Machine Learning Simplified: A New Resource to Introduce Advanced Concepts in a High School Approachable Format

Eshaan Debnath

15 September 2022

1 Abstract

Initially, a problem was identified regarding the lack of machine learning resources available to high schoolers that provide a mathematical and algorithmic background to the subject. A set of educational articles fulfilling this goal was created to resolve this issue. In this study, the efficacy of the aforementioned series was evaluated. This was done by a pre test-post test design, along with a small control group, applied to each article. The scores of the control group on the pre tests and the post tests show that the post tests are harder than the pre tests in terms of difficulty. The mean and median values of the pre tests and the post tests were then compared, and a quantitative improvement in scores was confirmed.

2 Introduction

In recent times, Computing Education Research (CER) has been a growing field (3)(4)(16)(17). Rather than focusing on physically teaching students computer science, the purpose of the field is to figure out the most effective *way* to teach computer science (Ko). Overall, it helps students to not only learn faster, but also to gain a deeper understanding of the subject within computer science that they are trying to learn. However, the majority of effort in the field of CER is spent on trying to find the best way to teach people to code, and not finding the best way to teach machine learning. Where there is research in CER on machine learning, it typically has been limited to a college-based audience such as undergraduates, not high schoolers (16).

My goal in the series of articles I created is to introduce the subject of machine learning from a mathematical and algorithmic perspective, as opposed to just an operational one. The purpose of doing so is in the hope that other teens will also attain a deeper understanding of machine learning and its inner mechanics.

2.1 Personal Motivation

I was first introduced to machine learning in 6th grade, and I invested time into learning it in 7th grade. However, the majority of the resources that I had available back in 2018 did one of three things:

- Jumped right into how to apply machine learning using libraries such as keras, tensorflow, scikit-learn, or some other machine learning library (1)
- Introduced the topic conceptually at a very high level, without the smaller details necessary for a mathematical intuition or an algorithmic implementation (2)
- Dived deep into the mathematics behind machine learning with statistics, calculus, matrices, and more as it assumed (at the very least) an undergrad audience (15)(20)

I realized that I could learn how to use a library any day, and decided that it wasn't worth my time. I wanted to know how machine learning actually worked, instead of blindly trusting libraries. Unfortunately, this endeavor lead me only to the second and third bullet points above. Here I was at an impasse, as one option wasn't specific enough, and the other required years of prerequisite knowledge. However, I persisted with this endeavor for a couple of years, and made it a goal to make my own learning resource, one which gave intuition appealing to both common sense and mathematics and explained exactly how to implement machine learning concepts.

My goal in this research paper is to formally establish that my article series does exactly that, in a manner that a high schooler or an everyday person can interpret and understand with a reasonable amount of effort without having an extensive background in machine learning, programming, or mathematics.

3 Summary of Each Article

The overall content of each article in concern is detailed below. You can find the links to my articles under the References section. Note that each article but the first (i.e. Gradient Descent) had a hands-on lab associated with it, which was linked at the end of the respective article. Within these labs, the only major library used to train the model is numpy, as everything is done from scratch.

The majority of the media present in these articles was generated via code. This was done in order to maintain visual consistency throughout the series, while also allowing me fine control over any diagrams and layouts present. The other images that do appear have been hand-curated by me and have a source linked below them within the article. All of the code and other media can be found at the GitHub repository for this project (5).

3.1 Gradient Descent (8)

Explains gradient descent in an intuitive fashion before diving into somewhat formalized mathematics explaining the specifics, while still maintaining the level of intuition with analogies and images. Also introduces new math concepts from scratch with explanations and diagrams, so as to specifically define them in an low-level and understandable manner. In addition, the article offers resources for more advanced types of gradient descent, along with other external resources that might be helpful (ex. basics of taking a derivative).

