تحلیل داده‌های حجیم

مدرس : دکتر ایمان غلامی

[پاییز ۱۴۰۰]

تمرن سری ۳: سوال ۲ نگارنده: امیرمحمد شعبانی

با استفاده از الگوریتم A-Priori پیش می‌رویم و ابتدا مجموعه‌های پرتکرار به طول یک را بدست می‌آوریم و سپس به طول دو و همین‌طور ادامه می‌دهیم.

$$\forall i \in \{1, \cdots, 10\} \quad \frac{1}{i} \geq \frac{1}{10} \Rightarrow \text{all set with length one is frequent itemset}$$

بنابراین با توجه معادله بالا تمام مجموعه‌های به طول یک، مجموعه پرتکرار هستند. حال مجموعه‌های پرتکرار به طول دو را بدست می‌آوریم:

$$\forall i, j \in \{1, \cdots, 10\} \quad \{i, j\} \quad \text{is frequent} \Rightarrow \frac{1}{i \times j} \geq \frac{1}{10} \Rightarrow i \times j \leq 10$$

$$\text{frequent itemset with length two} = \{\{\{1, i\} \quad | \quad i \in 2, \cdots, 10\}, \quad \{\{2, i\} \quad | \quad i \in 3, \cdots, 5\}\}$$

برای مجموعه های به صورت $\{\{j, i\} \quad | \quad i \in j+1, \cdots, k, \quad \max_k kj \leq 10, \quad k \geq j+1\}$ تنها $j \in \{1, 2\}$ تهی نیستند. بنابراین کل مجموعه‌های پرتکرار به طول دو برابر است با:

$$\{1,2\}, \{1,3\}, \{1,4\}, \{1,5\}, \{1,6\}, \{1,7\}, \{1,8\}, \{1,9\}, \{1,10\}, \{2,3\}, \{2,4\}, \{2,5\}$$

می‌دانیم طبق الگوریتم A-Priori مجموعه‌ای برای طول ۳ کاندید است که از اجتماع دو مجموعه پرتکرار به طول دو که باهم در یک عضو اختلاف دارند تشکیل شده‌باشد و تمام زیرمجموعه‌های به طول ۲ آن پرتکرار باشد. در ابتدا مجموعه‌هایی که تنها در یک عضو اختلاف دارند را بدست می‌آوریم:

$$\{\{(\{1, i\}, \{1, j\}) \quad | \quad i < j, \{1, i\}, \{1, j\} \in \text{frequent itemset length two}\}$$
$$\{\{(\{2, i\}, \{2, j\}) \quad | \quad i < j, \{2, i\}, \{2, j\} \in \text{frequent itemset length two}\}$$

تمام این مجموعه‌ها تنها در یک عضو اختلاف دارند. بنابراین تمام $\{1, i, j\}, \{2, i, j\}$ها کاندید مجموعه‌های پرتکرار به طول سه هستند. حال به شمارش هرکدام از این کاندیدها می‌پردازیم. همانند نتیجه‌ای که برای مجموعه‌های به طول دو بدست آوردیم، ضرب همه عضوهای مجموعه‌های پرتکرار به طول سه کمتر مساوی ۱۰ خواهد شد. بنابراین مجموعه‌های پرتکرار به طول سه عبارتند از:

$$\{1,2,3\}, \{1,2,4\}, \{1,2,5\}$$

با توجه به اینکه هیچ مجموعه چهارتایی از اعداد ۱ تا ۱۰ نداریم که ضربشان کمتر ۱۰ شود؛ بنابراین الگوریتم به پایان می‌رسد.