Data Science Individual Project Interim Report
Differentially Private Location-based Histogram Publication

1. Introduction

With prevalence of applications collecting data about life patterns of users, it is made possible for analysts to look into both open datasets and private datasets for intelligence. However, it is of prime importance to maintain individual privacy while distributing any of the data and findings. In this project, the problem of releasing one-dimension dataset in its histogram representations under differential privacy policies will be addressed.

Once sufficient privacy enforced, analyst can then publish their findings without risking any sensitive data of both firms and users. Such work will contribute not only to daily life of everyone, but also to development strategies of firms. Taking restaurant reviews and check-in data as an example, analyst can then segment users into different preferences and provide them with suggestions of new restaurants when they are travelling to a new place; analysts can also look into the positioning of their firm and adopt strategies to either consolidate their reputation or mitigating to the desired market segments.

With the possible benefits said, this project covers over a broad domain. By modifying differential private algorithm, our work would be applicable for applications which uses histograms as query answers. Our work would hopefully provide insights on clustering and classification as well. Among all the datasets, our work will be focusing on one-dimension datasets and specifically on datasets containing check-in records.

2. Preliminaries

2.1. Histogram

Given an attribute θ with a set of values V in a one dimensional dataset D, we can aggregate the count (or the frequency), f for each value $v \in V$. With the set of θ and its corresponding count f, a histogram, H can be generated. A histogram H regarding the attribute θ can then be denoted as $H = \{\mathrm{H}_1, \mathrm{H}_2, \dots, \mathrm{H}_n\}$ with H_i representing the count of values of θ covered in the bin i. Under most circumstances, bin i and bin j do not overlap for every $i \neq j$ (i.e. $range(\mathrm{H}_i) \cap range(\mathrm{H}_j) \neq \emptyset$). With such a histogram, we would be able to answer to different range queries.

2.2. Differential Privacy

In this article, we will look into algorithms generating output from databases while enforcing sufficient privacy such that released data. Assume there exists histograms H_1 and H_2 from database D_1 and D_2 which differs by exactly one record. H_1 and H_2 can then be referred as neighbours. Base on these assumptions, differential privacy can then be defined as follows:

Definition 1. (J. Xu, 2102) A randomised algorithm $\mathcal{A}$ is said to be ϵ-differentially private if for any two neighbour histograms H_1, H_2 and any subset of output value $S \subseteq Range(\mathcal{A})$,

$$\Pr(\mathcal{A}(\mathrm{H}_1) \in S) \leq \exp(\epsilon) \cdot \Pr(\mathcal{A}(\mathrm{H}_2) \in S)$$

where $\epsilon > 0$.

With the parameter ϵ, users can specify the level of privacy to be enforced. The smaller ϵ, the higher privacy protection will be applied. Normally, ϵ will be set as a small value (e.g. $\epsilon < 1$) such that removal of one individual data prior to running the algorithm will not significantly affect the output (i.e. the histogram).

It is often that a series of algorithms $\mathcal{A} = \{\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_N\}$ will be run to generate the output. The algorithms will each be ϵ-differentially private, and thus can come up with a privacy budget of $\epsilon = \sum_{i=1}^{N} \epsilon_i$ that will be allocated to each of the algorithm $\mathcal{A}_i$.

In the literatures, one of the most well-developed algorithms for enforcing differential privacy is the Laplace mechanism. This mechanism masks the original data with random noises generated from the Laplace distribution, $Lap(\sigma)$ (i.e. a Laplace distribution with mean 0 and scale σ). The scale is determined based on the concept of a sensitivity of a function f over the histogram.

Definition 2. Denote $f(H)$ as a function of the histogram H which generates a vector in $\mathbb{R}^d$. The Laplace mechanism will then give output of $Lap(H) = f(I) + \boldsymbol{z}$, where $\boldsymbol{z}$ is a d-length vector and that each $z_i \sim Lap(\Delta f/\epsilon)$. The constant Δf is the sensitivity of f and is defined as the maximum difference in f between 2 neighbouring histograms H_1, H_2 (i.e. $\Delta f = max_{H_1,H_2} \|f(\mathrm{H}_1) - f(\mathrm{H}_2)\|_1$.