3.2 Linear Regression (9)

Explains linear regression by building everything from scratch. Builds from the prediction function, which is the "result" that is presented after a linear regression model is trained. Explains the bias term, different types of loss functions, and the cost function. Also discusses bias and variance, and how to mitigate those problems. The lab includes performing linear regression on a fish dataset with MSE. Accuracy is judged via R^2 score.

3.3 Training Faster And Better (12)

Introduces matrices and basic operations with them. Also discusses data normalization, various ways of transforming data, and how to bucket data to model variables with nonlinear relationships. In the lab, demonstrates how matrices are used to shorten code, speed up training, and make everything more readable and logical. The lab includes a linear regression activity on a house pricing dataset, but this time data is processed manually by the programmer before the training process (nonlinear transformations, regularization, etc.), and the entire process is sped up by matrices.

3.4 Logistic Regression (10)

Discusses how the prediction function is modified by wrapping it by a sigmoid function and how that affects it. Also discusses how the loss function and cost function are modified due to the prediction function's output being between 0 and 1, and why regularization is applied. Clarifies the purpose of the bias term in this new scenario, and goes over two ways to measure "accuracy" - by prediction matched percentage and by F1 score. The lab performs logistic regression with log loss on a heart attack dataset and calculates accuracy with both afforementioned methods.

3.5 Neural Networks (11)

Explains the origin of the motivation of a neural network by connecting it with biology, before explaining how it works. Then, explains the cost function, regularization, and the backpropagation algorithm when updating the weights of the neural network. Also mentions various extensions of a neural network, such as CNNs and RNNs, and explains in-depth the practical benefits of a neural network over explicit programming or manually accounting for different cases. The lab implements a 3-layer neural network structure (with bias nodes) from scratch. The model is trained using the MNIST handwritten digits dataset, and backpropagation is implemented from scratch by the reader/programmer as well.

4 Methodology

Motivated by the pre test-post test design popularly seen in CER research (18)(19)(14), this study also employs a similar approach. Each article contains a survey linked as a Google Form. In the first section, anyone filling out the survey can provide optional demographical data and contact information, or they can remain anonymous. Next, in the second section, a respondent fills out their current background in school, calculus, and statistics. The third section is a small pre test consisting of five or six questions, and each question has four options. After that, the respondent is directed to read the article before coming back. After reading the article, they fill out the fourth section which asks them to subjectively rate parts of the article and judge it from an overall standpoint. Finally, the last section contains a post test which is different and slightly more difficult than the pretest.

Results were also taken from a control group to account for the variability in difficulty between pre and post tests, which allowed for direct comparisons between the control and the sample groups without any sort of experimental bias. The control group's survey had a copy of the first and second sections present in each survey of the sample group. Then, it had collections of the same pre and post test questions present in the sample group's five surveys. These questions were grouped by the name of the article, and each article had a separate section devoted to it. Finally, the last section asked the respondent to rank the confidence level in their answers. Note that the respondent did not read the article before or during filling out the control form.

A key noteworthy detail is that there weren't enough survey results for the articles other than Gradient Descent to make a statistically significant observation with each individually. This issue regarding a limited dataset will be addressed in the Discussion section.

5 Results of Study

The results from the surveys are summarized below in table form. Note that the same respondent number in two tables may not be the same person. Data about demographics is not included as the sample size is too small to perform an analysis with them. For each survey, the pre test and the post test contain the same number of questions. The surveys for "Gradient Descent" and "Training Faster And Better" contained 6 survey questions for each test, whereas the other surveys contained 5 survey questions for each test.

