دانشگاه تهران
دانشکده‌ی مهندسی کامپیوتر

تمرین سوم

داده کاوی

خانم دکتر شاکری

محمد امین عرب خراسانی
۸۱۰۱۰۲۲۰۵

بهار ۱۴۰۳

بخش تشریحی

۱ سوال اول

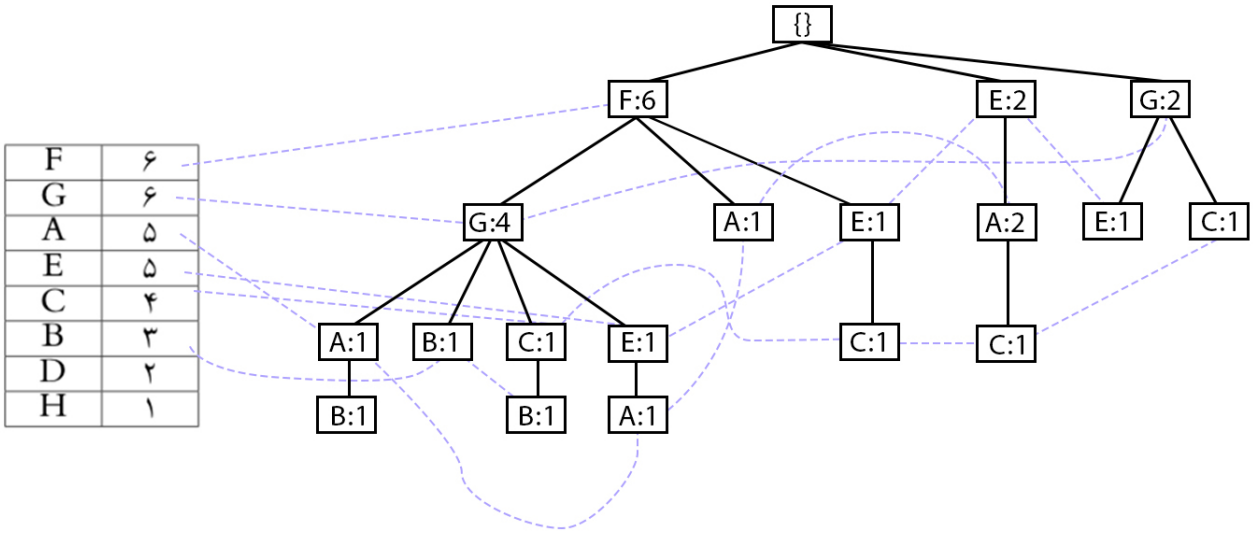
الف) برای رسم FP-tree ابتدا تعداد دفعات تکرار هر item-bought در محاسبه شده و به ترتیب تعداد دفعات تکرار از بیشترین به کمترین دفعات تکرار مرتب می‌شود. در نتیجه F-list به فرم زیر خواهد بود.

F	۶
G	۶
A	۵
E	۵
C	۴
B	۳
D	۲
H	۱

همانطور که از جدول بالا مشخص است، D و H مقدار min-support را ارضا نمی‌کنند بنابراین در FP-tree لحاظ نمی‌شوند. در ادامه transaction ها بر اساس تعداد تکرار مرتب می‌شوند تا FP-tree رسم شود. نتیجه‌ی این مرتب‌سازی در جدول زیر آورده شده است.

Item-bought	rearranged
G، B، A، F	F، G، A، B
H، A، E	E، A
F، B، G	F، G، B
A، C، D، E	E، A، C
C، F، G، B	F، G، C، B
A، F، D	F، A
E، F، G، A	F، G، E، A
C، F، E	F، E، C
G، E	G، E
H، C، G	G، C

در نهایت FP-tree به شکل زیر خواهد بود.



