به نام خدا

داده کاوی

تمرین عملی دوم

محمد ناصری

۸۱۰۱۰۰۴۸۶

بهار ۱۴۰۱

تمرین‌های تشریحی

سوال اول

یک پایگاه داده، 4تراکنش دارد که در جدول زیر نشان داده شده‌اند. با فرض آن که min_sup = 60 وmin_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 ،Aprioriتمام itemsetهای مکرر را پیدا کنید.

C1		L1		C2		L2		C3		L3	
A	4	A	4	AB	4	AB	4	ABD	3	ABD	3
B	4	B	4	AD	3	AD	3				
C	2	D	3	BD	3	BD	3				
D	3										
E	2										
K	1										

با پیاده سازی الگوریتم priori به جداول بالا میرسیم.در جدول موارد پر تکرار هایلایت شده اند. برای بدست آوردن جدول بالا در ابتدا مقادیر آیتمهای تکی بدست می آید و با کمک min support آنها را prune میکنیم و این عمل را برای آیتم‌های دوتایی و سه‌تایی تکرار میکنیم. نتیجه جداول بالا خواهند شد و نتایج frequent patternها به شرح زیر است:

{A:4, B:4, D:3, AB:4, AD:3, BD:3, ABD:3}

ب)

تمام Association Ruleهای 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 $imet_i$ بیانگر متغیرهایی است که آیتم‌ها را نشان میدهند.

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

دراین سوال منظور metarule آورده شده به زبان ساده این است که اگر فردی آیتم های X_1, X_2 را خریداری کرده باشد آیتم X_3 را نیز خریداری کرده است. برای بدست آوردن این دسته Associationها باید آیتم ست های ۳ تایی را مورد بررسی قرار بدهیم. با بررسی به موارد زیر میرسیم:

X1, X2	X3	Support	Confidence
AB	D	75%	75%
AD	B	75%	100%
BD	A	75%	100%

همانطور که در بالامشاهده میشود ۳ رابطه بدست می آید که از میان آنها مورد اول از min-confidence مقدار کمتری دارد و نمیتوان گفت که رابطه قوی هست.

سوال دوم

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 $\overline{hot\ dogs}$ تراکنشهای فاقد hot dogs اشاره میکند. همچنین، hamburgers به تراکنش های شامل hamburgers و $\overline{hamburgers}$ تراکنشهای فاقد hamburgers اشاره دارد.

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

الف)

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

در یادگیری قاعده انجمنی در داده کاوی، Lift معیار عملکرد برای هدف قرار دادن مدل (قاعده انجمنی) و در پیش‌بینی یا طبقه‌بندی موارد برای بدست آوردن پاسخ درست، افزایش یافته (با توجه به کل جمعیت) است، که برای مقایسه و انتخاب هدفمند تصادفی این مدل اندازه‌گیری می‌شود. در صورتی که نتیجه درون هدف، بسیار بهتر از متوسط برای کل جامعه باشد یعنی یک مدل هدف گذاری درستی انجام می‌دهد. Lift نسبت به این مقادیر می‌باشد: پاسخ هدف تقسیم بر میانگین پاسخ.

$$Lift(A,B) = \frac{P(A \cap B)}{P(A) * P(B)}$$

با توجه به جدول خواهیم داشت:

$$P(hot\ dogs) = \frac{3000}{5000}$$
$$P(hamburgers) = \frac{2500}{5000}$$
$$P(hot\ dogs, hamburgers) = \frac{2000}{5000}$$
$$LIFT(hot\ dogs, hamburgers) = \frac{\frac{2}{5}}{\frac{1}{2} * \frac{3}{5}} = \frac{4}{3}$$

با توجه به مقدار بدست امده برای lift که $4/3 > 1$ میتوانیم بگوییم بر اساس معیار لیفت خرید این ۲ آیتم با یکدیگر همبستگی (correlation) مثبت دارند.

