Advanced Topics in Machine Learning 2024

Rasmus Pagh

Home Assignment 2

Deadline: 23.59, Thursday September 26, 2024

The assignments must be submitted individually – each student must write and submit a personal solution but we do not prevent you from discussing ***high-level*** *ideas in small groups. If you use any LLM tool such as ChatGPT, please specify the* ***purpose*** *and* ***manner*** *in which you have utilized it.*

We are interested in how *you solved the problems, and not just in the final answers. Please explain your solutions and ideas as clearly as you can.*

Late Penalty and multiple Submissions *Late submissions will incur a penalty of 10% of the total marks for every hour of delay (rounded up) with a maximum allowed delay of 5 hours after which the submission server will close. If you submit multiple submissions, only the last submission will be considered relevant both for grading answers as well as late penalty.*

Submission format: *Please upload your answers in a single* `.pdf` *file. If you have created code please include at least the key parts in the PDF as well as a link to the full code (on Github, Colab, or similar).*

Suppose we are given a dataset of n vectors in d dimensions where each vector has Euclidean norm at most 1. Two datasets $\mathbf{x}$ and $\mathbf{x}'$ are neighboring if they differ in at most one of the n inputs. We consider the following private version of the k-means algorithm that starts with random clusters and iteratively updates the cluster centers in t rounds. Parameters σ and σ' are used to control the level of privacy. We use $[n]$ to denote the set $\{1,\ldots,n\}$, and let $\|v\|_2$ denote the Euclidean norm of vector v.

Input: $\mathbf{x} = x_1,\ldots,x_n \in \mathbf{R}^d$ where $\forall i : \|x_i\|_2 \leq 1$
Parameters: $k \in [n]$, $t \in \mathbf{N}$, $\sigma, \sigma' > 0$

1. Randomly sample disjoint sets $S_1^{(0)},\ldots S_k^{(0)} \subseteq [n]$ such that $\sum_{i=1}^k n_i^{(0)} = n$ where $n_i^{(0)} = |S_i^{(0)}|$
2. For $\ell = 1,\ldots,t$ do the following for $i = 1,\ldots,k$:
   (a) Let $c_i^{(\ell)} = \frac{1}{\max(1,n_i^{(\ell-1)})}\left(z_{\ell,i} + \sum_{j\in S_i^{(\ell-1)}} x_j\right)$ where $z_{\ell,i} \sim \mathcal{N}(0,\sigma^2)^d$
   (b) Let $S_i^{(\ell)}$ be the set of indices $j \in [n]$ for which (among $c_1^{(\ell)},\ldots,c_k^{(\ell)}$) the vector $c_i^{(\ell)}$ is the closest to x_j (in case of ties, choose the lower index i or use any other tiebreaking rule)
   (c) Let $n_i^{(\ell)} = |S_i^{(\ell)}| + z'_{\ell,i}$ where $z'_{\ell,i} \sim \mathcal{N}(0,\sigma'^2)$
3. Output $\mathcal{M}_{\text{means}}(\mathbf{x}) = c_1^{(t)},\ldots,c_k^{(t)}$

Some differential privacy facts. The notion of ρ-zero-concentrated differential privacy (ρ-zCDP) due to Bun and Steinke will be used. The following facts about zCDP suffice for answering this hand-in.

Proposition 1 (ε-DP to zCDP)**.** *If $\mathcal{M}$ satisfies ε-differential privacy then $\mathcal{M}$ satisfies $(\frac{1}{2}\varepsilon^2)$-zCDP.*

Definition 2 (Sensitivity)**.** *$f : \mathcal{X}^n \to \mathbb{R}^d$ has (L_2)* sensitivity Δ *if for all $\mathbf{x},\mathbf{x}' \in \mathcal{X}^n$ differing in a single entry, we have $\|f(\mathbf{x}) - f(\mathbf{x}')\|_2 \leq \Delta$.*

Proposition 3 (Gaussian Mechanism)**.** *Let $f : \mathcal{X}^n \to \mathbb{R}^d$ have sensitivity Δ. The mechanism $\mathcal{M}(\mathbf{x}) = f(\mathbf{x}) + z$ where $z \sim \mathcal{N}(0,\sigma^2)^d$ satisfies $(\Delta^2/2\sigma^2)$-zCDP.*

