Draft

27. 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.

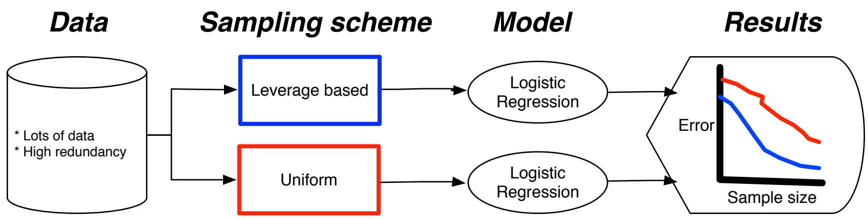
2 Motivation

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



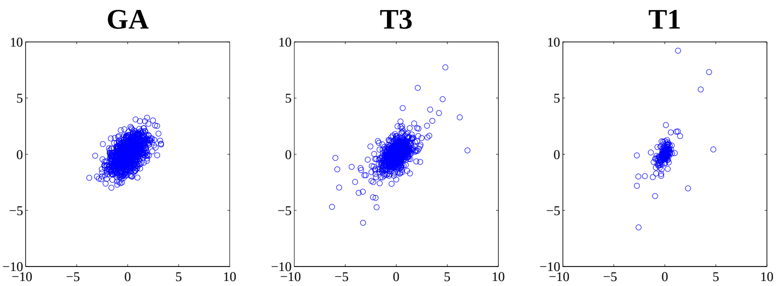
4 Research Questions

- 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



Figur 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 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 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

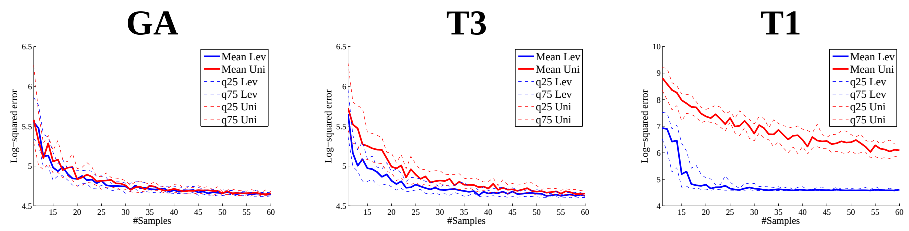
$$\mathbf{H} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T \tag{6.1}$$

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

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

7 Validation of the results Ma et al.

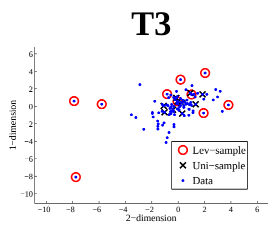
We have empirically tested and validated the results shown by Ma et al. [?].



Figur 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.

Figur 3: Comparison of sampling methods



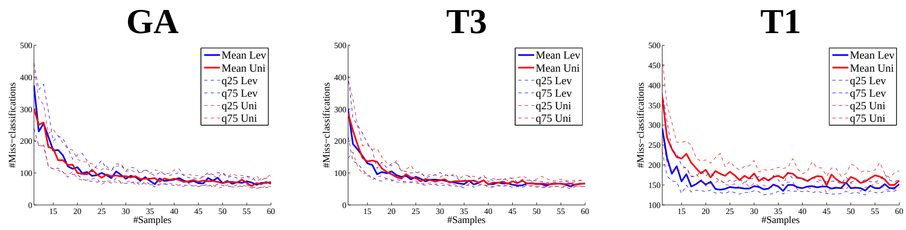
There results are consistent when varying N and d, although the level of improvement varies.

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 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.

