاوی

# تمرین سوم

| | |
|---|---|
| طراحان | علی ادیبی<br>adibialii@ut.ac.ir |
| تاریخ بارگذاری | ۱۴۰۳/۰۱/۲۳ |
| مهلت ارسال | ۱۴۰۳/۰۲/۰۸ |

# فهرست

# جدول‌ها

# بخش تشریحی

## سوال اول

پایگاه‌داده‌های زیر را در نظر بگیرید. با فرض آن که برای پایگاه‌داده اول مقادیر min_support=۰٫۳ و min_confidence=۰٫۶ و برای پایگاه‌داده دوم مقادیر min_support=۰٫۶ و min_confidence=۰٫۶ باشد، به سوالات زیر پاسخ دهید.

جدول 1. پایگاه‌داده تراکش‌های اول

| Transaction_id | Item_bought |
|---|---|
| 01 | {G, B, A, F} |
| 02 | {H, A, E} |
| 03 | {F, B, G} |
| 04 | {A, C, D, E} |
| 05 | {C, F, G, B} |
| 06 | {A, F, D} |
| 07 | {E, F, G, A} |
| 08 | {C, F, E} |
| 09 | {G, E} |
| 10 | {H, C, G} |

جدول 2. پایگاه‌داده تراکنش‌های دوم

| Transaction_id | Item_bought |
|---|---|
| 01 | {A, B, C, D} |
| 02 | {D, F, B, C, A, G} |
| 03 | {H, F, A, G} |
| 04 | {F, D, E, G} |
| 05 | {A, D, F, G} |

برای پایگاه‌داده اول، به سوالات زیر پاسخ دهید.

الف) درخت FP را رسم کنید. توجه کنید برای ساختن F-list نزولی در صورت برابر بودن تعداد تکرار، اولویت را به آیتمی بدهید که زودتر در الفبا آمده باشد.

ب) پایگاه‌داده B’s conditional را نشان دهید.

ج) تمام closed patternها و max-patternها را بیابید.

د) تمام Association rule های قوی را بیابید.

برای پایگاه داده دوم، به سوالات زیر پاسخ دهید.

ه) الگوریتم Apriori را روی پایگاه‌داده مدنظر پیاده‌سازی کنید.

و) بزرگ‌ترین itemset ممکن به نام S را بیابید به‌طوری‌که خود itemset مکرر نباشد، اما تمام زیرمجموعه‌های آن (به‌استثنای itemset تهی) مکرر باشد.

ز) تمام Association rule های قوی که با meta rule زیر مطابقت دارند را بیابید و مقادیر support و confidence را برای آن‌ها محاسبه کنید.

$$x \in transaction, buys(x, item1) \wedge buys(x, item2) \Rightarrow buys(x, item3)[s, c]$$

## سؤال دوم

الف) دو itemset مکرر s و l را در نظر بگیرید. فرض کنید که Association rule قوی $s \Rightarrow l$ برقرار است. با دلیل در مورد درستی یا نادرستی عبارت‌های زیر تصمیم‌گیری کنید.

- به‌ازای a itemset به‌طوری‌که $a \subseteq s$، Association rule قوی $a \Rightarrow l$ برقرار است.
- به‌ازای a itemset به‌طوری‌که $s \subseteq a$، Association rule قوی $a \Rightarrow l$ برقرار است.
- به‌ازای a itemset به‌طوری‌که $a \subseteq l$، Association rule قوی $s \Rightarrow a$ برقرار است.
- به‌ازای a itemset به‌طوری‌که $l \subseteq a$، Association rule قوی $s \Rightarrow a$ برقرار است.

ب) فرض کنید به‌ازای ۳ a itemset و b و c، رابطه $a \subseteq b \subseteq c$ برقرار است. ثابت کنید مقدار confidence $a \Rightarrow (c - a)$ از مقدار confidence $b \Rightarrow (c - b)$ کمتر نیست.

## سوال سوم

الف) تفاوت الگوهای مکرر پیدا شده در sequential pattern mining و Association ruleهای قوی در frequent pattern mining را توضیح دهید.

ب) توضیح دهید الگوریتم PrefixSpan از نوع الگوریتم‌های DFS است یا BFS؟

ج) با استفاده از الگوریتم PrefixSpan الگوهای مکرر را به دست بیاورید. توجه کنید که در اینجا min_support برابر با ۳ است.

