Centro de Investigación en Matemáticas, A.C.

Optimización

Tarea 7

José Miguel Saavedra Aguilar

Abstract

In this homework we perform a logistic regression on the MNIST dataset for number 1 images via Stochastic gradient descent and Stochastic modified Newton's method.

1 Introduction. The logistic regression

Let $x_1, \dots x_n \in \mathbb{R}^m$ and $y_i \in \{0, 1\}$ be a dataset. We define the logistic regression function:

$$h(\beta, \beta_0) = -\frac{1}{n}\sum_{i=1}^{n} y_i \log(\pi_i) + (1 - y_i)\log(1 - \pi_i)$$

$$\pi_i(\beta, \beta_0) = \frac{1}{1 + \exp(-\beta_0 - \beta^\top x_i)}$$

We say β, β_0 fit the dataset best, if they are the minimizer the function h. Note the gradient of the logistic regression is given by

$$\nabla h(\beta, \beta_0) = \frac{1}{n}\sum_{i=0}^{n}(\pi_i - y_i)z_i$$

where $z_i \in \mathbb{R}^{785}$ is given by

$$z_i = \begin{bmatrix} 1 \\ x_i \end{bmatrix}$$

The hessian matrix of the logistic regression function is then given by

$$\nabla^2 h = \frac{1}{n}\sum_{i=0}^{n}\pi_i(1 - \pi_i)z_i z_i^\top$$

1.1 Stochastic gradient

Given n samples of $x_i \in \mathbb{R}^m$ and a decomposable function:

$$h(z) = \frac{1}{n}\sum_{i=1}^{n} f(z, x_i)$$

where $f(z, x_i)$ are denoted loss functions (the loss of z on the sample x_i. For this kind of problems a descent directions algorithm may take a long time to compute the gradient (and the Hessian) for each iteration of the algorithm. For this reason, one may turn to Stochastic gradient descent methods. In this methods, one chooses a subsample $x_{i_1}, \dots, x_{i_k}$ with $k < n$ randomly each iteration and updates z with the rule:

$$z_{t+1} = z_t - \alpha_t \nabla h(z_t, x^{(i_k)})$$

where $\nabla h(z_t, x^{(i_k)})$ indicates the sum of the decomposable functions inside the gradient is over the subsample indices. For stability and convergence results of Stochastic Gradient Method (SGM) we suggest consulting [3]. We may also present a Newton's point based descent directions method by taking a subsampled Hessian:

$$H_t = \nabla^2 h(z_t, x^{(i_k)})$$

and following the update rule:

$$z_{t+1} = z_t - \alpha_t (\phi H_t + (1-\phi)\frac{\mathbb{I}}{k})^{-1} \nabla h(z_t, x^{(i_k)})$$

for some $\phi \in [0, 1)$ so $\phi B + (1-\phi)\frac{\mathbb{I}}{k}$ is symmetric and positive definite.

2 Algorithm

The stochastic methods we shall use are given by:

Algorithm 1: Stochastic gradient method with fixed step size.

```
Input: $f, z_0, x_1, \dots, x_n, \alpha$
Output: $z^*$
$k \leftarrow 0$;
while $\|\nabla f(x_k)\| > 0$ do
    Take a subsample of $x_1, \dots, x_n$ of size $k$;
    Compute $d_k = -\nabla h(z_t, x^{(i_k)})$;
    Update $x_{k+1} \leftarrow x_k + \alpha d_k$;
    $k \leftarrow k + 1$;
end
```

Algorithm 2: Stochastic modified Newton's method with fixed step size.

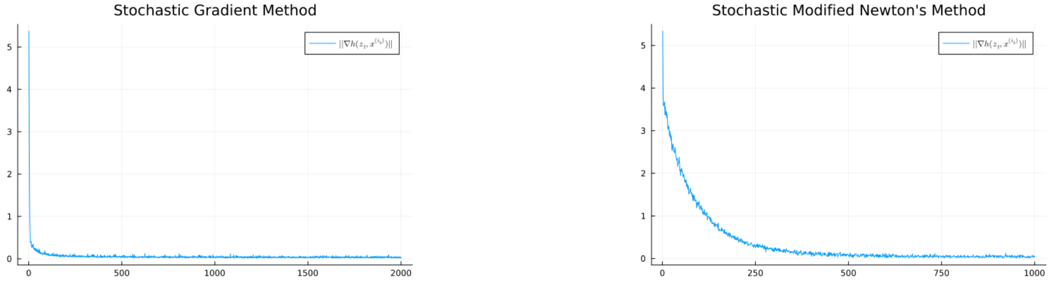
```
Input: $f, z_0, x_1, \dots, x_n, \alpha, \phi$
Output: $z^*$
$k \leftarrow 0$;
while $\|\nabla f(x_k)\| > 0$ do
    Take a subsample of $x_1, \dots, x_n$ of size $k$;
    Compute $B = \phi \nabla^2 h(z_t, x^{(i_k)}) + (1-\phi)\frac{\mathbb{I}}{k}$;
    Solve $B d_k = -\nabla h(z_t, x^{(i_k)})$;
    Update $x_{k+1} \leftarrow x_k + \alpha d_k$;
    $k \leftarrow k + 1$;
end
```

3 Results

Algorithms 1 and 2 were implemented in Julia[1]. We shall fit the data from the `MNIST` dataset's [2] training subset via logistic regression using this algorithms. Then, we test the fitment of the data on the testing subset of the dataset.
For the Stochastic Gradient Method, we take subsamples of size $k = 500$, $\alpha = 10^{-1}$ and 2000 iterations, meanwhile for the modified Newton based method we take subsamples of size $k = 500$, $\alpha = 10^{-2}$, $\phi = 0.9$ and 1000 iterations. We present the results on table 1 and plots of the subsampled gradient for both methods on figure 1.

Algorithm	1	2
Error	0.0105	0.0089
$\left\lVert \nabla h(z_t, x^{(i_k)}) \right\rVert$	0.02397	0.02190

Table 1: Error and $\left\lVert \nabla h(z_t, x^{(i_k)}) \right\rVert$ for the Stochastic Gradient Method and the Stochastic modified Newton's Method



(a) Stochastic Gradient Method

(b) Stochastic Newton's Method

Figure 1: Evolution of the logistic regression function gradient's norm $\left\lVert \nabla h(z_t, x^{(i_k)}) \right\rVert$

4 Results discussion and conclusions

We see both methods work well considering every sample is less than 1% of the total size of the dataset. The error for logistically fit β, β_0 is very good on both cases, with about 1% error on both cases. We may conclude the logistic regression is very efficient for the `MNIST` dataset.

References

[1] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, "Julia: A fresh approach to numerical computing," *SIAM Review*, vol. 59, no. 1, pp. 65–98, 2017. [Online]. Available: https://epubs.siam.org/doi/10.1137/141000671

[2] L. Deng, “The mnist database of handwritten digit images for machine learning research,” *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[3] M. Hardt, B. Recht, and Y. Singer, “Train faster, generalize better: Stability of stochastic gradient descent,” in *Proceedings of The 33rd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. F. Balcan and K. Q. Weinberger, Eds., vol. 48. New York, New York, USA: PMLR, 20–22 Jun 2016, pp. 1225–1234. [Online]. Available: https://proceedings.mlr.press/v48/hardt16.html