Benchmarking Machine Learning Utility on k-Anonymized Data:

The Impact of Anonymization Parameters and Quasi-Identifier Selection

Çağrı Bilginer, İbrahim Cebecioğlu, Utku Çubukçu

SUMMARY

This project will investigate how k-anonymity affects the utility of machine learning models trained on anonymized data. We will benchmark the accuracy of classification across three datasets (Adult, Bank Marketing and German Credit) using two classifiers (Logistic Regression and Random Forest) under varying anonymization parameters. We will also compare two QI selection strategies. We will compare treating all attributes as QIs against only a subset of the attributes as QIs. We expect having fewer QIs to reduce the strictness of anonymization thus improving the utility.

MOTIVATION AND PROBLEM STATEMENT

The Data Sharing Dilemma

Organizations need to share or publish large datasets to facilitate research and machine learning development. These datasets often contain sensitive personal information. To comply with privacy regulations like GDPR and protect individual identities, data must be anonymized before being published.

The Utility-Privacy Tradeoff

k-anonymity is one of the most widely adopted privacy models. As we learned in the lectures, it works by ensuring each record in a released dataset is indistinguishable from at least k - 1 other records based on a set of QIs. This is achieved by data transformation techniques like generalization and suppression. These naturally cause information loss and degrades the quality of the data for machine learning tasks.

The Problem Statement

While there are prior studies on this trade off, we would like to look into two important practical questions:

1. **Which attributes should be designated as quasi-identifiers?**
   Practitioners must choose which attributes to protect. Protecting all attributes maximizes privacy but will likely destroy utility. A realistic approach (such as protecting only demographic attributes) preserves more information but may leave linkage risks.

2. **How do different ML models respond to anonymization?**
   Simple linear models may be more robust to generalization than complex models that rely on fine-grained feature interactions.

TECHNICAL APPROACH

Datasets

We use three well-established classification datasets from the UCI Machine Learning Repository:

- **Adult Census Income** (48,842 instances, 14 attributes): Predict whether income exceeds $50K/year.
- **Bank Marketing** (45,211 instances, 17 attributes): Predict whether a client subscribes to a term deposit.
- **German Credit** (1,000 instances, 20 attributes): Predict credit risk (good/bad).

Preprocessing Steps:

- **Data Cleaning:** Handle missing values either by filling them in with the median, dropping them entirely if many are missing or let the model learn about missing data as well.
- **Feature Encoding:** Convert categorical attribute values into numeric format using one-hot encoding.
- **Normalization:** Scale numeric attribute values to similar ranges so features with larger numbers don't dominate the model.

Anonymization Tool and Parameters

We use the recommended ARX Data Anonymization Tool because it supports k-anonymity through generalization hierarchies. We plan to test four k values (2, 5, 10, 20). This range covers minimal to strong anonymity and will allow us to observe how accuracy degrades with more strict privacy requirements.

Quasi-Identifier Selection Strategies

Our key experimental variable is the choice of quasi-identifiers. We compare two strategies:

1. **Aggressive (All Non-Target):** All attributes except the target variable are treated as QIs. This represents a worst-case assumption where any attribute could be used for linkage attacks.

2. **Realistic (Demographic Only):** Only demographic/personal attributes (age, sex, race, education, occupation, etc.) are designated as QIs. Behavioral/transactional attributes are left ungeneralized.

Machine Learning Pipeline

We train two classifiers using scikit-learn:

- **Logistic Regression:** A linear model that may be more robust to generalization.

- **Random Forest:** A non-linear model that relies on feature splits and may suffer more from generalized values.

Experimental Procedure

1. Preprocess each dataset (handle missing values, encode categorical variables for baseline).
2. Train baseline models on non-anonymized data.
3. For each dataset, QI strategy, k-value combination, anonymize the data using ARX.
4. Train both classifiers on anonymized data and evaluate performance.
5. Generate comparison plots showing accuracy vs. k-value for each configuration.

DELIVERABLES

- **Project Report:** Detailed documentation of methodology, experimental results with tables and graphs, analysis of findings.
- **Source Code:** Python scripts for the ML pipeline and Java/Python code for ARX integration with documentation.
- **Experimental Results:** CSV files containing all experimental measurements and Jupyter notebooks for visualization.

TIMELINE

The timeline may shift, it's difficult to predict how much time and effort each task will take.

Task	Dec 27 - Jan 2	Jan 3-9	Jan 10-16	Jan 17-19	Jan 20-22
Setup & ARX learning	All				
Data preprocessing & baseline models	Çağrı				
ARX anonymization pipeline	İbrahim	İbrahim			
Define QI strategies per dataset	Utku				
Run experiments (all configs)		All	All		
Analysis & visualization			Çağrı	Çağrı	
Write final report				İbrahim, Utku	All

BIBLIOGRAPHY

[1] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," Technical Report, SRI International, 1998.

[2] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian Multidimensional K-Anonymity," in *Proceedings of the 22nd International Conference on Data Engineering (ICDE)*, 2006.

[3] F. Kohlmayer, F. Prasser, and K. A. Kuhn, "The Cost of Quality: Implementing Generalization and Suppression for Anonymizing Biomedical Data with Minimal Information Loss," *Journal of Biomedical Informatics*, vol. 58, pp. 37-48, 2015.

[4] ARX Data Anonymization Tool. Available: https://arx.deidentifier.org

[5] UCI Machine Learning Repository. Adult Dataset. Available: https://archive.ics.uci.edu/dataset/2/adult

[6] UCI Machine Learning Repository. Bank Marketing Dataset. Available: https://archive.ics.uci.edu/dataset/222/bank+marketing

[7] UCI Machine Learning Repository. Statlog (German Credit Data). Available: https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data

[8] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.