تمرین‌های درس داده‌کاوی

مصطفی سبیلو (۴۰۱۰۸۲۵۴)

مهدی محمدی (۴۰۱۱۲۰۹۴)

فصل ۴

تمرین ۱

الف)

مرحله ۱:

تعداد تکرار هر آیتم در تراکنش‌ها را نسبت به تعداد کل تراکنش‌ها محاسبه می‌کنیم.

ردیف	آیتم	پشتیبان (Support)	درصد	
۱	A	$\frac{۱}{۵}$	۲۰	
۲	C	$\frac{۲}{۵}$	۴۰	
۳	D	$\frac{۱}{۵}$	۲۰	
۴	E	$\frac{۴}{۵}$	۸۰	✓
۵	I	$\frac{۱}{۵}$	۲۰	
۶	K	$\frac{۵}{۵}$	۱۰۰	✓
۷	M	$\frac{۳}{۵}$	۶۰	✓
۸	N	$\frac{۲}{۵}$	۴۰	
۹	O	$\frac{۳}{۵}$	۶۰	✓
۱۰	U	$\frac{۱}{۵}$	۲۰	
۱۱	Y	$\frac{۳}{۵}$	۶۰	✓

مرحله ۲:

از مرحله ۱ اقلامی که پشتیبان ۶۰ درصد و بیشتر از ۶۰ درصد دارند را انتخاب می‌کنیم. مجموعه اقلام ۲ تایی از آن‌ها می‌سازیم و مجدد پشتیبان را برای این مجموعه جدید محاسبه می‌کنیم.

ردیف	آیتم	پشتیبان (Support)	درصد	
۱	E, k	۴/۵	۸۰	✓
۲	E, M	۲/۵	۴۰	
۳	E, O	۳/۵	۶۰	✓
۴	E, Y	۲/۵	۴۰	
۵	K, M	۳/۵	۶۰	✓
۶	K, O	۳/۵	۶۰	✓
۷	K, Y	۳/۵	۶۰	✓
۸	M, O	۱/۵	۲۰	
۹	M, Y	۲/۵	۴۰	
۱۰	O, Y	۲/۵	۴۰	

مرحله ۳:

با اقلام ۲تایی از مرحله ۲ که پشتیبان برابر یا بیشتر از ۶۰ درصد دارند و اقلامی که از مرحله ۱ قبلا انتخاب شده‌اند، مجموعه اقلام ۳تایی می‌سازیم و پشتیبان را برای این مجموعه اقلام ۳تایی محاسبه می‌کنیم.

ردیف	آیتم	پشتیبان (Support)	درصد	
۱	E, K, M	۲/۵	۴۰	
۲	E, K, O	۳/۵	۶۰	✓
۳	E, K, Y	۲/۵	۴۰	
۴	E, M, O	۱/۵	۲۰	
۵	E, O, Y	۲/۵	۴۰	
۶	K, M, O	۱/۵	۲۰	
۷	K, M, Y	۲/۵	۴۰	
۸	K, O, Y	۲/۵	۴۰	

مرحله ۴:

با اقلام ۳تایی از مرحله ۳ که پشتیبان برابر یا بیشتر از ۶۰ درصد دارند و اقلامی که از مرحله ۱ قبلا انتخاب شده‌اند، مجموعه اقلام ۴تایی می‌سازیم و پشتیبان را برای این مجموعه اقلام ۴تایی محاسبه می‌کنیم.

ردیف	آیتم	پشتیبان (Support)	درصد	
۱	E, K, M, O	$\frac{۱}{۵}$	۲۰	
۲	E, K, O, Y	$\frac{۲}{۵}$	۴۰	

در این مرحله چون هیچ مجموعه اقلام ۴تایی پشتیبان برابر یا بیشتر از ۶۰ درصد ندارد، بنابراین الگوریتم متوقف می‌شود.

