Journal

Hugo Simon
hugo.simon@telecom-paris.fr

1 Introduction

Let $\mathcal{X} = \{x_1, \dots, x_n\}$ be a set of n datapoints. Let Θ be a space of parameters, and θ an element of Θ. We consider cost functions of the form:

$$L(\theta) = \sum_{x \in \mathcal{X}} f(x, \theta)$$

Let $\mathcal{S} = \{x_{s_1}, \dots, x_{s_m}\}$ be a subset of $\mathcal{X}$ (possibly with repetitions). To each element $x \in \mathcal{S}$, associate a weight $\omega(x) \in \mathbb{R}^+$. Define the estimated cost associated to the weighted subset $\mathcal{S}$ as:

$$\hat{L}(\theta) = \sum_{x \in \mathcal{S}} \omega(x) f(x, \theta).$$

Definition 1.1 (Coreset)**.** Let $\varepsilon \in]0, 1[$. The weighted subset $\mathcal{S}$ is a ε-coreset for L if, for any parameter θ, the estimated cost is equal to the exact cost up to a relative error:

$$\forall \theta \in \Theta \quad \left| \frac{\hat{L}(\theta)}{L(\theta)} - 1 \right| \leq \varepsilon$$

An important consequence of the coreset property is the following

$$(1 - \varepsilon) L\left(\theta^{\text{opt}}\right) \leq (1 - \varepsilon) L\left(\hat{\theta}^{\text{opt}}\right) \leq \hat{L}\left(\hat{\theta}^{\text{opt}}\right) \leq \hat{L}\left(\theta^{\text{opt}}\right) \leq (1 + \varepsilon) L\left(\theta^{\text{opt}}\right)$$

See Bachem et al. 2017.

2 Variance argument

2.1 Multinomial case

Multinomial case $\mathcal{S} \sim \mathcal{M}(m, q)$ i.e. m independent categorical sampling where $\mathbb{P}(x_i) = q(x_i)$

$$\operatorname{Var}[\hat{L}(\theta)] = \frac{1}{m} \operatorname{Var}\left[\frac{f_\theta(x)}{q(x)}\right] = \frac{1}{m} \sum_{x \in \mathcal{X}} \frac{f_\theta(x)^2}{q(x)} - \frac{1}{m} L(\theta)^2$$

For any query $\theta \in \Theta$, the variance is reduced to 0 by

$$q_\theta(x) := \frac{f_\theta(x)}{\sum_{x' \in \mathcal{X}} f_\theta(x')}$$

2.2 DPP case

DPP case where $\mathcal{S} \sim \mathcal{DPP}(K)$, $\pi_i := K_{ii}$. We have

$$\operatorname{Var}[\hat{L}(\theta)] = \sum_{i,j} \mathbb{E}\left[\varepsilon_i \varepsilon_j\right] \frac{f_\theta(x_i) f_\theta(x_j)}{\pi_i \pi_j} - L(\theta)^2 \quad \text{with} \quad \mathbb{E}\left[\varepsilon_i \varepsilon_j\right] = \begin{cases} \det\left(\mathrm{K}_{ij}\right) = \pi_i \pi_j - \mathrm{K}_{ij}^2, & \text{if } i \neq j \\ \mathbb{E}\left[\varepsilon_i\right] = \pi_i, & \text{if } i = j \end{cases}$$

Introducing $\Pi = \mathrm{diag}(\pi)$ and $\tilde{K} = \Pi^{-1} K^{\odot 2} \Pi^{-1}$, we can rewrite

$$\mathbb{V}\mathrm{ar}[\hat{L}(\theta)] = \sum_i \left(\frac{1}{\pi_i} - 1\right) f_\theta(x_i)^2 - \sum_{i \neq j} \frac{\mathrm{K}_{ij}^2}{\pi_i \pi_j} f_\theta(x_i) f_\theta(x_j) = f_\theta^\top (\Pi^{-1} - \tilde{K}) f_\theta$$

For a Bernoulli process where $\mathbb{P}(x_i \in \mathcal{S}) = \pi_i$ independently, $K = \Pi$ then $\tilde{K} = I$. The DPP variance beats uniformly the Bernoulli process variance if $\tilde{K}$ dominates the identity i.e.

$$\forall f_\theta,\ \mathbb{V}\mathrm{ar}[\hat{L}_K(\theta)] < \mathbb{V}\mathrm{ar}[\hat{L}_\Pi(\theta)] \iff \tilde{K} > I$$

But $\tilde{K}$ is a symmetric positive definite matrix and by Hadamard inequality $\det(\tilde{K}) \leq \prod_i \tilde{K}_{ii} = 1$. Therefore at least one of its eigenvalue is lower than 1, hence $\tilde{K} \not> I$.

3 Sensitivity

Definition 3.1 (Sensitivity)**.** The sensitivity σ_i of a datapoint $\boldsymbol{x}_i$ and the total sensitivity $\mathfrak{S}$ of $\mathcal{X}$ are

$$\begin{cases} \sigma_i = \sup_{\theta \in \Theta} q_\theta(x_i) = \sup_{\theta \in \Theta} \frac{f_\theta(\boldsymbol{x}_i)}{L(\theta)} & \in [0,1] \\ \mathfrak{S} = \sum_{i=1}^n \sigma_i \end{cases}$$

Let s be an upper bound on sensitivity σ i.e. $\forall i, s_i \geq \sigma_i$, and $S := \sum_{i=1}^n s_i$. Furthermore, let sample $\mathcal{S} \sim \mathcal{M}(m, s/S)$, the multinomial sampling case. Define $g_\theta(x_i) := \frac{q_\theta(x_i)}{s_i} \quad \in [0,1]$

By Hoeffding's inequality, we thus have for any $\theta \in \Theta$ and $\varepsilon' > 0$

$$\mathbb{P}\left[\left|\mathbb{E}\left[g_\theta(x)\right] - \frac{1}{m} \sum_{x \in \mathcal{S}} g_Q(x)\right| > \varepsilon'\right] \leq 2 \exp\left(-2m\varepsilon'^2\right).$$

By definition, $\mathbb{E}\left[g_\theta(x)\right] = \frac{1}{S}$ and $\frac{1}{m} \sum_{x \in C} g_Q(x) = \frac{\mathrm{cost}(C,Q)}{S\ \mathrm{cost}(\mathcal{X},Q)}$. As such, for any $Q \in \mathcal{Q}$

$$\mathbb{P}\left[\left|\ \mathrm{cost}(\mathcal{X}, Q) - \mathrm{cost}(C, Q)\right| > \varepsilon' S\ \mathrm{cost}(\mathcal{X}, Q)\right] \leq 2 \exp\left(-2m\varepsilon'^2\right)$$

Hence, the set C satisfies the coreset property in (2.2) for any single query $Q \in \mathcal{Q}$ and $\varepsilon > 0$ with probability at least $1 - \delta$, if we choose

$$m \geq \frac{S^2}{2\varepsilon^2} \log \frac{2}{\delta}$$

4 SGD Paper

References

Bachem, Olivier et al. (2017). *Practical Coreset Constructions for Machine Learning*. DOI: `10.48550/ARXIV.1703.06476`. URL: `https://arxiv.org/abs/1703.06476`.