Leverage Score Sampling for Function Fitting

Aarshvi Gajjar

NYU Tandon
Classical Regression Problem

- Given a data matrix $X \in \mathbb{R}^{n \times k}$ and a vector representing labels, $y \in \mathbb{R}^n$, the least squares objective is to find a vector $w^*$ such that:

$$w^* = \arg \min_{w \in \mathbb{R}^k} \|Xw - y\|^2_2$$

- Sometimes, it is expensive to get access to all the labels

- So instead we sample $m \ll n$ rows from $X$ using a sampling matrix $S \in \mathbb{R}^{m \times n}$ and we hope that the problem is approximately solved

- Formally, if $\tilde{w} = \arg \min_{w \in \mathbb{R}^k} \|SXw - Sy\|^2_2$, then, we want:

$$\|X\tilde{w} - y\|^2_2 \in (1 \pm \varepsilon)\|Xw^* - y\|^2_2$$
Naive Method

- Sample the rows uniformly with replacement over $[n]$ and solve the Empirical Risk Minimizer

$$\min_{\mathbf{w} \in \mathbb{R}^k} \frac{1}{m} \sum_{i=1}^{m} (\mathbf{X}_i \mathbf{w} - y_i)^2$$

- From law of large numbers, the Monte Carlo estimate converges to the expected loss.
- However, the variance of this estimator can be very high
- If one row is orthogonal to all others, then it has to be included in the sample, making $m$ very large
Importance Sampling

- Method to emphasize the **important** data points such that the variance of the estimator is reduced.
- Suppose the rows of $X$ are samples generated according to the probability distribution, $p$.
- And, it is expensive to sample from $p$
- Basic Idea: Generate samples from another distribution which is easy to sample from, encourages the important data points and controls the variance.
If $q$ is a probability density function such that $q(X_iw) > 0$ wherever $p(X_iw)(X_iw - y_i)^2 > 0$, then the importance sampling algorithm involves generating $m \ll n$ samples according to $q$

The estimator is:

$$\frac{1}{m} \sum_{i=1}^{m} \frac{p(X_iw)}{q(X_iw)}(X_iw - y_i)^2$$

Regression problem is to minimise (1) over $w \in \mathbb{R}^k$

Question: How to choose $q$?
Leverage Score Sampling

• Define a score for each point being sampled

Definition

The leverage score, \( l(i) \) of the \( i \)th row of a matrix \( X \in \mathbb{R}^{n \times k} \) is:

\[
l(i) := \max_{\beta \in \mathbb{R}^k} \frac{(X\beta)_i^2}{\|X\beta\|_2^2}
\]

• Sample points proportional to the score
Known Results

**Theorem**

Given a data matrix, $X \in \mathbb{R}^{n \times k}$ and a vector $y \in \mathbb{R}^n$. Let $w^* = \arg \min_{w \in \mathbb{R}^k} \|Xw - y\|_2$. For any $\varepsilon < 1$, suppose $S$ is a sampling matrix that selects $m = O \left( d \log d + \frac{d}{\delta \varepsilon} \right)$ rows of $X$ via leverage score sampling.

- Let $\tilde{w} = \arg \min_{w \in \mathbb{R}^k} \|SXw - Sy\|_2$
- Then, w.p. $\geq 1 - \delta$,

$$\|Xw^* - y\|_2^2 \in (1 \pm \varepsilon)\|X\tilde{w} - y\|_2^2$$

- Thus leverage score sampling is powerful for active regression
- What if we want to approximately solve $\min_{w \in \mathbb{R}^k} \|p(Xw) - y\|_2$, where $p$ is a polynomial of degree $d$?
Generalisation for Active Linear Regression

- Suppose we are given access to $s$ samples of a function $g : \mathbb{R}^k \to \mathbb{R}$.
- Let $\mathcal{F}$ be a function class containing functions $f$ that map $\mathbb{R}^k$ to $\mathbb{R}$.
- Let $p$ be some density over $\mathbb{R}^k$.
- We want to find a function $\tilde{f} \in \mathcal{F}$ such that:

$$\int_{\mathbb{R}^k} (\tilde{f}(x) - g(x))^2 p(x) dx \in (1 \pm \varepsilon) \min_{f \in \mathcal{F}} \int_{\mathbb{R}^k} (f(x) - g(x))^2 p(x) dx$$

where $\varepsilon > 0$ is fixed.
Sensitivity and Sampling

Definition (General Leverage Scores (Sensitivity))

Let $\mathcal{F}$ be a family of functions, $f : \mathbb{R}^k \rightarrow \mathbb{R}$ and let $p$ be a probability density over $\mathbb{R}^k$. The leverage score of any $x \in \mathbb{R}^k$ is given by

$$
\tau_{\mathcal{F}}(x) = \sup_{f \in \mathcal{F}} \frac{f(x)^2 p(x)}{\int_{x \in \mathbb{R}^k} f^2(x) p(x) dx}
$$

- The total sensitivity, $T_{\mathcal{F}} = \int_{\mathbb{R}^k} \tau_{\mathcal{F}}(x) d(x)$ represents the number of samples required to fit a function.
Our work

- We aim to find an upper bound on the total sensitivity of function classes of high dimensional functions
- Example. ReLU, polynomials
- Finally applying this to get sample complexity for nonlinear active regression