Advantages of Prior Mathematical Knowledge for Studying Machine Learning

Differences in Knowledge Gain between Computer Science and Physics Students

Oisín Hageman[1]

Supervisor(s): Gosia Migut[1]**, Ilinca Rențea**[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
January 26, 2025

Name of the student: Oisín Hageman
Final project course: CSE3000 Research Project
Thesis committee: Gosia Migut, Ilinca Rențea, Jesse Krijthe

Abstract

With the growing need for machine learning knowledge for many different expertises and positions, comes a growing need for machine learning education for non-computer scientists. Teaching machine learning concepts to non-majors comes with the added challenge of dealing with different levels of prior mathematical knowledge. Existing research is inconclusive on the correlation between this prior knowledge and topic-specific machine learning knowledge gain. This paper evaluated this via an experiment conducted on Computer Science and Physics students without prior machine learning education. We find that there is no clear correlation between general math knowledge and knowledge gain. There is however a clear correlation of proficiency in probability and statistics, and algorithm heavy machine learning topics. The experiment also concluded that most students struggled most with these math-heavy topics, as well as understanding abstract systems such as perceptrons.

1 Introduction

Machine learning (ML) is one of the fastest growing expertises, both in academia and in job listings requirements [1]. As the use cases for ML-based systems range from harmless to very sensitive situations, there is a growing need to understand the pitfalls of ML education. Especially the rise in ML and Artificial Intelligence related education for non-computer scientists brings the need for new insights [2], as these students do not have the same mathematical baseline that a traditional computer science curriculum would bring. There is however a knowledge gap in regards to pedagogical challenges in this topic [3]. The research question we aim to answer is as follows:

”How does prior mathematical knowledge influence students to learn specific machine learning topics between Computer Science and Physics majors?”

This study looks into the target group of physics students, as they have more more mathematics courses in their curricula, and thus likely a different, more math-focused approach to learning than their computer science peers. The research question can be answered by measuring two attributes for a sample of Computer Science and Physics students and checking correlation between these two attributes: mathematical proficiency and ability to learn ML topics. We do this by means of a user survey that measures these statistics.

The paper is set out in the following sections: Section 2 compiles the existing work surrounding the research question and introduces the knowledge gap. Section 3 presents the setup for the experiment, and details how results will be quantified. Results of the experiment are laid out in section 4. And sections 6 and 7 put the results into a broader perspective and present conclusions respectively. Additionally, section 5 presents some of the ethical considerations made to properly conduct the study.

2 Related Work

Several frameworks are commonly used for defining topic understanding, such as Bloom’s taxonomy [4] and Structure of Observed Learning Outcomes (SOLO) taxonomy [5]. These frameworks are used for identifying different levels of topic-specific skill development. There is however growing concern over students’ mathematical proficiency. Studies show that students’ mathematical abilities have been steadily declining [6], and this pattern has been correlated in the Netherlands to the Covid-19 pandemic [7]. Mathematical proficiency is however a difficult concept to define, let alone measure [8]. Studies emphasize the significance of mathematics in machine learning, highlighting how foundational mathematical concepts are to understanding and applying machine learning techniques [9]. On the other hand, teachers of ML courses state that it is possible, and even beneficial, to teach ML to non-majors without requiring them to possess a strong background in mathematics [10][11][2].

It thus becomes apparent that there are many papers that state there is an implicit, albeit not necessarily crucial link between prior mathematical proficiency or knowledge and ML education. The main obstacles for teaching ML to non-majors are also well-documented. However, a quantifiable, comparative study on the correlation of (prerequisite) mathematical knowledge, and ML knowledge gain between computer science and non-majors is still missing.

3 Methodology

To answer the research question, we set up an experiment consisting of three parts. Firstly, we asked the participants to take a survey and math `test`, for measuring their prior proficiency in mathematics. The survey part was used to obtain information about former math courses taken at different levels in their academic career, as well as any experience with machine learning topics. Secondly, students followed an ML `tutorial`, introducing three entry-level topics. Finally, they completed an ML `quiz`, which gives info on what knowledge was gained from the tutorial. With the measurements of both prior mathematical knowledge and topic-specific knowledge gain, we then checked for correlation between the two. The test, tutorial and quiz were created with an iterative approach, where draft versions of each part were tried by others in the research group. This approach gave the necessary insights to evaluate the range of difficulty of the questions and the overall quality of the tutorial.