ب)

با توجه به اطلاعات جدول بالا، دو معیار all-confidenceو cosineرا برای خریدنhot dogs, hamburgers محاسبه نمایید.

$$alll_{confidence}(A,B) = \frac{\sup(A,B)}{\max\{\sup(A), \sup(B)\}} = \frac{2000}{\max\{2500, 3000\}} = \frac{2}{3} > 0.5$$
$$Cosine(A,B) = \frac{\sup(A,B)}{\sqrt{\sup(A)\sup(B)}} = \frac{2000}{\sqrt{2500 * 3000}} \sim \frac{2000}{2738} > 0.5$$

تمامی معیارهای محسابه شده همبستگی مثبت را برای این دو متغیر نشان میدهند.

سوال سوم

مجموعهی تراکنش‌ها و ارزش آیتم‌های مربوط به آنها در جداول زیر گزارش شده‌اند. میخواهیم همه‌ی itemsetهای مکرری را بیابیم که محدودیت $\min(\text{value}(s)) \leq 2000$ برایشان برقرار است. با فرض این که $\text{min\_sup} = 2$ باشد، itemsetهای 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

در مرحله اول لیست آیتم‌ها و ساپرت آنها را استخراج میکنیم.

Item	Support	Value
Milk	3	**3000**
Tea	1	**3000**
Butter	2	**2500**
Peanut	5	**2300**
Chips	4	**2000**
Cake	2	**1500**
Cheese	3	**1200**
Water	1	**1000**

سپس آیتم‌ها را بر اساس support مرتب میکنیم:

Item	Support	Value
Peanut	5	**2300**
Chips	4	**2000**
Milk	3	**3000**
Cheese	3	**1200**
Butter	2	**2500**
Cake	2	**1500**

سپس تراکنش‌ها را بر اساس ترتیبی بدست آمده از آیتم‌ها مرتب میکنیم:

{Peanut, Milk, Butter, Cake}

{Peanut, Chips, Cake}

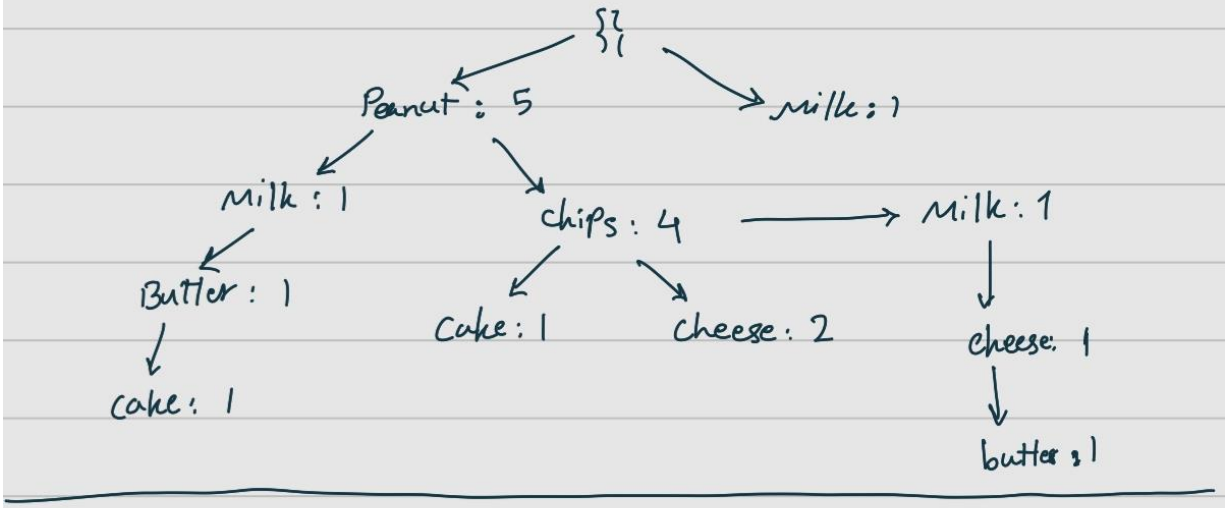
{Peanut, Chips, Cheese}

{Peanut, Chips, Milk, Cheese, Butter}

{Milk}

{Peanut, Chips, Cheese}

سپس درخت اولیه را رسم میکنیم:



conditional Pattern (Cake) : { { Peanut, Milk, butter : 1 } , { Peanut, chips : 1 } }

(value = 1500)

