Leverage based sampling for classification

Julian Kopka Larsen and Jesper Løve Hinrich .
Superviser Lars Kai Hansen

DTU Compute · Technical University of Denmark
Kgs. Lyngby, Denmark

Technical University of Denmark

DTU Compute
Department of Applied Mathematics and Computer Science

Abstract

We validate the results of leverage based sampling for Least-squares regression, shown by Ma et al. [1]. We explore the possibility of using the leverage based sampling distribution from LS-regression on 2 class classification, and introduce a new approach for sampling from an leverage distribution (important points).

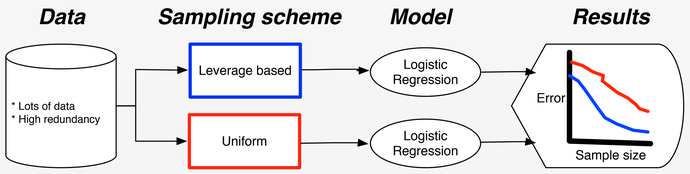
Motivation

For video the importance of sampling methods is initiated by very large and high-dimensional datasets where

- It is not feasible to use all of the available data.
- There is a high redundancy between datapoints (25 fps).
- Time taken in fitting a model is rarely linear to the input.

We therefore want to explore alternative sampling methods, and try to identify which datapoints are important when fitting a model.

Concept



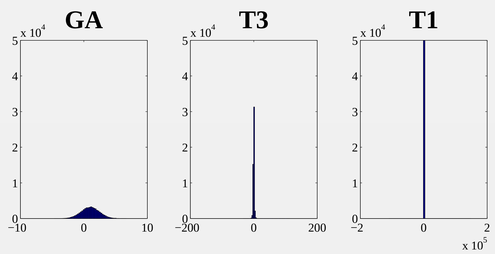
Research Questions

- Can we validate the results for least-squares regression shown by Ma et al. ?
- Will a linear regression based sampling distribution improve our performance in classification?
- Can leverage based sampling be generalized and used for classification?

Datasets

These datasets are drawn from distributions defined in Ma et al. [1] and characterised by

- GA: Nearly uniform leverage-scores
- T3: Mildly non-uniform leverage-scores
- T1: Very non-uniform leverage-scores



Figur 2: The three distributions considered.

Leveraging for least-squares regression

When fitting a model, we know that some datapoints are more important that others, leveraging is based on the idea that we can determine the importance of these point beforehand.

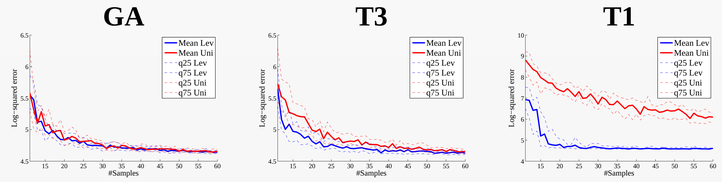
1. A leverage-score is calculated for each datapoint (its importance).
2. These scores are normalised into a distribution $\mathbf{H}$to sample from.

Ma. et al. [1] show that the leverage-scores for least-square regression is the diagonal elements of $\mathbf{H}$

$$\mathbf{H} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{y} \tag{1}$$

Validation of the results Ma et al.

We have empirically tested and validated the results shown by Ma et al. [1].



Figur 3: Comparison of uniform vs. leverage based sampling schemes for least-squares regression. $N = 1000$, $d = 10$.

- GA: The leverage score are approximately uniform, and thus there is no significant difference between the two sampling schemes.
- T3: Leveraging consistently provides slightly better results compared to uniform sampling.
- T1: With *very non-uniform* leverage-scores, leveraging clearly outperforms uniform sampling.

There results are consistent when varying N and d, although the level of improvement using leveraging compared to uniform varies.

