

Draft

26. maj 2014

1 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 2-class classification, and introduce a new leverage distribution based on a generalization of the idea.

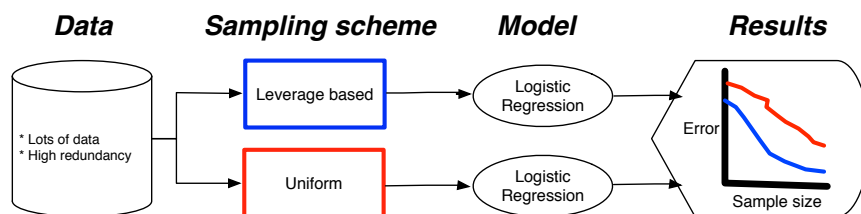
2 Motivation

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 (25 fps).
- 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.

3 Concept



4 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?

5 Datasets

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

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

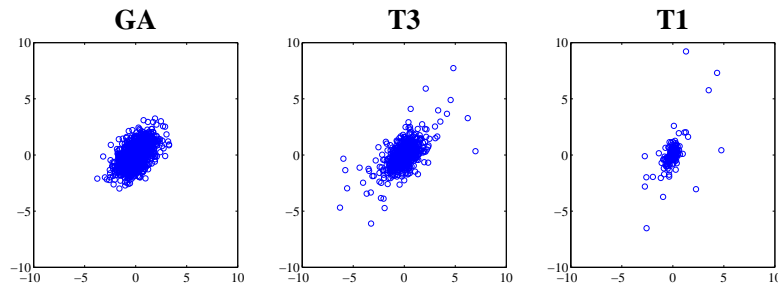


Figure 1: The three distributions considered standardized for comparison

6 Leveraging for least-squares regression

When fitting a model, we know that some datapoints are more important than 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 normalized into a distribution π to sample from.

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

$$\mathbf{H} = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \quad (6.1)$$

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

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

7 Validation of the results Ma et al.

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

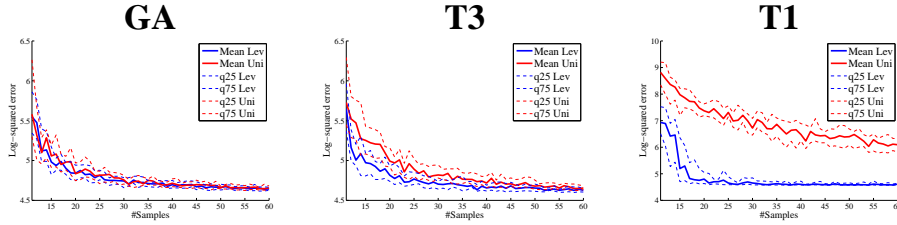
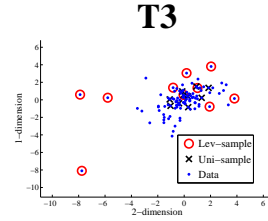


Figure 2: Comparison of uniform (red) vs. leverage (blue) based sampling schemes for least-squares regression. $N = 1000$, $d = 10$.

- 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 varies.

Figure 3: Comparison of sampling methods

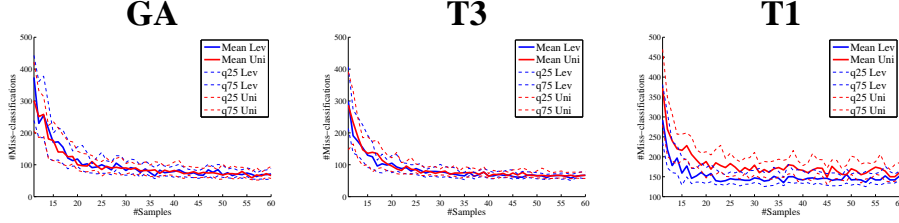


8 LS-based Distribution for Classification

We sample from the same distribution (6.1) as for least-squares regression. We use these samples to train a logistic regression model for 2 class classification, with equal class size.

9 Test Results

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.



- Sampling from the LS-distribution is no better than 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 .

10 Sensitivity Based Distribution

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

$$\frac{\delta \hat{\mathbf{y}}_n}{\delta \mathbf{y}_n} = \text{Diag}(H) \quad (10.1)$$

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

$$\hat{\mathbf{y}}_n = p(y|\bar{\mathbf{x}}_n, \bar{\mathbf{w}}) \quad \bar{\mathbf{w}} \text{ s.t. } \frac{\delta L}{\delta \bar{\mathbf{w}}} = 0 \quad (10.2)$$