3.1 Subject Group

The study contained two target groups: One consisting of 12 BSc Computer Science and Engineering (CSE), the other 5 BSc Applied Physics (AP) students. The study focused specifically on students who have not yet taken an ML course, and have little to no experience with the subject. This ensured that the knowledge gain of both target groups was compared without little underlying bias. The choice of physics students as non-major group, is because we expect this group to have a more math oriented outlook on the presented subject matter, and many students of Applied Mathematics do a double bachelors in conjunction with Physics. Which in turn, could

Topic	Learning Goal	Tutorial Chapters	Quiz Questions
ML Pipeline	Understand the general machine learning pipeline	1	-
	Explain the purpose of training, test and validation sets	3.1; 3.2; 4.1; 4.2	1
	Compare the performance of a model trained on different dataset splits	3.3; 4.3	2, 5
	Identify overfitting and underfitting of a machine learning model	2	3, 4
Bayes' Rule	Apply Bayes' rule to solve probability problems	1	-
	Understand the relationship between prior, likelihood, and posterior probabilities	1; 2	6
	Calculate conditional probabilities in real-world scenarios	1; 2	8, 9, 10
	Evaluate decisions with Bayesian reasoning	3	7
Perceptrons	Understand the pipeline of training artificial neural networks	1; 2.2; 3.1	12, 15
	Calculate the output of a single perceptron	2.1	13
	Explain the advantages/disadvantages of neural networks	3.2	11, 14

Table 1: Constructive Alignment learning goals

make teaching from a more math-heavy explanation of ML topics be beneficial for their knowledge gain.

3.2 Survey and Math Test

For the first part, the participants filled in an online form via Microsoft Forms. To ensure anonymity during the research, each participant was firstly prompted to create an anonymous identifier to link the different components of the research. Further details about this process are provided in section 5. Next, the participants answered some open ended questions about any math subjects taken throughout their academic career, and whether they have had any second-hand experience with ML topics. Following this, there was a general, university-level mathematics test, with 9 questions. The questions are based on three common undergraduate math courses: probability & statistics, linear algebra, and calculus, that link to different parts of the tutorial as explained in the next section.

As mathematical *proficiency* is a difficult statistic to measure, we have opted to test for mathematical knowledge, by way of calculation questions [8]. This method does not fully capture proficiency in its entirety. However, doing a complete study would take a lot more time for our participants, which is already in limited supply due to the multiple parts of the experiment setup.

For purpose of grading, the questions are all worth the same value on the final score, giving 0 or 1 points to the total. Question 3 and could also get 0.5 point, in case of a partially correct answer due to a calculation error, and question 8 could get the same for a correct, but incomplete answer. The complete survey form can be found in Appendix A.1.

3.3 ML tutorial

The tutorial consisted of a webpage with explanations, images and examples. The participants were expected to spend about one hour reading through and studying, or about 20 minutes per part. For constructing both the tutorial and quiz, the constructive alignment method introduced by Biggs was followed [12]. The different topics and their respective learning goals can be found in Table 1. This also displays the link between the intended learning goals to its respective tutorial sections and final ML quiz questions.

	μ Math score	μ ML score	r	ρ
CSE	0.458	0.756	-0.205	-0.167
AP	0.467	0.667	0.606	0.671

Table 2: Aggregated scores and correlation coefficients

For each part, the topics were explained without any computer science related or other technical prerequisite knowledge, except for some parts which either required some mathematical knowledge to follow, and other parts which were easier to follow with this knowledge. More specifically, the first part about the ML Pipeline, required no mathematics. The second part, Bayes' Rule, relies on probability and statistics. And the third Perceptrons part was set up that it is able to be followed without mathematics, but knowledge of Linear Algebra and certain notations will make it easier to understand. This distribution was created so testing was possible for not just specific ML topics, but topics that range from abstract, high-level concepts to specific algorithm or formula-based knowledge. The tutorials can be found in Appendix A.2.

3.4 ML quiz

The machine learning quiz was the final part of the experiment. This step aimed to measure the knowledge gain of students after following the tutorial for about one hour. There were 15 multiple-choice questions total, with 5 questions per topic. The questions are tightly coupled to the learning goals as described earlier. At the end of the quiz, there were some extra questions about perceived ease/difficulty of both tutorial topics and quiz and time spent on studying the tutorial. This was to allow some qualitative insights to be gained in supposed patterns or outliers in the data. Grading was the same as the math test, with 0 or 1 point to be gained per question. The full list of questions and answers can be found in Appendix A.3.

