Influence Sampling for Generalized Linear Models

Jinzhu Jia

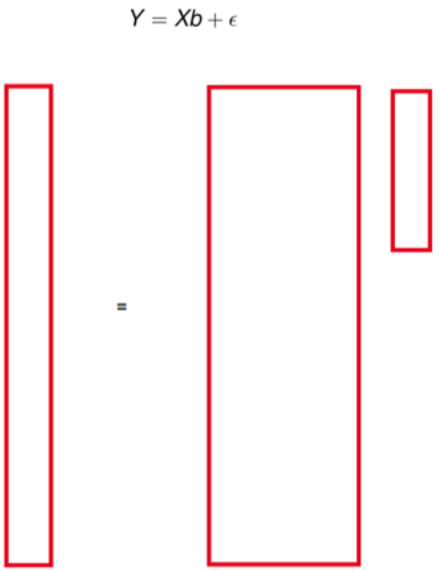
Peking Universtiy

June 17 2014

Joint work with Michael Mahoney, Petros Drineas and Bin Yu.

Motivations

Consider a very BIG problem:



One way to estimate the coefficient b is via least squares:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 = \arg\min_{\beta} \| Y - X\beta \|_2^2$$

Motivations (2)

Sampling and reweight:

$$\tilde{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (w_i (y_i - x_i^T \beta))^2 = \arg\min_{\beta} \|W(Y - X\beta)\|_2^2$$

Goal: To design a "good" sampling scheme, such that the distance between $\hat{\beta}$ and $\tilde{\beta}$ is small "enough" !

Leverage score sampling for least squares

There are many works on this sampling scheme. See for example:

- Drineas, Mahoney, and Muthukrishnan (2006)
- Sarlos (2007); Drineas, Mahoney, Muthukrishnan, and Sarlos (2007)
- Drineas, Magdon-Ismail, Mahoney, and Woodruff (2011)

Quick review of Leverage score sampling

Definition (Leverage scores)

Given any orthonormal basis U for the range(X) (where recall X is of size $n \times p$ with $n \gg p$), the *leverage scores* of X are the squared ℓ_2 norms of U's rows:

$$L_i = \|U_{(i)}\|_2^2, i = 1, \ldots, n.$$

Sampling Scheme:

- Let $(w_1, w_2, \ldots, w_n) = 0$
- for $k = 1, 2, \ldots, r$
- randomly select $i_k \in \{1, 2, \ldots, n\}$ from multi-nominal distribution with parameters $(p_1, p_2, \ldots, p_n)$, where $p_i \propto L_i$.
- Update $w_{i_k} = \frac{1}{\sqrt{r p_{i_k}}}$

$$\tilde{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} [w_i(x_i^T \beta - y_i)]^2 := \arg\min_{\beta} \|W(Y - X\beta)\|_2^2, \qquad (1)$$

Quick review of Leverage score sampling

We use n to denote the number of observations (rows) and p for the number predictors (features, columns).

