Comparison of Machine Learning Algorithms

Project Plan

Maciej Roszkowski

Supervised by Dr. Michail Fasolakis
Department of Computer Science
Royal Holloway, University of London

1. Abstract

The objective of machine learning is to make computers learn; that is for them to improve over time by themselves. A good definition of machine learning is: learning from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E [1]. This project will involve the research and implementation of machine learning algorithms to investigate their behaviour on artificial and later benchmark datasets.

Initially I will implement 1-NN and K-NN (first and K nearest neighbour) algorithms also testing differing tie breaking strategies. I will also implement the decision trees algorithm using different measures of uniformity. The next phase will involve the implementation and testing of more advanced algorithms such as multi class support vector machines and kernel based nearest neighbours[2].

Once all the algorithms are implemented I will cross test them and evaluate their performance based on the chosen test set. I will ensure that in the project I will adhere to modern software engineering standards by unit testing code, using version control to back up my code and communicating regularly with my supervisor.

2. Timeline

This is a rough timeline for the project, of course during all of this I will ensure to be communicating with my supervisor, to ensure that my project meets the standards of the college. I will regularly make updates in my diary as well as keep rough informal notes of my work to aid my own memory of what has been completed and what has not and at what time including any problems that may be faced throughout the project. This will be covered in more detail in the risks and mitigations section.

Week (Refers to teaching week)	Key Tasks	Deliverable
Term 1 Week: 1-3 (W/C 29/09/25 - W/C 20/10/25)	Background research and complete project plan. Read key chapters from project reading list including other relevant sources Initialise project repository and begin first steps of programming	Completed project plan by 10/10/25 Completed reading of important sections of required reading books and other relevant sources to the project Initialised readme and git repo including pipeline
Term 1 Week: 3-6 (W/C 20/10/25- W/C 10/11/25)	Select algorithms and datasets based on research Research and decide exact technologies to use e.g programming language, libraries and datasets	Completed readme of techstack and description of technologies and resources used
Term 1 Week: 6-10 (W/C 10/11/25- W/C 1/12/2025)	Implement baseline models (Decision Tree, KNN)	Working basic implementation of basic algorithms
Term 1 Week 10 - 11 (W/C 1/12/25- W/C 8/12/25)	Evaluate overall performance and progress, use this as extra overflow time if earlier objectives are not completed. Complete and submit the interim report during this time.	Completed interim report
Term 1 Week 11 - Term 2 Week 1 (W/C 8/12/25-	Ensure all term one objectives are completed and begin planning work for term 2	Concrete plan for term 2 Completed evaluation of mistakes made and mitigations of mistakes

W/C 13/01/26)		
Term 2 Week 1 - 4 (W/C 12/01/25- W/C 09/02/25)	Research and implement real data sets with algorithms already implemented	Completed research into which real datasets to use along with their output with basic algorithms
Term 2 Week 4 - 8 (W/C 09/02/25- W/C 09/03/25)	Begin implementation of Kernel-based KNN algorithm	Fully implemented and working algorithms on real dataset
Term 2 Week 8 - 11 (W/C 09/03/25- W/C 06/04/25)	Complete whole project and finish all writeup and documentation	Completed project report and fully working code

3. Bibliography

[1]

Mitchell, T.M. (1997). *Machine learning*. [online] New York: Mcgraw-Hill. Available at: https://www.cs.cmu.edu/~tom/files/MachineLearningTomMitchell.pdf.

[2]

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical learning, Second Edition : Data mining, inference, and Prediction*. 2nd ed. New York: Springer.

4. Risks and Mitigations

This section will detail the potential risks associated with the project along with mitigation strategies for them. Obviously there are always risks to anything but my objective in this section is to identify risks early on so that I can reduce their potential future impact.

4.1 Implementation errors or bugs from algorithms

- If there are logical mistakes or errors in programming it may lead to wasted time debugging

Mitigations

- Develop and test algorithms incrementally
- Use unit testing (and include a pipeline with gitlab) to ensure pushed code works correctly
- Use gitlab to ensure changes are safe and revert back to old version if necessary

4.2 Computational resource constraints

- Certain algorithms and datasets may consume large amounts of computational resources

Mitigations

- Test algorithms on smaller datasets initially to ensure that they work properly
- Place a large focus on optimising code
- Use university or cloud infrastructure if possible

4.3 Data quality issues

- If datasets used are low quality then that may affect output and effectiveness of algorithm, may also contain unclean data

Mitigations

- Inspect datasets before use
- Implement basic python scripts to clean and check validity of data

4.4 Time management and scope creep

- The project could expand beyond its original scope (e.g., trying too many algorithms or datasets), leading to missed deadlines.

Mitigations

- Define clear milestones for early deliverables (1-NN, k-NN, Decision Trees).
- Move to advanced models (kernel k-NN, SVM) only after core components are validated.
- Allocate weekly goals and maintain a progress log.

4.5 Potential data loss

- Loss of code, datasets, or experiment results could delay project completion.

Mitigations:

- Use cloud-based version control (GitHub or GitLab).
- Maintain regular local and cloud backups.
- Store final results and documentation in multiple locations.