Problem Set 1

September 19, 2012

Due date: Wed, October 3 2012 at 4pm; before class.

Exercise 1 (20 points): You are given a set V consisting of n integers. The task is to report all n products of the n distinct $(n-1)$-cardinality subsets of V. Your algorithm should run in linear time and it should not use division.

Exercise 2 (20 points): Every time you go to Espresso Royal Coffee (ERC) and you buy a latte, the barista is providing you with a sticker. Every such sticker names one of the 100 different coffee types that ERC has served throughout the years. Once you collect all 100 *distinct* stickers you will earn an espresso machine. If the coffee types are uniformly assigned to stickers find the *expected* number of lattes you need to drink before you get your own espresso machine.

Exercise 3 (20 points): For each of the following measures, determine whether it is monotone, anti-monotone, or non-monotone (i.e., neither monotone nor anti-monotone).

For example, a measure s over itemsets is anti-monotone (resp. monotone) if for two itemsets X and Y we have that $s(X) \geq s(Y)$ whenever $X \subset Y$ (resp. $X \supset Y$).

A characteristic rule is a rule of the form $\{p\} \rightarrow \{q_1, \ldots q_n\}$, where p, $q_1, \ldots, q_n$ are items and the rule antecedent contains only a single item. An itemset of size k can produce up to k characteristic rules. Let ξ be the minimum confidence c of all characteristic rules generated from a given itemset $\{p_1, \ldots, p_k\}$. That is,

$$\begin{aligned} \xi(\{p_1, \ldots, p_k\}) \quad = \min \quad & \{c(\{p_1\} \rightarrow \{p_2, \ldots, p_k\}) \\ & c(p_k \rightarrow \{p_1, \ldots, p_{k-1}\})\}. \end{aligned}$$

(5 points): Is ξ monotone, anti-monotone or non-monotone?
(5 points): Repeat the above where instead of min we use max.

A discriminant rule is a rule of the form $\{p_1, \ldots, p_n\} \rightarrow \{q\}$, where the rule consequent contains only a single item. An itemset of size k can produce up to k discriminant rules. Let η be the minimum confidence of all discriminant rules generated from a given itemset:

$$\begin{aligned} \eta(\{p_1, \ldots, p_k\}) \quad = \min \quad & \{c(\{p_2, \ldots, p_k\} \rightarrow \{p_1\}) \\ & c(\{p_1, \ldots, p_{k-1}\}) \rightarrow \{p_k\}\}. \end{aligned}$$

(5 points): Is η monotone, anti-monotone or non-monotone?
(5 points): Repeat the above where instead of min we use max.

Exercise 4 (20 points): Consider the association rule $A \rightarrow B$ and the interestingness measure $M = \frac{P(B|A)-P(B)}{1-P(B)}$.

1. (5 points) What is the range of this measure? When does the measure attain its minimum and maximum values?
2. (5 points) How does M behave when $P(A, B)$ is increased, while $P(A)$ and $P(B)$ remain unchanged?
3. (5 points) What is the value of the measure when A and B are statistically independent?
4. (5 points) How does the measure behave under the inversion operation (in the $0-1$ table representing the data 0s become 1s and vice versa).

Exercise 5 (20 points): Let D be a transaction dataset, and D' another dataset formed from D by independently erasing items from the transactions in D; every item from every transaction in D is erased with probability p. Provide answer to the following questions:

1. (10 points) For an itemset S, compute its expected support in D' as a function of its support in D. (Hint: the expected support will depend on the size of S as well as the probability p).
2. (10 points) Compute the probability that a frequent itemset in D will remain frequent in D' for the same *minsup* threshold.