UNIVERZA V LJUBLJANI
FAKULTETA ZA MATEMATIKO IN FIZIKO

Matematika – 1. stopnja

Luka Lodrant

O geometriji diferencirane zasebnosti

Delo diplomskega seminarja

Mentor: doc. dr. Aljoša Peperko

Ljubljana, 2018

Kazalo

O geometriji diferencirane zasebnosti

POVZETEK

On the Geometry of Differential Privacy

ABSTRACT

Math. Subj. Class. (2010): 52-02
Ključne besede: diferencirana zasebnost
Keywords: differential privacy

1. Uvod

2. Priprava splošnega okolja

$\updownarrow$ Če želimo rigurozno analizirati zasebnost podatkov moramo najprej postaviti okolje v katerem bomo lahko to počeli.

Podatkovno bazo bomo predstavili kot vektor $x \in \mathbb{R}^n$, *poizvedbo* na taki podatkovni bazi pa z linearno kombinacijo členov x. Natančneje lahko d poizvedb združimo v linearno preslokavo $F : \mathbb{R}^n \to \mathbb{R}^d$, kjer omejimo vse koeficiente v $d \times n$ matriki na interval $[-1, 1]$. V celotnem delu bomo predpostavili tudi $d \leq n$.

Mehanizem bo v tem primeru naključen algoritem, ki kot vhod vzame podatkovno bazo $x \in \mathbb{R}^n$ ter poizvedbo $F : \mathbb{R}^n \to \mathbb{R}^d$, vrne pa razultat v obliki $a \in \mathbb{R}^d$. Tak mehanizem lahko analitik uporablja za izvajanje analiz na podatkovni bazi x. Neformalno bi tak mehanizem bil diferencirano zaseben, če bi se za dve dovolj podobni podatkovni bazi odgovori razlikovali za multiplikativen faktor največ $exp(\varepsilon)$. Tu je ε parameter, ki pove, kako močno zaseben je obravnavani mehanizem. Manjši ε pomeni višjo zasebnost. *Napaka* takega algoritma je pričakovana evklidska razdalja med pravilni odgovorom $F(x)$ in dejanskim odgovorom a.

Omenjeni naključen algoritem je tak algoritem, ki v svojem delovanju uporabi stopnjo naključnosti. Razultati z istimi vhodnimi podatki je zato načeloma različen vsakič, ko ta algoritem izvedemo. Eden najosnovnejši primerov takega algoritma je metoda Monte Carlo.

Definicija 2.1. *Podatkovna baza* = vektor v Rn

Definicija 2.2. *poizveda* = linearna preslikava

3. Verjetnosta mera

V naslednjem razdelku želimo še bolj strogo definirati pojme iz prejšnjega razdelka, a bomo za to potrebovli nekaj dodatnih teoretičnih osnov iz teorije mere. Tukaj bomo navedli le definicije uporabljenih pojmov, vse uporabljene izreke in leme pa bomo navedli sproti.

Definicija 3.1. Naj bo Ω neprazna množica. Družino podmožic $\mathcal{F}$ množice Ω imenujemo *σ-algebra*, če ima naslednje tri lastnosti:

(1) $\Omega \in \mathcal{F}$
(2) za vsako podmnožico $S \in \mathcal{F}$ je tudi $S^c \in \mathcal{F}$
(3) za vsako števno družino $\{F_i : i \in \mathbb{N}\}$ elementov iz $\mathcal{F}$ je tudi unija $\bigcup_{i \in \mathbb{N}} F_i$ v $\mathcal{F}$

Definicija 3.2. Elemente družine $\mathcal{F}$ imenujemo *merljive množice*, par $(\Omega, \mathcal{F})$ pa imenujemo *merljiv prostor*.

Definicija 3.3. *Pozitivna mera* (ponavadi kar *mera*) na merljivem prostoru $(\Omega, \mathcal{F})$ je funkcija:

$$\mu : \mathcal{F} \to [0, \infty],$$

ki zadošča pogojema

(1) $\mu(\emptyset) = 0$
(2) $\mu(\bigcup_{n=1}^{\infty} F_n) = \sum_{n=1}^{\infty} \mu(F_n)$

za vsako zaporedje disjunktnih množič $F_n \in \mathcal{F}$. Trojko $(\Omega, \mathcal{F}, \mu)$ bomo imenovali prostor z mero.

