Homework 1

Association Rule

Association Rules are frequently used for Market Basket Analysis (MBA) by retailers to understand the purchase behavior of their customers. This information can be then used for many different purposes such as cross-selling and up-selling of products, sales promotions, loyalty programs, store design, discount plans and many others.

Evaluation of item sets: Once you have found the frequent itemsets of a dataset, you need to choose a subset of them as your recommendations. Commonly used metrics for measuring significance and interest for selecting rules for recommendations are:

1. **Confidence** (denoted as $\text{conf}(A \to B)$): *Confidence* is defined as the probability of occurrence of B in the basket if the basket already contains A:

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

   where $\Pr(B|A)$ is the conditional probability of finding item set B given that item set A is present.

2. **Lift** (denoted as $\text{lift}(A \to B)$): *Lift* measures how much more "A and B occur together" than "what would be expected if A and B were statistically independent":

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

   where $S(B) = \frac{\text{Support}(B)}{N}$ and N = total number of transactions (baskets).

3. **Conviction** (denoted as $\text{conv}(A \to B)$): *Conviction* compares the "probability that A appears without B if they were independent" with the "actual frequency of the appearance of A without B":

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

(a) [15pts]

A drawback of using *confidence* is that it ignores $\Pr(B)$. Why is this a drawback? Explain why *lift* and *conviction* do not suffer from this drawback.

(b) [15pts]

A measure is *symmetrical* if measure($A \to B$) = measure($B \to A$). Which of the measures presented here are symmetrical? For each measure, please provide either a proof that the measure is symmetrical, or a counterexample that shows the measure is not symmetrical.

(c) [20pts]

Perfect implications are rules that hold 100% of the time (or equivalently, the associated conditional probability is 1). A measure is *desirable* if it reaches its maximum achievable value for all perfect implications. This makes it easy to identify the best rules. Which of the above measures have this property? You may ignore 0/0 but not other infinity cases. Also you may find it easy to explain by an example.

Application in product recommendations: The action or practice of selling additional products or services to existing customers is called *cross-selling*. Giving product recommendation is one of the examples of cross-selling that are frequently used by online retailers. One simple method to give product recommendations is to recommend products that are frequently browsed together by the customers.

Suppose we want to recommend new products to the customer based on the products they have already browsed online. Write a program using the *A-priori* algorithm to find products which are frequently browsed together. Fix the support to s =100 (*i.e.* product pairs need to occur together at least 100 times to be considered frequent) and find itemsets of size 2 and 3.

Use the online browsing behavior dataset from `browsing.txt` in `data`. Each line represents a browsing session of a customer. On each line, each string of 8 characters represents the ID of an item browsed during that session. The items are separated by spaces. Some lines contain duplicate items. Removing or ignoring duplicates should not impact your results.

Note: for parts (d) and (e), the writeup will require a specific rule ordering but the program need not sort the output. We are not giving partial credits to coding when results are wrong. However, two sanity checks are provided and they should be helpful when you progress: (1) there are 647 frequent items after 1st pass ($|L_1| = 647$), (2) the top 5 pairs you should produce in part (d) all have confidence scores greater than 0.985. See detailed instructions below.

(d) [20**pts**]

Identify pairs of items (X, Y) such that the support of $\{X, Y\}$ is at least 100. For all such pairs, compute the *confidence* scores of the corresponding association rules: $X \Rightarrow Y$, $Y \Rightarrow X$. Sort the rules in decreasing order of *confidence* scores and list the top 5 rules in the writeup. Break ties, if any, by lexicographically increasing order on the left hand side of the rule. (You need not use Spark for parts d and e)

(e) [30**pts**]

Identify item triples (X, Y, Z) such that the support of $\{X, Y, Z\}$ is at least 100. For all such triples, compute the *confidence* scores of the corresponding association rules: $(X, Y) \Rightarrow Z$, $(X, Z) \Rightarrow Y$, $(Y, Z) \Rightarrow X$. Sort the rules in decreasing order of *confidence* scores and list the top 5 rules in the writeup. Order the left-hand-side pair lexicographically and break ties, if any, by lexicographical order of the first then the second item in the pair.