LEVERAGE BASED SAMPLING FOR CLASSIFICATION

Julian Kopka Larsen *Jesper Løve Hinrich*

DTU Compute
Technical University of Denmark
Kgs. Lyngby, Denmark

ABSTRACT

Ma et al. [1] has shown leverage sampling to outperform uniform sampling for Least-Squares regression. We explore the possibility of using the same sampling distribution on binary classification, and introduce a new leverage distribution based on a generalization of the idea.

1. MOTIVATION

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

- It is not feasible to use all of the available data at once.
- There is a high redundancy between datapoints (frames in video).
- Computational cost is rarely linear to the input size.

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

2. RESEARCH QUESTIONS

- Can we validate the results for least-squares regression shown by Ma et al. [1]
- Will a linear regression based sampling distribution improve our performance in classification?
- Can leverage based sampling be generalized and used for classification?

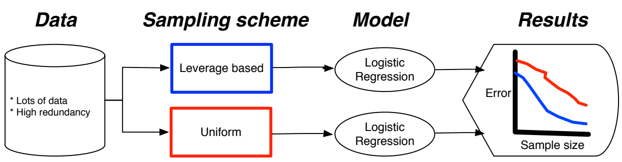

Fig. 1. The concept of leverage sampling

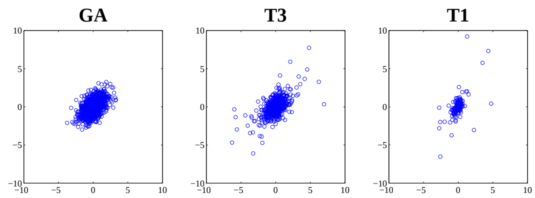

Fig. 2. The three distributions considered standardized for comparison

3. DATASETS

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

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

Samples from the three distributions are shown in **Fig. 2**

4. 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 a point beforehand and assign it a leverage-score to represent this.

1. A leverage-score is calculated for each datapoint.
2. These scores are normalized into a distribution π to sample from.

Ma. et al. [1] use the leverage-scores for least-square regression defined as the diagonal elements of (1)

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

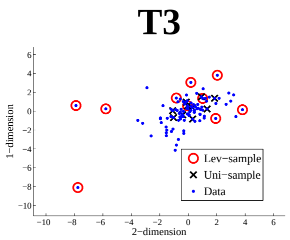

Fig. 3. Comparison of sampling methods

This comes from the closed form expression for predictions which is linear in y

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

After normalizing this to a probability distribution we can sample points that represent the structure better than a uniform sample. See **Fig. 3**

5. VALIDATION OF PREVIOUS RESULTS

We have empirically tested and validated the results shown by Ma et al. [1]. This is shown in **Fig. 4**

- 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.

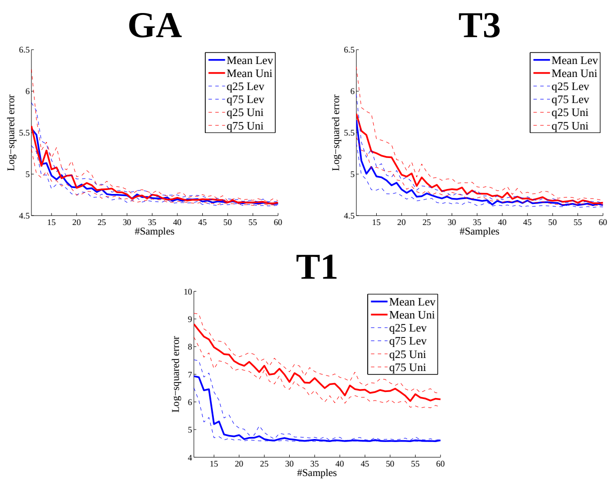

Fig. 4. Comparison of uniform **(red)** vs. leverage **(blue)** based sampling schemes for least-squares regression. $N = 1000$, $d = 10$.