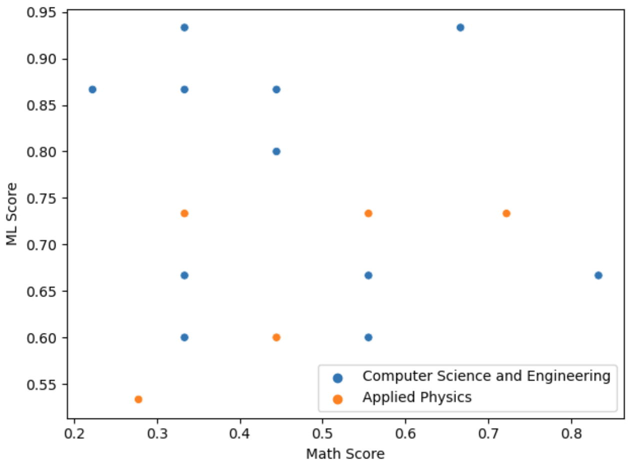


Figure 1: Distribution of math test and ML quiz scores

4 Results

Code used to generate these statistics and figures can be found here[1]. Table 2 contains aggregated data of math test and ML quiz mean scores, as well as both Pearson's (r) and Spearman's (ρ) correlation coefficients between these two scores. These statistics are calculated separately for both our target groups. Figure 1 shows the same total scores in a scatterplot. This figure illustrates the low correlation (r = -0.205) between the test scores for CSE students. Figure 2 displays more Pearson's correlations in a heatmap. These scores scores are calculated per individual topic. The math test was split up into Calculus (Calc, Q1 and Q2), Linear Algebra (LinAlg, Q3-5), and Probability and Statistics (ProbStat, Q6-8). For the ML Quiz the divisions are the ML Pipeline (Pipeline, Q1-5), Bayes' rule (Bayes', Q6-10) and Perceptrons (Percept., Q11-15). The questions per topic can be found in Appendix A. Cross-correlation between different ML scores or Math scores have been omitted, as these metrics are irrelevant to the research topic. This results in 4x4 matrices instead of 8x8 matrices with double values. And lastly, Table 3 contains an analysis of most frequent answers the ML quiz question "During the learning phase, which part did you find the most difficult to learn/hardest to understand? Why do you think that is?".

5 Responsible Research

The major contribution of this research is the experiment with human subjects. This type of research comes with additional ethical challenges, for our case mainly in regard to privacy. To execute this in a structured manner, we followed the procedure and guidelines from TU Delfts Human Research Ethics Committee (HREC). This included submitting an Informed Consent Statement as detailed below, a Data Management Plan, and an Ethics Risk and Mitigation Plan[2]. Additionally, multiple measures were taken to ensure privacy and anonymization at different steps of the study.

	Perceptrons	Bayes'	Difficulty w/ Probability
CSE	67%	33%	8%
AP	0%	100%	100%

Table 3: Most common perceived difficulties and reasons during learning phase

5.1 Participant Recruiting

For recruiting participants, a modified HREC informed consent form was used. The main points of this statement are:

- What participating in the research entails, describing the three parts of the experiment.
- The participants have the right to drop out at any point in the study.
- The fact that there is always the possibility of a data breach, and what steps have been taken to minimize this chance, and mitigate the risks this would bring.

After the participants have agreed to this, their contact information was gathered. This information was only used to communicate about where the forms/surveys could be found, when these should be completed, and to answer any further questions. The contact information was not linked to any research data, and was destroyed a week after the study concludes.

5.2 Data Anonymization

The second part of ensuring privacy of participants is in the surveys. Because the experiment included two tests at different times, an identifier was necessary for linking the math test to the quiz. To ensure there is no way to link research data back to individual participants, each participant generated an identifier at the start of the first quiz according to a predefined code. It also had the added benefit of lowering the chance that a participant would forget the identifier by the time of the second quiz.

The last part of anonymization lies in the choice of not releasing raw research data. The only data presented in the paper is aggregated data. For this particular research the individual raw data is not necessary or useful, so this is the right choice for improving privacy of participants even further.

6 Discussion

6.1 Interpretation of Correlation Coefficients

The means in Table 2 line up with the original assumptions: AP students score slightly higher on mathematics than CSE students. This can be explained due to physics students having a higher affinity with mathematics. The ML scores for CSE students however lie significantly higher than the AP students.

This is likely due to the ML topics being similar to the rest of the CSE curriculum, and thus easier to pick up on for CSE students. The first anomalies that can be observed are the total Math to ML correlation scores of CSE students,

[1] https://github.com/oisinhageman/learning-machine-learning

[2] https://www.tudelft.nl/en/about-tu-delft/strategy/integrity-policy/human-research-ethics

