ASSIGNMENT 2
DUE DATE: Monday 4 October, 2010 11:59 PM

In this assignment you will take the role of a data publisher. You will be required to modify a given data set such that it meets some privacy requirement. While simply enforcing a privacy requirement is not difficult, data publisher’s often face difficulty in creating one that is useful too. Hence, I will act as a legitimate user of the modified data set and perform some classification task using it. Please read each section thoroughly to understand the different parts of this assignment.

A. The data set:

The data set to use in this assignment is available from the assignment page of the course website. It is a downscaled version of the Census Income data set available from the UCI Machine Learning repository. http://archive.ics.uci.edu/ml/datasets/Adult

The provided data set has the following quasi-identifiers:

1. Age: numerical attribute with a value between 17 and 90

2. Education: categorical ordinal with one of the following values
   Preschool, 1st-4th, 5th-6th, 7th-8th, 9th, 10th, 11th, 12th, HS-grad, Some-college, Assoc-acdm, Assoc-voc, Bachelors, Masters, Prof-school and Doctorate

3. Gender: categorical attribute with a value of either “Male” or “Female”

Besides these three QI-attributes, there is also a sensitive attribute called SalaryClass. The value in this attribute is either “<=50K” or “>50K” signifying the salary information for a record.

B. The privacy requirement to enforce:

You will be required to modify the QI attributes in the data set such that the modified data becomes k-anonymous. Any value of k greater than 25 is acceptable.

C. How to modify the data:

There are multiple ways to modify the data. You will have to decide how you do it. The only constraints are as follows:

1. Age: You can group different age values into ranges. For e.g. [15-20], [20-25], ..., [85-90] is a valid modification for this attribute. You can have bigger or smaller
ranges as well. Each range also need not be of the same size. I would suggest you decide on a few range sizes (say 5, 10, and 15) and work with them.

2. Education: This is a categorical ordinal. So, a certain order can be imposed on the values (as already shown in section A). You can modify this attribute by clubbing together consecutive values. For e.g. you can generalize Preschool through HS-grad as “School”; but generalizing Preschool and HS-grad to one category, and 1st to 12th grade as another is not allowed.

3. Gender: You can keep the value in this attribute as it is, or completely suppress it.

Following these constraints, you will have to come up with a modification scheme that makes the data privacy preserving as stated in section B.

D. How to evaluate a modification:

Once you decide on a set of modification schemes to try out, you should have a method to evaluate the goodness of a certain scheme. You can do this using the following method.

Modify the data using a scheme. This will result in the formation of QI-groups.

   a) For each QI-group, determine which SalaryClass value (“<=50K” or “>50K”) appears more number of times (called the majority class label). If both values occur the same number of times, then choose one randomly as the majority class label.

   b) For each record in a QI-group, compare the SalaryClass value with the majority label. If they are different, then assign a penalty of 1 to the record; zero otherwise.

   c) Add up all the penalties assigned to records in the QI-group. This will give you the total penalty in the QI-group.

   d) Once you have computed the total penalty for every QI-group, add them up to get the total penalty for the entire data set (this is also known as the classification metric).

The final modification method you submit (out of the ones you would try out) should be the one that has the minimum value for the classification metric.

E. How should you approach this assignment?

Since there are more than 30000 records in the data set, it will not be possible for you to manually modify each record, determine the QI-groups, find the smallest QI-group size (i.e. the value of k) and compute the classification metric value. You will have to write some kind of a program or a script that does all that for you. Write a program that lets you check different modification schemes (check means it tells
you the size of the smallest QI-group and the value of the classification metric). Once you have such a program, you can easily test all the schemes you want to try out and find the one with the lowest classification metric value.

F. What to submit:

Create a zip file with the following two files and submit the zip file.

1) **The generalized data.** This will have the same structure as the data set provided, except that the Age, Education and Gender values should be the modified ones. This is essentially the data set to be published by you.

2) **The modification scheme** used by you to create the generalized data. For example:

17   (15-25]  
18   (15-25]  
....
89   (85-95]  
90   (85-95]  
Preschool  School
1st   School
...  
HS-grad  School
Some-college  College
...  
Doctorate  University
Male   *  
Female  *

Format your scheme similar to the example shown above. This file lists the modified value for every possible value that may appear in the Age, Education and Gender attributes.

Your are **NOT** required to submit any source code.

G. How will I use the anonymized data to perform classification?

Using the modified data submitted by you, I shall build a classifier that predicts the SalaryClass based on Age, Education and Gender information. I also have another test data set (not known to you) which has the same attributes as the one provided to you. This set has records in their original form (not generalized). The learned classifier will predict the SalaryClass label of each record in the test set. I can then compare this predicted label with the actual label of the record and determine if the prediction was correct. The percentage of records for which I am able to make a
correct prediction will determine the usefulness of the anonymized data generated by your modification scheme.

**H. How will you be graded?**

This assignment will account for 8 points towards your final grade. You can get up to 4 points if most of the SalaryClass predictions are correct. The remaining 4 points will be based on uniqueness of your generalization method. If you are the only one who suggested the method you submitted, then you will get the full 4 points. Otherwise, points will be deducted depending on how many others also came up with the same method. You **must** work alone on the assignment.