ML1819 Research Assignment 1

Team 50

102 - “Dataset Pruning: What is the effect on Machine Learning Performance?”

Liam Greene: 15320482 - Author

Ciaran McGarr: 14321016 – Co-author

Andrew Knox: 14320552 – Contributor / Editor

**Word Count:** 987

**h. URL to source code repository:** https://github.com/liamogreene/CS4044-Machine-Learning-Team-50-Project-

**URL to source code repository activity:** https://github.com/liamogreene/CS4044-Machine-Learning-Team-50-Project-/graphs/contributors

j. Screenshot showing the commit activity of all students, e.g.

Dataset Pruning: What is the effect on Machine Learning Performance?

|  |  |  |  |
| --- | --- | --- | --- |
| Liam Greene Trinity College Dublin Ireland greenel3@tcd.ie | | Ciaran McGarr Trinity College Dublin Ireland cmcgarr@tcd.ie | Andrew Knox? Trinity College Dublin Ireland knoxshia@tcd.ie |
|  |

# Abstract

Data is now seen as one the most valuable resources in the world, even surpassing the value of oil [1] (Economist)

# Introduction

Introduce a) the background of your work, and clearly state b) the research problem and c) your research question/goal (do not just copy & paste the question from our project list. Think carefully about the problem and question and phrase it accordingly).

# Related Work

Explain and discuss critically what other researchers did to answer the research question and what their results were. Do not spend too much time in finding related work but focus on describing properly the work you found. We will not mark you down when some important related work is missing (as long as you describe some related work).

# methodology

Briefly describe the dataset used, including the format of the data, and how the data was processed. Explain how and why you (pre-)processed the data to make it suitable for your analysis. Describe the machine learning algorithms selected and how you went about selecting appropriate values for the algorithm parameters. Present plots justifying your choices and discuss your decisions. Given the limited time you have, we do not expect a perfectly tuned system. Rather it is the critical discussion here that is important, and this should cover the major issues affecting your choices plus the level of uncertainty that your analysis indicates for the parameter choices. Also, explain and justify how you evaluate your work (e.g. chosen metrics, how training and test data was split …).

# Results & Discussion

Present, explain and discuss the results that you obtained. Include tables and figures where appropriate. Finally, answer clearly the research question. The answer must be based only on your own experiments and results. Discuss your results also under consideration of the related work. For instance, are your results confirming the results of related work or contradicting it?

# limitations & outlook

Discuss the limitations of your work, and what steps you would undertake next if you were to continue the project.

ACKNOWLEDGMENTS

This analysis was conducted as part of the 2018/19 Machine Learning module CS7CS4/CS4404 at Trinity College Dublin [2].

REFERENCES

|  |  |
| --- | --- |
| [1] | Patricia S. Abril and Robert Plant. 2007. The patent holder’s dilemma: Buy, sell, or troll? *Commun. ACM* 50, 1 (Jan. 2007), 36–44. DOI: http://dx.doi.org/10.1145/1188913.1188915 |
| [2] | Joeran Beel and Douglas Leith. Machine Learning (CS7CS4/CS4404). Trinity College Dublin, School of Computer Science and Statistics. 2018. |