CS 6604 Assignment #2
Assigned: Sep 25, 2007
Date Due: Oct 3, 2007, in class, before class starts

---

1. (5 points) Present the Venn diagram showing inter-relationships between the following classes of constraints: monotonic, anti-monotonic, succinct, convertible anti-monotone, convertible monotone, loosely anti-monotonic, loosely monotonic, and others.

2. (10 points) Assume that you are given a 2-class dataset where instances of each class correspond to strings of length 2. The first character of the string is binary and the second character of the string has six possible values, from 'A' through 'F.' The instances of the first class are: '1A', '0E', '0B', '1B', '1F', '0D'. The instances of the second class are: '0A', '0C', '1C', '0F', '0B', and '1D'. Using the two characters of the instances as 'features' for classification, construct decision trees in at least two ways, using different choices of impurity measures. You must choose the impurity measures such that one decision tree branches on the first character at the root node, and the other decision tree branches on the second character at the root node. In light of your answer to the above, present general guidelines on which impurity measure is better in which types of datasets.

3. (10 points) In class we presented three different impurity measures. Verify whether each of them is a concave function. If so, prove it. If not, provide a counter example. Why is it important that impurity be a concave measure?

4. (20 points) Typically frequent itemset mining algorithms are run with a minimum support constraint. However, it is difficult for a data miner to 'guess' what a reasonable minimum support constraint might be for a dataset. Hence there is growing interest in mining 'top-k' patterns which refer to, in this case, finding the k itemsets that have the highest support. Can you think of an anti-monotonicity or monotonicity guiding principle for mining such patterns? You are free to make any additional assumptions to the formulation if it helps in identifying an efficient constraint and suggests a suitable algorithm.

5. (25 points) In class we discussed how to mine itemsets with constraints on their average value over some attribute or variance over the attribute. Explain how you will mine itemsets X that have the following constraints with respect to a numeric attribute a (e.g., price): (i) Median (X.a) $\leq$ or $\geq \theta$, (ii) Mode(X.a) $\leq$ or $\geq \theta$.

6. (30 points) Read the paper 'How to quickly find a witness' which presents an algorithm for finding itemsets with constraints on the variance. Compare the ideas in this paper to the one by Bonchi and Lucchese presented in class 'Pushing Tougher Constraints in Frequent Pattern Mining' which gives an alternate algorithm for finding such itemsets. Determine whether the notions of 'witness' presented in the former and the 'loose anti-monotonicity' presented in the latter are compatible and if so, identify the relationships between them.

Turnin a typed (not handwritten) paper copy giving answers to the questions, plots, including a brief description of how you solved each question. Write enough to convince us that you completed the assignment independently.