ب) برای به دست آوردن B's conditional تمامی شاخه‌هایی که به B ختم می‌شوند از FP-tree به دست می‌آیند. برای B's conditional داریم:

B's conditional database: FGA:۱، FG:۱، FGC:۱

ج) برای پیدا کردن closed pattern و max-pattern به صورت sequential هر itemset محاسبه می‌شود.

۱-D itemset

item	count
F	۶
G	۶
A	۵
E	۵
C	۴
B	۳
D	۲
H	۱

با توجه به min-support و جدول بالا، H و D کمتر از ۳ بار تکرار شده‌اند بنابراین در ادامه‌ی itemset ها لحاظ نمی‌شوند. بقیه‌ی item-bought ها این شرط را ارضا می‌کنند.

۲-D itemset

item	count
F، G	۴
F، A	۳
F، E	۲
F، C	۲
F، B	۳
G، A	۲
G، E	۲
G، C	۲
G، B	۳
A، E	۳
A، C	۱
A، B	۱
E، C	۲
E، B	۰
C، B	۱

با توجه به شرط minimum-support سطرهایی از جدول بالا که صورتی هستند حذف می‌شوند.

F(۶) => FG(۴) ، FA(۳) ، FB(۳) ، FE(۲) ، FC(۲)

با توجه به آن که F(۶) از تمام زیرمجموعه‌ها بزرگتر است بنابراین F، closed می‌باشد.
از آن‌جایی که زیرمجموعه‌ای وجود دارد که شرط minimum-support را ارضا می‌کند بنابراین F، maximal نمی‌باشد.

G(۶) => FG(۴) ، GA(۲) ، GE(۲) ، GC(۲) ، GB(۳)

با توجه به آن که G(۶) از تمام زیرمجموعه‌ها بزرگتر است بنابراین G، closed می‌باشد.
از آن‌جایی که زیرمجموعه‌ای وجود دارد که شرط minimum-support را ارضا می‌کند بنابراین G، maximal نمی‌باشد.

A(۵) => FA(۳) ، GA(۲) ، AC(۱) ، AB(۱)

با توجه به آن که A(۵) از تمام زیرمجموعه‌ها بزرگتر است بنابراین A، closed می‌باشد.

از آن‌جایی که زیرمجموعه‌ای وجود دارد که شرط minimum-support را ارضا می‌کند بنابراین A، maximal نمی‌باشد.

E(۵) => FE(۲) ، GE(۲) ، AE(۳) ، EC(۲) ، EB(۰)

با توجه به آن که E(۵) از تمام زیرمجموعه‌ها بزرگتر است بنابراین E، closed می‌باشد.

از آن‌جایی که زیرمجموعه‌ای وجود دارد که شرط minimum-support را ارضا می‌کند بنابراین E، maximal نمی‌باشد.

C(۴) => FC(۲) ، GC(۲) ، AC(۱) ، EC(۲) ، CB(۱)

با توجه به آن که C(۴) از تمام زیرمجموعه‌ها بزرگتر است بنابراین C، closed می‌باشد.

از آن‌جایی که زیرمجموعه‌ای وجود ندارد که شرط minimum-support را ارضا کند بنابراین C، maximal می‌باشد.

B(۳) => FB(۳) ، GB(۳) ، AB(۱) ، EB(۰) ، CB(۱)

با توجه به آن که B(۳) از تمام زیرمجموعه‌ها بزرگتر نیست بنابراین B، closed نمی‌باشد.

از آن‌جایی که زیرمجموعه‌ای وجود دارد که شرط minimum-support را ارضا می‌کند بنابراین B، maximal نمی‌باشد.

۳-D itemset

item	count
F، G، A	۲
F، G، B	۳
F، A، B	۱
F، E، A	۱

از آن‌جایی که ۱۰ تا از زیرمجموعه‌های itemset قبلی frequent نیستند فقط برای زیرمجموعه‌های frequent بررسی می‌شوند.

FG(۴) => FGA(۲) ، FGB(۳)

با توجه به آن که FG(۴) از تمام زیرمجموعه‌ها بزرگتر است بنابراین FG، closed می‌باشد.

