差分プライバシー, 数式メモ

Hirotsugu Seike

2025 年 8 月 12 日

1 導入

差分プライバシーでは, データベースへの問合せに擾乱 (random noise) を付加 (perturbation) することで, 問合せした結果である数値を確率的に振舞わせる. この処理を乱択関数 (randomized function) K を用いて実施する. 例えば, データベースを D で表現し, データベース D から決定的に出力が定まる関数を f とすれば, 乱択関数は次のように表せる.

$$K(D) = f(D) + n. \tag{1}$$

ここで, n は平均 0, 分散 σ^2 の正規分布に従う確率変数である ($n \sim \mathrm{N}(0, \sigma^2)$). ノイズ項 n により, $K(D)$ も確率変数として取り扱えるため, 確率質量関数 (または確率密度関数) が定義可能である. 確率密度関数を ρ で定義し, 2 つのデータベース D, D', その問合せ結果 t を考えた時, 次の不等式が成立すれば, それを $\epsilon-$ 区別困難性 (ϵ-indistinguishable) [1] と呼ぶ.

$$\frac{\rho(K(D) = t)}{\rho(K(D') = t)} \leq \exp(\epsilon). \tag{2}$$

例えば, $\epsilon = 1/100$ の場合, $\exp(\epsilon) \approx 1.0100...$ となる. つまり, $\exp(\epsilon)$ を用いることで, 任意の問合せ結果 t における確率密度関数の近さの最悪値を, 見積もることができるようになる. このようにプライバシーの問題を, 確率密度関数を用いた数学の問題に定式化することで, 証明可能なプライバシー (Provable Privacy) [2] を実現することができる.

1.1 ラプラス機構 (Laplace Mechanism) による ϵ-差分プライバシーの実装

上記では, 一つの問合せ結果 t に関する差分プライバシーを定量的に評価した. より一般的な複数の問合せに対応するため, ランダムノイズが付加された問合せ結果である関数 K の出力の部分空間 S を考えることで, 以下のように (2) 式をアップデートできる.

$$\frac{\mathrm{P}r[K(D) \in S]}{\mathrm{P}r[K(D') \in S]} \leq \exp(\epsilon). \tag{3}$$

(3) の関係式が成立することを, $\epsilon-$ 差分プライバシー (ϵ-Differential Privacy) が保証されると呼ぶ [3]. 任意の部分空間 S (全てのデータベースへの問合せ) について, (3) 式が成立すれば, 常に ϵ-区別困難になるということである.

ここで, 尺度母数 b が等しく, 期待値 μ のみが異なる 2 つのラプラス分布 (平均は μ_1, μ_2) の確率密度関数の比を考える. ラプラス分布の確率密度関数は (4) 式で与えられる. 一般性が失われな

いため, $\mu_1 < \mu_2$ とする. 絶対値で 3 つの区間に場合分けする (図 1) と, 各区間の確率密度関数の比 $R(x)$ は (5) 式で与えられる.

$$f(x \mid \mu, b) = \frac{1}{2b} \exp\left(-\frac{|x - \mu|}{b} \right), \quad -\infty < x < \infty. \tag{4}$$

$$R(x) = \frac{f(x \mid \mu_1, b)}{f(x \mid \mu_2, b)} = \begin{cases} \exp\left(\frac{\mu_2 - \mu_1}{b}\right), & x < \mu_1, \\ \exp\left(\frac{2x - \mu_1 - \mu_2}{b}\right), & \mu_1 \le x \le \mu_2, \\ \exp\left(\frac{\mu_1 - \mu_2}{b}\right), & x > \mu_2. \end{cases} \tag{5}$$

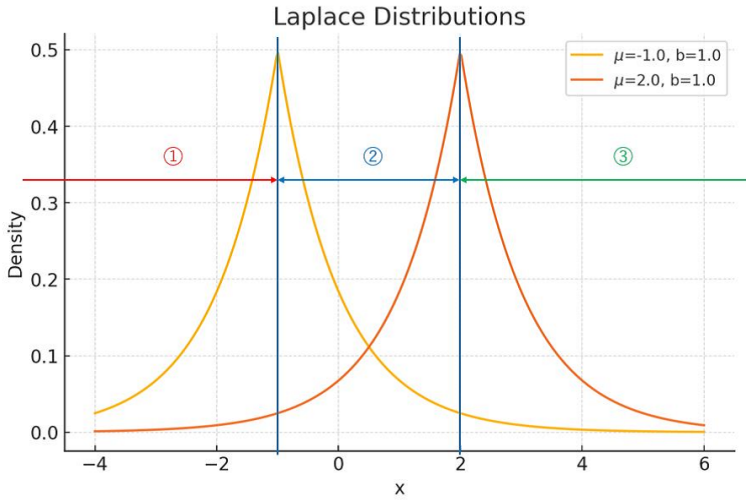

図 1: Laplace distributions with $\mu_1 = -1.0$, $\mu_2 = 2.0$, and $b = 1.0$.