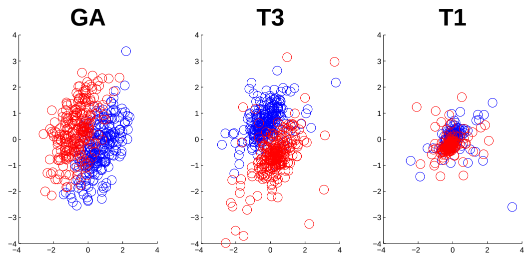

Fig. 5. The three distributions for binary classification standardized for comparison

There results are consistent when varying N and d, although the level of improvement varies.

6. DATASETS BINARY CLASSIFICATION

We define three datasets GA, T3 and T1, which resemble the ones from Ma et al. only for classification.

The data is generated by as for regression, but then split into two sets by adding a random unit vector scaled by the variance and a distance constant of 1.3. See Fig 5.

7. LEAST-SQUARES-BASED DISTRIBUTION FOR CLASSIFICATION

We sample from the same distribution (1) as for least-squares regression. We use these samples to train a logistic regression model for binary classification, with equal class size.

8. TEST RESULTS FOR LEAST-SQUARES-BASED SAMPLING

We compared the LS-distribution **(blue)** to a uniform-distribution **(red)** in sampling for a logistic regression. The mean, 25th and 75th quantile are plotted. See **Fig. 6**.

- Sampling from the LS-distribution is no better that uniform 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.

9. SENSITIVITY BASED DISTRIBUTION

We generalize the leverage scores to other models by seeing that they can be described as:

$$\frac{\delta \hat{\mathbf{y}}_n}{\delta \mathbf{y}_n} = Diag\left(H\right) \tag{2}$$

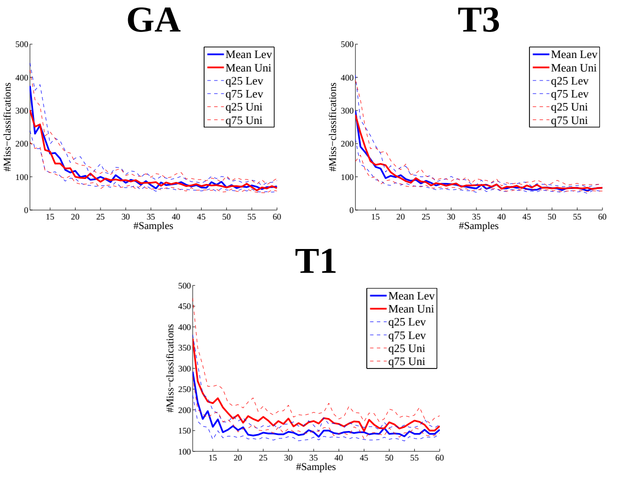

Fig. 6. Comparison of uniform **(red)** vs. leverage **(blue)** based sampling schemes for classification. $N = 1000$, $d = 10$.

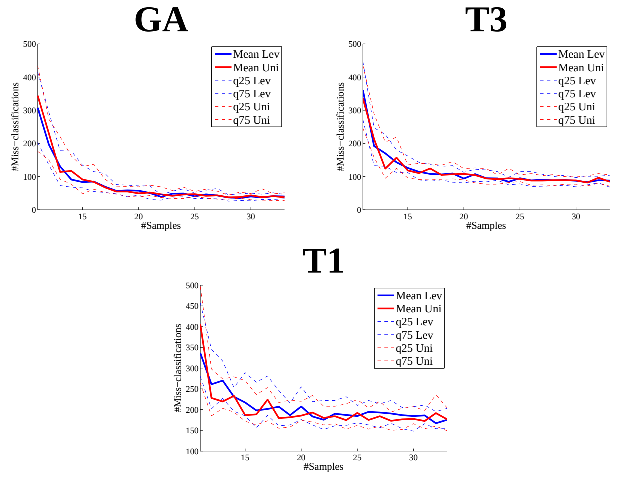

Fig. 7. Comparison of uniform **(red)** vs. sensitivity **(blue)** based sampling schemes for logistic regression.

Which we call the sensitivity of the model to a specific datapoint. For a general probabilistic discriminative model this requires the following:

$$\hat{\mathbf{y}}_n = p(y|\bar{\mathbf{x}_n}, \bar{\mathbf{w}}) \quad \bar{\mathbf{w}} \text{ s.t. } \frac{\delta L}{\delta \bar{\mathbf{w}}} = 0 \tag{3}$$

Since (3) depends both directly and indirectly on y we see that

$$\frac{\delta}{\delta \mathbf{y}} \frac{\delta \mathcal{L}}{\delta \mathbf{w}} = 0 \Rightarrow \frac{\delta^2 \mathcal{L}}{\delta \mathbf{y} \delta \bar{\mathbf{w}}} + \frac{\delta^2 \mathcal{L}}{\delta \bar{\mathbf{w}} \delta \bar{\mathbf{w}}^T} \frac{\delta \bar{\mathbf{w}}}{\delta \mathbf{y}} = 0 \tag{4}$$

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

$$\begin{aligned} \frac{\delta \hat{\mathbf{y}}_n}{\delta \mathbf{y}_n} &= \frac{\delta p(y|\bar{\mathbf{x}}_n, \bar{\mathbf{w}})}{\delta \bar{\mathbf{w}}^T} \frac{\delta \bar{\mathbf{w}}}{\delta \mathbf{y}} \\ &= -\frac{\delta p(y|\bar{\mathbf{x}}_n, \bar{\mathbf{w}})}{\delta \bar{\mathbf{w}}^T} \left[\frac{\delta^2 \mathcal{L}}{\delta \bar{\mathbf{w}} \delta \bar{\mathbf{w}}^T} \right]^{-1} \frac{\delta^2 \mathcal{L}}{\delta \mathbf{y} \delta \bar{\mathbf{w}}} \end{aligned} \tag{5}$$

When using this model, initial weights are found by fitting a small uniform sample. This is expected outperform LS-based sampling since it introduces dependence on class information.

10. SENSITIVITY FOR LOGISTIC REGRESSION

In the generalised sensitivity expression (5) we insert logistic regression and the log-likelihood function. Where we need the following expressions:

$$\begin{aligned} -\frac{\delta p(y|\bar{\mathbf{x}}_n, \bar{\mathbf{w}})}{\delta \bar{\mathbf{w}}^T} &= X_n \frac{e^{-X * w_0^T}}{e^{-X * w_0^T} + 1} \\ \left[\frac{\delta^2 \mathcal{L}}{\delta \bar{\mathbf{w}} \delta \bar{\mathbf{w}}^T} \right]^{-1} &= \sum_{n=1}^{N} t_n X_n^T X_n \frac{e^{-X_n w_0^T}}{e^{-X_n w_0} + 1} \\ &\quad + (1 - t_n) \frac{X_n^T X_n}{e^{-X_n w_0} + 1} \\ \frac{\delta^2 \mathcal{L}}{\delta \mathbf{y} \delta \bar{\mathbf{w}}} &= \sum_{n=1}^{N} X_n \end{aligned}$$

11. TEST RESULTS FOR SENSITIVITY BASED SAMPLING

We see that the *sensitivity based sampling* gives us a performance equivalently to that of uniform sampling. Shown in **Fig. 7**

12. FUTURE WORK

From our work several new question arise.

- How large show the initial sampling size be for sensitivity-based sampling?
- Should the non-linear sensitivity based leverage scores be transformed? and how?
- Should all points be sampled from the initial weights found, or should the process be iterative?

13. 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 is slightly better with very non-uniform leverage-scores, T1 data.

We have generalized the concept of leverage-based scores to classification with logistic regression and it has shown no improvements. However further analysis and tweaking might improved this approach.

14. REFERENCES

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