به نام خدا

دانشگاه صنعتی امیرکبیر

(پلی تکنیک تهران)

دانشکده مهندسی کامپیوتر

گزارش تمرین اول درس کلان داده‌ها

استاد درس: دکتر مصطفی حقیر چهرقانی

دانشجو: فاطمه غلام‌زاده

۹۹۱۳۱۰۰۳

نیم سال دوم ۱۳۹۹-۱۴۰۰

درس کلان داده ها ـ تمرین اول ـ فاطمه غلام زاده ـ ۹۶۱۲۶۰۰۳

سوال ۱ الف ـ ۱) چه تعداد association rule متفاوت (حتی با support صفر) می توان از این سبدهای خرید استخراج کرد؟

پاسخ:

تعداد آیتم های متفاوت ۶ عدد است. تعداد افرازهای یک مجموعه ۶ عضوی را می یابیم و ضربدر ۲ می کنیم تا تعداد rule های متفاوت به دست آید. برای راحتی در محاسبات آیتم ها را متغیر با اعداد ۱ تا ۶ نشان می دهیم:

حالت اول: قانون هایی به شکل $۲\to ۱$ یا $۵\to ۳$:

$$\binom{۶}{۲}\times ۲ = ۱۵\times ۲ = \boxed{۳۰}$$

حالت دوم: قانون هایی به شکل $۱\to ۲,۳$

$$\binom{۶}{۳}\times\binom{۳}{۱}\times ۲ = ۲۰\times ۳\times ۲ = \boxed{۱۲۰}$$

حالت سوم: قانون هایی به شکل $۱\to ۲,۳,۴$

$$\binom{۶}{۴}\binom{۴}{۱}\times ۲ = ۱۵\times ۴\times ۲ = \boxed{۱۲۰}$$

حالت چهارم: قانون هایی به شکل $۱,۲\to ۳,۴$

$$\binom{۶}{۴}\binom{۴}{۲} = ۱۵\times ۶ = \boxed{۹۰}$$

حالت پنجم: قانون هایی به شکل $۱\to ۲,۳,۴,۵$

$$\binom{۶}{۵}\binom{۵}{۱}\times ۲ = ۶\times ۵\times ۲ = \boxed{۶۰}$$

حالت ششم: $۱,۲\to ۳,۴,۵$

$$\binom{۶}{۵}\binom{۵}{۲}\times ۲ = ۶\times ۱۰\times ۲ = \boxed{۱۲۰}$$

حالت هفتم: $۱\to ۲,۳,۴,۵,۶$

$$\binom{۶}{۶}\binom{۶}{۱}\times ۲ = ۱\times ۶\times ۲ = \boxed{۱۲}$$

• حالت هشتم: $1,2 \rightarrow 3,4,5,6$

$$\binom{6}{4}\binom{6}{2} \times 2 = 15 \times 2 = \boxed{30}$$

حالت نهم: $1,2,3 \rightarrow 4,5,6$

$$\binom{6}{4}\binom{6}{3} = 1 \times 20 = \boxed{20}$$

در حالت هایی که تعداد اعضای آیتم ست ها در دو طرف قانون یکسان است ضریب ۲ نداریم چون حالت تکراری ایجاد می شود.

مجموع ۹ حالت:

$$30 + 120 + 120 + 90 + 60 + 120 + 12 + 30 + 20$$

$$= \boxed{602}$$ ← تعداد association rule های متفاوت

الف ـ ۲) بیشترین اندازه‌ی آیتم ستی که می توانیم استخراج کنیم؟

با توجه به جدول داده شده، آیتم ست های شماره ۶ و ۹ دارای بیشترین اندازه هستند که ۴ است ← پس بیشترین اندازه‌ی Itemset = $\boxed{4}$

الف ـ ۳) عبارتی برای بیشینه تعداد Itemset های با اندازه ۳ که می توانیم استخراج کنیم؟