Since 10.2 depends both directly and indirectly on y we see that

$$\frac{\delta}{\delta \mathbf{y}} \frac{\delta \mathcal{L}}{\delta \bar{\mathbf{w}}} = 0 \Rightarrow \frac{\delta^2 \mathcal{L}}{\delta \mathbf{y} \delta \bar{\mathbf{w}}} + \frac{\delta^2 \mathcal{L}}{\delta \bar{\mathbf{w}} \delta \bar{\mathbf{w}}^T} \frac{\delta \bar{\mathbf{w}}}{\delta \mathbf{y}} = 0 \quad (10.3)$$

and from this we can get our leverage-score (10.1)

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

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

11 Test results

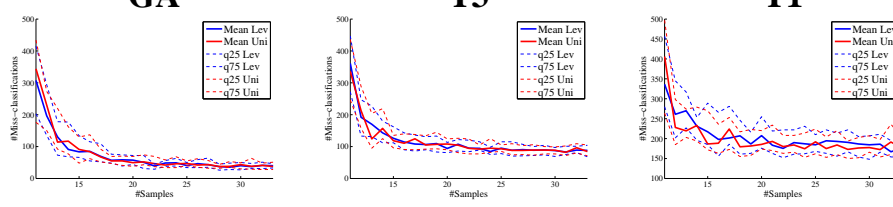


Figure 4: Comparison of sensitivity vs. uniform -based sampling for logistic regression.

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

12 Future work

From our work several new question arise.

- How large show the initial sampling size be for sensitivity-based sampling?
- How should the non-linear sensitivity based leverage scores be normalised?
- 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 Uncertainty of prediction

For a number of weight-vectors \bar{w} , we take the mean of predictions based on these weight-vectors;

$$\langle p(y|\bar{x}, \bar{w}) \rangle \approx p(y|\bar{x}, \hat{\bar{w}}) = p(y|\bar{x}, \mathbf{E}(\bar{w})) \quad (.1)$$

We now look at a small change in the prediction Δp , caused by a change of $\Delta \bar{w}$ in true weight vector \bar{w}_0 .

$$\Delta p = p(y|\bar{x}, \bar{w}_0 + \Delta \bar{w}) - p(y|\bar{x}, \bar{w}_0) \approx \left. \frac{\delta p}{\delta \bar{w}} \right|_{\bar{w}_0} \cdot \Delta \bar{w} \quad (.2)$$

The variance of Δp , can then be computed as

$$\langle (\Delta p)^2 \rangle = Tr \left[\frac{\delta p}{\delta \bar{w}} \left(\frac{\delta p}{\delta \bar{w}} \right)^T \langle \Delta \bar{w} \Delta \bar{w}^T \rangle \right] = \frac{1}{N} \left(\frac{\delta p}{\delta \bar{w}} \right)^T \bar{\bar{H}}^{-1} \frac{\delta p}{\delta \bar{w}} \quad (.3)$$

.1.1 For a linear model with known σ^2

The prediction in a linear model is:

$$p(y|\bar{x}, \bar{w}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-f(\bar{x}|\bar{w}))^2}{2\sigma^2}} \quad (.4)$$

Where y is the target and $f(\bar{x}|\bar{w})$ is the prediction. Differentiating (??) with respect to \bar{w} : (*Hvorfor er det vi gør det??*)

$$\frac{\delta p}{\delta \bar{w}} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-f(\bar{x}|\bar{w}))^2}{2\sigma^2}} - (y - f(\bar{x}|\bar{w})) \frac{\delta f(\bar{x}|\bar{w})}{\delta \bar{w}} \quad (.5)$$

We let $y = f(\bar{x}|\bar{w}) + \epsilon$. (Targets kan beskrives som en approximativ funktion + en fejl ..)

$$\frac{\delta p}{\delta \bar{w}} = \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\epsilon^2}{2\sigma^2}}}_{\text{const. w.r.t. } \bar{x}} \epsilon^2 \frac{\delta f(\bar{x}|\bar{w})}{\delta \bar{w}}^T \bar{\bar{H}}_0^{-1} \frac{\delta f(\bar{x}|\bar{w})}{\delta \bar{w}} \quad (.6)$$

A Logbook

Week 1 (11): 10.03.2014 - 16.03.2014

Project meeting

Questions:

asda

Implementation:

Overall project goals

Week 2 (12): 17.03.2014 - 23.03.2014

Week 3 (13): 24.03.2014 - 30.03.2014

Week 4 (14): 31.03.2014 - 06.04.2014

Week 5 (15): 07.04.2014 - 13.04.2014

Week 6 (16): 14.04.2014 - 20.04.2014

Week 7 (17): 21.04.2014 - 27.04.2014

Week 8 (18): 28.04.2014 - 04.05.2014

Week 9 (19): 05.05.2014 - 11.05.2014

Week 10 (20): 12.05.2014 - 18.05.2014

Week 11 (21): 19.05.2014 - 25.05.2014

Week 12 (22): 26.05.2014 - 01.06.2014