ASSIGNMENT 3  
**DUE DATE: Wednesday 20 October, 2010  11:59 PM**

A data publisher is often faced with the difficulty of choosing the value of a privacy parameter while sharing the data. For instance, what value of k in k-anonymity would give a good balance between privacy and utility? Although this is a difficult question to answer, the data publisher can certainly do some analysis before deciding on a value for k.

In this assignment you will perform some analysis on the trade-off behavior between privacy and utility using the Mondrian algorithm.

**Task 1:**

Your first task will be to implement the Mondrian algorithm to obtain k-anonymous generalizations for a given value of k. You can download the data set to use from the assignments page of the course website. This is the same data set as used in Assignment 2, with the difference that categorical values have been assigned numbers.

Unlike in Assignment 2 where you randomly checked a few modification schemes, in this assignment you will use your Mondrian implementation to find a scheme that achieves k-anonymity for a given value of k.

I do not want to constrain you on the usage of a certain language. But please do not make me buy a compiler just to execute your code!

**Task 2:**

In order to perform the trade-off analysis, you will run the Mondrian implementation for k = 2, 5, 10, 15, 25, 50, 100, 300, 500 and 1000 (you can run it for more values if you want). For each of the runs, compute the classification metric value from the generalized data. Finally, create a plot showing the k values on the X-axis and the classification metric value on the Y-axis.

This plot shows the data publisher how the classification utility of the data set changes as the privacy enforcement is changed. There are better methods to do this analysis; but let’s not make life harder than it is now.
What to submit:

There are two components you need to submit (preferably as a zip file).

1) Your **source code** for the Mondrian implementation, fully commented, and along with instructions on how to compile and execute it. Do not forget to put your name in the beginning of source code files.

2) The **plot** generated in Task 2. Also include a **short paragraph** saying what you learned about the trade-off behavior of k-anonymity and the classification metric (on the given data set) from the plot. What do you think would be a good approach to select k, given this plot?

Requirements:

You **must** work alone on this assignment. Each one of you must implement Mondrian individually.

How will you be graded?

This assignment will account for 8 points towards your final grade. You can get up to 5 points for the implementation. You will certainly lose most points here if I find code fragments to be unjustifiably similar. The remaining 3 points can be accrued by completing Task 2. Start early!