LS-based Distribution

We want to generalise the leveraging concept, that is to find the effect that a datapoints class has on the predicted class for that datapoint:

$$\frac{\delta\hat{\mathbf{y}}_n}{\delta\mathbf{y}_n} \tag{2}$$

There is a closed form solution which is linear in y

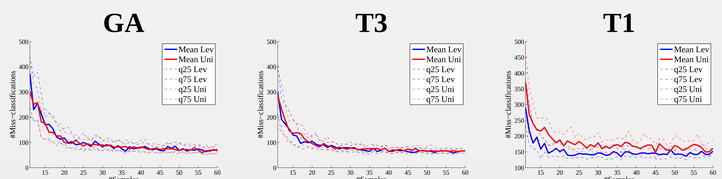
$$\hat{\beta}_{OLS} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \quad \text{where} \quad \hat{\mathbf{y}}_n = \mathbf{X}_n * \hat{\beta}$$

Therefore the leverage-score (2) is the coefficient

$$\frac{\delta\hat{\mathbf{y}}_n}{\delta\mathbf{y}_n} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$$

Test Results

We compared logistic regression when sampling from a LS-distribution (blue) vs. a uniform-distribution (red). The mean, 25th and 75th quantile are plotted.



Figur 4: LS- vs. uniform-distribution based sampling for logistic regression.

- A LS-distribution sample scheme is no better that uniform sampling on datasets of type GA and T3.
- With very non-uniform leverage scores, T1, the LS-distribution slightly outperforms uniform sampling.

The results shown are for dimension $p = 10$ and $N = 1000$ datapoints, but it is consistent when varying p and N.

Sensitivity Based Distribution

Vi generalisere til noget andet, og derfor kan WLS ikke bruges, s**r**a vi laver bare LS. The target is again (2)

$$\hat{\mathbf{y}}_n = p(y|\bar{\mathbf{x}_n}, \bar{\mathbf{w}}) \quad \bar{\mathbf{w}} \text{ s.t. } \frac{\delta L}{\delta\bar{\mathbf{w}}} = 0$$

Since depends both directly and indirectly on y we see that

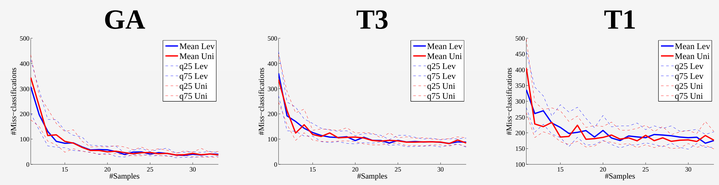
$$\frac{\delta}{\delta\mathbf{y}}\frac{\delta\mathcal{L}}{\delta\bar{\mathbf{w}}} = 0$$
$$\Downarrow$$
$$\frac{\delta^2\mathcal{L}}{\delta\mathbf{y}\delta\bar{\mathbf{w}}} + \frac{\delta^2\mathcal{L}}{\delta\bar{\mathbf{w}}\delta\bar{\mathbf{w}}^T}\frac{\delta\bar{\mathbf{w}}}{\delta\mathbf{y}} = 0$$

and from this we can get our leverage-score (2)

$$\frac{\delta\hat{\mathbf{y}}_n}{\delta\mathbf{y}_n} = \frac{\delta p(y|\bar{\mathbf{x}}_n, \bar{\mathbf{w}})}{\delta\bar{\mathbf{w}}^T}\frac{\delta\bar{\mathbf{w}}}{\delta\mathbf{y}} = -\frac{\delta p(y|\bar{\mathbf{x}}_n, \bar{\mathbf{w}})}{\delta\bar{\mathbf{w}}^T}\left[\frac{\delta^2\mathcal{L}}{\delta\bar{\mathbf{w}}\delta\bar{\mathbf{w}}^T}\right]^{-1}\frac{\delta^2\mathcal{L}}{\delta\mathbf{y}\delta\bar{\mathbf{w}}}$$

When evaluating this model, weights trained on a small training-set is used. This is expected to be better than LS-based sampling since it introduces dependence on class information.

Test results



Figur 5: Comparison of sensitivity vs. uniform -based sampling for logistic regression.

New questions??????

From our work several new question arise.

- How large show the initial sampling size be for sensitivity-based sampling?
- Should all points be sampled from the initial weights found, or should the process be iterative?

Conclusion

In the case of linear regression, leverage-based sampling provides a improvement over uniform sampling when the leverage-scores are mildly or very non-uniform.
Using the LS-based sampling for classification shows know improvement on datasets *GA* and *T3*, but for *T1* with very non-uniform leverage-scores, the approach is slightly better.
We have generalised the concept of leverage-based scores to classification, but for logistic regression it has not led to improvements.

References

Litteratur

[1] Ma et al. A statistical perspective on algorithmic leveraging. *arXiv:1306.5362v1 [stat.ME]*, June 2013.