10 Sensitivity Based Distribution

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

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

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

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

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

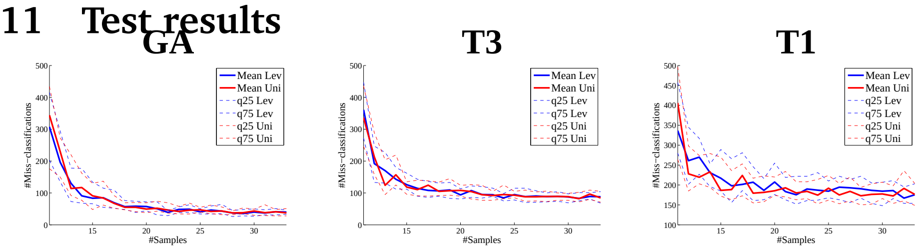
$$\frac{\delta}{\delta \mathbf{y}} \frac{\delta \mathcal{L}}{\delta \mathbf{w}} = 0 \Rightarrow \frac{\delta^2 \mathcal{L}}{\delta \mathbf{y} \delta \bar{\mathbf{w}}} + \frac{\delta^2 \mathcal{L}}{\delta \bar{\mathbf{w}} \delta \bar{\mathbf{w}}^T} \frac{\delta \bar{\mathbf{w}}}{\delta \mathbf{y}} = 0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

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 outperform LS-based sampling since it introduces dependence on class information.

11 Test results



Figur 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;

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

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

$$\Delta p = p\left(y|\bar{x},\bar{w}_0+\Delta\bar{w}\right) - p\left(y|\bar{x},\bar{w}_0\right) \approx \left.\frac{\delta p}{\delta \bar{w}}\right|_{w_0} \cdot \Delta\bar{w} \tag{.2}$$

The variance of Δp, can then be computed as

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

.1.1 For a linear model with known σ^2

The prediction in a linear model is:

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

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??)*

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

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

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

A Logbook

Learning objectives

Overall Project Goals/Delimitation/Hypotheses

General stuff

Week 1 (9): 24.02.2014 - 02.03.2014

Project meeting

No project meeting was possible this week, and we had yet to decided between

1. Randomized algorithms:
   A Statistical Perspective on Algorithmic Leveraging, Ping Ma, Micheal W. Mahoney, Bin Yu. $http://arxiv.org/abs/1306.5362$

2. Spectral learning of HMMs:
   A Method of Moments for Mixture Models and Hidden Markov Models. A. Anandkumar, D. Hsu, and S.M. Kakade. Preprint, Feb. 2012 : $http://newport.eecs.uci.edu/anandkumar/pubs/AnandkumarEtal_mixtures12.pdf$

We spend the week getting an overview of the articles and the projects.

Week 2 (10): 03.03.2014 - 09.03.2014

Project meeting

Questions:

- What is the idea behind leveraging for least-squares regression?
- Can we generalise the idea to general?
- Can leveraging improve performance in video screen classification?
- Video classification e.g. faces, emotions, gender.

Implementation:
No implementation at this point.
Results:
No results at this time.
Decisions:
Gain a better understanding of the underlying idea of leveraging, by watching a talk on *Statistical Leverage and Improved Matrix Algorithms* by M. W. Mahoney ($http://videolectures.net/icml09_mahoney_itslima/$). And analyses the results for

Updated Project Goals and Delimitation

- Validation of the results shown by Ma. et al.
- Can we generalise the idea of leveraging for a general likelihood funktion?

Week 3 (11): 10.03.2014 - 16.03.2014

Project meeting

Questions:

- How does the leverage scores look for LS-regression? (Plotting $H_{n,n}$ vs. $||x_n||$)

Implementation:
No implementation at this point.

Results:

- The general idea of leveraging is to identify how the estimated value $\hat{\mathbf{y}}$ relates to the targeted value $\mathbf{y}$. Which for LS-regression is $\hat{\mathbf{y}} = H\mathbf{y}$.

Decisions:

Updated Project Goals and Delimitation

- Can we generalise the expression $\hat{\mathbf{y}} = H\mathbf{y}$ to logistic regression?

Week 4 (12): 17.03.2014 - 23.03.2014

Project meeting

Questions:

- How are the distributions used by Ma. et al. calculated?
- Finding emotional faces datasets.