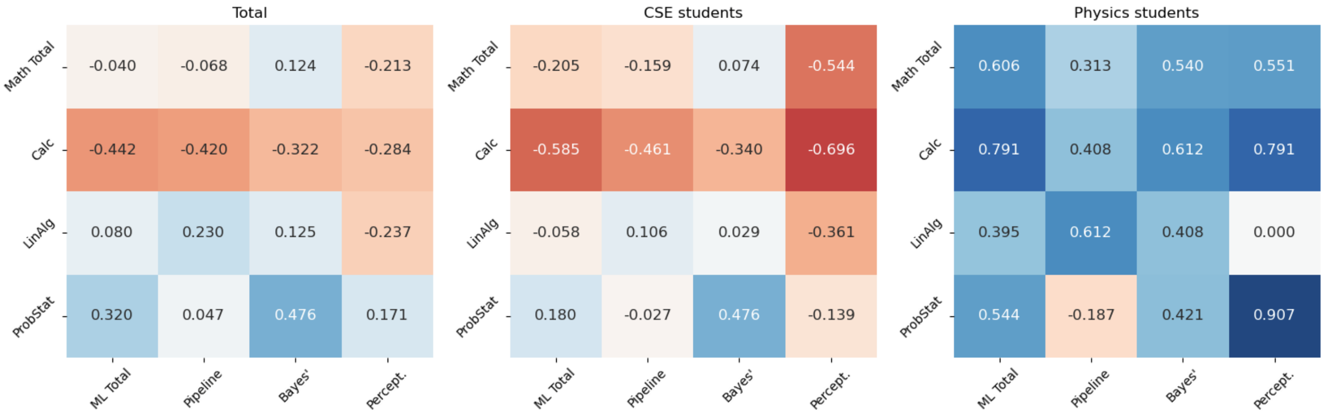


Figure 2: Correlation matrix between math and ML scores, over different topics

which show a weak negative correlation. For the AP students a moderately strong correlation can be observed. This is a possible indication that majors with more affinity for mathematical courses, are also able to more effectively incorporate this in their learning behaviors.

The correlation matrix in Figure 2 show more unexpected results: First of all the strong negative correlations for CSE students of Calculus scores to ML scores across topics, and same for correlation of Perceptron scores to Math scores. As it is unlikely that an understanding of Calculus has a negative effect on learning machine learning, this can be explained by one of two factors. The first being the fact that our sample size is too low for finding a stable correlation, as discussed in section 6.3. Or secondly, the calculus part of the math test being insufficient for testing calculus understanding. As the calculus part only has 2 questions as opposed to 4 and 3 for the other parts, the range of scores is very limited. Adding to this is that this part was very easy for most students, and only students not familiar with the notation or formulation getting the questions wrong. Combined this caused extremes in the data that do not represent what the test originally tried to measure.

Another visible pattern in the matrices is the moderately strong correlation between Probability and Statistics score and the ML Bayes' rule topic (r = 0.48), across both subject groups. This aligns with the original assumption that topics relying on understanding of formulas and math-heavy theorems do benefit greatly from a strong background in the aligned math topics.

6.2 Perceived Difficulties

Table 3 highlights the most frequent perceived difficulties of both subject groups. Interestingly, all students from the physics group mention that they had most difficulty with the Bayes' rule part, and they all indicate that this is due to difficulty or dislike for probability. This perceived difficulty lines up with our findings that the largest correlation of knowledge gain and math is between these two topics. Of the CSE students, a third considers this topic to be difficult as well, but the majority found perceptrons to be harder, with the most common reason being unfamiliar with the concepts or finding them too abstract.

6.3 Limitations of Subject Group

While our study is an indication of correlation between specific math and ML topics, there are some underlying problems with our data:

To start with, the choice of TU Delft CSE students limits understanding of the math to ML knowledge gain dependency. This is due to the timing of the study and the CSE curriculum. The average first-year CSE participants had not finished any mathematics courses during their program yet. Meanwhile, most second-year students had just finished the Machine Learning course and thus were not suitable for participating.

Additionally, the sample sizes for both groups were too small for any conclusive correlation analysis [13]. For subsequent research, we suggest taking not only more students into account, but also at a more opportune time for measuring mathematical proficiency. For Delft CSE students this would be year 1, quarter 4, when the majority have completed more math courses, but still have not done a ML course. Furthermore, combining datasets to also include other non-major subject groups would give more insight into the influence of different mathematical backgrounds on ML learning.

7 Conclusion

The proposed experiment gives us valuable, although inconclusive insights into how different mathematical proficiencies affect the process of learning machine learning. The tests showed significant correlation between proficiency in probability theory and the learning of Bayes' rule. Besides this correlation, students' perceived difficulty of math-heavy machine learning topics further demonstrates this fact. This is a clear indication that fundamental topics of machine learning require this prerequisite foundation, and would require a significant extra time investment for students with a weaker background in mathematics.

For more conclusive results, a similar study as this could be conducted. This experiment ideally boasts a larger sample size for both groups to perform a stable correlation analysis. Additionally, the math test would benefit from a larger scope, but more importantly a wider range of question difficulty. A broader definition of the concept of mathematical *proficiency* also should be considered, such as problem analyzing skills.

Acknowledgements

I would like to specifically thank Justin Jo and Junwon Yoon for their collaboration. The tests and tutorial of our participant survey were created and conducted in parallel with their respective research on different target groups. The responsibility of different topics was divided to create a single baseline for our studies. Specifically linear algebra + perceptrons sections were my specific responsibility. Whereas the calculus + ML pipeline, and probability + Bayes' rule theirs respectively.

I would also like to thank the others in our research group, Carina Oprean and Ghiyath Alaswad, for their peer review, which increased the quality of our experiment significantly. And lastly I would like to thank both my supervisors Ilinca Rențea and Gosia Migut, as well as my academic counselor Jitske van der Laan. Without their insights and support this paper would not exist.