از آن‌جایی که زیرمجموعه‌ای وجود دارد که شرط minimum-support را ارضا می‌کند بنابراین FG، maximal نمی‌باشد.

FA(۳) => FGA(۲) ، FAB(۱) ، FEA(۱)

با توجه به آن که FA(۳) از تمام زیرمجموعه‌ها بزرگتر است بنابراین FA، closed می‌باشد.
از آن‌جایی که زیرمجموعه‌ای وجود ندارد که شرط minimum-support را ارضا کند بنابراین FA، maximal می‌باشد.

FB(۳) => FGB(۳) ، FAB(۱)

با توجه به آن که FB(۳) از تمام زیرمجموعه‌ها بزرگتر نیست بنابراین FB، closed نمی‌باشد.
از آن‌جایی که زیرمجموعه‌ای وجود دارد که شرط minimum-support را ارضا می‌کند بنابراین FB، maximal نمی‌باشد.

GB(۳) => FGB(۳)

با توجه به آن که GB(۳) از تمام زیرمجموعه‌ها بزرگتر نیست بنابراین GB، closed نمی‌باشد.
از آن‌جایی که زیرمجموعه‌ای وجود دارد که شرط minimum-support را ارضا می‌کند بنابراین GB، maximal نمی‌باشد.

AE(۳) => FEA(۱)

با توجه به آن که AE(۳) از تمام زیرمجموعه‌ها بزرگتر است بنابراین AE، closed می‌باشد.
از آن‌جایی که زیرمجموعه‌ای وجود ندارد که شرط minimum-support را ارضا کند بنابراین AE، maximal می‌باشد.

۴-D itemset

item	count
F، G، B، A	۱
F، G، C، B	۱

FGB(۳) => FGBA(۱) ، FGCB(۱)

با توجه به آن که FGB(۳) از تمام زیرمجموعه‌ها بزرگتر است بنابراین FGB، closed می‌باشد.
از آن‌جایی که زیرمجموعه‌ای وجود ندارد که شرط minimum-support را ارضا کند بنابراین FGB، maximal می‌باشد.

در نهایت جدول زیر closed pattern و max-pattern آورده شده است.

closed pattern: F، G، A، E، C، FG، FA، AE، FGB

max-pattern: C، FA، AE، FGB

د) با توجه به itemset های حاصل در بخش قبل، و فرمول زیر جدول مربوط به association rule تکمیل می‌شود.

$$\{X\} \to \{Y\} : \frac{\text{support}\,(\{X,\ Y\})}{\text{support}\,(\{X\})}$$

$\{A\} \to \{E\} : ۰/۶$	$\{A\} \to \{F\} : ۰/۶$	$\{G\} \to \{F\} : ۰/۶۷$	$\{F\} \to \{G\} : ۰/۶۷$
$\{F,G\} \to \{B\} : ۰/۷۵$	$\{B\} \to \{G\} : ۱$	$\{B\} \to \{F\} : ۱$	$\{E\} \to \{A\} : ۰/۶$
$\{B,G\} \to \{F\} : ۱$	$\{G\} \to \{F,B\} : ۱$	$\{F,B\} \to \{G\} : ۱$	$\{B\} \to \{F,G\} : ۱$
			$\{F\} \to \{B,G\} : ۱$

همانطور که از نتایج مشخص است association rules قوی با فرض minimum-confidence ارائه شده حاصل می‌شود.

ه) ابتدا یک جدول تشکیل داده می‌شود که شامل support count های هر آیتم می‌باشد. ترتیب این جدول بر اساس تعداد دفعات تکرار است.

۱-D itemset

itemset	support-count
A	۴
D	۴
F	۴
G	۴
B	۲
C	۲
E	۱
H	۱

با توجه به minimum-support داده شده، item هایی که frequent نمی‌باشند حذف می‌شوند.