Definicija 3.4. Meri μ pravimo *verjetnostna mera*, če velja $\mu(\Omega) = 1$.

4. Diferencirana zasebnost

Definicija 4.1. *Mehanizem* M je družina verjetnostnih mer $M = \{\mu_x : x \in \mathbb{R}^n\}$, kjer je vsak μ_x definiran na $\mathbb{R}^d$.

Definicija 4.2. Mehanizem je *ε-diferencirano zaseben*, če za vse $x, y \in \mathbb{R}^n$ za katere je $\|x - y\|_1 \leq 1$ velja $sup_{S \subseteq \mathbb{R}^d} \frac{\mu_x(S)}{\mu_y(S)} \leq exp(\varepsilon)$, kjer supremum teče čez vse merljive podmnožice $S \subseteq \mathbb{R}^n$.

Pogosta je tudi šibkejša oblika ε-diferencirane zasebnosti.

Definicija 4.3. Mehanizem je *δ-približno ε-diferencirano zaseben*, če za vse $x, y \in \mathbb{R}^n$ za katere je $\|x - y\|_1 \leq 1$ velja $\mu_x(U) \leq exp(\varepsilon)\mu_y(S) + \delta$.

Za obravnavo diferencirano zasebnih mehanizmov sta pomembna tudi pojma napake in občutljivosti.

Definicija 4.4. *Napako* mehanizmo M po normi ℓ befiniramo kot $err_\ell(M, F) = \sup_{x \in \mathbb{R}^n} \mathbb{E}_{a \sim \mu_x}(\|a - Fx\|_\ell)$. Tu je kot prej $F : \mathbb{R}^n \to \mathbb{R}^d$.

Definicija 4.5. Mehanizem je *δ-približno ε-diferencirano zaseben*, če za vse $x, y \in \mathbb{R}^n$ za katere je $\|x - y\|_1 \leq 1$ velja $\mu_x(U) \leq exp(\varepsilon)\mu_y(S) + \delta$.

5. Spodnja meja prek ocene volumna

Definicija 5.1. Množica točk $Y \subseteq \mathbb{R}^d$ se imenuje *r-pakiranje* , če je $\|y - y'\|_2 \geq r$ za vsak $y, y' \in Y, y \neq y'$.

Trditev 5.2. *Naj bo $K \subseteq \mathbb{R}^d$ in $R = Vol(K)^{1/d}$. Potem K vsabuje $\Omega(R\sqrt{d})$-pakiranje velikosti vsak* $\exp(d)$.

Izrek 5.3. *Naj bo $\varepsilon > 0$, $F : \mathbb{R}^n \to \mathbb{R}^d$ linearna preslikav in $K = FB_1^n$. Potem ima vsak ε-zaseben mehanizem M napako vsaj $err(M, F) \geq \Omega(\varepsilon^-1d\sqrt(d) \cdot Vol(k)^{1/d}$.*

Dokaz. Naj bo $\lambda \geq 1$ in $R = Vol(K)^{1/d}$. □

6. K-norma mehanizem

To je

Slovar strokovnih izrazov

diferencirana zasebnost differential privacy
približna diferencirana zasebnost approximate differential privacy
podatkovna baza database
poizvedba query
naključen algoritem random algorithm

Literatura

[1] M. Hardt in K. Talwar, *On the Geometry of Differential Privacy*, 9.11. 2009, dostopno na `https://arxiv.org/abs/0907.3754`.

[2] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, L. Zhang, *Deep Learning with Differential Privacy*, 25. 10. 2016, dostopno na `https://arxiv.org/abs/1607.00133`.

[3] B. Magajna, *Osnove teorije mere*, Podiplomski seminar iz matematike 27, DMFA – založništvo, Ljubljana, 2011