Implementation:

- Finding leverage-scores for LS-regression
- Solving LS-regression when comparing uniform- to leverage-based sampling.

- Illustrating leverage scores ($H_{n,n}$ vs. $||x_n||$)

Results:
Initial results promising, but only single run performance between uniform- and leverage-based sampling.
Decisions:

Updated Project Goals and Delimitation

- We want to validate the results of Ma et. al. empirically.
- In video classification we want to do binary classification of *happy* and *sad* faces.

Week 5 (13): 24.03.2014 - 30.03.2014

Project meeting

Questions:

- Will using the leverage-scores for LS-regression improve our performance in binary classification?

Implementation:

- The three distributions GA,$T3$ and $T1$ are implemented, and tested for linear regression.
- Learning curves and test-framework for LS-regression, used for testing the results show by Ma et al.

Results:

- We get comparable results on LS-regression to those shown by Ma et al.
- A leverage-based sampling does not improve for GA-type data, as the leverage scores are approximately uniform, thus there are no "important"datapoints that can be sampled.
- A leverage-based sampling for T3-type data consistently performs better or equal to a uniform sampling. Although the performance increase modest.

- A leverage-based sampling for T1-type data also consistently outperforms a uniform-based sampling, this is expected as the T1 data have very non-uniform leverage scores i.e. ”important”datapoints.

Decisions:
Generalisation of the leverage-based sampling scheme $\frac{\delta \hat{\mathbf{y}}}{\delta \mathbf{y}}$ to logistic regression, as well as a sampling distribution based on the uncertainty of the predictions (asymptotic theory) is to be done by Lars Kai.

Updated Project Goals and Delimitation

- Will using the leverage-scores for LS-regression improve our performance in binary classification?
- We have validated the results of Ma et al. for LS-regression on GA,$T3$ and $T1$ distributed data.

Week 6 (14): 31.03.2014 - 06.04.2014

Project meeting

Implementation:

- Three distributions for binary classification data, also named GA, $T3$ and $T1$ which represent respectively classification data with nearly uniform, moderately non-uniform and very non-uniform leverage-scores.
- Learning curves for logistic regression based on uniform or LS-regression leverage-scores.

Results:
Initial results using leverage-scores based on LS-regression shows no improvement on GA-type (expected) and performs significantly worse on T3- and T1-type data.

Lars Kai has derived a generalised expression $\frac{\delta \hat{\mathbf{y}}}{\delta \mathbf{y}}$ for a general likelihood function. As well as the uncertainty based sampling approach.

Decisions:
Lars Kai gathers his scribles on the back of some insignificant article in a form that is easier to read and follow.

Our full focus is now on midterm preparation.

Updated Project Goals and Delimitation

- Compare uniform sampling to a leverage based distribution (generalisation) and a uncertainty based distribution.

Week 7 (15): 07.04.2014 - 13.04.2014

Project meeting

Discussion about the midterm and improvements that should be done.

Week 8 (16): 14.04.2014 - 20.04.2014

Easter, no project meeting, but sporadic work was done, mostly clarification and bug-finding.

Week 9 (17): 21.04.2014 - 27.04.2014

Project meeting

Results:
Lars Kai gives us a copy and explains the general concepts behind the generalisation of $\frac{\delta \hat{\mathbf{y}}}{\mathbf{y}}$ and uncertainty-based sampling.

Decisions:
We are to understand and digitalise the results derived.

Week 10 (18): 28.04.2014 - 04.05.2014

Project meeting

Questions:

Results:

Decisions:

Week 11 (19): 05.05.2014 - 11.05.2014

Project meeting

Questions:

Implementation:

Results:

Decisions:

Updated Project Goals and Delimitation

Week 12 (20): 12.05.2014 - 18.05.2014

Week 13 (21): 19.05.2014 - 25.05.2014

Week 14 (22): 26.05.2014 - 01.06.2014