MLSD: Assignment 1
Frequent itemsets and association rules
Similar items

– Due date: April 05, 2024 –

For each of the following exercises, you should implement the solutions using Spark. Use small samples of the dataset for developing and initial testing, then run on the full data.

What to submit
For each exercise, submit a documented Jupyter notebook (or alternatively a documented python script to run through spark-submit) and the results of the algorithm. If the results are too large, submit a download link instead.

The documentation/comments should explain the main steps of the solution with sufficient detail.

1. The file 'conditions.csv.gz' (available on the shared folder) lists conditions for a large set of patients. Our purpose is to find associations between conditions.

   The file contains the following fields, with multiple non-consecutive entries for each patient:

   START,STOP,PATIENT,ENCOUNTER,CODE,DESCRIPTION

   PATIENT is the patient identifier

   CODE is a condition identifier

   DESCRIPTION is the name of the condition

   You may prefer to reorganize the data before applying the algorithms.
   Try to use Spark for this as well.

   1.1. Implement the A-Priori algorithm in Spark.

   1.2. Apply the implemented algorithm with a support threshold of 1000 to obtain the frequent itemsets for sizes $k = 2$ and $k = 3$. Include in your results the lists of the 10 most frequent itemsets for $k = 2$ and $k = 3$.

   1.3. Obtain associations between conditions by extracting rules of the forms $(X) \to Y$ and $(X, Y) \to Z$, with minimum standardised lift of 0.2.
   Write the rules to a text file, showing the standardised lift, lift, confidence and interest values, sorted by standardised lift.

2. TBD