به نام خدا

دانشگاه تهران

دانشکدگان فنی

دانشکده مهندسی برق و کامپیوتر

درس داده کاوی

تمرین عملی ۲

فروردین ماه ۱۴۰۲

* فهرست

تمرین‌های تشریحی

سؤال ۱

مجموعه داده در جدول زیر را در نظر بگیرید. ویژگی اول پیوسته بوده، در حالی که دو ویژگی باقی‌مانده به صورت دودویی نامتقارن هستند. یک rule در این سؤال، قوی بوده اگر مقدار support آن از ۱۵ درصد و مقدار confidence آن از ۶۰ درصد بیشتر باشد. مجموعه داده زیر شامل دو rule قوی هست:

A	B	C
1	1	1
2	1	1
3	1	0
4	1	0
5	1	1
6	0	1
7	0	0
8	1	1
9	0	0
10	0	0
11	0	0
12	0	1

این دو rule به صورت زیر هستند:

$$\{(1 \leq A \leq 2), B = 1\} \rightarrow \{C = 1\}$$
$$\{(5 \leq A \leq 8), B = 1\} \rightarrow \{C = 1\}$$

الف) مقدار support و confidence را برای هر دو rule محاسبه کنید.

ب) برای بدست آوردن این ruleها با استفاده از الگوریتم سنتی Apriori، باید ویژگی A را گسسته‌سازی کنیم. فرض کنید که می‌خواهیم از روش equal width binning برای گسسته‌سازی این ویژگی، با مقادیر bin-width=۲,۳,۴ استفاده کنیم. برای هرکدام از bin-width، گزارش دهید که هرکدام از ruleهای فوق توسط Apriori کشف می‌شوند یا خیر(توجه داشته باشید که ruleها در این قسمت ممکن است دقیقاً به فرمت فوق نباشند، زیرا ممکن است شامل بازه‌های بزرگتر یا کوچکتری از A باشند). برای هر rule متناسب با ruleهای فوق، مقدار support و confidence را محاسبه کنید.

ج) درباره مؤثر بودن استفاده از روش equal width برای دسته‌بندی مجموعه داده فوق توضیح دهید. آیا bin-widthای وجود دارد که به شما اجازه دهد هردو rule را بیابید؟ اگر نه، چه روش‌های جایگزینی را می‌توان استفاده کرد تا مطمئن شویم هردو rule پیدا می‌شوند؟

سؤال ۲

یک پایگاه داده دارای ۵ تراکنش است که در جدول زیر فهرست شده است. با فرض آن‌که min_support=۰.۶ و min_confidence=۰.۷ باشد، به سوال‌های زیر پاسخ دهید.

transaction_id	item_bought
001	{H, A, D, B, C}
002	{D, A, E, F}
003	{C, D, B, E}
004	{B, A, C, H, D}
005	{B, G, C}

الف) k-itemset مکرر را برای بزرگترین k فهرست کنید.

ب) یک مجموعه itemset به نام S بیابید که شروط زیر را داشته باشد.

- $S_0 \ ، \forall S_0 \subset S \ (S_0 \neq \phi)$ مکرر باشد.
- S مکرر نباشد.

ج) تمامی closed pattern ها را بیابید.

د) تمامی max-pattern ها را بیابید.

ه) تمامی association rule های قوی که با metarule زیر مطابقت دارند را بیابید و مقادیر support و confidence آن‌ها را نیز محاسبه کنید.

$$x \in \{001,002,\dots,005\}, buys(x, item_1) \wedge buys(x, item_2) \Rightarrow buys(x, item_3). [s, c]$$

و) اکنون می‌خواهیم الگوهای مکرر را با استفاده از FP-Growth استخراج کنیم. ما ابتدا الگوهای مکرر single را پیدا می‌کنیم و آنها را به ترتیب نزولی مرتب می‌کنیم. برای شکستن پیوندها، ترتیب را B-C-D-A فرض می‌کنیم. FP-tree را بسازید.

ز) پایگاه داده A's conditional را نشان دهید.

