EECS700: Data Privacy and Security

Homework 1

Due Date: 11:59 PM on September 22, 2025

Background

In this homework, you will practice implementing privacy-preserving data publishing techniques **on the UCI Adult dataset. The full dataset description is available at: https://archive.ics.uci.edu/ml/machine-**learning-databases/adult/old.adult.names **(Note that the dataset has missing values. If so, it is considered as "Generalized to the top of the hierarchy".**

You will work with a simplified setting:

Quasi-identifiers (QIs): attributes that could potentially re-identify individuals. Here we only use:

- Age (continuous)
- Education (categorical)
- Marital-Status (categorical)
- Race (categorical)

Sensitive Attribute (SA): for the l-diversity task, use **Occupation**.

Goal: You will design hierarchies for QIs, implement heuristic anonymization algorithms, and evaluate privacy–utility trade-offs.

Task 1: Define Generalization Hierarchies (30 points)

Before anonymization, each attribute needs a **generalization hierarchy**. A hierarchy describes how raw values can be gradually generalized into broader categories.

- **Example for Age (continuous):**

    - Level 0: exact ages (23, 24, 25, …)
    - Level 1: 5-year ranges (20–24, 25–29, …)
    - Level 2: 10-year ranges (20–29, 30–39, …)
    - Level 3: Life stage (Young, Adult, Senior, Elder)
    - Level 4: * (completely generalized)

Write down the hierarchies (as tables or trees). Be explicit about levels from the lowest to top level.

Task 2: Personalized k-Anonymity (40 points)

(a) Assigning Personalized "k" (10 points)

We assume each record belongs to a specific user and each user has a privacy requirement "**k**". Thus, write the code to randomly assign a "k" privacy requirement of each record from {4, 5,7}. Please print out the counts of how many records got k1=4, k2=5, k3=7 with your code. Please have the screenshot of your code and result in the report.

(b) Achieving Personalized k-Anonymity (20 points)

In the personalized k-anonymity, each record does not share the same privacy requirement. Instead, every individual record has its own k value. Then each record's k value should be satisfied within an equivalence class. Then please design and implement an algorithm to achieve the personalized k with the "k" assignment of step (a). (Hint: you can extend the datafly algorithm to achieve this. When we form an equivalence class by grouping records, the size of this group must be at least as large as the strictest requirement among its members.) In your report, please include the pseudocode of your algorithm and important part of your algorithm (such as the condition to check the privacy requirement "k") with explanation why your algorithm guarantee the personalized k.

(c) Evaluations (10 points)

Based on your algorithm, set the {k1=4, k2=5, k3=7}, please choose distortion or precision to evaluate your output of (b) with the programming code. Also, setting the {k1=2, k2=3, k3=4}, please have the distortion or precision and compare it with {k1=4, k2=5, k3=7}'s evaluation. Print out the evaluation result and have the screenshot in your report.

Task 3: l-Diversity (30 points)

(a) Ignoring the personalized "k" and we assume the "k=6" is the same for each record, now please using generalization/suppression to achieve the 3-diversity. The sensitive attribute is the "occupation". please include the pseudocode of your algorithm and important part of your algorithm (such as the condition to check the privacy requirement "k" and "l") with explanation why your algorithm guarantee the privacy strictly. (20 points)

(b) Based on your algorithm, setting the k=6 and l=3, please choose distortion or precision to evaluate your output with the programming code. Also, setting the {k=4, l=2}, please have the distortion or precision and compare it with the setting of k=6 and l=3. Print out the evaluation result and have the screenshot in your report. (10 points)