Lemma 4 (Composition). *Let $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ and $\mathcal{M}' : \mathcal{X}^n \times \mathcal{Y} \to \mathcal{Z}$ be randomized algorithms that satisfy ρ-zCDP and ρ'-zCDP, respectively. (For $\mathcal{M}'$ the neighboring relation is changing the first part of the input in one coordinate, inputs that differ in $\mathcal{Y}$ are not neighboring.) Then $\mathcal{M}''(\mathbf{x}) = \mathcal{M}'(\mathbf{x}, \mathcal{M}(\mathbf{x}))$ satisfies $(\rho + \rho')$-zCDP. In particular, if $\mathcal{M}'(\mathbf{x}, y) = p(y)$ for some "post-processing" function p then $\mathcal{M}''(\mathbf{x})$ satisfies ρ-zCDP.*

1 Analyzing $\mathcal{M}_{\text{means}}$ (60 points)

To be able to refer to intermediate results in the algorithm we define $f_\ell : \mathcal{X}^n \to (\mathbb{R}^d)^k$ and $g_\ell : \mathcal{X}^n \to \mathbb{R}^k$ where, for $i \in [k]$, $f_\ell(\mathbf{x})_i = \sum_{j \in S_i^{(\ell-1)}} x_j$ and $g_\ell(\mathbf{x})_i = |S_i^{(\ell)}|$. Note that if we interpret $f_\ell(\mathbf{x})$ as a vector in $\mathbb{R}^{dk}$ we have $\|f_\ell(\mathbf{x})\|_2^2 = \sum_{i=1}^k \|f_\ell(\mathbf{x})_i\|_2^2$. Also consider the noise vectors $z_\ell = (z_{\ell,1}, \ldots, z_{\ell,k})$ and $z'_\ell = (z'_{\ell,1}, \ldots, z'_{\ell,k})$

(a) Consider $\mathcal{M}'_{\text{means}}$, a variant of mechanism $\mathcal{M}_{\text{means}}$ that changes the last step (3) to output $f_\ell(\mathbf{x}) + z_\ell$ and $g_\ell(\mathbf{x}) + z'_\ell$ for all $\ell \in [t]$. Argue that if $\mathcal{M}'_{\text{means}}$ satisfies ρ-zCDP then $\mathcal{M}_{\text{means}}$ satisfies ρ-zCDP.

(b) To analyze iteration ℓ of $\mathcal{M}'_{\text{means}}$ let be $S_1^{(\ell-1)}, \ldots, S_k^{(\ell-1)}$ be any collection of disjoint sets (i.e., we ignore that these sets in fact depend on the input $\mathbf{x}$). Argue that for neighboring $\mathbf{x} \sim \mathbf{x}'$ we have $\|f_\ell(\mathbf{x}) - f_\ell(\mathbf{x}')\|_2 \leq 2$.

(c) Argue that $\mathcal{M}'_{\text{means}}$ satisfies ρ-zCDP with $\rho = 2t/\sigma^2 + t/\sigma'^2$.

2 Implementing $\mathcal{M}_{\text{means}}$ (40 points)

The *cost* of a clustering with centers $c_1, \ldots, c_k$ is defined as the average squared Euclidean distance between points and their nearest cluster center, i.e., $\frac{1}{n} \sum_{j_1}^n \min_{i \in [k]} \|c_i - x_j\|_2^2$. We have provided you with a code template that loads a set of points, runs a non-private version of k-means, and plots the cost as a function of the number of iterations.

(a) Modify the code such that it implements the algorithm $\mathcal{M}_{\text{means}}$ given above with one change: Instead of taking σ and σ' as parameters, it should take a single privacy parameter ρ and automatically choose suitable parameters such that the algorithm satisfies ρ-zCDP.

(b) Plot the cost of the private clustering obtained by $\mathcal{M}_{\text{means}}$ when varying ρ between 0.001 and 1 and using $k = 5$ clusters, $t = 5$ iterations. (It is ok to compute the cost exactly based on the cluster centers provided by $\mathcal{M}_{\text{means}}$ and the input vector $\mathbf{x}$ even though this is not a private computation.)

Good luck!

Rasmus