۲-D itemset

itemset	support-count
F، G	۴
A، D	۳
A، F	۳
A، G	۳
D، F	۳
D، G	۳

در این مرحله تمام زیر مجموعه‌ها حفظ می‌شوند.

۳-D itemset

itemset	support-count
D، F، G	۳
A، F، G	۳
A، D، F	۲
A، D، G	۲

برای تشکیل مراحل فوق برای مجموعه‌های چهارتایی فقط ۲ آیتم‌ست سه تایی موجود است و بنابراین یک مجموعه ۴ تایی از ادغام آن‌ها می‌توان ساخت که به صورت A، D، F، G است، که از زیرمجموع‌های همین مجموعه چهارتایی، موردی مثل A، D، G در لیست پیشین وجود ندارد یا Frequent نبوده و تاییدی است بر اینکه مراحل در همین جا متوقف شده و هیچ subset دیگری نمی‌توان ساخت.

و) همانطور که در مراحل قبل اشاره شد با توجه به مراحل الگوریتم، بزرگترین itemset ممکن مربوط به A، D، F و A، D، G می‌باشد. زیرا در A، D، F، G که خود مکرر نمی‌باشد آیتم‌ست‌های دیگری نظیر A، D، F وجود دارد که frequent نمی‌باشد. اما در A، D، F و A، D، G خود frequent نمی‌باشند و آیتم‌ست دیگری غیر از تهی ندارند.

ز) با توجه به meta rule ی که در صورت سوال اشاره شده است، فقط آن آیتم‌ست‌هایی لحاظ می‌شوند که به فرم زیر باشند.

$$\{X, Y\} \rightarrow \{Z\}$$

در نتیجه برای association rules قوی با توجه به minimum-confidence عنوان شده داریم:

$$\{F,\ G\} \rightarrow \{A\} : \frac{support\left(\{A,\ F,\ G\}\right)}{support\left(\{F,\ G\}\right)} = \frac{۳}{۴} \times ۱۰۰ = ۷۵٪$$

$$\{A,\ F\} \rightarrow \{G\} : \frac{support\left(\{A,\ F,\ G\}\right)}{support\left(\{A,\ F\}\right)} = \frac{۳}{۳} \times ۱۰۰ = ۱۰۰٪$$

$$\{A,\ G\} \rightarrow \{F\} : \frac{support\left(\{A,\ F,\ G\}\right)}{support\left(\{A,\ G\}\right)} = \frac{۳}{۳} \times ۱۰۰ = ۱۰۰٪$$

$$\{F,\ G\} \rightarrow \{D\} : \frac{support\left(\{D,\ F,\ G\}\right)}{support\left(\{F,\ G\}\right)} = \frac{۳}{۴} \times ۱۰۰ = ۷۵٪$$

$$\{D,\ F\} \rightarrow \{G\} : \frac{support\left(\{D,\ F,\ G\}\right)}{support\left(\{D,\ F\}\right)} = \frac{۳}{۳} \times ۱۰۰ = ۱۰۰٪$$

$$\{D,\ G\} \rightarrow \{F\} : \frac{support\left(\{D,\ F,\ G\}\right)}{support\left(\{D,\ G\}\right)} = \frac{۳}{۳} \times ۱۰۰ = ۱۰۰٪$$

---

۲ سوال دوم

الف) جمله‌ی اول: درست است. از آنجایی که این association قوی بین s و l برقرار است به این معناست که شرط min-confidence ارضا شده است.

$$\{s\} \to \{l\} : \frac{support\,(\{s,\ l\})}{support\,(\{s\})} \geqslant min - confidence$$

برای آیتم‌ست a نیز داریم:

$$\{a\} \to \{l\} : \frac{support\,(\{a,\ l\})}{support\,(\{a\})}$$

از آنجایی که a زیرمجموعه‌ی s است بنابراین هر ترنزاکشنی از a در s وجود دارد پس داریم:

$$support(a) \leqslant support(s) \to support(a, l) \leqslant support(s, l)$$

$$\frac{support\,(\{a,\ l\})}{support\,(\{a\})} \leqslant \frac{support\,(\{s,\ l\})}{support\,(\{s\})}$$

جمله‌ی دوم: الف) ۱. نادرست است.

