Подготовил: Безносиков А.Н.

Домашнее задание 1 по курсу «Дифференциальная приватность»

Дедлайн: 19 ноября 18:59.

Задание 1. Равномерный шум

Пусть мы хотим сделать приватным следующий алгоритм от датасета $X = \{x_i\}_{i=1}^n$:

$$f(X) = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

где $x_i \in \{0; 1\}$. На занятии мы уже обсуждали, как сделать такой алгоритм приватным с помощью Лапласовского шума:

$$A(X) = f(X) + \xi,$$

где $\xi \sim \text{Lap}(\alpha)$.

- Как следует выбрать α, чтобы алгоритм A был ε-DP?
- Возьмем вместо распределения Лапласа равномерное распределение $U[-\alpha; \alpha]$. Можно ли подобрать α так, чтобы A был ε-DP? Если да, то найдите как можно меньшее α.

Задание 2. Групповая приватность

В определении ε-DP важным является понятие соседних датасетов. В классическом варианте соседние датасеты отличаются ровно на один объект. Пусть теперь k-соседние датасеты отличаются на $1 \leq k$ объектов. Покажите, что если для $k = 1$ алгоритм A является ε-DP, то для произвольного k верно:

$$\mathbb{P}\{A(D) \in E\} \leq \exp(k\varepsilon) \cdot \mathbb{P}\{A(D') \in E\},$$

где E – любое множество исходов алгоритма A, а датасеты D и D' являются k-соседними.

Задание 3. Сравнение экспоненциального подхода и подхода зашумленного максимума

На занятии мы изучили два подхода к DP в случае, когда на множестве ответов задана некоторая функция порядка/стоимости. Это экспоненциальный подход и подход зашумленного максимума. В теории эти подходы во многом эквивалентны. Сравните эти подходы на практике:

- Выберите один из предложенных на лекции примеров: с выборами, ценой на товар или предложите свой. Придумайте и реализуйте случайную генерацию одного из этих сюжетов. Необязательно генерировать цены/голоса равномерно, можно сразу внести некоторое смещение.
- Реализуйте два подхода: экспоненциальный и зашумленного максимума, для выбранной задачи.

- Сравните два этих подхода в зависимости от уровня приватности ε. Какой/какие критерий/критерии сравнения будете использовать? Стоит ли для каждого критерия повторят эксперимент много раз и строить доверительные интервалы? Сравнительный анализ лучше представить в виде графиков. Сделайте вывод.

Задание 4. Сравнение Лапласовского и Гауссовского шума

Сгенерируйте датасет $X = \{x_i\}_{i=1}^n$, где $x_i \in \{0;1\}^d$. Рассмотрим уже привычный алгоритм

$$f(X) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Данный алгоритм можно сделать приватным с помощью добавления шума:

$$A(X) = f(X) + \xi.$$

Цель данного задания – сравнить два шума с распределениями Лапласа и нормальным распределением.

- Укажите параметры шумов в зависимости от ε (положите $\delta = \frac{1}{n}$ для нормального шума), чтобы оба подхода были эквивалентны с точки зрения DP. Обратите внимание на важный факт с занятия – как определяется чувствительность для этих шумов.

- Сравните данные шумы с точки зрения теории в зависимости от размера задачи d, уровня приватности ε и количества данных n. Какой/какие критерий/критерии сравнения будете использовать?

- Сравните данные шумы с точки зрения теории в зависимости от размера задачи d, уровня приватности ε и количества данных n. Какой/какие критерий/критерии сравнения будете использовать?

- Реализуйте два подхода для описанной задачи. Саму задачу сгенерируйте случайно.

- Сравните два подхода на практике в зависимости от размера задачи d, уровня приватности ε и количества данных n. Какой/какие критерий/критерии сравнения будете использовать? Стоит ли для каждого критерия повторят эксперимент много раз и строить доверительные интервалы? Сравнительный анализ лучше представить в виде графиков. Обязательно исследуйте случаи, когда $d = 1$ и $d \geq 100$. Базово предлагается построить графики зависимости критерия, который выбран, от n при фиксированных d и ε. Но никто не ограничивает в проведении других исследований. Сделайте вывод.