References

[1] R. B. Shapiro and R. Fiebrink, "Introduction to the special section: Launching an agenda for research on learning machine learning," *ACM Trans. Comput. Educ.*, vol. 19, no. 4, Oct. 2019. [Online]. Available: https://doi.org/10.1145/3354136

[2] R. Fiebrink, "Machine learning education for artists, musicians, and other creative practitioners," *ACM Trans. Comput. Educ.*, vol. 19, no. 4, Sep. 2019. [Online]. Available: https://doi-org.tudelft.idm.oclc.org/10.1145/3294008

[3] A. J. Ko, "We need to learn how to teach machine learning," 2017.

[4] B. S. Benjamin, B. Bloom, and D. Krathwohl, "Taxonomy of educational objectives," *New York: McKey New York*, 1956.

[5] J. Biggs and K. Collis, *Evaluating the Quality of Learning: The SOLO Taxonomy (structure of the Observed Learning Outcome)*, ser. Educational psychology. Academic Press, 1982. [Online]. Available: https://books.google.nl/books?id=0kQmAQAAIAAJ

[6] G. A. Nortvedt and A. Siqveland, "Are beginning calculus and engineering students adequately prepared for higher education? an assessment of students' basic mathematical knowledge," *International Journal of Mathematical Education in Science and Technology*, vol. 50, no. 3, pp. 325–343, 2019. [Online]. Available: https://doi.org/10.1080/0020739X.2018.1501826

[7] Ministerie van Onderwijs, Cultuur en Wetenschap, "Rekenen-wiskunde einde leerjaar 2 voortgezet onderwijs 2021-2022 - themarapport - inspectie van het onderwijs," Ministerie van Onderwijs, Cultuur en Wetenschap, rapport, 2024. [Online]. Available: https://www.onderwijsinspectie.nl/documenten/themarapporten/2024/02/27/rekenen-wiskunde-einde-leerjaar-2-voortgezet-onderwijs-2021-2022

[8] A. H. Schoenfeld, "What is mathematical proficiency and how can it be assessed," *Assessing mathematical proficiency*, vol. 53, pp. 59–73, 2007.

[9] N. Lavesson, "Learning machine learning: A case study," *IEEE Transactions on Education*, vol. 53, no. 4, pp. 672–676, 2010.

[10] E. Sulmont, E. Patitsas, and J. R. Cooperstock, "Can you teach me to machine learn?" in *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*, ser. SIGCSE '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 948–954. [Online]. Available: https://doi.org/10.1145/3287324.3287392

[11] ——, "What is hard about teaching machine learning to non-majors? insights from classifying instructors' learning goals," *ACM Trans. Comput. Educ.*, vol. 19, no. 4, Jul. 2019. [Online]. Available: https://doi.org/10.1145/3336124

[12] J. Biggs, "Enhancing teaching through constructive alignment," *Higher education*, vol. 32, no. 3, pp. 347–364, 1996.

[13] M. A. Bujang and N. Baharum, "Sample size guideline for correlation analysis," *World*, vol. 3, no. 1, pp. 37–46, 2016.

A Appendix

A.1 Survey and Math Test Questions

Welcome, and thank you for participating in this research! This study investigates how prior mathematics knowledge influences learning outcomes when students are introduced to foundational machine learning (ML) topics. You will be asked to provide information about your past maths-related coursework and activities by asking you questions about your experience and with mathematics questions. Afterwards, you can continue with the tutorials for the ML topics. After you study it and become familiar with it, we will test your learning outcome with a set of quiz in Week 7. The information you provide in this form, combined with quiz results, will help us analyze trends across students with different academic backgrounds.

Important Notes:

- Participation is entirely voluntary, and you may withdraw at any time.
- Your responses will remain anonymous, and no personal identifiers will be collected.
- Data will be used only for academic purposes and published in aggregated form in a master's thesis and the university repository.
- Please answer the questions honestly to ensure the reliability of the research.

1. Do you agree to participate in this research, knowing that your responses will remain anonymous and be used only for academic purposes?

- Yes
- No

2. Please generate your unique participant id using the following formula: Your favorite color + favorite dessert + favorite animal.(e.g. Orange Stroopwafel Cat) Please remember this id, as you will use it again for the quiz in Week 7.

Your answer here

3. What is your current degree program?

- Computer Science and Engineering
- Industrial Design
- Applied Mathematics
- Aerospace Engineering
- Applied Physics
- Electrical Engineering
- Other (Please specify)

4. What is your current academic level?

- BSc
- MSc
- Other (Please specify)

5. Please list any maths-related courses you took in high school (e.g., IB Mathematics HL, AP Mathematics, Dutch Mathematics A).

Your answer here

6. Please list any maths-related courses you took Bachelor's degree (if applicable).

Your answer here

7. Please list any maths-related courses you took during your Master's degree (if applicable).

Your answer here

8. Please list any extracurricular maths-related activities or courses you have participated in (e.g., online courses, maths clubs).

Your answer here

9. Please list any prior experience in machine learning (e.g., highschool course, online course, YouTube tutorial).If you have any, please list the topics you have learnt.