نتیجه نهایی:

{{E}, {K}, {M}, {O}, {Y}, {E, K}, {E, O}, {K, M}, {K, O}, {K, Y}, {E, K, O}}

ب)

مرحله ۱:

برای هر آیتم، پشتیبان (support) را محاسبه می‌کنیم:

ردیف	آیتم	پشتیبان (support)	درصد
۱	M	$\frac{۳}{۵}$	۶۰
۲	O	$\frac{۳}{۵}$	۶۰
۳	N	$\frac{۲}{۵}$	۴۰
۴	K	$\frac{۵}{۵}$	۱۰۰
۵	E	$\frac{۴}{۵}$	۸۰
۶	Y	$\frac{۳}{۵}$	۶۰
۷	D	$\frac{۱}{۵}$	۲۰
۸	A	$\frac{۱}{۵}$	۲۰
۹	U	$\frac{۱}{۵}$	۲۰
۱۰	C	$\frac{۲}{۵}$	۴۰
۱۱	I	$\frac{۱}{۵}$	۲۰

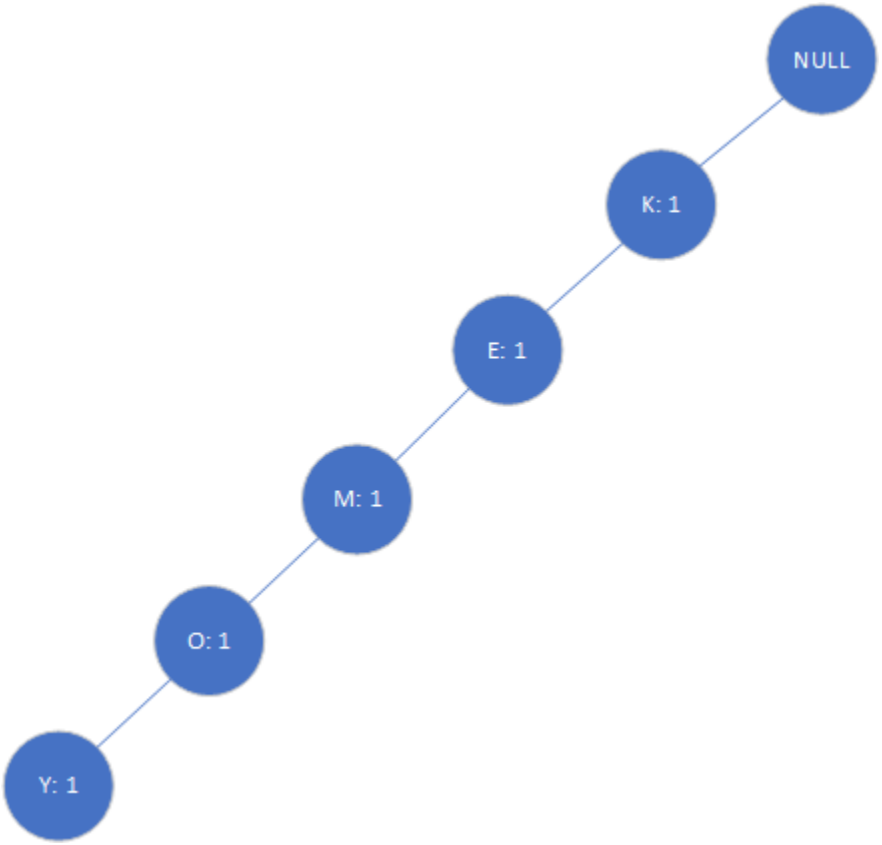
مرحله ۲:

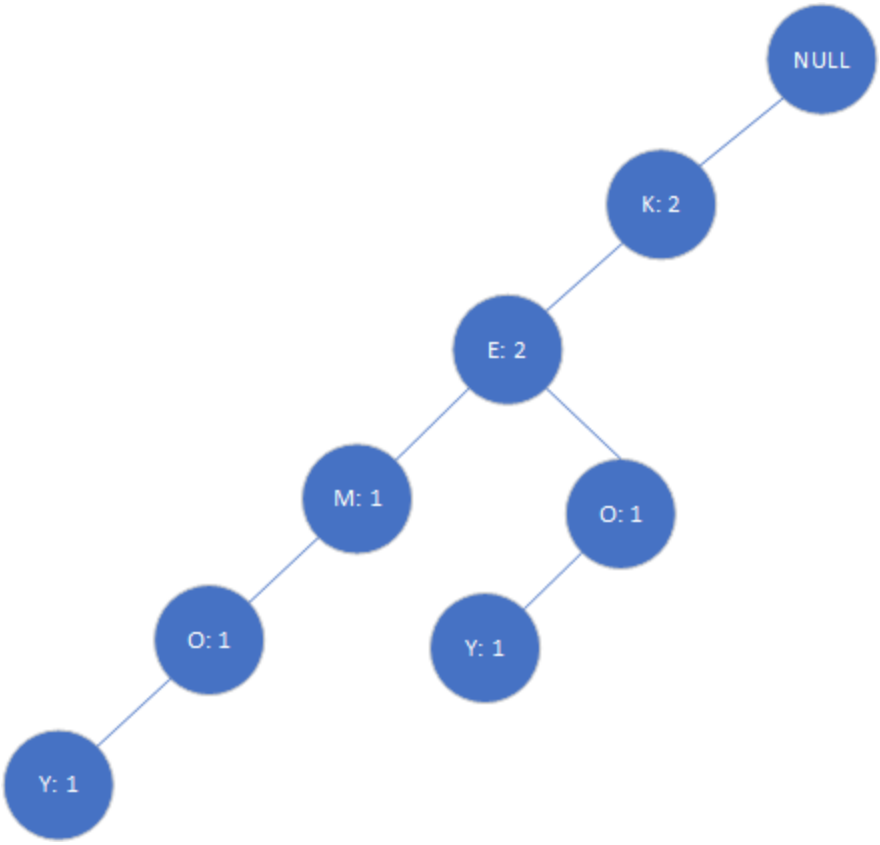
در هر تراکنش آیتم‌هایی که مقدار پشتیبان زیر ۶۰ درصد دارند را حذف می‌کنیم و آیتم‌های باقی مانده را به ترتیب مقدار پشتیبان محاسبه شده در مرحله قبل به صورت نزولی مرتب می‌کنیم:

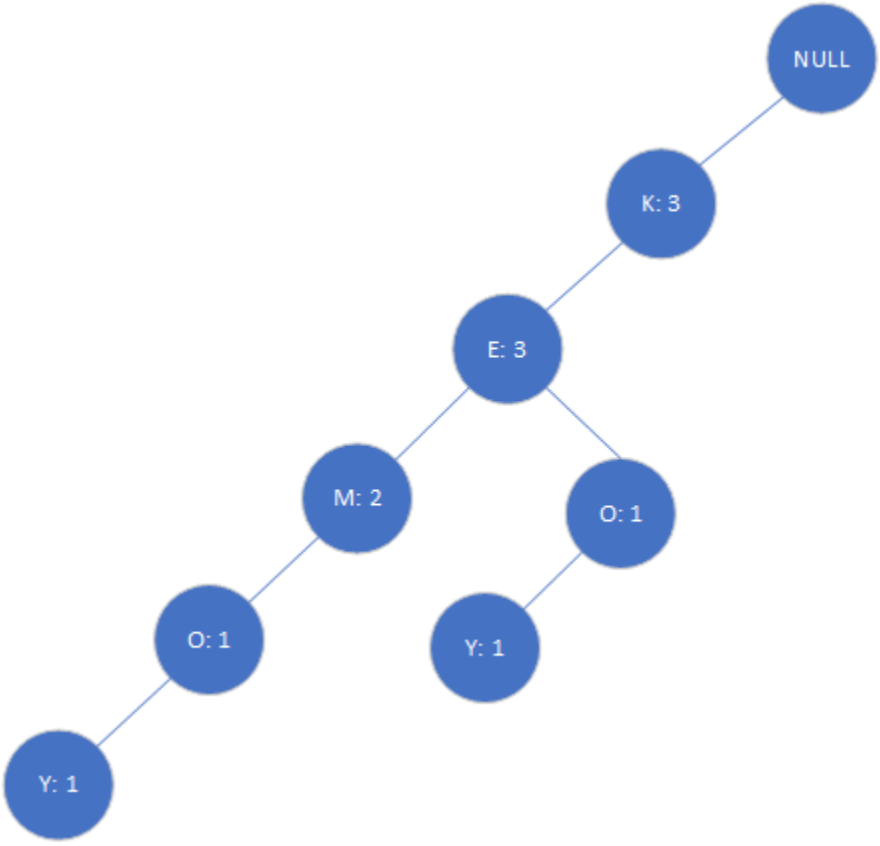
#	TID	Items_bought
1	T100	{K, E, M, O, Y}
2	T200	{K, E, O, Y}
3	T300	{K, E, M}
4	T400	{K, M, Y}
5	T500	{K, E, O}