سؤال ۳

فرض کنید ما فقط به الگوهای مکرر که محدودیت های خاصی را برآورده می‌کنند علاقه‌مندیم. به عنوان مثال، در جدول سوال قبل، هر کالا قیمت خود را دارد. اطلاعات قیمت در جدول زیر آمده است. همچنین min_support=۰.۶ می‌باشد.

item	A	B	C	D	E	F	G	H
price	10	20	40	30	90	90	30	50

الف) تمامی الگوهای مکرر مانند S در جدول سوال ۲ را بیابید که محدودیت $sum(S.price) \geq 45$ برای‌شان برقرار باشد. (مجموع قیمت تمام اقلام در S کمتر از ۴۵ نباشد).

ب) آیا محدودیت $sum(S.price) \geq 45$ monotonic است یا anti-monotonic ؟ $sum(S.price) \leq 45$ چطور؟ آیا می‌توان روشی کارآمد برای استخراج الگوهای مکرر با $sum(S.price) \leq 45$ پیدا کرد؟

ج) بگذارید روی دو محدودیت دیگر $avg(S.price) \geq 30$ (یعنی میانگین قیمت همه اقلام در S کمتر از ۳۰ نیست) و $avg(S.price) \leq 30$ بحث کنیم. آیا آن‌ها convertible هستند؟ اگر بله، چگونه می‌توان آن‌ها را به موارد anti-monotonic تبدیل کرد؟

تمرین‌های عملی

تحلیل لیست مشاهده فیلم و سریال کاربران در سرویس‌های استریم فیلم مثل netflix، یکی از کلیدی‌ترین تکنیک‌ها برای درک رفتار کاربران، تعیین اینکه کدام فیلم‌ها معمولاً باهم دیده می‌شوند و ایجاد پیشنهادات شخصی به کاربران براساس تاریخچه مشاهدات است. این تحلیل معمولاً با جست‌وجو در مجموعه داده مرتبط و پیدا کردن روندهای موجود در آن(مثل فیلم‌هایی که معمولاً باهم ظاهر می‌شوند)، صورت می‌پذیرد.

فایل‌های movies.csv و rating.csv، حاوی اطلاعات فیلم‌ها و نظرات کاربران به آن‌هاست. هر سطر از rating.csv، شامل اطلاعات نظر یک کاربر به یک فیلم است(شناسه کاربر، شناسه فیلم و نظری که آن کاربر به آن فیلم داده است).

شما در این تمرین، ابتدا به بررسی داده‌ها می‌پردازید و سپس به سراغ استخراج الگوهای مکرر و Association ruleها می‌روید. به این منظور می‌توانید از کتاب‌خانه‌های apyori و MLxtend برای الگوریتم‌های مدنظر استفاده کنید.

سؤال ۱

در ابتدا به پیش‌پردازش داده‌ها بپردازید و اقدامات خود را به صورت دقیق در گزارش شرح دهید(راهنمایی: علاوه بر کارهای معمول در پیش‌پردازش، احتمالاً لازم باشد که دو مجموعه داده را یکی کنید. همچنین احتمالاً لازم باشد که فیلم‌هایی را برای ادامه‌ی کار در نظر بگیرید که از یک تعداد حداقلی نظر برخوردار باشند). سپس در قالب یک نمودار مناسب، میزان تعداد نظرات هر فیلم را نمایش دهید و نمودار بدست آمده را تفسیر کنید.

سؤال ۲

اطلاعات زیر را از مجموعه داده بدست آورید و در گزارش ذکر کنید:

- تعداد نظرات
- تعداد فیلم‌های متمایز
- ۱۰ فیلمی که بیشترین نظر را داشته‌اند

- تعداد کاربرانی که به فیلم Forrest Gump نظر داده‌اند

سؤال ۳

الف) به ازای هر کدام از موارد پایین، itemset های مکرر را با استفاده از الگوریتم Apriori کاوش کنید و تعداد آن‌ها را در گزارش بیاورید. سپس نتایج چهار حالت را با هم مقایسه کنید.

- min-support=۰.۱ و min-lenght=۲
- min-support=۰.۲ و min-lenght=۲
- min-support=۰.۳ و min-lenght=۲
- min-support=۰.۵ و min-lenght=۲

ب) مناسب‌ترین حالت را از میان موارد بالا انتخاب نمایید و دلیل انتخاب خود را شرح دهید.

