Machine Learning ICS3206, Course Project 2021

Very important – Read before starting

- The deadline for **completing and submitting** your assignment is strictly Friday 21st January 2022.
- **VLE will be set up to not accept late submissions** meaning that you will get zero marks if late.
    - Please plan ahead (it is recommended that you **upload and verify your work a day before**).
    - Technical problems, internet connectivity issues, lost backups, cats eating laptops, etc… are not valid excuses.
- You must complete a plagiarism declaration form and include it in your report. **Submissions without the form will not be considered.**
- **Projects must be submitted using VLE only.** Physical copies or projects (including parts of) sent by email will not be considered.
- For your convenience, a draft and final submission area will be set up in VLE. **Only projects submitted in the final submission area will be graded.** Projects submitted to the draft area are not considered.
- It is suggested that after submitting your project, you re-download it and check it just in case. **It is your responsibility to ensure that your upload is complete, valid, and not corrupted.** You can re-upload the assignment as many times as you wish **within the deadline.**
- **Your project must be submitted in ZIP format without passwords** or encryption. Project submitted in any other archiving format (e.g. RAR, 7Z, etc…) will not be considered.
- The total size of your ZIP file should not exceed 99 megabytes.
- Your submission should include your report in PDF format, your source code, and executable file(s).
- It is expected that you submit a quality report with a proper introduction, discussion, evaluation of your work, and conclusions. Also, make sure you properly cite other people's work that you include in yours (e.g. diagrams, algorithms, etc…).
- In general, I am not concerned with which programming language you use to implement this project. However, unless you develop your artifact in BASIC, C, C++, Objective C, Swift, Go, Pascal, Java, C#, Matlab, or Python, please consult with me to make sure that I can correct it properly.
- This is not a group project.
- Plagiarism will not be tolerated.

Project

- This project is about the **ID3 decision tree learning** algorithm.
- Obtain **two or more *classification*** datasets from https://archive-beta.ics.uci.edu/ml/datasets.
    - It is up to you to choose whichever datasets you like but **choose them wisely**.
    - Make sure that at least one of the datasets you choose has at least one attribute with **continuous** values.
    - Make sure that the target attribute (label) of at least one of the datasets you choose can have **more than two possible values** (not simply binary yes/no classification). For example, the instances in the *wine* datasets belong to one of three different classes.
    - You will need to split the datasets into **training sets** and **validation sets**; make sure that there is enough data to do this.
- You are required to implement the ID3 algorithm **yourself** – do not use an existing implementation (or copy someone else's work).
- Your implementation needs to **support continuous-valued attributes**.
- **Experiment** with your implementation on the datasets you have chosen and discuss your results.
- In your implementation make sure to include a method (whichever one you like) to deal with **overfitting**.
- Experiment with this overfitting countermeasure and discuss your results.
- If you need to, feel free to use any external libraries help you to import (read) the datasets. The datasets are plain text files, so reading them yourself shouldn't be a big deal.

Report:

- You do not need to extensively discuss how ID3 works. However, describe the methods you used to select attribute nodes, and how you deal with continuous values and overfitting.
- Please write a good report. Describe the datasets you chose and why, describe your methodology, conclusions, etc…
- In your report, briefly discuss one alternative approach which is suitable for the task. Speculate on whether you think it would perform better or worse than ID3.
- I am expecting a good evaluation and discussion regarding the results you obtained. Use a proper experimental procedure discussing your setup (e.g., training/validation split), expected outcomes, results, and discuss.

Statement of completion – MUST be included in your report

Item	Completed (Yes/No/Partial)
Dataset selection and import	
ID3	
Support continuous attributes	
Overfitting management	
Good discussion on an alternative method	
Experiments and evaluation	
If partial, explain what has been done	

Marking Breakdown

Description	Marks allocated
Dataset selection and import	5%
ID3	35%
Support continuous attributes	10%
Overfitting management	10%
Discussion on an alternative method	10%
Experiments and evaluation	20%
Overall report quality	10%