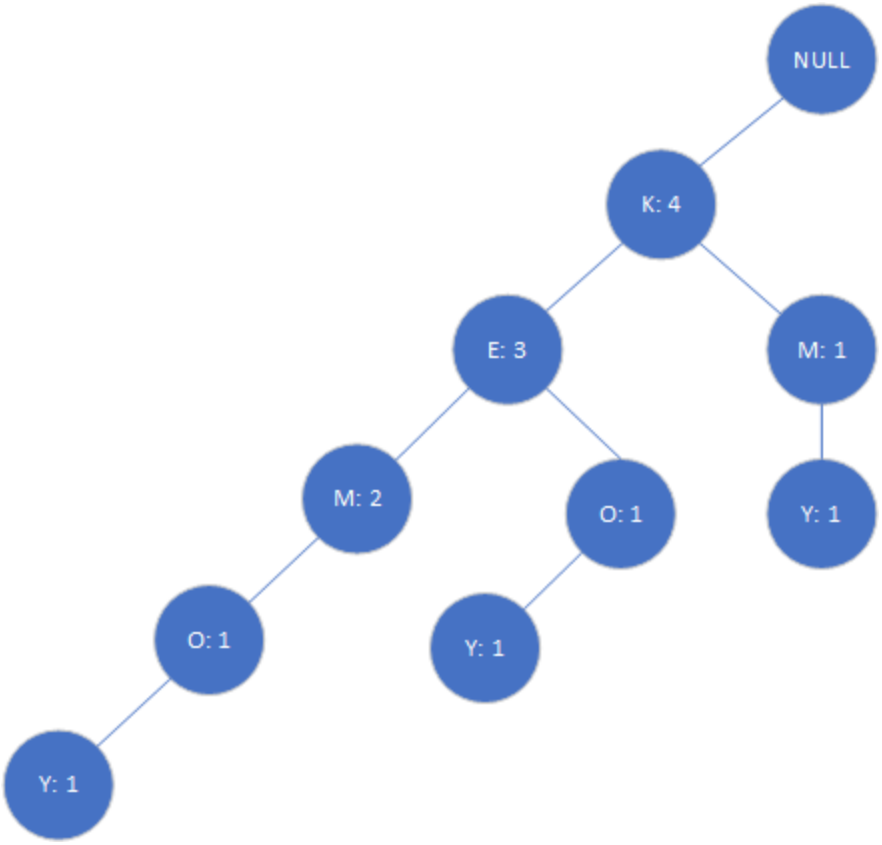
مرحله ۳:

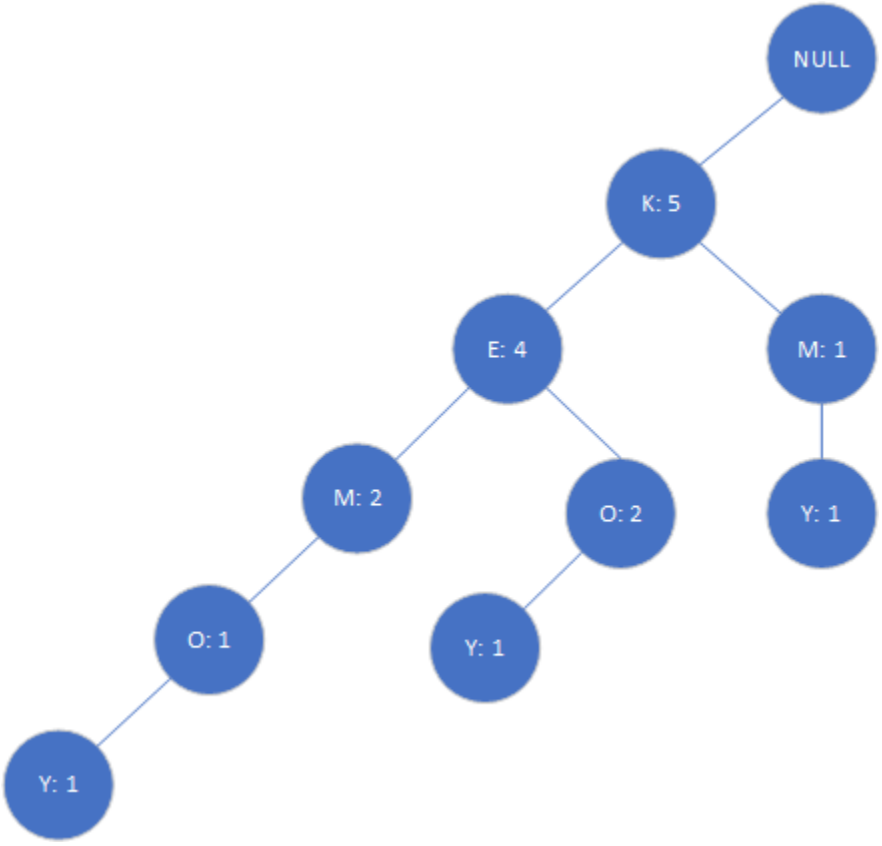
درخت FP-tree را ایجاد می‌کنیم؛ به این صورت که یک گره با مقدار null ایجاد می‌کنیم. سپس برای هر تراکنش یک شاخه به گره null اضافه می‌کنیم:




NULL
K: 2
E: 2
M: 1
O: 1
O: 1
Y: 1
Y: 1


NULL
K: 3
E: 3
M: 2
O: 1
O: 1
Y: 1
Y: 1


NULL
K: 4
E: 3
M: 1
M: 2
O: 1
Y: 1
O: 1
Y: 1
Y: 1


NULL
K: 5
E: 4
M: 1
M: 2
O: 2
Y: 1
O: 1
Y: 1
Y: 1

مرحله ۴:

پایگاه الگوهای شرطی را ایجاد می‌کنیم.

برای هر آیتم Conditional Pattern Base را ایجاد می‌کنیم:

#	Item	Conditional Pattern Base (CPB)
1	Y	{K, E, M, O: 1}, {K, E, O: 1}, {K, M: 1}
2	O	{K, E, M: 1}, {K, E: 2}
3	M	{K, E: 2}, {K: 1}
4	E	{K: 4}

در پایگاه داده جدید ساخته شده مجدد برای هر هر تراکنش، بازای هر آیتم در ستون CPB، پشتیبان را محاسبه کرده و آیتم‌هایی که پشتیبان کمتر از ۶۰ درصد دارند را حذف می‌کنیم:

#	Item	Conditional Pattern Base (CPB)	Conditional FP-tree
1	Y	{K, E, M, O: 1}, {K, E, O: 1}, {K, M: 1}	K: 3
2	O	{K, E, M: 1}, {K, E: 2}	K: 3, E: 3
3	M	{K, E: 2}, {K: 1}	K: 3
4	E	{K: 4}	K: 4