ج) به ازای هر یک از موارد زیر، با استفاده از الگوریتم FP-Growth، itemset های مکرر را کاوش کنید و هر کدام از آن‌ها را به همراه مقدار support آن گزارش کنید.

- min-support = 0.1
- min-support = 0.2
- min-support = 0.3
- min-support = 0.5

سوال ۴

الف) تمام Association Rule ها با min-support=۰.۳ و min-confidence=۰.۶ را استخراج نمایید و تعداد این قوانین را ذکر کنید. سه قانونی که بالاترین lift را دارند، بنویسید.

ب) تمام Association Rule ها با min-support=۰.۳ و min-confidence=۰.۸ را استخراج نمایید و تعداد این قوانین را ذکر کنید. تعداد قوانین نسبت به حالت قبل چه تغییری کرد؟ علت این تغییر را توضیح دهید.

تمرین تشریحی امتیازی

سؤال ۱

جدول زیر تعداد transactionهایی که شامل شیر و/یا خرما هستند را میان ۱۰۰۰ transaction نمایش می‌دهد. مقادیر معیارهای χ^2، $lift$ و $all-confidence$ را براساس این جدول محاسبه کنید. براساس معیارهای محاسبه شده، رابطه‌ی بین خریدن شیر و خرما را نتیجه بگیرید.

	شیر	بدون شیر	کل
خرما	50	800	850
بدون خرما	150	9000	9150
کل	200	9800	10000

راهنمایی: برای محاسبه معیارهای فوق، می‌توانید از فرمول‌های زیر استفاده کنید:

- $\chi^2 = \Sigma \frac{(observed-expected)^2}{expected}$
- $lift(A,B) = \frac{P(A \cup B)}{P(A)P(B)}$
- $all-conf(X) = \frac{sup(X)}{max-item-sup(X)} = \frac{sup(X)}{max\{sup(i_j)|i_j \in X\}}$

که در فرمول آخر، $max\{sup(i_j)|i_j \in X\}$ برابر با بیشترین item support در بیان تمامی itemهای X است.

ملاحظات (حتما مطالعه شود)

تمامی نتایج شما باید در یک فایل فشرده با عنوان DM_CA2_StudentID تحویل داده شود.

- این فایل فشرده، باید حاوی یک فایل با فرمت PDF (گزارش تایپ شده) و یک پوشه به نام Codes باشد که کدهای نوشته شده را شامل می‌شود. در صورتی که از Jupyter Notebook استفاده می‌کنید نیازی به ارسال جداگانه کدها و گزارش بخش عملی نیست و هر دو را می‌توانید در یک فایل Notebook قرار دهید. حتما خروجی html فایل Notebook خود را نیز همراه فایل Notebook ارسال کنید.
- 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.
- گزارش به صورت تایپ شده در قالب PDF شامل شرح آزمایش‌های انجام شده، پارامترهای آزمایش، نتایج و تحلیل‌ها باشد. دقت داشته باشید که در تمامی تمرین‌ها، نمره‌ی اصلی به تفسیر و تحلیل شما تعلق می‌گیرد.
- مهلت تحویل تمرین به هیچ عنوان تمدید نخواهد شد. مجموعا ۱۴ روز برای تمامی تمرین‌ها و پروژه‌ی درس به عنوان Grace day در نظر گرفته می‌شود و پس از پایان مجموعا ۱۴ روز، برای هر تمرینی که پس از زمان اختصاص یافته ارسال شود روزی ۱۵ درصد از نمره آن تمرین کسر خواهد شد.
- توجه کنید این تمرین باید به صورت تک نفره انجام شود و پاسخ‌های ارائه شده باید نتیجه فعالیت فرد نویسنده باشد (همفکری و به اتفاق هم نوشتن تمرین نیز ممنوع است). در صورت مشاهده تقلب به همه افراد مشارکت کننده، نمره تمرین صفر و به استاد نیز گزارش می‌گردد.
- در صورت بروز هرگونه مشکل با ایمیل زیر در ارتباط باشید:

mailto:taha.fakharian@gmail.com

mailto:mohammad.saadati80@gmail.com

مهلت تحویل بدون جریمه: ۰۱ / ۰۲ / ۱۴۰۲