Table 1: Control Survey Results

	Respondent #	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	Mean
Gradient Descent	Pre Test Score	0	0	4	2	1	3	4	2	4	4	2.4
	Post Test Score	1	0	1	2	0	0	1	3	1	1	1
Linear Regression	Pre Test Score	2	1	0	1	2	2	2	2	2	2	1.6
	Post Test Score	1	0	2	1	2	1	2	2	0	1	1.2
Training Faster And Better	Pre Test Score	3	2	4	2	2	3	5	5	3	0	2.9
	Post Test Score	3	3	1	1	1	4	1	2	1	2	1.9
Logistic Regression	Pre Test Score	2	2	1	2	2	3	5	0	3	2	2.2
	Post Test Score	1	2	2	1	0	2	2	3	1	1	1.5
Neural Networks	Pre Test Score	1	1	2	4	2	2	3	3	4	2	2.4
	Post Test Score	2	1	2	2	1	0	1	2	1	2	1.4

In table 2 below, note that respondents 1, 2, and 3 are not from a lower education setting (grades 12 and below). These columns are highlighted in purple for extra clarity. In this table, the mean is calculated over all 10 survey respondents. For the other tables, there is too little data to pick and choose during analysis.

Table 2: Gradient Descent Survey Results

	Respondent #	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	Mean
Gradient Descent	Pre Test Score	5	1	3	6	2	5	1	4	1	5	3.3
	Post Test Score	4	0	3	4	5	6	1	2	5	5	3.5

Table 3: Linear Regression Survey Results

	Respondent #	#1	#2	Mean
Linear Regression	Pre Test Score	0	3	1.5
	Post Test Score	3	4	3.5

Table 4: Training Faster And Better Survey Results

	Respondent #	#1	#2	Mean
Training Faster And Better	Pre Test Score	5	5	5
	Post Test Score	3	4	3.5

Table 5: Logistic Regression Survey Results

	Respondent #	#1	#2	Mean
Logistic Regression	Pre Test Score	0	0	0
	Post Test Score	2	5	3.5

Table 6: Neural Networks Survey Results

	Respondent #	#1	#2	Mean
Neural Networks	Pre Test Score	1	4	2.5
	Post Test Score	2	4	3

These tables only represent part of the results from the study. The complete results can be found on Google Spreadsheets (6)(7). Note that some response data has been stripped as it is not crucial to this study. This includes, but is not limited to: data about response timestamps, respondent emails, and more.

6 Discussion

Due to the limited dataset, one can better analyze this by splitting the data points into two parts - one with the scores involving the "Gradient Descent" article, and the other with the scores involving the other articles.

6.1 Gradient Descent

One can measure this article's effectively by comparing the means of pre test and post test scores between the study sample and the control sample. This can also be done by comparing the median scores of the two tests.

The table below gives a visual comparison between the aggregate scores between the two samples. Note that the study sample consists of all 10 submissions, including submissions from respondents beyond high school.

Type	Pre Test Mean	Post Test Mean	Pre Test Median	Post Test Median
Control	2.4 (40%)	1 (16.67%)	2.5 (41.67%)	1 (16.67%)
Study	3.3 (55%)	3.5 (58.33%)	3.5 (58.33%)	4 (66.67%)

Within this table, one can see that the study sample initially starts out with a higher pre test score than the control sample. This indicates that the respondents in the study sample might have been slightly more knowledgeable in this topic than the control sample was.

When looking at the post test, one can see that the score of the control sample significantly dips in comparison to the pre test results, thus empirically proving that the post test is more difficult than the pre test. In addition, when comparing the average score on the post tests between the control sample and study sample, there is clearly an appreciable difference between the two groups. On average, the study sample was able to get 2.5 more questions correct than the control sample on the post test, whereas in the pre test this difference was only 0.9. This difference can be seen more vividly in the median scores, where the pre test difference is 1 and the post test difference is 3. In addition to this, one can also see that the post test scores of the study sample goes up, even with the more difficult post test.

These results lead to the conclusion that the "Gradient Descent" article is indeed effective in teaching the reader about the subject.

One can also filter out scores from only high school students. Such a constraint would result in the table below.

