دانشکده مهندسی برق

تمرین سری سوم آنالیز داده های حجیم

سال تحصیلی ۱۴۰۱-۱۴۰۰

سوال اول)

Confidence احتمال رخداد B در سبد است اگر سید قبلا شامل A باشد:

$$\mathrm{conf}(A \to B) = \Pr(B|A),$$

Lift به معنی احتمال رخداد A ,B با یکدیگر است، با این پیش فرض که A ,B از یکدیگر مستقل هستند. مقدار S(B) برابر ساپورت B تقسیم بر تعداد کل سبدهاست.

$$\mathrm{lift}(A \to B) = \frac{\mathrm{conf}(A \to B)}{S(B)},$$

الف) مقدار conviction به صورت زیر تعریف میشود.مفهوم آن را شرح دهید(چه رخدادی را توصیف میکند)

$$\mathrm{conv}(A \to B) = \frac{1 - S(B)}{1 - \mathrm{conf}(A \to B)}.$$

ب) کدام یک از سه تعریف بالا نسبت به A و B متقارن اند. تقارن یا عدم تقارن را اثبات کنید.

ج) ضعف تعریف Confidence نسبت به دو تعریف دیگر در چیست. در واقع چه المانی در این تعریف در نظر گرفته نشده است. با ذکر مثال شرح دهید.(راهنمایی: احتمال رخداد B را یکبار زیاد و یک بار کم در نظر بگیرید).

سوال دوم)

فرض کنید تعداد بسیار زیادی سبد و ۱۰ آیتم موجود است که با شمارهی یک تا ده شماره گذاری شده اند. هر سبد به احتمال 1/i شامل آیتم i ام است. مثلا هر سید به احتمال ۰.۵ آیتم شماره ۲ را شامل میشود. اگر ترشولد ساپورت برابر ۰.۱باشد مجموعه های پر تکرار(frequent itemset) را نام ببرید.

سوال سوم)

در این سوال قصد داریم الگوریتم های A-priori و SON را بر روی یک دیتا واقعی تست کنیم. در سایت cw لینک درایو مجازی با عنوان **اطلاعات پروژه و تمرین درس** موجود است. توضیحات و نمونه کد برای پردازش اطلاعات در کولب نیز در این درایو وجود دارد. این اطلاعات مربوط به تردد خودروهای سطح شهر تهران است. توجه کنید که برای حل سوال نمی توانید از الگوریتم های frequent item اسپارک استفاده کنید. همچنین می توانید از دیتاست Sample_Data.zip استفاده کنید. برای حل این سوال در اول گزارش فرض خود را برای ساپورت و یا دیگر پارامتر ها وارد کنید توجه داشته باشید که هدف شناسایی مسیر های پرتکرار است.

الف) مسیر عبوری هر خودرو را به تفکیک روز در یک rdd مشخص کنید. به عنوان مثال key= (Plate,Date) و value=[Device Code list]

ب) مسیر های پرتکرار را بر مبنای rdd قسمت الف و با استفاده از الگوریتم A-Priori بدست آورید.

ج) مسیر های پرتکرار را بر مبنای rdd قسمت الف و با استفاده از الگوریتم SON بدست آورید. (راهنمایی: rdd را به ۲ یا ۳ بخش تقسیم کنید)