University of Essex
School of Computer Science and Electronic Engineering
CE802 Machine Learning

Combined coursework reassessment for CE802

Set by: Prof. Luca Citi <lciti@essex.ac.uk>
Dr Delaram Jarchi <delaram.jarchi@essex.ac.uk>
Submission mode: as instructed by the School Office

Assignment objectives

This document specifies a combined reassessment coursework assignment for CE802. The assignment aims to assess your ability to evaluate the suitability and limitations of different machine learning algorithms for a particular practical problem, and then describe the correct procedure to assess the performance of these different models and undertake a comparative evaluation. It also assesses the general understanding of the theoretical foundations behind machine learning methods and of various machine learning techniques covered during the module.

Assignment description

1. General Questions on Machine Learning

Question 1

Define "Machine Learning" and in particular describe:

- some examples of problems for which it is used successfully;
- some examples of problems for which it is unfit;

Question 2

Explain the difference between supervised and unsupervised learning. Then name two examples of supervised learning algorithms and two examples of unsupervised learning algorithms.

Question 3

Describe the hierarchical agglomerative clustering algorithm and state two of its limitations.

Question 4

Explain "overfitting" in the context of machine learning and in particular discuss the factors contributing to it. Also explain the relationship between the bias-variance tradeoff and overfitting.

Question 5

Explain the learning procedure known as boosting.

2. Exercise on identifying machine learning techniques appropriate for a particular practical problem

A university wants to develop a tool to support the applicant admission process and automatically determine whether a prospective student is likely to succeed in their undergraduate program or not. The university has collected data on 3,500 applicants, including their demographic information, academic history, extracurricular activities and application details.

The dataset contains the following features for each applicant:

- Demographic information: age, gender, ethnicity, socio-economic background (as determined by postal code), first language
- Academic history: high school grades, standardized test scores, courses taken, rank in class
- Extracurricular activities: number of clubs/societies, volunteer hours, participation in sports
- Application details: essay score, letters of recommendation, date and time of the application

The target variable is whether the applicant is *HighPotential* or *LowPotential*. As an expert on machine learning, you are asked to assist the university in developing a system that can predict whether an applicant is likely to succeed in their undergraduate program. Describe how you would set about trying to solve this problem.

Your answer should include:

1. Discussion of the type of problem to be solved.
2. Selection of a small set of learning procedures, with an explanation of why they may be suitable.
3. A brief description of a comparative evaluation of the selected machine learning procedures.
4. Detailed description of how you would estimate the success of the final chosen system.
5. An account of how you would use the selected procedure to predict whether a new applicant is likely to succeed in their undergraduate program.

Important note: In your solution, please discuss any potential biases in the features used and how they might be contrary to ethical AI practices. For example, you may want to consider whether using some features could disproportionately affect certain demographics (e.g., low-income individuals or racial minorities). How would you address these concerns?

This second part should consist of approximately 500–1000 words of narrative.

3. Coding Exercise

Create a function that builds a Decision Tree predictor using scikit-learn, optimising hyperparameters through grid search. The function should take in two parameters: a matrix `X` and a vector `y`. Implement the necessary steps to train the model using the given `X` and `y`, optimizing the hyperparameters `max_depth` and `min_samples_split` through a grid search.

Requirements:

- Develop a Python function that accepts a matrix `X` and a vector `y` as input.
- Implement a Decision Tree model using scikit-learn's `tree` module.
- Utilise grid search to optimise the hyperparameters `max_depth` and `min_samples_split`.
- The function should return a trained Decision Tree predictor using the best hyperparameters discovered via grid search.

Grid Search Parameters:

- Values of `max_depth` to explore: {2, 4, 6, 8, 10, None}
- Values of `min_samples_split` to explore: {2, 5, 10, 20, 50}

Marking criteria

This assignment will be assessed based on:

- General Questions on Machine Learning
  - Question 1–5 (each) 8%
- Exercise on identifying machine learning techniques appropriate for a particular practical problem
  - Correctness of identified type of predictive task 3%
  - Appropriateness of learning procedures suggested 8%
  - Correctness of the comparative evaluation methods suggested 9%
  - Quality of the discussion on ethical AI practices 9%
  - Overall clarity of presentation 6%
- Coding Exercise
  - Correctness and completeness of the implementation 18%
  - Quality of the code and comments 7%

Late Submission and Plagiarism

Please refer to the Postgraduate Students' Handbook for details of the Departmental policy regarding late submission and University regulations regarding academic offences and responsible use of AI. In particular, it is essential to note that this assignment requires original work and thought. While you are encouraged to discuss ideas, learn from online resources, or use AI tools responsibly to prepare for your assignment, the submission of work resulting from AI generation, collaborative efforts, or other forms of academic misconduct will not be tolerated. All submitted work must be the result of your own understanding, analysis, and synthesis of the topic.

Revision 1.0
18/06/2025
Luca Citi