انتخاب ۳ عضو از ۶ آیتم متفاوت ← $\binom{6}{3} = 20$

الف-۴) شبیه الگوریتم apriori عمل می‌کنیم. ابتدا ساپورت Itemset ها

تک عضوی را می‌یابیم:

sup(شیر) = ۶

sup(کره) = ۵

sup(نوشابه) = ۴

sup(بیسکویت) = ۴

sup(دستمال) = ۷

sup(نان) = ۴

حال به سراغ Itemset های ۲ عضوی می‌رویم، سعی می‌کنیم از آن‌هایی شروع کنیم که ترکیب Itemset های تک عضوی با ساپورت بالا هستند:

sup(شیر، دستمال) = ۵

با توجه به اینکه ساپورت {شیر، دستمال}

عدد ۵ به دست آمده می‌توانیم آیتم ست‌های ۲ تایی که یکی از اعضای آن ساپورت کمتر از ۵ دارند را به طور کلی حذف کنیم و شمارش نکنیم چون قطعا ساپورت آن‌ها از ۵ کمتر خواهد شد.

پس ترکیب‌های ۲ تایی که بیسکویت، نان و نوشابه را در خود دارند در نظر نمی‌گیریم.

sup(کره و شیر) = ۳

sup(دستمال، کره) = ۳

پس Itemset با اندازه‌ی ۲ که بیشترین ساپورت را دارد همان {شیر، دستمال} است.

آیتم ست‌های با اندازه‌ی بزرگتر از ۲ هم قطعا ساپورت کمتری از {شیر، دستمال} دارند چون

فقط یک کاندیدای ۳ عضوی داریم: {شیر، کره، دستمال} که ساپورت آن هم ۳ است.

پاسخ نهایی: {شیر، دستمال}

الف-۵)

$$Conf(A \to B) = \frac{sup(A \cup B)}{Sup(A)}$$

$$Conf(B \to A) = \frac{sup(B \cup A)}{sup(B)}$$

$$conf(A \to B) = Conf(B \to A) \Rightarrow \frac{Sup(A \cup B)}{Sup(A)} = \frac{Sup(A \cup B)}{sup(B)}$$

$$\Rightarrow Sup(A) = sup(B)$$

پس کافی است سایورت آیتم ست های A و B با هم برابر باشد. مثال:

بیسکویت ← نان

$$conf(\text{نان} \to \text{بیسکویت}) = \frac{sup(\{\text{نان، بیسکویت}\})}{sup(\text{نان})} = \frac{۱}{۴} \checkmark$$

$$conf(\text{بیسکویت} \to \text{نان}) = \frac{sup(\{\text{نان، بیسکویت}\})}{sup(\text{بیسکویت})} = \frac{۱}{۴} \checkmark$$

ب-۱) خیر معیار confidence متقارن نیست. مثال نقض:

D	۱
C	۲
A, B	۳
A	۴
A, B	۵

$$Conf(A \to B) = \frac{sup(A \cup B)}{sup(A)} = \frac{۲}{۳}$$

$$Conf(B \to A) = \frac{sup(A \cup B)}{sup(B)} = \frac{۲}{۲} = ۱$$

$$\frac{۲}{۳} \neq ۱$$

ب-۲) بله Lift متقارن است. تعداد کل تراکنش‌ها را N در نظر می‌گیریم.

اثبات:

$$Lift(A \to B) = \frac{Conf(A \to B)}{S(B)}$$

$$= \frac{\frac{sup(A \cup B)}{sup(A)}}{\frac{sup(B)}{N}} = \frac{sup(A \cup B) \times N}{sup(A) \times sup(B)}$$

$$= \frac{\frac{sup(A \cup B)}{sup(B)}}{\frac{sup(A)}{N}} = \frac{Conf(B \to A)}{S(A)} = Lift(B \to A)$$

ب-۳) Conviction ← متقارن نیست. مثال نقض:

D/B	۱
A,B	۲
A/B	۳
A	۴
B	۵

$$Conf(A \to B) = \frac{۲}{۳}$$

$$Conv(A \to B) = \frac{۱ - ۴/۵}{۱ - ۲/۳} = \frac{۳}{۵}$$

$$Conf(B \to A) = ۲/۴$$

$$conv(B \to A) = \frac{۱ - ۳/۵}{۱ - ۲/۴} = \frac{۴}{۵}$$

$$\frac{۳}{۵} \neq \frac{۴}{۵} \Rightarrow conv(A \to B) \neq conv(B \to A)$$

سوال۲) کاربرد Association rules

الف) ۵ قانون ابتدایی به صورت زیر است که در آن مجموعه ['A','B'] به معنای قانون A→ B می‌باشد و عدد نوشته شده در روبه‌روی هر قانون، میزان confidence برای آن قانون می‌باشد.

```
[[['DAI93865', 'FRO40251'], 1.0],
 [['GRO85051', 'FRO40251'], 0.999176276771005],
 [['GRO38636', 'FRO40251'], 0.990654205607476 6],
 [['ELE12951', 'FRO40251'], 0.9905660377358491],
 [['DAI88079', 'FRO40251'], 0.9867256637168141]]
```