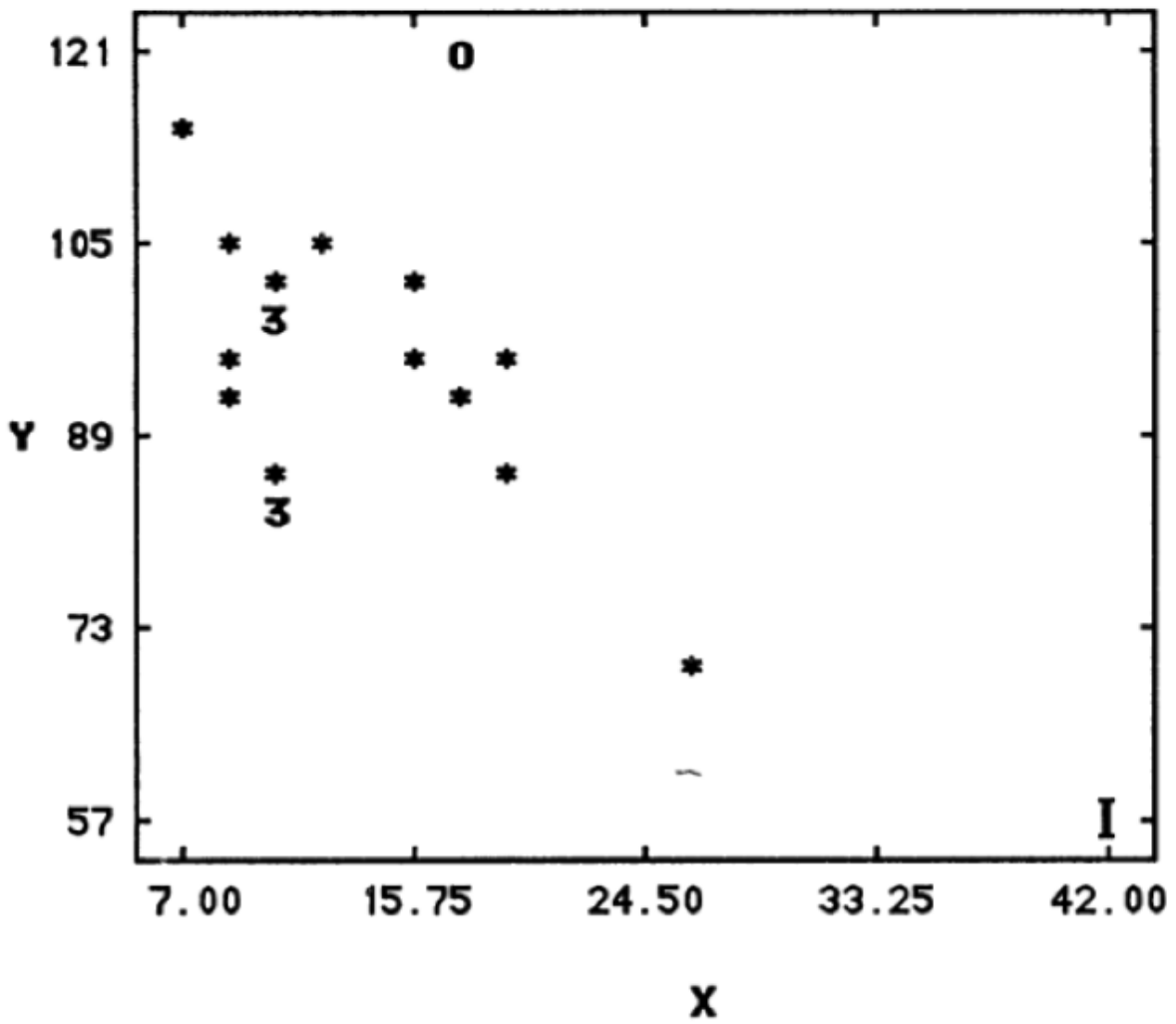
Theorem (Drineas et al (2011))

With leverage score sampling, if $r = O(\frac{p}{\epsilon}\log(np))$, *then with probability greater than 0.8,*

$$\|Y - X\hat{\beta}\|_2^2 \leq \|Y - X\tilde{\beta}\|_2^2 \leq (1+\epsilon)\|Y - X\hat{\beta}\|_2^2$$

Ref: P.Drineas, M. W. Mahoney, S. Muthukrishnan, and T. Sarlos, Faster Least Squares Approximation,Numerische Mathematik, 117(2), pp. 217-249, 2011.

Leverage score or Influence score?



Influence score or Leverage score ?

$$IF_i = L_i + \frac{\hat{e}_i^2}{\sum_{i=1}^{n} \hat{e}_i^2},$$

where $\hat{e}_i$ is the residue of least square problems.

Lemma

The influence score is equal to the leverage score of the extended matrix $Z = [X, Y]$.

Ref: Samprit Chatterjee and Ali S Hadi. Influential observations, high leverage points, and outliers in linear regression. Statistical Science, 1(3):379 - 393, 1986.

Influence sampling scheme

We consider the following exact r-sampling procedure.

