به نام خدا



تحلیل داده های حجیم

تمرین سری دوم

استاد: دکترغلامپور

دانشجو:سجاد هاشم بیکی

پاییز1401

سوال1)

الف:

is the probability that B does not appear in a basket.

Conviction compares the probability that A appears without B if they were dependent with the actual frequency of the appearance of X without B. Unlike confidence, conviction factors in both and and always has a value of 1 when the relevant items are completely unrelated. In contrast to lift, conviction is a directed measure because it also uses the information of the absence of the consequent. Hence, conviction is monotone in confidence and lift.

ب:

Lift is not sensitive to rule direction:

ج:

ضعف پارامتر confidence بدین صورت است که ممکن است یک rule مقدار confidence بالایی داشته باشد ولی اصلا interest نباشد.مثلا B را ایتم شیر در نظر بگیریم،احتمالا confidence بالایی به ازای A های مختلف داشته باشد چرا که،ایتم شیر پرفروش است(در اکثرسبدها است) و احتمالا مستقل از A خریداری میشود،که در این صورت confidence بالاست اما interest نیست.

اما در دیگر پارمترها مانند interest , lift , conviction این موضوع درنظرگرفته شده است.بطوریه احتمال B (ساپورت B تقسیم بر تعداد سبدها) به رابطه انها اضافه شده است.

سوال2)

ایتم های 1تا20 هرکدام حداقل در 5 سبد وجود دارند،اما ایتم های 21 تا 100 در کمتر از 5 سبد وجود دارند.پس صرفا ایتمست های شامل ایتم های 1تا20 را بررسی میکنیم.

-ک م م هر ایتمست را محاسبه میکنیم.سبدهایی که بر ک م م آن ایتم ست بخش پذیر هستند،آن ایتمست را دارند.

100 تقسیم بر ان ک م م،تعداد سبدهایی که ان ایتمست را دارند را نشان میدهد.

ایتم ست های frequent (100/lcm >=5) و n عضوی را بدست میاوریم.

از ایتم ست های 20 عضوی شروع میکنیم وسپس 19 تایی،18 تایی و....

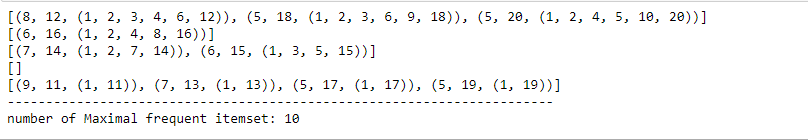
میدانیم که ایتمست های تک عضوی هیچ کدام ماکسیمال نیستند.

صرفا ایتم ست های 2 عضوی و بالاتر را بررسی میکنیم.

ایتم ست های 7 عضوی وبالاتر ساپورت کمتر از 5 دارند.مثلا برای ایتم ست 7تایی،مینیمم ک م م برابر 24 میباشد،که ایتم ست مربوط به ان ساپورت4 دارد.

بنابراین ایتم ست های 6و5و4و3و2 عضوی را بررسی و ماکسیمال ها را بدست میاوریم.

**ایتم ست های 6عضوی که ساپورت بزرگتریامساوی5 دارند ماکسیمال اند.ایتم ست های 5 عضوی که ساپورت بزرگتریامساوی5 دارند و زیر مجموعه ایتم ست های 6 عضوی ماکسیمال نیستند،ماکسیمال میباشند.با همین استدلال ایتم ست های ماکسیمال 4و3و2 عضوی را بدست میاوریم.**

در نهایت ایتم ست های ماکسیمال بدین صورت میباشد:

Maximal frequent itemset:

(1, 2, 3, 4, 6, 12), (1, 2, 3, 6, 9, 18), (1, 2, 4, 5, 10, 20)

(1, 2, 4, 8, 16)

(1, 2, 7, 14), (1, 3, 5, 15)

(1, 11), (1, 13), (1, 17), (1, 19)

سوال3)

روند کلی کد:

ابتدا موارد تکراری در هر ریکورد را حذف میکنیم و سپس هر ایتم را بصورت key value دراورده و سپس میشماریم.

ایتم های پرتکرار را با توجه به ترشولد داده شده بدست میاوریم.

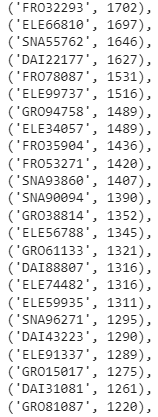
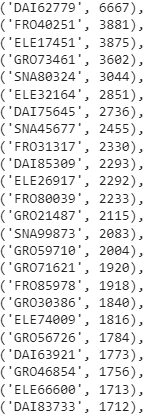
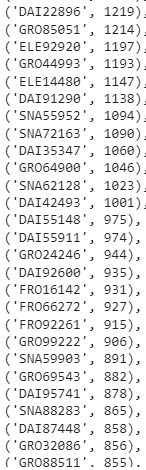
تمام زیرمجموعه های دوعضوی را از ایتم های پرتکرار بدست میاوریم و ساپورت هر کدام دا محاسبه میکنیم.

از میان ست های دو عضوی،پرتکرار ها را جدا میکنیم.

