Leverage Score Sampling for Function Fitting

Aarshvi Gajjar

NYU Tandon

Classical Regression Problem

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

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

- Sometimes, it is expensive to get access to all the labels
- So instead we sample $m \ll n$ rows from $\mathbf{X}$ using a sampling matrix $\mathbf{S} \in \mathbb{R}^{m \times n}$ and we hope that the problem is approximately solved
- Formally, if $\tilde{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathbb{R}^k} \|\mathbf{S}\mathbf{X}\mathbf{w} - \mathbf{S}\mathbf{y}\|_2^2$, then, we want:

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

Naive Method

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

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

- 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 $\mathbf{X}$ 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.

Importance Sampling Defined

- If q is a probability density function such that $q(\mathbf{X}_i\mathbf{w}) > 0$ wherever $p(\mathbf{X}_i\mathbf{w})(\mathbf{X}_i\mathbf{w} - \mathbf{y}_i)^2 > 0$, then the importance sampling algorithm involves generating $m \ll n$ samples according to q
- The estimator is:

$$\frac{1}{m}\sum_{i=1}^{m}\frac{p(\mathbf{X}_i\mathbf{w})}{q(\mathbf{X}_i\mathbf{w})}(\mathbf{X}_i\mathbf{w} - \mathbf{y}_i)^2 \tag{1}$$

- Regression problem is to minimise (1) over $\mathbf{w} \in \mathbb{R}^k$
- Question: How to choose q?

Leverage Score Sampling

- Define a score for each point being sampled

Definition

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

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

- Sample points proportional to the score

Known Results

Theorem

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

- *Let* $\tilde{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathbb{R}^k} \|\mathbf{S}\mathbf{X}\mathbf{w} - \mathbf{S}\mathbf{y}\|_2$
- *Then, w.p.* $\geq 1 - \delta$,

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

- Thus leverage score sampling is powerful for active regression
- What if we want to approximately solve $\min_{\mathbf{w} \in \mathbb{R}^k} \left\|p(\mathbf{X}\mathbf{w}) - \mathbf{y}\right\|_2$, 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, $f : \mathbb{R}^k \to \mathbb{R}$ and let p be a probability density over $\mathbb{R}^k$. The leverage score of any $\mathbf{x} \in \mathbb{R}^k$ is given by

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

- The total sensitivity, $\mathrm{T}_{\mathcal{F}} = \int_{\mathbb{R}^k} \tau_{\mathcal{F}}(\mathbf{x}) d(\mathbf{x})$ 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