Theorem 1. (J. Xu, 2102) For any function $f : H \rightarrow \mathbb{R}^d$, the algorithm set $\mathcal{A}$ is said to be ϵ-differentially private if

$$\mathcal{A}(H) = f(H) + [Lap_1(\Delta f/\epsilon), \dots, Lap_d(\Delta f/\epsilon)]$$

where each $Lap_i(\Delta f/\epsilon)$ are i.i.d. Laplace variables with scale $\sigma = \Delta f/\epsilon$.

As given any database, adding one new record will lead to an increase by exactly 1, the sensitivity constant $\Delta f = 1$. In other words, the mechanism can be simplified as adding random noises of $Lap(1/\epsilon)$, where ϵ is the privacy budget allocated to the Laplace mechanism.

2.3. Threshold Function

Given an attribute X with a value set V, and a function f, such that $f(X) = MX + e$, the output value of the function will be masked with external noises such as Laplace noise. A threshold function can be used to recover the unknown value $v \in V$ from the noisy output.

In the literature, it is common that threshold function for differential privacy is implemented as a high-pass filter (J. Xu, 2102) such that

$$\widehat{H} = \begin{cases} \widehat{H} & if\ \widehat{H} \geq \theta \\ 0 & otherwise \end{cases}$$

where $\theta = \eta \log(n)/\epsilon$ with ϵ as the privacy budget allocated onto this threshold function, and $\eta > 0$ as the tuning parameter. This will then trim of negligible small counts while leaving the more significant ones.

3. Latest Literature Review

3.1. Sparse Vector Threshold

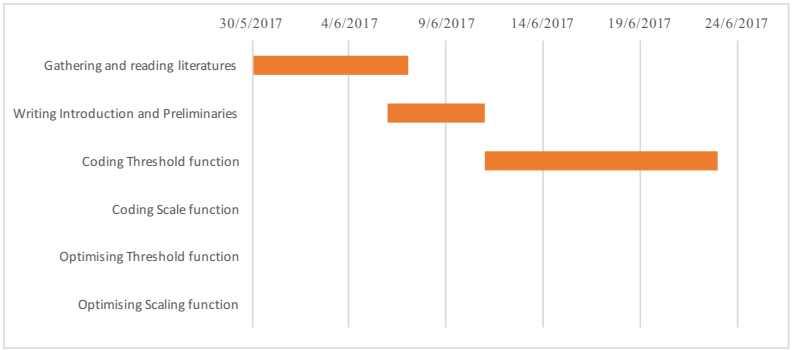
Apart from the high-pass filter threshold function, another possible approach is via sparse vector technique (Min Lyu, 2016) where the privacy budget allocated for threshold function will further split down into two parts. With the two smaller budgets, they will be responsible for altering the threshold value, on top of the Laplace mechanism. The mechanism is proved to be a ϵ-differentially private Laplace mechanism. On top of that, they have looked into the optimal privacy budget allocation within the threshold function. Based on their findings, we should further apply the technique to see if it actually fits for answering range queries with histograms.

3.2. Iterative Threshold

Threshold functions can also be defined with an iterative approach (Ulugbek S. Kamilov, 2016). This is implemented with a technique of backpropagation for evaluating the gradient. With such technique, the threshold function will run through a fixed number of iterations to minimalise the mean-squared-errors in the noisy data. The technique can be further extended with online learning, offline learning etc. This implementation would however require significantly more computational resources. A possible way to apply this concept to our research would be to combine the iterative approach to the sparse vector technique such that privacy can be enforced with minimal error incurred.

4. Latest Progress

Currently, snippets are coded to look into the negative impacts of incorporating iterative approach into the sparse-vector threshold function. A break-off point should be located and derived as the next step to generalise the approach. Apart from that, multilevel threshold function and dynamic programming are also some extra perspectives on the scope of threshold function. In light of the scope of threshold function, scope function may be postponed.



As presented above, the first iteration for introduction and preliminaries sections are completed. This would however still subject to future change as the project progress and as more literatures are included.

5. Bibliography

Xiaojian Zhang, R. C. (2014). Towards Accurate Histogram Publication under Differential Privacy. *2014 SIAM International Conference on Data Mining* , 9.

Min Lyu, D. S. (2016). Understanding the Sparse Vector Technique for Differential Privacy. *Proceedings of the VLDB Endowment , 10* (6), 637-648.

Ulugbek S. Kamilov, H. M. (2016). Learning optimal nonlinearities for iterative thresholding algorithms. *IEEE Signal Processing Letters , 23* (5), 1-9.