Type	Pre Test Mean	Post Test Mean	Pre Test Median	Post Test Median
Control	2.4 (40%)	1 (16.67%)	2.5 (41.67%)	1 (16.67%)
Study	3.43 (57.15%)	4 (66.67%)	4 (66.67%)	5 (83.33%)

Once again, one can make the same observations as to how not only the post test is harder and the study sample performs significantly better than the control sample, but also the study sample is able to achieve a higher score even with the more difficult test after reading the article.

6.2 Other Articles

The other articles were a bit harder to judge, due to the limited data set. So, I decided to combine all of their results. This was difficult, given that they contained an inconsistent number of questions. To solve this issue, I decided to convert each respondent's score into a percentage of how much they answered correctly. Then, with that modified data set, I computed the mean and median for pre tests and post tests, for each group. The results are summarized in the table below.

Type	Pre Test Mean	Post Test Mean	Pre Test Median	Post Test Median
Control	43.08%	28.42%	40.00%	20.00%
Study	40.83%	64.58%	40.00%	63.33%

When comparing the pre test scores in this table, one can clearly see that the two groups scored very similarly: the means differed only by 2.25%, and the median scores were exactly the same. This indicates that the control and study samples were comparable. Now looking at the post test, it can be seen that the control scores drop significantly. However, the scores in the study sample increase by more than one and a half times. The significant difference between the two group once again shows that the series of articles were quite effective, even when working with very limited data.

7 Discussion of Subjective Data From Study

Each of the surveys in the study group also included subjective questions for the respondent to answer. Since only the "Gradient Descent" survey received a significant number of responses, only that will be discussed.

The tables below contain the average rating of respondents on various subjective questions.

Understanding of Calculus	Understanding of Statistics	Intuitive Understanding
5.3/10	5.2/10	3.3/5

Mathematical Understanding	Rating (0 = other articles preferred, 5 = this article preferred)
2.5/5	3.4/5

In the data above, one can see that the respondents were somewhat experienced with calculus and statistics, but not a whole lot. In addition, the "Gradient Descent" article gave the reader a decent intuition of the concept, and conveyed the same concept mathematically somewhat well. While these are open to subjective interpretation, the rating measure clarifies that the article is indeed better than other articles in terms of approachability and understanding for the general public.

8 Conclusion

Overall, these articles do a good job of teaching machine learning from the basics. They help the reader gain both an intuitive and mathematical understanding, and build everything from the basics in order to be accessible to everyone. However, it does have certain flaws.

Many respondents provided feedback that the articles were too mathematically involved, and that they weren't able to understand all of it. In addition, due to the variety of audience, some were left confused, whereas others said that they would have liked more details. In addition, I received in-person feedback from few of the respondents who said that they would have liked more images. In the end, my opinion is that while this series is certainly an improvement over other resources available on the internet, there is still vast room for improvement.

In the future, I will be looking to create a more comprehensive and well-organized resource that addresses these flaws while also being even more friendly to beginners and stimulating towards those with more experience.

9 Acknowledgments

I would like to thank Polygence for providing me a platform to conduct research. I would also like to thank my mentor, Jesse Stern, for helping me edit my articles, catch assumptions, and point out places that needed clarifications. Jesse also helped me throughout my research and guided me on how to write a good research paper, along with proofreading everything.

10 Bibliography

[1] Aflahi, T. F. (2019). How to get started with machine learning in about 10 minutes. Available at https://www.freecodecamp.org/news/how-to-get-started-with-machine-learning-in-less-than-10-minutes-b5ea68462d23/.

[2] Brown, S. (2021). Machine learning, explained. Available at https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained.

[3] Cooper, S., Forbes, J., Fox, A., Hambrusch, S., Ko, A., and Simon, B. (2016). The importance of computing education research. DOI: 10.48550/ARXIV.1604.03446, available at https://arxiv.org/abs/1604.03446.

[4] Cooper, S., Grover, S., Guzdial, M., and Simon, B. (2014). A future for computing education research. *Communications of the ACM*, 57(11):34–36. Available at https://sci-hub.se/10.1145/2668899.