Peanut	2	✓
milk	1	✗
butter	1	✗
chips	1	✗

FP: { { Cake, Peanut : 2 } , { Cake : 2 } }

conditional Pattern (Butter) : { { Peanut, Milk : 1 } , { Peanut, chips, Mik, cheese : 1 } }

(value = 2500)

Peanut	2	✓
milk	2	✓
chips	1	✗
cheese	1	✗

→ Min مقدار هرده از value

بالا آمدیت ، پایه نیز 2500 می باشد

پس Prune می شوند.

conditional Pattern (Cheese) : { { Peanut, chips : 2 } , { Peanut, chips, milk : 1 } }

(value = 1200)

Peanut	3	✓
chips	3	✓
milk	1	✗

FP: { { Peanut, cheese : 3 } , { chips, cheese : 3 } , { chips, cheese, Peanuts 3 } }

نیازی به بررسی باقی conditional patternها نیست زیرا شرط محدودکننده مارا ندارند و همه بالاتر از 2000 ارزش هستند.

تمرین‌های عملی

سوال اول

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.

در مرحله پیش پردازش داده‌ها در ابتدا لیستی از لیست‌های تراکنش‌ها ایجاد میکنیم و ۲ عمل بر روی این لیست انجام میدهیم. اول آنکه ساپرت یا همان تعداد تکرار هر آیتم را در کل تراکنش‌ها بدست می آوریم و اقدام دوم هم به کمک کتابخانه mlextend دیتافریمی به صورتی ایجاد میکنیم که ستون‌های آن متشکل از آیتم‌ها و هر ردیف نشان‌دهنده تراکنش باشد و در هر خانه دیتافریم وجود یا عدم وجود آیتم در تراکنش به صورت True, False ثبت شود.

در مرحله بعد به کمک دیکشنری بدست آمده که شامل تعداد تکرار آیتم در تراکنش‌هاست نمودار barplot رسم میکنیم. در این نمودار تعداد کمی آیتم بسیار پرتکرار (مانند آب معدنی) و تعداد زیادی آیتم با تعداد تکرار متوسط و پایین وجود دارند. با توجه به این شکل میتوان انتظار داشت برای مثال آیتم آب معدنی در اکثر خریدهای افراد وجود داشته باشد و افراد هنگام خرید خود این ایتم را نیز خریداری میکنند. همینطور مورد قابل مشاهده دیگر این است که در این فروشگاه آیتم‌های سوپرمارکتی فروش بیشتری نسبت به باقی آیتم‌ها دارند.

سوال دوم

موارد خواسته شده سوال در فایل نوت‌بوک مربوطه بدست‌اورده و نمایش داده شده‌اند.

سوال سوم

الف

موارد خواسته شده سوال در فایل نوت‌بوک مربوطه بدست‌اورده و نمایش داده شده‌اند.

ب)

بهترین انتخاب از بین این ۳ مورد، مورد دوم است زیرا در حالت اول تعداد آیتم‌ست‌های پرتکرار بسیار زیادی برگردانده میشود و بطور عکس در حالت سوم هیچ آیتم‌ستی بازگردانده نمیشود ولی در حالت دوم تعداد مناسب آیتم‌ست پرتکرار برای تحلیل بدست می‌آوریم.

ج)

موارد خواسته شده سوال در فایل نوت‌بوک مربوطه بدست‌اورده و نمایش داده شده‌اند.

سوال چهارم

الف)

موارد خواسته شده سوال در فایل نوت‌بوک مربوطه بدست‌اورده و نمایش داده شده‌اند.

ب)

موارد خواسته شده سوال در فایل نوت‌بوک مربوطه بدست‌اورده و نمایش داده شده‌اند.

معیار confidence به نوعی نشان‌دهنده قدرت ارتباط آیتم‌هاست. این معیار نشان میدهد که در صورت وجود آیتم A در یک تراکنش چقد احتمال دارد که آیتم B نیز در تراکنش مشاهده شود. این معیار در کنار معیار support با یکدیگر Strong-association-rulهها را مشخص میکنند و باید در کنار یکدیگر بررسی شوند. در این سوال تعدادی از آیتم‌ست‌های پرتکرار با بالا بردن confidence حذف شدند و آیتم‌ها با رابطه قوی‌تر باقی ماندند.