Your answer here

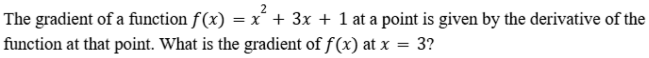
10. How confident do you feel in your mathematics skills? Rate from 1 being not confident at all to 5 being very confident.

- 1
- 2
- 3
- 4
- 5

11. Following questions are mathematics questions. Please give solutions to them without using any extra materials (e.g., calculators, textbooks, etc). You may use pen and paper. If you are unsure, you can write "Unsure" and move on to the next question.

- I understand

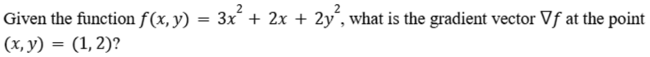
12. Q1



The gradient of a function $f(x) = x^2 + 3x + 1$ at a point is given by the derivative of the function at that point. What is the gradient of $f(x)$ at $x = 3$?

Your answer here

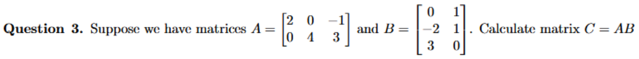
13. Q2



Given the function $f(x, y) = 3x^2 + 2x + 2y^2$, what is the gradient vector ∇f at the point $(x, y) = (1, 2)$?

Your answer here

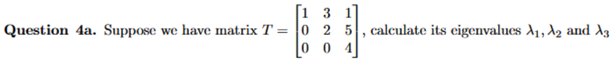
14. Q3



Question 3. Suppose we have matrices $A = \begin{bmatrix} 2 & 0 & -1 \\ 0 & 4 & 3 \end{bmatrix}$ and $B = \begin{bmatrix} 0 & 1 \\ -2 & 1 \\ 3 & 0 \end{bmatrix}$. Calculate matrix $C = AB$

Your answer here

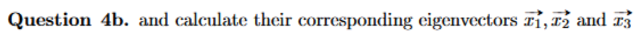
15. Q4a



Question 4a. Suppose we have matrix $T = \begin{bmatrix} 1 & 3 & 1 \\ 0 & 2 & 5 \\ 0 & 0 & 4 \end{bmatrix}$, calculate its eigenvalues λ_1, λ_2 and λ_3

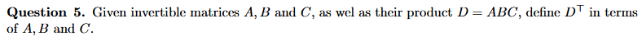
Your answer here

16. Q4b


Question 4b. and calculate their corresponding eigenvectors $\vec{x_1}, \vec{x_2}$ and $\vec{x_3}$

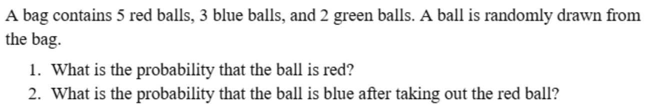
Your answer here

17. Q5


Question 5. Given invertible matrices A, B and C, as wel as their product $D = ABC$, define $D^\top$ in terms of A, B and C.

Your answer here

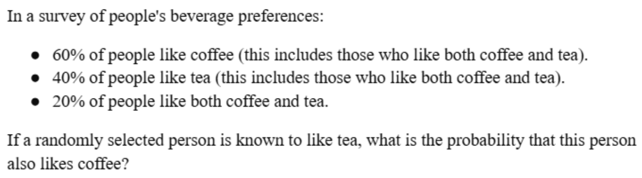
18. Q6


A bag contains 5 red balls, 3 blue balls, and 2 green balls. A ball is randomly drawn from the bag.

1. What is the probability that the ball is red?
2. What is the probability that the ball is blue after taking out the red ball?

Your answer here

19. Q7

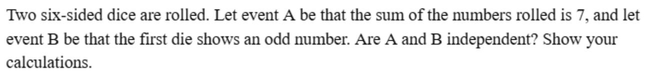

In a survey of people's beverage preferences:

- 60% of people like coffee (this includes those who like both coffee and tea).
- 40% of people like tea (this includes those who like both coffee and tea).
- 20% of people like both coffee and tea.

If a randomly selected person is known to like tea, what is the probability that this person also likes coffee?

Your answer here

20. Q8


Two six-sided dice are rolled. Let event A be that the sum of the numbers rolled is 7, and let event B be that the first die shows an odd number. Are A and B independent? Show your calculations.

Your answer here

A.2 ML Tutorial

The full tutorials with images can also be found here[3]

Tutorial 1 - Machine Learning Pipeline
After this tutorial, you will be able to:

- Understand the general machine learning pipeline
- Explain the purpose of training, test and validation sets
- Compare the performance of a model trained on different dataset splits
- Identify overfitting and underfitting of a machine learning model

Chapter 1: Introduction to the Machine Learning Pipeline

1.1 What is a Machine Learning Pipeline? In order to create and use ML models such as facial recognition, voice-to-text, or Instagram filters, we need to train and test the model. A machine learning pipeline is the series of steps involved in building, training, and evaluating a machine learning model. It organizes the process to ensure the model learns effectively from the data and generalizes well to unseen situations. In short, we prepare the resource and give it to the model to learn, and we test how well can it apply it to other examples.