[5] Debnath, E. Polygence project. Available at https://github.com/Endothermic-Dragon/Polygence.

[6] Debnath, E. Polygence survey data. Available at https://docs.google.com/spreadsheets/u/1/d/e/2PACX-1vQ4FkiR-W5BWaTlzzs4rtWSXPva6P8srhmtekCcg7pXdHynkoLrSi_AwWxOcSc1OkCRKB63UyjFu94s/pubhtml.

[7] Debnath, E. Polygence survey data. Available at https://docs.google.com/spreadsheets/d/1cv4CYCXmYwg2qu2BBy-XX7fPDgA6nNRk-ZihCLghCgY/edit?usp=sharing.

[8] Debnath, E. (2022a). In-depth machine learning for teens: Gradient descent. Available at https://medium.com/@endothermic-dragon/in-depth-machine-learning-for-teens-gradient-descent-ce2d0370303c.

[9] Debnath, E. (2022b). In-depth machine learning for teens: Linear regression. Available at https://medium.com/@endothermic-dragon/in-depth-machine-learning-for-teens-linear-regression-d66db85c8f3a.

[10] Debnath, E. (2022c). In-depth machine learning for teens: Logistic regression. Available at https://medium.com/@endothermic-dragon/in-depth-machine-learning-for-teens-logistic-regression-3e9b86102482.

[11] Debnath, E. (2022d). In-depth machine learning for teens: Neural networks. Available at https://medium.com/@endothermic-dragon/in-depth-machine-learning-for-teens-neural-networks-ded1af6a84de.

[12] Debnath, E. (2022e). In-depth machine learning for teens: Training faster and better. Available at https://medium.com/@endothermic-dragon/in-depth-machine-learning-for-teens-training-faster-and-better-93f35bb263e3.

[Ko] Ko, A. J. Amy j. ko, ph.d.: Uw seattle. Available at https://faculty.washington.edu/ajko/cer.

[14] Maniktala, M., Cody, C., Barnes, T., and Chi, M. (2020). Avoiding help avoidance: Using interface design changes to promote unsolicited hint usage in an intelligent tutor. *CoRR*, abs/2009.13371. Available at https://arxiv.org/abs/2009.13371.

[15] Saini, A. (2021). Conceptual understanding of logistic regression for data science beginners. Available at https://www.analyticsvidhya.com/blog/2021/08/conceptual-understanding-of-logistic-regression-for-data-science-beginners/.

[16] Shapiro, R. B., Fiebrink, R., and Norvig, P. (2018). How machine learning impacts the undergraduate computing curriculum. *Communications of the ACM*, 61(11):27–29. Available at https://sci-hub.se/10.1145/3277567.

[17] Tedre, M., Toivonen, T., Kahila, J., Vartiainen, H., Valtonen, T., Jormanainen, I., and Pears, A. (2021). Teaching machine learning in k-12 computing education: Potential and pitfalls. DOI: 10.48550/ARXIV.2106.11034, available at https://arxiv.org/abs/2106.11034.

[18] Threekunprapa, A. and Yasri, P. (2020). Unplugged coding using flowblocks for promoting computational thinking and programming among secondary school students. *International Journal of Instruction*, 13:207–222. DOI: 10.29333/iji.2020.13314a, available at https://files.eric.ed.gov/fulltext/EJ1259514.pdf.

[19] Türker, P. and Pala, F. (2020). The effect of algorithm education on students' computer programming self-efficacy perceptions and computational thinking skills. *International Journal of Computer Science Education in Schools*, 3:19. DOI: 10.21585/ijcses.v3i3.69, available at https://files.eric.ed.gov/fulltext/EJ1242561.pdf.

[20] Wade, C. (2021). Data science for teens. Available at https://medium.com/berkeley-coding-academy/data-science-for-teens-c213765f8678.