ب) ۵ قانون ابتدایی به صورت زیر است که در آن مجموعه ['A','B','C'] به معنای قانون A, B →C می‌باشد و عدد نوشته شده در روبه‌روی هر قانون، میزان confidence برای آن قانون می‌باشد.

```
[[['DAI23334', 'ELE92920', 'DAI62779'], 1.0],
 [['DAI31081', 'GRO85051', 'FRO40251'], 1.0],
 [['DAI55911', 'GRO85051', 'FRO40251'], 1.0],
 [['DAI62779', 'DAI88079', 'FRO40251'], 1.0],
 [['DAI75645', 'GRO85051', 'FRO40251'], 1.0]]
```

سوال ۳) LSH

الف)

میانگین زمان جستجو در حالت خطی :

```
Average LSH Time: 0.997 Seconds
```

میانگین زمان جستجو با استفاده از LSH :

```
Average Linear Search Time: 14.204 Seconds
```

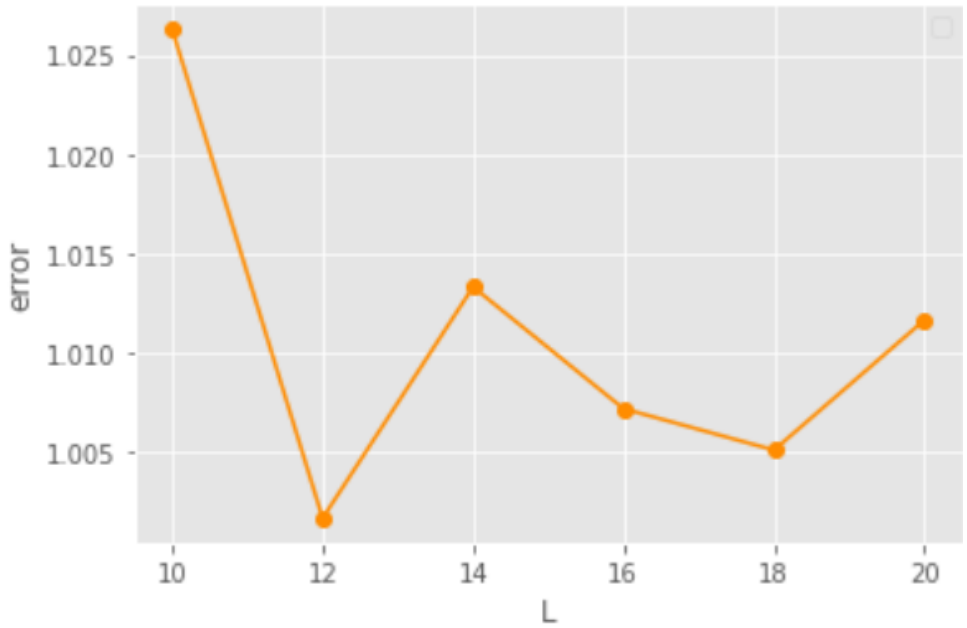
مقایسه : همانطور که مشاهده می شود میانگین زمان جستجو با استفاده از LSH نسبت به جستجوی خطی کاهش نسبتا زیادی داشته است.

ب) مقدار خطای به دست آمده:

```
Error: 1.023511
```

ج)

نمودار مقدار خطا به صورت تابعی از L :