برای آیتم‌ست a داریم:

$$\{a\} \to \{l\} : \frac{support\,(\{a,\ l\})}{support\,(\{a\})}$$

فرض می‌کنیم a و s برابر مجموعه‌ی زیر است.

a: {bread، egg، butter}

s: {bread، egg}

$$\{bread,\ egg,\ butter,\ l\} \to \{l\} : \frac{support\,(\{a,\ l\})}{support\,(\{a\})}$$

ب) برای مقایسه‌ی confidence این دو و اثبات عبارت موجود در سوال ابتدا مقادیر confi-dence محاسبه می‌شود. بنابراین داریم:

$$conf(a \to (c-a)) = \frac{sc - sa}{sa}$$

$$conf(b \to (c-b)) = \frac{sc - sb}{sb}$$

از آن جایی که $a \subseteq b$ و $b \subseteq c$ است بنابراین هر $sa \leqslant sb$ است زیرا همه‌ی زیرمجموعه‌های ممکن برای a در آیتم‌ست b می‌باشد. با توجه به این نکته confidence ها با یکدیگر مقایسه می‌شوند.

$$\frac{sc - sb}{sb} \leqslant \frac{sc - sa}{sa}$$

$$\frac{sc}{sb} - ۱ \leqslant \frac{sc}{sa} - ۱$$

$$\frac{sc}{sb} \leqslant \frac{sc}{sa}$$

همانطور که مشخص است در نهایت عبارت موجود در سوال اثبات می‌شود.

---

۳ سوال سوم

الف) استخراج الگوهای متداول در sequential pattern mining به معنای کشف دنباله‌هایی است که به طور قابل ملاحظه در داده‌ها ظاهر می‌شوند. این الگوها نشان‌دهنده رخدادهای مکرر و الگوهایی هستند که در داده‌ها به صورت پنهانی وجود دارند. در sequential pattern mining زمان ثبت یک ترنزاکشن ثبت نمی‌شود. این مورد در شرایطی که کشف الگوهای متوالی حیاتی است به کار می‌رود. از طرفی Association rule هدف اصلی شناسایی روابط یا ارتباطات جالب بین موارد مختلف در مجموعه داده است، و اغلب در داده‌های تراکنشی می‌بینیم. همچنین این مورد در درک ارتباطات بین موارد مهم موثر می‌باشد.

ب) الگوریتم PrefixSpan از نوع الگوریتم‌های DFS است. این الگوریتم به صورت بازگشتی

در گراف جستجوی دنباله حرکت می‌کند. به عبارت دیگر، ابتدا به عمق دنباله می‌رود و سپس به طور بازگشتی بازگشته و به حالت‌های دیگری از گراف حرکت می‌کند. این رویکرد برای کاهش زمان و حافظه مصرفی مفید است زیرا فقط دنباله‌هایی که قابلیت گسترش دارند را بررسی می‌کند و دنباله‌های غیرقابل گسترش را از لیست حذف می‌کند، بدون این که تمام فضای جستجو را بررسی کند. این ویژگی‌ها به الگوریتم PrefixSpan کمک می‌کند تا به صورت کارآمد الگوهای دنباله‌ای متداول را در مجموعه داده‌های بزرگی کشف کند. در قسمت ج یک مثال از این الگوریتم حل می‌شود.

ج) برای حل این قسمت با توجه به الگوریتم ذکر شده و با توجه به مقدار min-support مسئله حل می‌شود.

ID	Sequence
۱S	a ، b ، c ، b ، b ، c ، d
۲S	d ، c ، b ، a ، b ، c
۳S	c ، b ، b ، c ، d

برای هر مرحله min-support چک می‌شود. نتایج به شرح زیر است.

