Social Computing [CS60017]

Assignment 2

Deadline: 24th November, 11:55pm

K-Anonymity:

A dataset is said to have the k-anonymity property if the information for each person contained in the dataset cannot be distinguished from at least k - 1 individuals whose information also appear in the dataset. [wiki link for more info]
The records in a dataset are in this format: [QID → SA]. QID means quasi-identifier [the attributes which can be used together to uniquely identify a person] such as age, gender and birthday; SA means sensitive information such as disease information. Suppose a person is in a group with other people, all of whom have the same QID, then nobody can infer their sensitive information (SA) from this group using QID.
One way to achieve this is by generalization, in which we segregate the attributes into generalized buckets so that the QID becomes the same for a group of people. For example, we can convert the age to buckets of [Age<18, 18<=Age<25, 25<=Age]

Mondrian Algorithm:

Mondrian is an algorithm to convert a dataset to have k-anonymity property, based on generalization. Here is the basic workflow of Mondrian:

1. Partition the raw dataset into k-groups using kd-tree. [k-groups means that each group contains at least k records]
2. Generalization of each k-group such that each record in a group has the same QID.

For more explanation of Mondrian,
Read the description at: [link to explanation of Mondrian] and the paper: [link to paper].

Assignment Overview:

You have to implement the Mondrian algorithm to output a version of the **Adult** dataset that has K-anonymity property. You can get the dataset, “adult.data” for the Data Folder at: [Link to data]. It contains various attributes like age, work-class and education; with the final column as the label, which predicts if a person earns less or more than 50k per year. A description of the dataset is also given in “adult.names” in the Data Folder.

Submission instructions:

Create a folder with your <roll number>_A2 [for example: 18CS60021_A2]. The folder should contain a python script (or cpp file) named main.py (or main.cpp) from where the execution will start. [you may code in different files, but there should be a main file!].
Your code should take the value of ***k*** as an argument , so that we can run your code with the command: “python main.py 10” (or for cpp: “./main 10”), for ***k*** = 10.

Keep the dataset “adult.data” in the root folder itself, and output the k-anonymous records in a file “adult.out” into the same folder, grouped together by the equivalence classes [records in the same equivalence class should be stored in consecutive lines].
Extract and work with only the following features (the indices correspond to the attribute columns in the original dataset):

0	1	4	5	6	8	9	13	14
Age,	work_class,	education_num,	marital_status,	occupation,	race,	sex,	native_country,	class

Discard any record that has any of these attributes missing.

The output file should have the grouped attribute values separated by a ‘~’ (tilde). For example, for continuous values, you can print “*23~37*”, with the smaller value first, and for a group of categorical values you should print the list separated by ‘~’:
“*Adm-clerical~Exec-managerial~Handlers-cleaners*”, with the values sorted in lexicographic order. Each of the attributes should be comma separated, just like the input data file.

Your code should print out just the value of the Discernibility Metric [defined in the explanation of Mondrian]. You should **NOT** print anything else to stdout or take any input from stdin.
Also include a file instructions.txt which will have the python version you are using, any non-standard libraries you are using, and any other instructions you want to convey.
For example, one way to group the categorical values is simply by converting it to consecutive integers, and then finding the median, as you would do with continuous variables, but you may use other strategies too. Write this in instructions.txt

Finally, compress the folder into a zip or tar.gz file, for example, 18CS60021_A2.zip or 18CS60021_A2.tar.gz. Not following the given instructions for submission format will lead to marks being deducted!

For queries mail TA Soham Poddar at sohampoddar26 [at] gmail [dot] com