Below are the stages in the ML Pipeline:

Tutorial 2 - Bayes' Rule
After this tutorial, you will be able to:

- Apply Bayes' rule to solve probability problems
- Understand the relationship between prior, likelihood, and posterior probabilities
- Calculate conditional probabilities in real-world scenarios
- Evaluate decisions with Bayesian reasoning

Chapter 1: Introduction to Bayes' Rule

1.1 What is Bayes' Rule? Bayes' rule (also called Bayes' theorem) is a fundamental principle in probability theory that describes how to update our beliefs about events when we receive new evidence. It provides a mathematical framework for combining prior knowledge with new data.

The formula for Bayes' rule is:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where:

- P(A—B) is the posterior probability (probability of A given B)
- P(B—A) is the likelihood (probability of B given A)
- P(A) is the prior probability (initial probability of A)
- P(B) is the marginal probability (total probability of B)

1.2 Real-World Example: Medical Diagnosis Let's consider a doctor diagnosing a rare disease:

1. Prior knowledge: The disease affects 12. Test accuracy:
   - 95- 90

If a patient tests positive, what's the probability they have the disease?

Using Bayes' rule:

$$P(D|+) = \frac{P(+|D)P(D)}{P(+)}$$

$$= \frac{(0.95 * 0.01)}{(0.95 * 0.01) + (0.10 * 0.99)}$$

$$0.087, 8.7\%$$

Chapter 2: Components of Bayes' Rule

[3] https://github.com/oisinhageman/learning-machine-learning

2.1 Prior Probability Prior probability represents our initial belief about an event before seeing new evidence. It's based on:

- Historical data
- Previous experience
- General knowledge
- Initial assumptions

2.2 Likelihood Likelihood represents how probable the evidence is, given our hypothesis:

- Measures the compatibility of the evidence with different hypotheses
- Often based on empirical data or scientific models
- Can be updated as more data becomes available

2.3 Posterior Probability Posterior probability is our updated belief after considering the evidence:

- Combines prior knowledge with new evidence
- Becomes the new prior for future updates
- Represents our current best estimate

Chapter 3: Bayesian Classification and Error

3.1 Bayesian Classification Bayesian classification uses Bayes' rule to make decisions by comparing posterior probabilities:

- Choose class if P(—x) ¿ P(—x)
- Choose class otherwise

Where:

- , are possible classes
- x is the observed data/features

3.2 Classification Error Two types of errors can occur:

1. **Type I Error (False Positive)**
    - Incorrectly classifying as when true class is
    - Error probability: P(decide —)
2. **Type II Error (False Negative)**
    - Incorrectly classifying as when true class is
    - Error probability: P(decide —)

3.3 Bayes Error Rate The Bayes error rate is the theoretical minimum error achievable by any classifier:

- For two classes: Error = min[P(—x), P(—x)]
- Represents inherent overlap between classes
- Cannot be eliminated even with perfect classification

Example: Given two overlapping distributions:

- P(x—) = N(1, 1) // Normal distribution, mean=1, variance=1
- P(x—) = N(2, 1) // Normal distribution, mean=2, variance=1
- P() = P() = 0.5

Tutorial 3 - Perceptrons
After this tutorial, you will be able to:

- Understand the pipeline of training artificial neural networks
- Calculate the output of a single perceptron
- Explain the advantages/disadvantages of neural networks

Chapter 1: What is a Perceptron? A perceptron is like a decision-making machine that mimics how neurons in the brain work.

Biological Neurons:

A.3 ML Quiz

Welcome to the final step of our journey! Here, you will solve questions that are related to the topics that you have learned for the past weeks. This is not an assessment, and you will receive no penalties for your scores. If you are not sure about how to solve a question, choose "Not sure"; please do not refer to any external resource or try to guess an answer. You will solve 15 multiple choice questions, and at the end we prepared some questions for you to tell us about your experience during the learning program. Thank you for your time and cooperation in this study, and good luck!

1. Enter the same id you used for the math test. If you did not write it down, it was of the following format: Your favorite color + favorite dessert + favorite animal. (e.g. Orange Stroopwafel Cat)

Your answer here

2. What is your current degree program?

- Computer science and Engineering
- Industrial Design
- Applied Mathematics
- Aerospace Engineering
- Applied Physics
- Electrical Engineering
- Other

3. Why do we split data into training and test sets in a machine learning pipeline?

- To evaluate the model's performance on unseen data. V
- To improve the speed of the model.
- To reduce the size of the dataset.
- To optimize the hyperparameters of the model.
- Not sure

4. You are training a machine learning model with a dataset of 10,000 samples. If you use 80- The model may not have enough data to learn patterns properly.

- The model may suffer from overfitting during training.
- The test set might not fully represent the variability in the data. V
- The model's performance on the test set might overestimate its real-world accuracy.
- Not sure

5. What happens when a machine learning model overfits?

- The model becomes more efficient in processing data.

