Homework 1

Name:
CWID:

Pack all the files into a zip file "*yourname_hw1.zip*" (**Bonus: 2 Points**) and submit it on the Blackboard by **September 15, 2021 (11:59PM CDT)**.

1. k-Anonymity (40 Points)

Design and implement a heuristic algorithm to ensure (k_1, k_2)-anonymity for the Adult dataset in the UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Adult.

- The full dataset and description are available at: https://archive.ics.uci.edu/ml/machine-learning-databases/adult/
- For the attribute "Salary", there are only two distinct values ("$\leq 50K$" or "$> 50K$"). For adults with salary $\leq 50K$, they prefer a stronger protection $k_1 = 10$; For adults with salary $> 50K$, they are OK with $k_2 = 5$. "Salary" is only used to determine k_1 or k_2 for different users (no need to share them in the output).
- To simplify the problem, you only need to consider 4 attributes as quasi-identifiers (QIs) to implement the generalization and/or suppression:
  1. Age: positive integers
  2. Education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
  3. Marital-Status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
  4. Race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
- Consider "Occupation" as the sensitive attribute to share, which includes 15 distinct values (including "?"). Notice that,
  - The dataset has missing values. If so, it is considered as "Generalized to the top of the hierarchy".
  - All the remaining attributes (other than the 4 QIs and the sensitive attribute) can be suppressed in the data.
  - The algorithm should try to maximize the output utility. For instance, applying $k_1 = k_2 = 10$ can also satisfy the privacy demand of all the users. However, it is not an acceptable solution due to high utility loss.

– The distortion and precision can be different for different hierarchies and algorithms. If your hierarchy and algorithm are reasonable, the distortion should not be very large. Otherwise, for instance, if each hierarchy (for a QI) only includes 2 levels, the distortion might be very large.

Tasks:

(a) Define reasonable hierarchies for the 4 QIs. (**5 points**)

(b) Write a program for the heuristic algorithm, which generalizes/suppresses the data for (k_1, k_2)-anonymity while minimizing the utility loss. You can use any programming language, e.g., Java, Python and C++. You can also extend the DataFly or μ-Argus Algorithm. (**25 points**)

(c) Calculate the distortion and precision of the output ($k_1 = 10, k_2 = 5$) based on your designed hierarchies and algorithm. (**10 points**)

Submission Part I: output dataset (*QIs* and *sensitive attribute*) and source code files – all named with the prefix "hw1-1-" (e.g., *hw1-1-anonymity.java*).

2. ℓ-Diversity (60 Points)

Using the same setting as above (dataset, QIs, sensitive attribute, and hierarchies), design and implement a heuristic algorithm to ensure ℓ-diversity besides k-anonymity.

Tasks:

(a) Write a program for the heuristic algorithm (which generalizes/suppresses the data for "Entropy ℓ-diversity" while minimizing the utility loss). You can use any programming language, e.g., Java, Python and C++. (**20 points**)

(b) Write a program for the heuristic algorithm (which generalizes/suppresses the data for "Recursive (c, ℓ)-diversity" while minimizing the utility loss). You can use any programming language, e.g., Java, Python and C++. (**20 points**)

(c) Set $k = 5, \ell = 3$ for (a), and calculate the distortion and precision of the output. (**10 points**)

(d) Set $k = 5, \ell = 3$ and $c = 0.5, 1, 2$ for (b), and calculate the distortion and precision of the outputs (three outputs for different c). (**10 points**)

The algorithm can iteratively search generalization levels (of the hierarchies) and the corresponding combinations of equivalence classes (of the records). The key criterion is to check if the distribution of sensitive values in each equivalence class satisfies a specific ℓ-diversity, and if the distortion or precision is minimized after forming the equivalence classes via generalization and suppression.

Submission Part II: output datasets (*QIs* and *sensitive attribute*) and source code files – all named with the prefix "hw1-2-" (e.g., *hw1-2-entropy.java*).

Submission Part III: a PDF file *hw1-report.pdf* to include the hierarchies, and screenshots for all distortion/precision results (for both k-anonymity and ℓ-diversity). All the results should be well-marked (e.g., for Task 1(a), 1(c), 2(c), and 2(d)).