تمام زیرمجموعه های سه عضوی را از ست های دوعضوی پرتکرار بدست میاوریم و از میان انها ست های سه عضوی پرتکرار را جدا میکنیم.

ایتم های پرتکرار:

647تا ایتم پرتکرار داریم که چند مورد اول بدین صورت میباشد.(ایتم ها براساس ساپورت به صورت نزولی مرتب شده اند)



ایتم ست های دوتایی پرتکرار:

تعداد ایتم ست های دوتایی که از ایتم های پرتکرار بدست امده برابر208981 میباشد که در واقع همان candidate pairs ها هستند،که از این تعداد تنها 1334تا پرتکرار میباشد.تعداد ایتم های منحصر بفرد که جفتهای پرتکرار را ساخته اند برابر 222 میباشد.

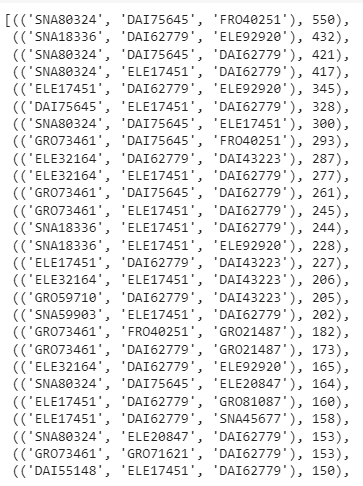
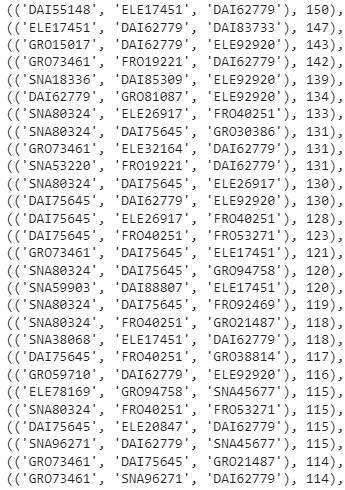
ایتم ست های دوتایی پرتکرار در فایل freq\_pairs.txt ذخیره شده است.(برای دفعات بعد فایل را میخوانیم چون هر بار ران کردن حدود2 ساعت طول میکشد به همین خاطر این قسمت از کد یکبار ران شده و سپس کامنت شده است)هر ایتم ست به همراه ساپورت ان نوشته شده است.بدین شکل:

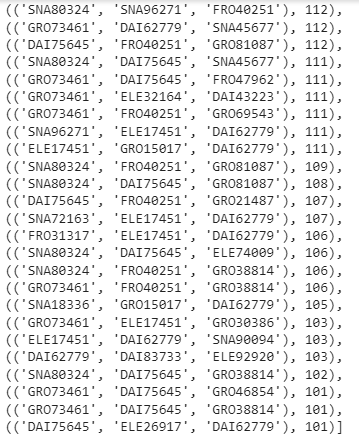
ایتم ست های سه تایی پرتکرار:

تعداد ایتم ست های سه تایی که از ایتم ست های دوتایی پرتکرار بدست امده برابر**1798940** میباشد که در واقع همان candidate triples ها هستند.

قبل از بدست اوردن سه تایی های پرتکرار، کاندیدها را فیلتر میکنیم،هر کاندید،سه عضو دارد،بنابراین سه زیر مجموعه دو عضوی دارد.هر کاندید در صورتی **میتواند** پرتکرار باشد، که هر سه زیرمجموعه دو عضوی ان پرتکرار باشد(دوتایی های پرتکرار در قسمت قبل بدست امد). پس از اعمال این شرط،تعداد کاندیدها به **1900** عدد میرسد که به طرز چشمگیری هزینه محاسبات کاهش میابد.

که از این تعداد تنها **79** تا پرتکرار میباشد.

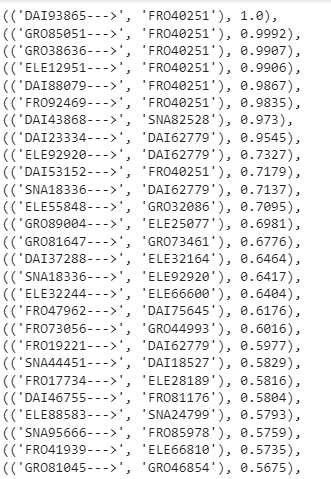
هر ایتم ست به همراه ساپورت ان نوشته شده است.(بصورت نزولی مرتب شده اند) بدین شکل:

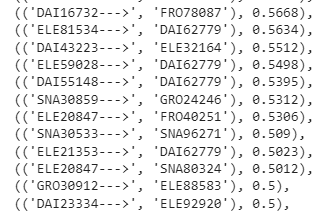


محاسبه confidence score :

پس از بدست اوردن ایتم ست های پرتکرار،rule های متناظر به هر ایتم ست را بدست اورده و کانفیدنس آن را حساب میکنیم.