- The model performs well on training data but poorly on new data. V
- The model fails to learn the patterns in the training data.
- The model performs equally well on training and test data.
- Not sure

6. A model achieves 95- The test set is not representative; adjust the train-test split.

- Underfitting; increase model complexity or use a larger test set.
- Balanced performance; no changes are needed.
- Overfitting; reduce model complexity or gather more training data. V
- Not sure

7. Why is it important to include a validation set when training a model?

- It ensures the test set remains untouched until final evaluation.
- It reduces the size of the training set, preventing overfitting.
- It simplifies the pipeline by eliminating the need for a test set.
- It allows the model to learn more patterns from the data.
- Not sure

8. Which of the following best describes the role of the prior probability in Bayes' Rule?

- It measures the compatibility of evidence with a hypothesis.
- It is the initial belief about an event before new evidence is considered. V
- It is the total probability of the evidence occurring.
- It represents the updated belief after considering new evidence.
- Not sure

9. In Bayesian classification, which of the following describes a Type I error?

- Incorrectly classifying an item as belonging to a class when it does not. V
- Failing to classify an item into a class when it belongs there.
- Miscalculating the prior probability of a class.
- Minimizing the posterior probability of an incorrect classification.
- Not sure

10. Spam Email Detection:

- 90¿ - 10¿ - 40¿ - 60

What is the Bayes error rate for this spam filter?

- 5- 10- 15- 20- Not sure

11. A company screens applicants for a job using a test. The test is designed such that:

- 80¿ - 30¿ - 60¿ - 40

If an applicant passes the test, what is the probability that they are actually qualified?

- 56- 64- 72- 82- Not sure

12. A factory uses a machine to sort defective items. The sorting system is imperfect:

- P(Detected Defective—Defective) = 0.9
- P(Not Detected Defective—Not Defective) = 0.85
- P(Defective) = 0.05
- P(Not Defective) = 0.95

The cost of classifying a defective item as not defective is 10. The cost of classifying a non-defective item as defective is 5.

Given this information, how should the system classify an item if the system detects it as defective?

- Defective
- Not Defective V
- Not sure

13. What is not a disadvantage of using an Artificial Neural Network?

- ANNs are not good at finding complex patterns in datasets. V
- It is difficult to figure out what made the model give a certain output.
- The training of ANNs generally requires large amounts of data.
- ANNs need a lot of computing power for training the model.
- Not sure

14. Given that the formula for updating weights during training is:

$$w_i \leftarrow w_i + \eta(y - \hat{y})x_i$$

What can we say about the learning rate η?

- The update of the weights are only dependent on whether the prediction is correct or wrong, not by how far is from the real expected output.
- The learning rate needs to be positive. V
- If the model predictions are 100- Not sure

15. Given a perceptron with weight vector [3,-1,1], bias -2, and activation function f(x) = -1 if x ¡ 0, f(x) = 1 if x ¿= 0. What would the perceptron output with input vector [-1,-2,3]?

- -1
- 0
- 1 V
- 2
- Not sure

16. Which of the following best describes a model acting as a black box?

- In a worst case scenario, the model will still function.
- The model requires a lot of data for computing its output.
- It is hard to find out how the model came to its output.
- The model has a large memory to store data.

17. Assume we have a Multi-Layer Perceptron with 3 input nodes, two hidden layers of 4 nodes (h1 h2), and an output layer of 2 nodes (out). What are the sizes of the weight matrices that can store this model?

- w$_h1 = 3x1, w_h2 = 4x1, w_out = 2x1w_h1 = 3x4, w_h2 = 4x4, w_out = 4x2V$
- w$_h1 = 4x1, w_h2 = 4x1, w_out = 2x1w_h1 = 3x2, w_h2 = 4x2, w_out = 4x2$
- Not sure

18. How did you find the difficulty of the topics you learned?

- Topic 1: ML pipelines
  - Very easy
  - Easy
  - Moderate
  - Difficult
  - Very difficult
- Topic 2: Bayes' Rule
  - Very easy
  - Easy
  - Moderate
  - Difficult
  - Very difficult
- Topic 3: Perceptrons
  - Very easy
  - Easy
  - Moderate
  - Difficult
  - Very difficult

19. How did you find the difficulty of the test?

- Very easy
- Easy
- Moderate
- Difficult
- Very difficult

20. How much time did you roughly take for studying the tutorial?

- Less than 30 minutes
- Between 30 minutes and 1 hour
- Between 1 hour and 2 hours
- More than 2 hours

21. During the learning phase, which part did you find the most comfortable to learn/easiest to understand? Why do you think that is?

Your answer here

22. During the learning phase, which part did you find the most difficult to learn/hardest to understand? Why do you think that is?

Your answer here

23. Were there any parts or formats of the tutorials that you found particularly helpful in learning and understanding new topics? If yes, what were they?

Your answer here

24. If you would teach the topics to students from your own study, how would you teach them? What kind of medium would you use?

Your answer here