RSI - Privacidade de Dados - 2022

Privacidade Diferencial
Mecanismos Laplace & Exponencial

Javam Machado

1 Objetivos

- Implementar os mecanismos de privacidade diferencial: Laplace e exponencial.
- Realizar consultas sobre uma base de dados aplicando os mecanismos e fornecer os resultados de forma correta e privada.

1.1 Mecanismo de Laplace - Especificação

- Carregue o conjunto de dados "covid.csv". Calcule a idade dos indivíduos representados em cada registro a partir da data de nascimento. Considere apenas datas de nascimento referentes aos séculos XX e XXI.
- Realizar um conjunto de três consultas como especificadas abaixo. As consultas deverão usar os seguintes valores de *budget*: $\varepsilon = 0.1$, $\varepsilon = 0.5$, $\varepsilon = 1.0$ e $\varepsilon = 10$.
- Calcule a sensibilidade global para cada consulta. Lembre que a sensibilidade global independe do conjunto de dados e é calculada com base na consulta. (Dica: faça premissas sobre os dados trabalhados quando for inviável calcular a sensibilidade global.)
- Consultas a serem realizadas sobre a totalidade dos dados:
  1. Q_1: Média da idade dos indivíduos representados no dataset;
  2. Q_2: Número de exames positivos (atributo *resultadoFinalExame*);
  3. Q_3: Total de exames realizados por município (atributo *municipioCaso*;
- Para as consultas Q_1 e Q_2, mostre um gráfico com o resultado da consulta para cada ε, comparando os valores originais e privados.

- Para a consulta Q_3, defina *bins* para apresentação dos resultados, macro região por exemplo, e apresente um gráfico para cada ε. Cada gráfico vai mostrar a frequência original e a frequência perturbada referente a cada *bin*. Neste caso, você pode apresentar graficamente o somatório dos resultados da consulta sobre os municípios referentes a cada macro região. Compare os resultados originais com os resultados da consulta adicionada de ruído.

1.2 Mecanismo Exponencial - Especificação

- Carregue o conjunto de dados "covid.csv".
- Realizar um conjunto de duas consultas como especificadas abaixo. As consultas deverão respeitar os seguintes valores de *budget*: $\varepsilon = 0.1$, $\varepsilon = 0.5$, $\varepsilon = 1.0$ e $\varepsilon = 10$.
- Considere décadas como por exemplo, década de 1970 são todos os nascidos entre 01/01/1970 a 31/12/1979. Considere igualmente o universo das décadas, todas aquelas cuja a data de nascimento (atributo *Nascimento*)tenha ocorrido a partir de 01/01/1900, portanto o universo é formado por 13 décadas ate o ano de 2022.
- Considere o universo dos municípios, todos os municípios do Ceará.
- Defina uma função de utilidade para os possíveis valores de resposta das consultas. Esta função de utilidade pode ser calculada a partir da frequência dos valores existentes no dataset, como por exemplo o número de casos negativos de cada década representada no dataset.
- Calcule a sensibilidade da função de utilidade para cada consulta.
- Consultas a serem realizadas sobre todo o dataset:
  1. Q_1: Qual o município (atributo *municipioCaso*) com o **menor** número de casos positivos (atributo *resultadoFinalExame*) de COVID-19;
  2. Q_2: Qual a década (atributo *Nascimento*) com o **maior** número de casos negativos de COVID-19;
- Você deve realizar por 10 vezes a mesma consulta e retornar como resposta o valor mais frequente dentre os 10 resultados. Construa um histograma indicando quantas vezes cada saída foi retornada dentre as 10 execuções.
- Para as consultas acima, mostre uma tabela comparativa entre o resultado da consulta para cada ε e o valor original.

2 Requisitos

- Linguagens: C++ ou Python
- Trabalho individual
- Meio de entrega: criar uma pasta zipada chamada "Trab_Priv_Diff_<nome>" contendo código, dataset de entrada e arquivos "csv" com os resultados das consultas. Escreva um Readme.txt descrevendo o projeto. Fazer o *upload* na plataforma *Classroom* da disciplina. O não cumprimento destes requisitos está sujeito a penalidade.
- O trabalho deverá ser entregue até as 23:59h do **segunda, 03/10/2022**.

3 Avaliação

Na avaliação serão considerados os seguintes indicadores:

- **Corretude** do programa para cada consulta;
- **Precisão** pela comparação do dataset original com o dataset anonimizado;
- **Pontualidade** da entrega e **documentação/qualidade** do código-fonte.