ایتم ست های دوتایی پرتکرار:

برای هر ایتم ست دو rule میتوان نوشت مثلا برای (a,b)،a🡪b , b🡪a را میتوان نوشت.باتوجه به اینکه 1334 ایتم ست دوتایی پرتکرار داریم،تعداد rule ها برابر 2668 خواهد بود که کانفیدنس متناظر با هر کدام محاسبه شده است.در ادامه کانفیدنس های بالاتر از 0.5 ذکر شده است.(مقادیر بصورت نزولی مرتب شده اند)



ایتم ست های سه تایی پرتکرار:

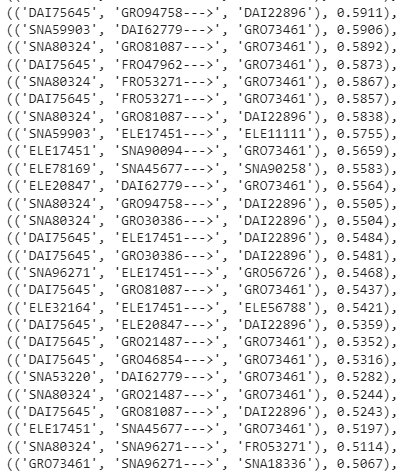
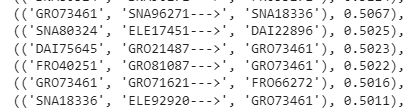
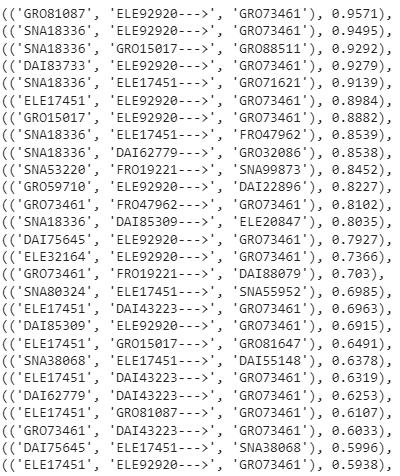
برای هر ایتم ست سه تایی،سه rule بصورت 2 به 1 میتوان نوشت.یعنی:

(a,b,c):

(a,b)🡪c

(a,c)🡪b

(b,c)🡪a

با توجه به اینکه 79 تا ایتم ست سه تایی پرتکرار داریم،تعداد ruleها برابر 237 خواهد بود.(79\*3=237).برای هر کدام از ruleهای گفته شده مقدار کانفیدنس محاسبه شده است.در ادامه مقادیر بالای 0.5 ذکر شده است. (مقادیر بصورت نزولی مرتب شده اند)

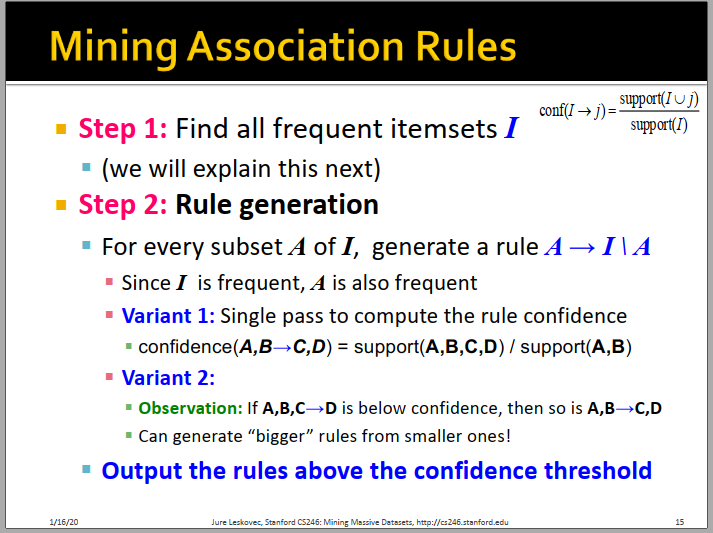
**بررسی یک مورد:**

**(('GRO81087', 'ELE92920--->', 'GRO73461'), 0.9571)**

برای مثال اگر یک مشتری دوکالای (**GRO81087**) و (**ELE92920**) را به سبد خود اضافه کرده(یا مثلا بازدید کرده)احتمالا کالای (**GRO73461**) برای بازدید ویا اضافه کردن به سبد خرید مناسب باشد،چرا که مقدار کانفیدنس این rule برابر **0.9571** این است که مقدار بالایی محسوب میشود، و هم چنین این rule از کالاهایی ساخته شده است که هر کدام پرتکرار هستند.

به همین صورت میتوان کالاها را به مشتری پیشنهاد داد و احتمالا یک ریکامندر سیستم با این منطق ساخت.

\*با توجه به ابهامی که در صورت سوال برای پیدا کردن rule ها و محاسبه confidence score وجود داشت،من طبق روشی که در اسلاید درس گفته شده یود عمل کردم.بدین صورت که پس از پیدا کردن frequent itemset ،rule های ممکن از انرا استخراج کرده و سپس مقدار confidence score را برای هر یک محاسبه کردم.\*در واقع هر ایتم ست را دو بخش کرده و rule را برای ان مینویسیم. یک بخش در سمت چپ و بخش دیگر سمت راست قرار میگیرد.



سوال 4: