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 for use in classification?

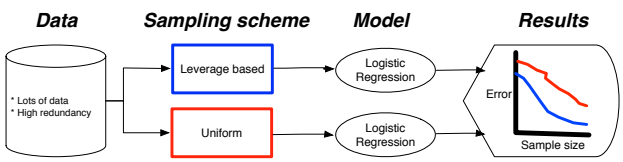

Fig. 1. The concept of leverage sampling

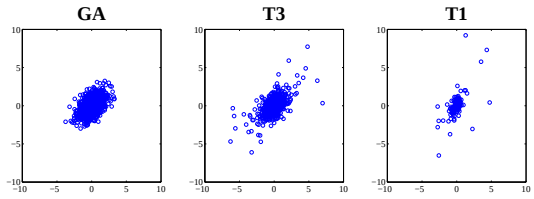

Fig. 2. The three distributions considered standardized for comparison

3. DATASETS

These datasets are drawn from distributions defined by 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)

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

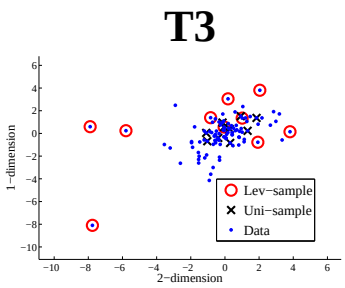

Fig. 3. Comparison of sampling methods

This comes from the closed form expression for predictions which is linear in $\mathbf{y}$

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

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.

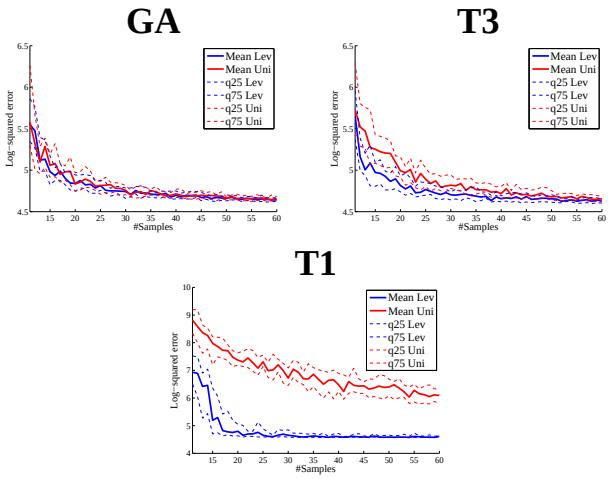

Fig. 4. Comparison of uniform **(red)** vs. leverage **(blue)** based sampling schemes for least-squares regression. $N = 1000$, $d = 10$.

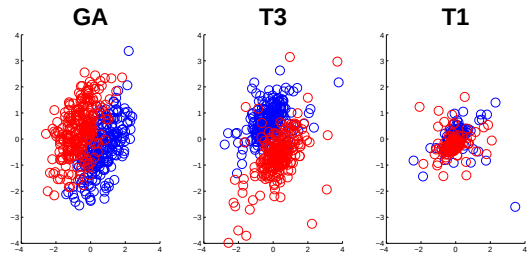

Fig. 5. The three distributions for binary classification standardized for comparison

- 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 varies.

6. DATASETS FOR BINARY CLASSIFICATION

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

The data is generated in the same way as for regression, but then split into two sets by adding a random vector to half of the datapoints. The vector added is a unit vector scaled by the covariance 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 used for least-squares regression. We use these samples to train a logistic regression model for binary classification.

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.

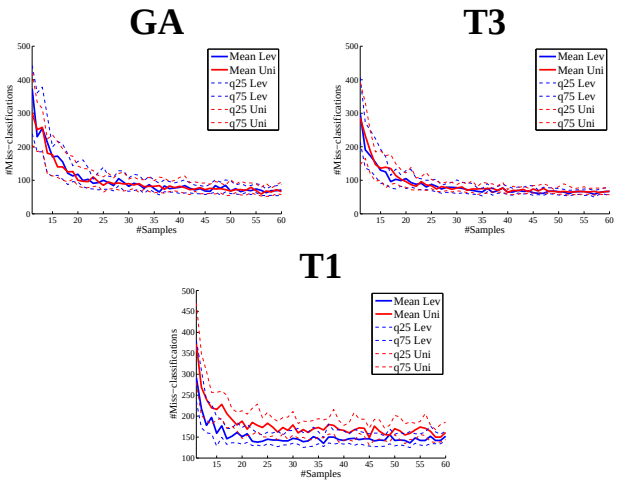

Fig. 6. Comparison of uniform **(red)** vs. leverage **(blue)** based sampling schemes for classification. $N = 1000$, $d = 10$.

