COSC–254 Data Mining
Homework 03 – Itemsets and Association Rules
Due: Wednesday, February 20, 2019, 1.59pm

Exercise 1 Prove the following Lemma or give a counterexample.

Lemma 1. *For any non-empty $A, B \subset \mathcal{I}$, such that $A \cap B = \emptyset$, it holds*

$$\mathsf{conf}_{\mathcal{D}}(A \cup \{a\} \Rightarrow B \setminus \{a\}) \geq \mathsf{conf}_{\mathcal{D}}(A \Rightarrow B)$$

for any $a \in B$.

Exercise 2 Consider the dataset $\mathcal{D}$ in Table 1. Answer the following questions, showing the derivation of your answers with sufficient level of detail.

- What is the maximum number of association rules that can be extracted from this data (including rules that have zero support)?
- What is the maximum size of $\mathsf{FI}(\mathcal{D}, \ell)$ that can be extracted (assuming $\ell > 0$), and for which value(s) of ℓ?
- Write an expression for the maximum number of size-3 itemsets (e.g., $\{a, b, c\}$) that can be derived from this dataset.
- Find all the itemsets with the largest support.
- Find a pair (a, b) of items such that the rules $\{a\} \Rightarrow \{b\}$ and $\{b\} \Rightarrow \{a\}$ have the same confidence. You can exploit this property to reduce the search space, please explain how.

tid	transaction
1	{Milk, Beer, Dog food }
2	{Bread, Butter, Milk }
3	{Milk, Dog food, Cookies }
4	{Bread, Butter, Cookies }
5	{Beer, Cookies, Dog food }
6	{Milk, Dog food, Bread, Butter }
7	{Bread, Butter, Dog food}
8	{Beer, Dog food }
9	{Milk, Dog food, Bread, Butter }
10	{Beer, Cookies }

Table 1: The dataset $\mathcal{D}$ for Exercise 2.

How to submit Submit your work at `https://www.cs.amherst.edu/submit` or via `cssubmit` from romulus/remus, as a *single* archive file with name `username.ext` where `username` is your user name and `ext` is one of `.zip`, `.tar.bz2`, or `.tar.gz` (no `.rar`, please).

The archive must contain a *single* directory with name `username`. This directory must contain a subdirectory with name `X` for each Exercise `X`. All files (source code or otherwise) for each exercise must be in the directory for that exercise. Directories containing source code should contain a `README.txt` file explaining how to run the code in that directory. For non-code answers, please submit a `.pdf` or a `.txt` (no `.doc(x)`, please). You can find an example archive at `http://bit.ly/DM19sub`.

Please post to the Moodle forum if you have problems with the submission.