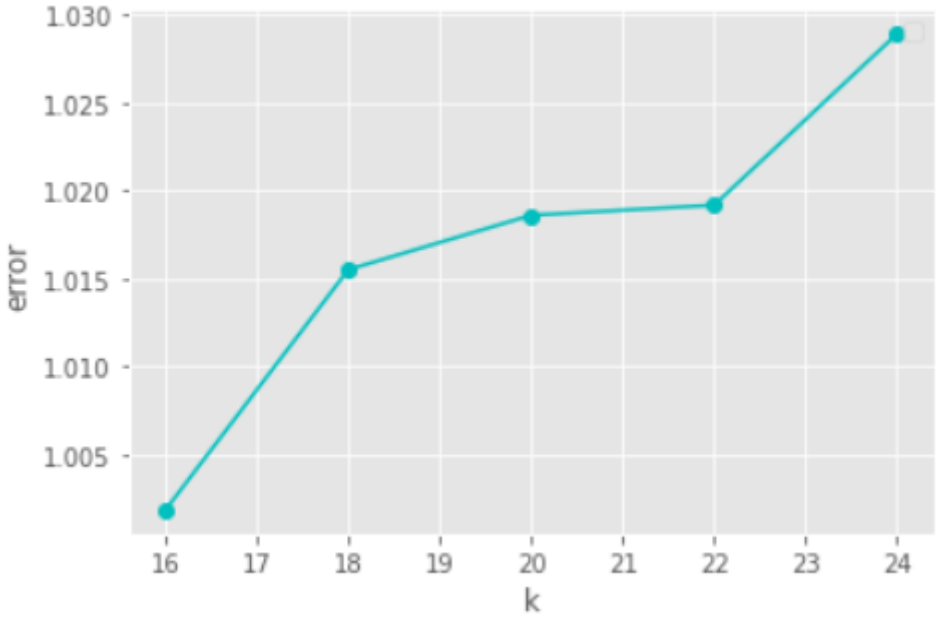
مقادیر خطا:

```
[1.0262867364053494, 1.0016595368482275,
1.0133086954891755, 1.0071734390323535,
1.0050886944625834, 1.0116438905081788]
```

تحلیل نمودار:

با افزایش مقادیر L مقدار خطا به طور کلی سیر نزولی دارد زیرا تعداد توابع hash افزایش می‌یابد.

نمودار مقدار خطا به صورت تابعی از k :



مقادیر خطا:

```
[1.0018693720851233, 1.015522122707 79,
1.018614931861614, 1.0191862558990539,
1.0289129370499819]
```

تحلیل نمودار:

با افزایش k خطا مقدار خطا افزایش یافته است چون در این حالت یافتن همسایه‌ها توسط LSH سخت‌تر می‌شود.

د)

تصویر مورد سوال : 

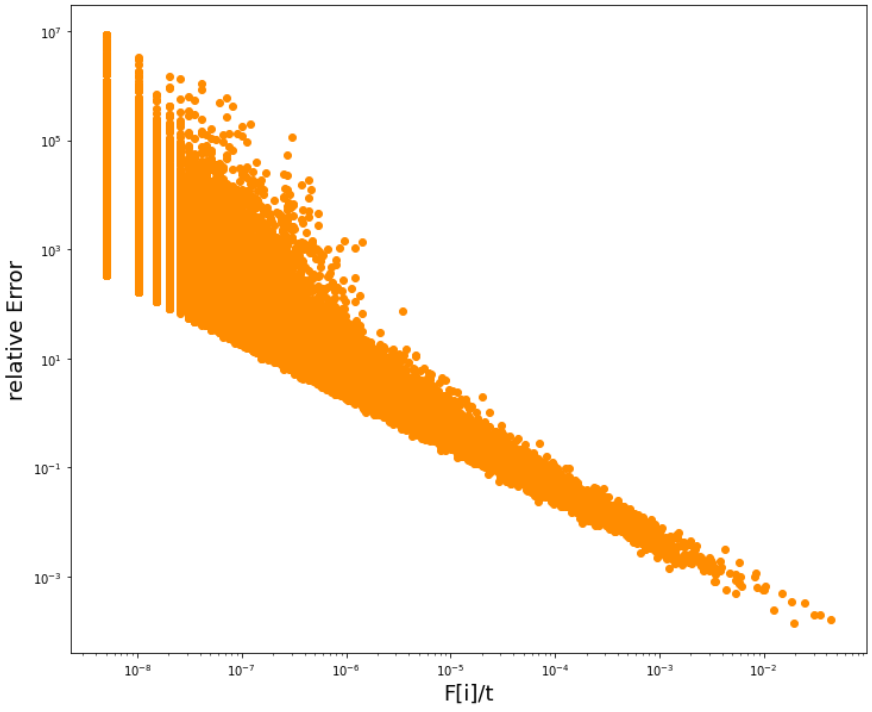
همسایه‌های به دست آمده با جستجوی خطی :