در آخر ترکیب اقلام ستون Conditional FP-tree را با ستون Item تولید می‌کنیم:

#	Item	CPB	Conditional FP-tree	Frequent Pattern Generated
1	Y	{K, E, M, O: 1}, {K, E, O: 1}, {K, M: 1}	K: 3	{K, Y}
2	O	{K, E, M: 1}, {K, E: 2}	K: 3, E: 3	{K, Y}, {E, Y}, {K, E, Y}
3	M	{K, E: 2}, {K: 1}	K: 3	{K, M}
4	E	{K: 4}	K: 4	{K, E}

جواب نهایی شامل اقلام محاسبه شده در مرحله اول و Frequent Pattern Generated هستند:

{{K}, {E}, {M}, {O}, {Y}, {E, K}, {E, O}, {K, M}, {K, O}, {K, Y}, {E, K, O}}

پ)

رابطه:

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

باید سه آیتمی را انتخاب کنیم که:

$$item_1, item_2 \rightarrow item_3$$

برای اینکه در رابطه بالا حداقل مقدار پشتیبان (minsup) را داشته باشیم، باید زیرمجموعه‌هایی ۳تایی از مجموعه ۱۱ آیتم را بیابیم که حداقل مقدار پشتیبان را داشته باشند. تنها مجموعه {E, K, O} حداقل مقدار پشتیبان (۶۰ درصد) را دارد. بنابراین مقدار minconf زیرمجموعه‌های دو عضوی از مجموعه {E, K, O} را محاسبه می‌کنیم:

$$conf(E, K \implies O) = \frac{3}{4} = 75\%$$

$$conf(K, O \implies E) = \frac{3}{3} = 100\%$$

$$conf(E, O \implies K) = \frac{3}{3} = 100\%$$

تمرین ۲

الف)

ابتدا تراکنش‌ها را براساس کالا بازنویسی می‌کنیم:

TID	Items_bought
T100	{Crab, Milk, Cheese, Bread}
T200	{Cheese, Milk, Apple, Pie, Bread}
T300	{Apple, Milk, Bread, Pie}
T400	{Bread, Milk, Cheese}

سپس الگوهای مکرر را براساس حداقل مقدار پشتیبان (minsup) یعنی ۶۰ درصد پیدا می‌کنیم (برای مثال از روش Apriori):

{{Milk: 100 %}, {Cheese: 75 %}, {Bread: 100 %}, {Milk, Cheese: 75 %}, {Milk, Bread: 100 %}, {Cheese, Bread: 75 %}, {Milk, Cheese, Bread: 75 %}}

مشاهده می‌شود که مجموعه اقلام ۳تایی بزرگ‌ترین مجموعه اقلام است؛ بنابراین برای حالت‌های مختلف مقدار پشتیبان و مقدار اطمینان را براساس قانون کلی زیر محاسبه می‌کنیم:

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

#	Rule	Support	Confidence
1	Milk, Cheese -> Bread	75 %	100 %
2	Milk, Bread -> Cheese	75 %	75 %
3	Cheese, Bread -> Milk	75 %	100 %

بنابراین دو قانون زیر حداقل مقدار پشتیبان (minsup) ۶۰ درصد و حداقل مقدار اطمینان (minconf) ۸۰ درصد را دارند:

Milk, Cheese -> Bread

Cheese, Bread -> Milk

ب)

ابتدا مقدار پشتیبان را برای هر برند-کالا محاسبه می‌کنیم:

{Best-Bread: 25%}, {Best-Cheese: 25%}, {Dairyland-Cheese: 50%}, {Dairyland-Milk: 50%}, {Goldenfarm-Apple: 25%}, {King’s-Crab: 25%}, {Sunset-Milk: 50%}, {Tasty-Pie: 50%}, {Westcoast-Apple: 25%}, {Wonder-Bread: 75%}

مشاهده می‌شود که فقط Wonder-Bread بیشتر از حداقل مقدار پشتیبان است.