جدول 3 . پایگاه‌داده تراکنش‌های ترتیبی

| ID | Sequence |
|---|---|
| $S_1$ | <{a},{b},{c},{b},{b},{c},{d}> |
| $S_2$ | <{d},{c},{b},{a},{b},{c}> |
| $S_3$ | <{c},{b},{b},{c},{d}> |

د) فرض کنید در روش‌های sequential pattern mining محدودیت طول کمتر از ۳ را اعمال می‌کنیم. این محدودیت anti-monotonic است یا monotonic؟

## سوال چهارم

جدول زیر را در نظر بگیرید. این جدول تعداد تراکنش‌های پیاز و/یا هویج در میان ۲۰۰۰ تراکنش را نشان می‌دهد. مقادیر $\chi^2$، $lift$ و $Kulczynski$ و $Imbalance\ ratio$ را بر اساس این جدول محاسبه کنید. بر اساس معیارهای محاسبه شده، رابطه‌ی بین خریدن پیاز و هویج را نتیجه بگیرید. در این مجموعه‌داده min_support=۰٫۲ و min_confidence=۰٫۵ است.

جدول 4. پراکندگی رده‌ها

| | پیاز | بدون پیاز | کل |
|---|---|---|---|
| هویج | ۴۰۰ | ۲۰۰ | ۶۰۰ |
| بدون هویج | ۴۰۰ | ۱۰۰۰ | ۱۴۰۰ |
| کل | ۸۰۰ | ۱۲۰۰ | ۲۰۰۰ |

برای محاسبه معیارهای ذکر شده از فرمول‌های زیر استفاده کنید:

- $\chi^2 = \sum \frac{(observed-expected)^2}{expected}$
- $lift(A,B) = \frac{P(A \cup B)}{P(A)P(B)}$
- $Kulczynski(A,B) = \frac{Support(A \cup B)}{2}(\frac{1}{Support(A)} + \frac{1}{Support(B)})$
- $IR(A,B) = \frac{|Support(A)-Support(B)|}{Support(A)+Support(B)-Support(A \cup B)}$
-

## سوال پنجم

فرض کنید مجموعه‌داده‌ای از تراکنش‌های یک مرکز خرید را در اختیار دارید و می‌خواهید الگوهای مکرر را بر اساس محدودیت‌های اعمال شده استخراج کنید. مشخص کنید هر کدام از این محدودیت‌ها از چه دسته‌ای هستند و چگونه می‌توان این الگوریتم‌ها را استخراج کرد؟ (توجه کنید قیمت کالاها بزرگ‌تر از ۰ است)

الف) محدودیت $sum(S.Price) \leq v$

ب) محدودیت $sum(S.Price) \geq v$

ج) محدودیت وجودداشتن حداقل یک پاکت شیر در سبد خرید

د) محدودیت $v \geq Avg(S.Price) \geq u$

# بخش عملی

## پیش‌نیازها و شرح دادگان

برای پاسخ به این تمرین عملی باید از Pyspark استفاده کنید. Pyspark، API زبان برنامه‌نویسی Python برای موتور apache spark است.

این مجموعه‌داده با نام Groceries.csv در فایل تمرین قرار داده شده است. مجموعه‌داده ذکر شده از ۳ ستون تشکیل شده است. ستون اول آیدی مشتری، ستون دوم تاریخ خرید و ستون سوم محصول خریداری شده را نشان می‌دهد.

## شرح وظایف

الف) مجموعه‌داده را با استفاده از دیتافریم Pyspark دریافت کنید.

- ۵ سطر اول را نمایش دهید.
- تعداد سطرهای دیتافریم را نمایش دهید.
- ستون تاریخ را حذف کنید.

ب) همان‌طور که مشاهده می‌کنید هر سطر نمایانگر خرید یک آیتم است.

- دیتافریم جدیدی بسازید که تمام خریدهای هر فرد را در یک سطر نمایش دهد. این کار به شما برای رسیدن به هدف قسمت بعد کمک می‌کند.
- آیدی خریدارانی که بیش از ۱۰ محصول جدا خریداری کرده‌اند را نشان دهید.

ج) با استفاده از تابع FPGrowth در Pyspark، itemsetهای مکرر را برگردانید. در اینجا min_support را برابر با ۰٫۱۵ در نظر بگیرید.

د) تمامی association ruleها با min_support=۰٫۱۵ و min_confidence=۰٫۴ را برگردانید.

## ملاحظات

تمامی نتایج شما باید در یک فایل فشرده با عنوان DM_CA3_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. لطفا 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 تمارین از چارچوب قرارداده شده در سامانه و کانال تلگرام استفاده کنید.
- در صورت قصد ارسال تمرین به صورت دیگر (انگلیسی، latex و ... )، لطفا پیش از ارسال با دستیارمسئول 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 تخلف برای همه‌ی افراد مشارکت کننده، نمره تمرین، صفر در نظر گرفته خواهد شد.
- در صورت بروز هرگونه مشکل با ایمیل زیر در ارتباط باشید:

mailto:adibialii@ut.ac.ir

مهلت تحویل: ۸ اردیبهشت ۱۴۰۳

مهلت تحویل با تاخیر: ۱۰ اردیبهشت ۱۴۰۳