(5) 式の値をプロットした結果を図 2 に示す. ラプラス分布の比の上限は $\exp(|\mu_2 - \mu_1|/b)$ で与えられていることが分かる.

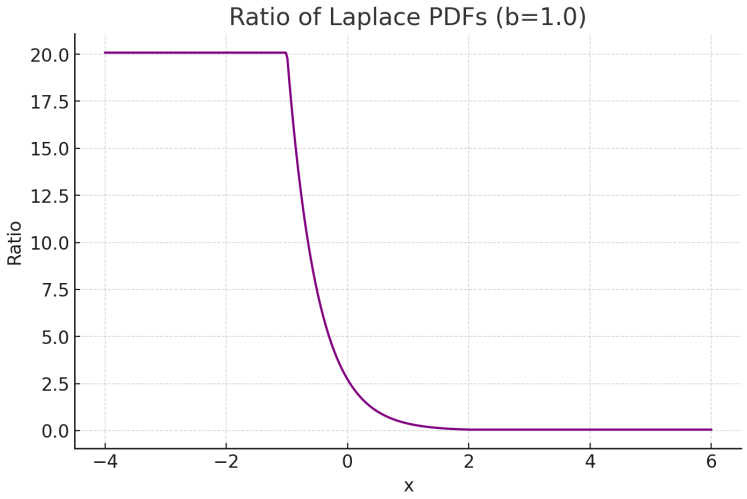

図 2: The ratio of two Laplace distributions with $\mu_1 = -1.0$, $\mu_2 = 2.0$, and $b = 1.0$.

従って, 以上の議論より, (1) 式に従って, 2 つのデータセット D, D' から出力される $f(D), f(D')$ として, その各成分に平均 0, 尺度母数 b のラプラス分布に従うノイズを付与すると, それら 2 つの確率変数の確率密度関数の比は, 次の関係式を満たす.

$$\frac{f(x \mid f(D), b)}{f(x \mid f(D'), b)} \leq \exp(|f(D) - f(D')|/b) \tag{6}$$

ここで, 考えられる全ての D, D' について, 最大となる $|f(D) - f(D')|$ を (6) 式の右辺に代入すれば, それが全ての D, D' に対して差分プライバシーを保証する上限となる. これを, ラプラス機構における関数 f の感度 (Sensitivity) Δf と呼び, 次式で定義される ($\mathcal{D}$ は考慮するデータベースの全体集合).

$$\Delta f \ = \ \max_{D,\, D' \in \mathcal{D}} \left| f(D) - f(D') \right| \tag{7}$$

n 次元ベクトルの出力 $f(D)$ に対して, ラプラス機構を適用させる場合, 次のような L_1 距離の最大値を考えればよい.

$$\Delta f \ = \ \max_{D,\, D' \in \mathcal{D}} \left|\left| f(D) - f(D') \right|\right|_1 \tag{8}$$

逆に言えば, 厳密にラプラス機構を適用させるためには, (8) 式を計算する必要がある. しかし, データセットの外れ値を考慮すると, それが困難なケースも存在する. データセットの生成モデルを考え, あまり生じない問合せである出力 $f(D)$ を無視する近似的な差分プライバシー (Approximate Differential Privacy) という手法も存在する.

$$\Pr[K(D) \in S] \leq \exp(\epsilon) \times \Pr[K(D') \in S] + \delta. \tag{9}$$

これは, 確率 $\Pr[K(D) \in S] < \delta$ について, 差分プライバシーを考慮しない手法であり, (ϵ, δ)-差分プライバシーと呼ばれる.

参考文献

[1] C. Dwork, "A firm foundation for private data analysis," in *Comm. ACM 54(1)*, pp.86-95, 2011.

[2] C. Dwork, "Differential Privacy," in *Proc. ICALP'06*, pp.1-12, 2006.

[3] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating Noise to Sensitivity in Private Data Analysis," in *Proc. TCC*, pp.265-284, 2006.

[4] 中島震 「AI リスク・マネジメント: 信頼できる機械学習ソフトウェアへの工学的方法論」丸善出版, 令和 4 年.