- Let $(w_1, w_2, \ldots, w_n) = 0$
- for $k = 1, 2, \ldots, r$
- randomly select $i_k \in \{1, 2, \ldots, n\}$ from multi-nominal distribution with parameters $(p_1, p_2, \ldots, p_n)$, where $p_i \propto IF_i$.
- Update $w_{i_k} = \frac{1}{\sqrt{r p_{i_k}}}$

$$\tilde{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} [w_i(x_i^T \beta - y_i)]^2 := \arg\min_{\beta} \|W(Y - X\beta)\|_2^2 \qquad (2)$$

Theoretical justification

Theorem

For any $0 < \epsilon < 1$ *and* $0 < \delta < 1$. *If*

$$r \geq \frac{96(p+1)}{\epsilon^2} \log\left(\frac{96(p+1)}{\epsilon^2\sqrt{\delta}}\right),$$

then with probability bigger than $1 - 2\delta$,

$$\|Y - X\tilde{\beta}\|_2 \leq \sqrt{\frac{1+\epsilon}{1-\epsilon}}\|Y - X\hat{\beta}\|_2,$$

$$\|\tilde{\beta} - \hat{\beta}\|_2 \leq \sqrt{\frac{2\epsilon}{(1-\epsilon)\sigma_{\min}(X)}}\|Y - X\hat{\beta}\|_2.$$

Beyond Least squares

- L_1 and $Lq(1 < q < 2)$ Regression [DDHKM09, CDMMMW13, CW 2013]

$$\hat{\beta} = \arg\min_{\beta} \|Y - X\beta\|^q$$

- Quantile regression [CDMMMMW13]

$$\rho_\tau(z) = \begin{cases} \tau z, & z \geq 0; \\ (\tau - 1)z, & z < 0. \end{cases}$$

$$\text{minimize}_{x \in \mathbb{R}^d} \quad \rho_\tau(Y - X\beta), \tag{3}$$

where $\rho_\tau(y) = \sum_{i=1}^n \rho_\tau(y_i)$, for $y \in \mathbb{R}^n$.

GLMs

A generalized linear model (GLM for short) consists of three elements:

- A probability distribution class of $Y|X$ with $Y \in \mathbb{R}$ and $X \in \mathbb{R}^p$

$$f(y;\theta,\psi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\psi)} + c(y;\psi)\right\}, \tag{4}$$

- A linear predictor $\eta = X\beta$
- A link function g such that $g(E(Y)) = \eta = X\beta$

Ref: John A Nelder and RJ Baker. Generalized linear models. Wiley Online Library, 1972.

Definition (Canonical Link Function)

For $Y \sim f(y;\theta,\psi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\psi)} + c(y;\psi)\right\}$, The link function $g(\cdot)$ is called the canonical link function, if $g(E(Y)) = \theta$.

$$E(Y) = \dot{b}(\theta) := \frac{d}{d\theta} b(\theta).$$

This result can be found from [John A Nelder and RJ Baker. Generalized linear models. Wiley Online Library, 1972.]

MLE for a GLM

the loglikelihood function is

$$\sum_{i=1}^{n}\left[\frac{y_i x_i^T\beta - b(x_i^T\beta)}{a(\psi)} + c(y_i;\psi)\right].$$

When ψ does not depends on β, we have the following property.

Lemma

The MLE (maximizied likelihood estimation) of β in a GLM with a canonical link function is:

$$\hat{\beta} = \arg\max_{\beta}\sum_{i=1}^{n}\left[y_i x_i^T\beta - b(x_i^T\beta)\right]. \tag{5}$$

Remarks:

- When $b(t) = t^2/2$, the above estimate is least squares estimate.
- When $b(t) = \log(1+e^t)$, the above estimate is logistic regression estimate.
- When $b(t) = e^t$, the above estimate is Possion regression estimate.

Sampling for a GLM

$$\tilde{\beta} = \arg\max_{\beta} \sum_{i=1}^{n} w_i \left[y_i x_i^T \beta - b(x_i^T \beta) \right], \tag{6}$$

Let $X = UR$, with $U^T U = I_{p\times p}$. If a row sampling matrix (with rescaling) $S \in \mathbb{R}^{r\times n}$ satisfies the following two conditions:

$$\|U^T S^T S\theta - U^T \theta\|_2 \leq \epsilon \|\theta\|_2/2, \text{ for any deterministic } \theta \in \mathbb{R}^{n\times 1}; \tag{7}$$

$$\sigma_{\min}(U^T S^T S U) \geq 1/2 \tag{8}$$

AND the following strong convexity condition holds

$$\text{For any } \beta, \frac{d^2 b(t)}{dt^2}|_{t=x_i^T\beta} \geq c > 0.$$

Define $W = S^T S$. Let $\hat{\beta}$ and $\tilde{\beta}$ are two estimators defined in Equations (5) and (6). Then

$$\|\hat{\beta} - \beta^*\|_2 \leq \frac{\epsilon}{c\sigma_{\min}(X)} \|Y - \hat{Y}\|_2.$$