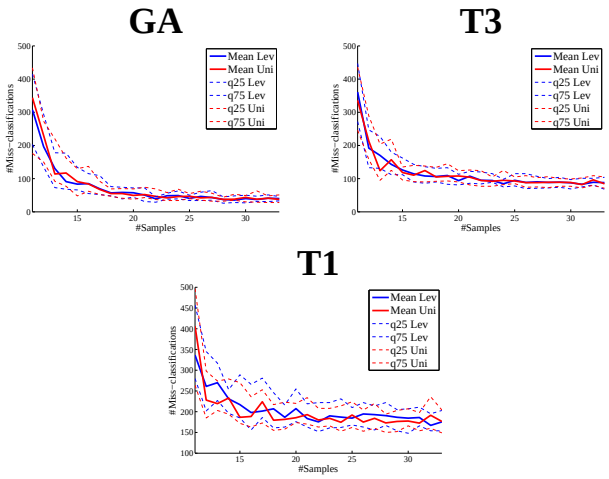

Fig. 7. Comparison of uniform **(red)** vs. sensitivity **(blue)** based sampling schemes for classification. $N = 1000$, $d = 10$.

9. SENSITIVITY BASED DISTRIBUTION

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

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

Which we call the sensitivity of the model w.r.t. 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

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

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

$$\begin{aligned} \frac{\delta \hat{\mathbf{y}}_n}{\delta \mathbf{y}_n} &= \frac{\delta p(y|\bar{\mathbf{x}}_n, \bar{\mathbf{w}})}{\delta \bar{\mathbf{w}}^T} \frac{\delta \bar{\mathbf{w}}}{\delta \mathbf{y}} \\ &= -\frac{\delta p(y|\bar{\mathbf{x}}_n, \bar{\mathbf{w}})}{\delta \bar{\mathbf{w}}^T} \left[\frac{\delta^2 \mathcal{L}}{\delta \bar{\mathbf{w}} \delta \bar{\mathbf{w}}^T} \right]^{-1} \frac{\delta^2 \mathcal{L}}{\delta \mathbf{y} \delta \bar{\mathbf{w}}} \end{aligned} \tag{4}$$

This can now be used to develop leveraging for any probabilistic discriminative method.

10. SENSITIVITY FOR LOGISTIC REGRESSION

In the generalised sensitivity expression (4) we insert the class probability and log-likelihood for logistic regression.

$$p(y|\bar{\mathbf{x}}_n, \bar{\mathbf{w}}) = \frac{1}{1 + e^{-X_n w^T}} = \hat{y}$$
$$\mathcal{L} = -\sum_{n=1}^{N} t_n \ln y_n + (1 - t_n) \ln(1 - y_n)$$

Where we need the following expressions:

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

This is unfortunately not a closed form solution so in order to implement this we need to first take a small uniform sample and based om this we can estimate w for use the expressions above. This is expected outperform LS-based sampling since it introduces dependence on class information.

An alternative to sensitivity was also developed.This method estimates the uncertainty of a prediction based on an second degree Taylor expansion of the likelihood around the optimal solution. But it provided similar results to the sensitivity based method.

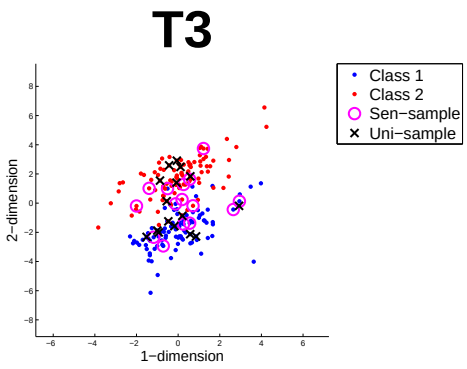

Fig. 8. Comparison of points sampled by uniform (crosses) to points sampled by a sensitivity (circles) based sample.

11. TEST RESULTS FOR SENSITIVITY BASED SAMPLING

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

In 2 dimensions with $N = 200$, we sample a total of 16 datapoints with the initial sampling size of 6 datapoints (3 from each class) and the rest from a sensitivity based sample. This is illustrated in Fig 8.

This is an example showing that it is of little importance whether one uses uniform or sensitivity based sampling. Although the points sampled from the sensitivity distribution are located more in the critical region between the two classes, it does not provide a benefit when classifying our data.

Normalizing is used to transform the scores into distributions in the shown results. Using a different transform for us the *soft-max* transformation

$$\pi_n = \frac{e^{\hat{\mathbf{y}}_n \cdot \mathbf{y}_n^{-1}}}{\sum_{j=1}^{N} e^{\hat{\mathbf{y}}_j \cdot \mathbf{y}_j^{-1}}}$$

does not improve on the results shown in **Fig. 7**

12. FUTURE WORK

From this work several new questions arise. Here are the most important ones.

- How large should the initial sample be for sensitivity-based sampling?
- How should the non-linear sensitivity based leverage scores be transformed?
- 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 shows no improvement on datasets *GA* and *T3*, but for *T1* with very non-uniform leverage-scores, the approach is slightly better.

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

14. REFERENCES

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