همسایه‌های به دست آمده با جستجوی LSH :

مقایسه : با توجه به جستجوی انجام شده و همسایه های یافت شده، با بررسی ایندکس همسایه ها مشخص گردید که در ۸ مورد LSH همسایه‌های یکسانی با linear پیدا کرده است و فقط در ۲ مورد تفاوت دارد. با توجه به تصاویر مشاهده می‌شود که آن ۲ مورد هم بسیار به هم شبیه هستند و همسایه نزدیکی برای تصویر مورد سوال توسط LSH گزارش شده است. بنابراین می‌توان نتیجه گرفت با توجه به نتایج خوبی که LSH دارد و زمان جستجوی بسیار کمتر آن نسبت به روش خطی، یک روش مناسب و به صرفه برای یافتن همسایه‌های نزدیک است.

سوال۴) جریان داده

نمودار خطا :



❖ **به نظر شما یک شرط تقریبی روی فرکانس واقعی عناصر که بتواند relative error را همواره کمتر از یک نگه دارد، چیست؟**

با توجه به رابطه‌ی relative error داریم:

$Er[i] = (F'[i] - F[i]) / F[i]$

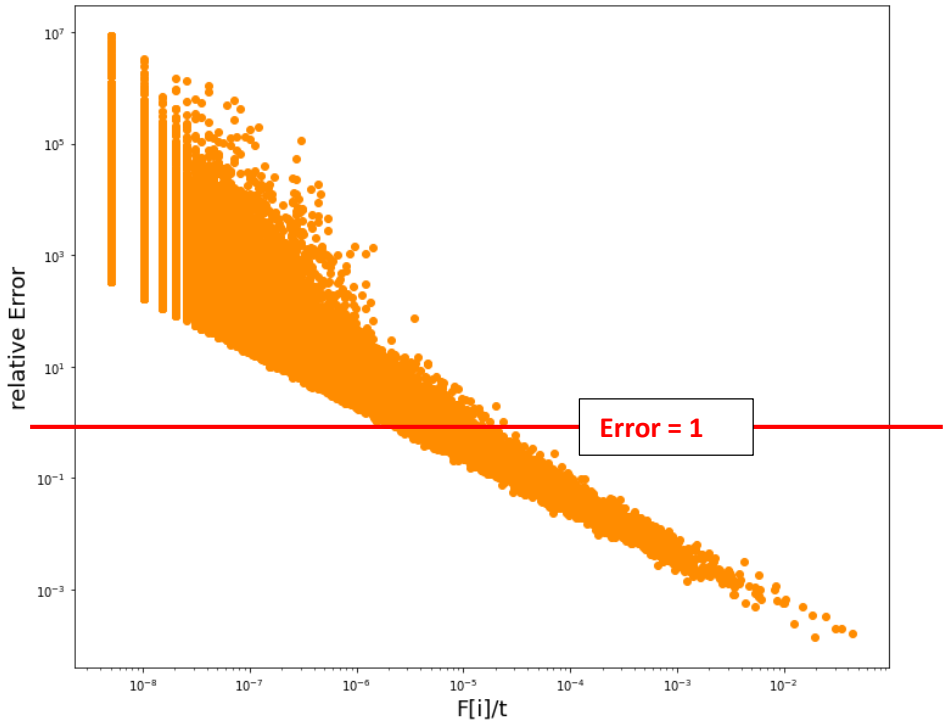
$Er[i] < 1 \Rightarrow (F'[i] - F[i]) / F[i] < 1$

$F'[i] - F[i] < F[i] \Rightarrow F'[i] < 2 * F[i]$

$F[i] > F'[i]/2$

بنابراین از روی رابطه به این شرط می‌رسیم که اگر فرکانس واقعی هر عنصر کوچکتر از نصف فرکانس تخمینی باشد خطا همواره کمتر از ۱ خواهد بود.

اگر بخواهیم از روی نمودار یک شرط پیدا کنیم خط error=1 را رسم میکنیم :



با توجه به خط رسم شده، اگر مقدار فرکانس واقعی هر عنصر حدودا بیشتر از 10^{-5} باشد مقدار خطا کمتر از ۱ خواهد شد.