Iteratively weighted least squares

Let $g(\beta) = \sum_{i=1}^{n}[y_i x_i^T \beta - b(x_i^T \beta)]$, then

$$\begin{aligned}
\beta^{(t+1)} &= \beta^{(t)} - [\ddot{g}(\beta^{(t)})]^{-1}\dot{g}(\beta^{(t)}) \\
&= \beta^{(t)} + (X^T B X)^{-1} X^T (Y - \hat{Y}^{(t)}) \\
&= (X^T B X)^{-1}[(X^T B X)\beta^{(t)} + X^T (Y - \hat{Y}^{(t)})] \\
&= (X^T B X)^{-1} X^T B[X\beta^{(t)} + B^{-1}(Y - \hat{Y}^{(t)})] \\
&= \arg\min_{\beta} \sum_{i=1}^{n} B_{ii}(\tilde{y}_i - x_i^T \beta)^2,
\end{aligned}$$

where $\tilde{Y}^{(t)} = X\beta^{(t)} + B^{-1}(Y - \hat{Y}^{(t)})$.

Sampling for weighted Least squares

Definition (Weighted leverage scores)

Given any orthonormal basis U for the range(X) (where recall X is of size $n \times p$ with $n \gg p$), the *weighted leverage scores* of X are the weighted squared ℓ_2 norms of U's rows: $\|w_i U_{(i)}\|_2^2$, $i = 1, \ldots, n$.

Define

$$p_i = \frac{\|w_i U_{(i)}\|_2^2}{\|WU\|_F^2}.$$

We consider the following exact r-sampling procedure.

- Let $(s_1, s_2, \ldots, s_n) = 0$
- for $k = 1, 2, \ldots, r$
- randomly select $i_k \in \{1, 2, \ldots, n\}$ from multi-nominal distribution with parameters $(p_1, p_2, \ldots, p_n)$.
- Update $s_{i_k} = \frac{1}{\sqrt{r p_{i_k}}}$

Let S be the sampling and re-scaling matrix, that is, S is a diagonal matrix with $S_{ii} := s_i$ defined in the above procedure. We have the following results

Theoretical Justification

Theorem

Let $X = UR$ *with* $U^T U = I_{p\times p}$. *Assume that*

$\sigma_{\min}(X), \sigma_{\max}(X), \sigma_{\min}(WX)$ *and* $\sigma_{\max}(WX)$*are all bounded from both and*

Suppose that the following conditions hold.

$$\|U^T W^T S^T SW(Y - X\hat{\beta})\|_2 \leq \epsilon \|Y - X\hat{\beta}\|_2. \tag{9}$$

$$\|U^T W^T S^T SWU - U^T W^T WU\| \leq \sigma^2_{\min}(WX)/(2\sigma^2_{\max}(X)) \tag{10}$$

We have

$$\|\tilde{\beta} - \beta\|_2 \leq \frac{\epsilon \|Y - X\hat{\beta}\|_2}{\sigma^2_{\min}(WX)/(2\sigma_{\max}(X))}$$

Theoretical Justification

Lemma

If $r > \frac{100p}{\epsilon^2}$, *then with probability at least 0.9,*

$\|U^T W^T S^T SW\theta - U^T W^T W\theta\|_2 \leq \epsilon \|\theta\|_2$, *for any given deterministic* $\theta \in \mathbb{R}^n$

Lemma (Concentration for Spectral Norm.)

For any $0 < \epsilon < 1$ *and* $0 < \delta < 1$. *If*

$$r \geq \frac{96p}{\epsilon^2} \log\left(\frac{96p}{\epsilon^2\sqrt{\delta}}\right).$$

we have

$$P(\|U^T W^T S^T SWU - U^T W^T WU\| \geq \epsilon) \leq \delta,$$

where $\|\cdot\|$ *denotes the spectral norm of matrix.*

Conclusions

- Influence sampling for Least squares
- Leverage sampling for GLMs
- weighted Leverage sampling for GLMs

Thanks!