به نام خدا

دانشگاه تهران

دانشکدگان فنی

دانشکده مهندسی برق و کامپیوتر

درس داده‌کاوی

تمرین عملی ۲

اردیبهشت ماه ۱۴۰۱

* فهرست

تمرین‌های تشریحی

سؤال ۱

یک پایگاه داده، ۴ تراکنش دارد که در جدول زیر نشان داده شده‌اند. با فرض آن که min_sup = 60% و min_conf = 80% باشد، به سؤالات زیر پاسخ دهید:

TID	items_bought
T100	{K, A, D, B}
T200	{D, A, C, E, B}
T300	{C, A, B, E}
T400	{B, A, D}

الف) با استفاده از الگوریتم Apriori، تمام itemset های مکرر را پیدا کنید.

ب) تمام Association Rule های قوی را که با metarule زیر مطابقت دارند، بیابید و مقادیر support و confidence آن‌ها را بنویسید. در metarule زیر، X متغیری است که مشتریان را نشان می‌دهد و $item_i$ بیانگر متغیرهایی است که آیتم‌ها را نشان می‌دهند.

$$\forall x \in transaction, buys(X, item_1) \wedge buys(X, item_2) \Rightarrow buys(X, item_3)$$

سؤال ۲

جدول زیر، خلاصه‌ای از داده‌های تراکنش یک سوپرمارکت را نشان می‌دهد که در آن، $hot\ dogs$ به تراکنش های حاوی hot dogs اشاره می‌کند و $\overline{hot\ dogs}$ به تراکنش‌های فاقد hot dogs اشاره می‌کند. همچنین، $hamburgers$ به تراکنش های شامل hamburgers و $\overline{hamburgers}$ به تراکنش‌های فاقد hamburgers اشاره دارد.

	$\boldsymbol{hot\ dogs}$	$\boldsymbol{\overline{hot\ dogs}}$	$\boldsymbol{\sum_{row}}$
$\boldsymbol{hamburgers}$	2000	500	2500
$\boldsymbol{\overline{hamburgers}}$	1000	1500	2500
$\boldsymbol{\sum_{col}}$	3000	2000	5000

الف) بر اساس داده‌های جدول، آیا خرید $hot\ dogs$ مستقل از خرید $hamburgers$ است؟ اگر خرید $hot\ dogs$ مستقل از خرید $hamburgers$ نیست، چه نوع رابطه‌ی همبستگی بین این دو وجود دارد؟ (با محاسبه‌ی معیار lift برای خریدن $hot\ dogs$ و $hamburgers$، به سؤالات بخش الف پاسخ دهید.)

ب) با توجه به اطلاعات جدول بالا، دو معیار all-confidence و cosine را برای خریدن $hot\ dogs$ و $hamburgers$ محاسبه نمایید.

سؤال ۳

مجموعه‌ی تراکنش‌ها و ارزش آیتم‌های مربوط به آن‌ها در جداول زیر گزارش شده‌اند. می‌خواهیم همه‌ی itemset های مکرری را بیابیم که محدودیت $\min(\text{value}(s)) \leq 2000$ برایشان برقرار است. با فرض این که $\text{min\_sup} = 2$ باشد، itemset های مکرر با این شرایط را با استفاده از الگوریتم FP-Growth بیابید.

TID	Items
100	Milk, Peanut, Butter, Cake
200	Cake, Chips, Peanut, Tea
300	Cheese, Chips, Peanut
400	Chips, Milk, Cheese, Butter, Peanut
500	Milk, Water
600	Chips, Peanut, Cheese

Item	Value
Milk	3000
Tea	3000
Butter	2500
Peanut	2300
Chips	2000
Cake	1500
Cheese	1200
Water	1000

تمرین‌های عملی

تحلیل سبد خرید، یکی از تکنیک‌های کلیدی است که برای کشف ارتباط بین اقلام مورد استفاده قرار می‌گیرد و اطلاعاتی را برای درک رفتار خرید مشتریان فراهم می‌کند. این تحلیل با جست‌وجوی ترکیباتی از اقلام که اغلب در تراکنش‌ها با هم حضور دارند، انجام می‌شود.

فایل Market_Basket.csv حاوی اطلاعاتی در مورد مشتریانی است که مواد غذایی مختلفی را از یک فروشگاه خریداری کرده‌اند. هر سطر از این فایل، یک تراکنش را نشان می‌دهد و شامل اقلامی هست که در آن تراکنش با هم توسط یک مشتری خریداری شده‌اند.

شما در این تمرین، ابتدا به بررسی داده‌ها می‌پردازید و سپس به سراغ استخراج الگوهای مکرر و Association Rule ها می‌روید. توجه نمایید که به منظور اجرای الگوریتم‌های Apriori و FP-Growth می‌توانید از کتاب‌خانه‌هایی مانند apyori و MLxtend استفاده کنید.

سؤال ۱

در ابتدا به پیش‌پردازش داده‌ها بپردازید و اقدامات خود را به صورت دقیق در گزارش شرح دهید. سپس، در قالب یک نمودار مناسب، میزان فروش هر آیتم را نشان دهید و نمودار به دست آمده را تفسیر کنید.

سؤال ۲

اطلاعات زیر را از مجموعه‌داده به دست آورید و در گزارش ذکر کنید:

الف) تعداد تراکنش‌ها

ب) تعداد آیتم‌های متمایز

ج) ۵ آیتمی که بیشترین فروش را داشته‌اند

د) تعداد تراکنش‌هایی که "black tea" در آن‌ها خریداری شده است

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 سه حالت را با هم مقایسه کنید.

- min-support = 0.003 و min-length = 2
- min-support = 0.03 و min-length = 2
- min-support = 0.3 و min-length = 2

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

ج) با در نظر گرفتن min-support = 0.05 و 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.

سؤال ۴

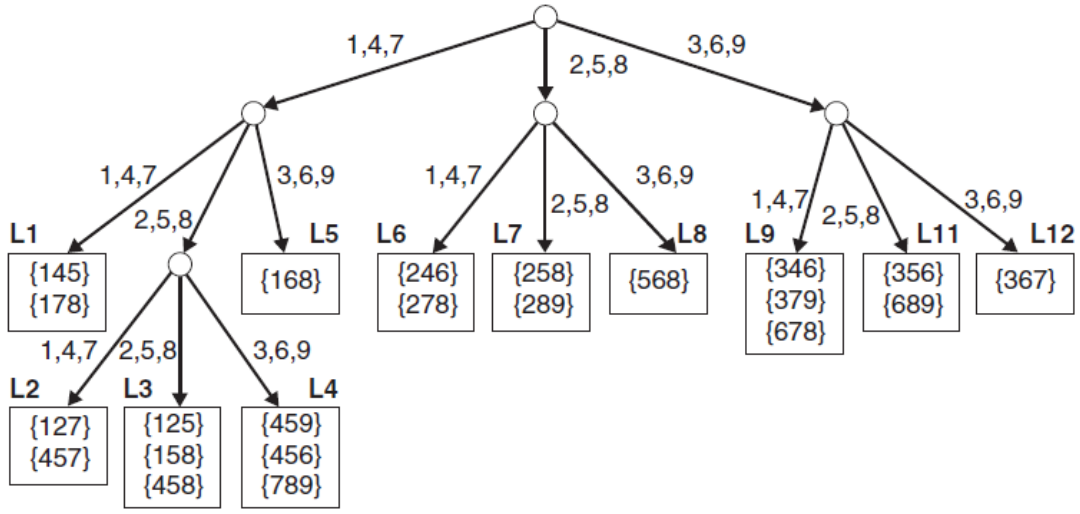
الف) تمام Association Rule ها با min-support = 0.03 و min-confidence = 0.2 را استخراج نمایید و تعداد این قوانین را ذکر کنید. سه قانونی که بالاترین lift را دارند، بنویسید.

ب) تمام Association Rule ها با min-support = 0.03 و min-confidence = 0.35 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

الگوریتم Apriori از ساختمان داده‌ی hash tree به منظور شمارش کارآمد support برای candidate itemset ها استفاده می‌کند. hash tree نشان داده شده برای candidate 3-itemset ها را در نظر بگیرید.



الف) با توجه به تراکنشی که شامل اقلام {1, 3, 4, 5, 8} است، کدام یک از گره‌های برگ hash tree هنگام یافتن candidate های تراکنش بازدید[۱] می‌شوند؟

ب) با استفاده از گره‌های برگ بازدید شده در بخش (الف)، candidate itemset های موجود در تراکنش {1, 3, 4, 5, 8} را تعیین نمایید.

[۱] visit

ملاحظات (حتما مطالعه شود)

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

- این فایل فشرده، بایستی حاوی یک فایل با فرمت PDF (گزارش تایپ شده) و یک پوشه به نام Codes باشد که کدهای نوشته شده را شامل 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.
- گزارش به صورت تایپ شده در قالبPDF 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. تمرین تا یک هفته بعد از مهلت تعیین شده با جریمه تحویل گرفته می‌شود که جریمه تاخیر تحویل تمرین تا یک هفته ۳۰ درصد است.
- توجه کنید این تمرین باید به صورت تک نفره انجام شود و پاسخ‌های ارئه 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 خانم اکبری امین از طریق ایمیل زیر در ارتباط باشید:

mailto:mahsan.a.a@gmail.com

مهلت تحویل بدون جریمه: ۱۸ اردیبهشت ۱۴۰۱

مهلت تحویل با تاخیر، با جریمه ۳۰ درصد: ۲۵ اردیبهشت ۱۴۰۱