<a>	۲
<b>	۳
<c>	۳
<d>	۳

b	c	(d
cbbcd	bbcd	(cbabc
abc	babc	)
bcd	bbcd	)

در ادامه ترکیبات مختلفی از هر جدول در نظر گرفته می‌شود تا جایی که min-suppoert ارضا نشود

---

۴ سوال چهارم

برای حل این سوال از روابط داده شده در صورت سوال استفاده می شود.

$$\chi^{۲} = \frac{(۴۰۰-۲۴۰)^{۲}}{۲۴۰} + \frac{(۲۰۰-۳۶۰)^{۲}}{۳۶۰} + \frac{(۴۰۰-۵۶۰)^{۲}}{۵۶۰} + \frac{(۱۰۰۰-۸۴۰)^{۲}}{۸۴۰} = ۲۵۳/۸$$

با توجه به عدد حاصل، همبستگی بالای بین این دو مولفه شناسایی می‌شود.

$$lift(carrot, onion) = \frac{\frac{۴۰۰}{۲۰۰۰}}{\frac{۶۰۰}{۲۰۰۰}\frac{۸۰۰}{۲۰۰۰}} = ۱/۶۶۷$$

با توجه به آن که مقدار lift بیشتر از یک به دست آمد بنابراین این دو با یک دیگر همبستگی مثبت دارن و این بدان معناست که با خرید هویج احتمال خرید پیاز زیاد است.

$$Kulczynski(carrot, onion) = \frac{۴۰۰}{۲}(\frac{۱}{۸۰۰} + \frac{۱}{۶۰۰}) = ۰/۵۸$$

عدد به دست آمدع نشان دهنده‌ی وابستگی این دو به یکدیگر می‌باشد.

$$IR(carrot, onion) = \frac{|۸۰۰-۶۰۰|}{۸۰۰+۶۰۰-۴۰۰} = ۰/۲$$

با توجه به نزدیکی این عدد به صفر می‌توان بالانس بودن این دو مولفه را نتیجه گرفت.

---

۵ سوال پنجم

الف) anti-monotone است. زیرا اگر یک عضو در مجموعه‌ی S محدودیتی را نقض کند، هر عضو جایگزین 's آن نیز محدودیت را نقض خواهد کرد. به عبارت دیگر، هرگاه یک مورد از مجموعه S قانونی را نقض کند، هر عضو جایگزین آن نیز قانونی را نقض می‌کند. علاوه بر این، اگر مجموع قیمت‌های مورد s برابر یا بیشتر از یک حداقل تعیین شده باشد، مجموع قیمت‌های مورد 's هم برابر یا بیشتر از این حد خواهد بود. در نتیجه، اگر تراکم قیمت‌های s بالاتر از حداقل تعیین شده باشد، تراکم قیمت‌های 's نیز بالاتر از آن خواهد بود.

ب) monotone است. به عبارت دیگر، اگر مجموعه s این ویژگی را از خود نشان دهد، یعنی اگر ترتیب مقادیر درون آن افزایشی یا کاهشی باشد، در این صورت برای هر یک از 's ها نیز این خاصیت حفظ خواهد شد. به این معنا که مجموع قیمت‌های 's همواره برابر یا بیشتر از حداقل مقدار تعیین شده خواهد بود.

ج) monotone است.اگر حداقل یک شیر در مجموعه s وجود داشته باشد، این نشان می‌دهد

که در هر مجموعه ’s نیز حداقل یک شیر وجود دارد و شرایط محدودیت همچنان برقرار است. به عبارت دیگر، اگر در یک مجموعه شامل حداقل یک شیر باشد، این ویژگی نسبت به هر مجموعه مشابه دیگری هم حفظ می‌شود و محدودیت ها همچنان رعایت می‌شود.

د) succinct است. اگر تنها داده‌هایی که محدوده‌ی قیمت آن‌ها بین مقادیر u و v است را در نظر بگیریم، این محدودیت اعمال می‌شود.