



Computer Science Division

For this assignment, you have a choice of either exploring ensemble learning or unsupervised learning. Machine Learning
CS441/741

Assignment 4: Ensemble Learning

Due: 22 October 2025, 08:00

Instructions

For the purposes of this assignment, please note the following:

- You may select any one of the options listed below.
- Please submit your own work.
- You may program in any language, provided that your code compiles and execute on Linux. While you are allowed to make use of machine learning libraries, note that you learn most about the machine learning algorithms when you implement these algorithms yourself.
- Submit only your pdf report. Make sure to name your pdf file as `???????RWxxxassignment4.pdf`, replacing the question marks with your student number and `xxx` with the module code that you are registered for.
- **NB:** You have to follow the file naming conventions as stated above. I make use of scripts to extract your pdf files. If you give your files arbitrary names, yours will not be pulled out and will therefore not be evaluated.
- Not that the report must be a formal technical report wherein you report all that you have done, your results, and a discussion of these results. For guidelines on the format of a technical report and a mark rubric, see the last sections.
- You are not allowed to make use of any generative AI tools.
- Note: No deadline extensions, and no late submissions will be accepted.

Option 1: Random Forests – Does Tree Size Matter?

The main objective of this assignment is to explore the impact of maximum tree depth of the decision trees that make up a random forest, as well as the impact of the the number of randomly selected features when deciding on a node split. Starting from the smallest possible trees, investigate the performance of a random forest with increasing decision tree depths. Then do the same with the number of randomly selected features to split on. The idea is to explore the performance of the random forest for very simple ensemble members which individually underfit the training data, moving to a random forest with individual members that overfit the training data. As a starting point, do this analysis for a fixed random forest size, i.e. a fixed number of decision trees in the forest. Also fix the number of descriptive features that are randomly selected for each split decision, as well as the bag size. When this is done, explore any relationship between decision tree depth and the number of decision trees in the random forest. Then fix the tree depth at the best value from the preceding experiment and explore the impact of different numbers of randomly selected descriptive features. For the final experiment, explore the performance of the random forest with decision trees of different depths and different numbers of randomly selected descriptive features, to include decision trees that overfit and decision trees that underfit.

For the purpose of this assignment, select at least three classification datasets that differ in complexity.

Your report should include the necessary background, detail on how your random forest has been implemented, your opinion on what you expect to observe with justifications thereof, detail on the empirical process followed, and then finally your results and a discussion of the results. Conclude on whether the results support your expectations.

Option 2: Logistic Regression Ensemble

The main objective of this assignment is to explore the performance of an ensemble of logistic regression models. Start with an ensemble of linear logistic regression models, where each model is constructed from a bagged subset of the original training set using a randomly selected subset of descriptive features. Explore the impact of the number of members in the ensemble, the number of randomly selected features, and the size of the bags.

Then proceed to add non-linear logistic regression models which makes use of basis functions to allow the formation of non-linear decision boundaries. Add different nonlinear logistic regression models, which differ in the basis functions used (from lower order to higher order basis functions), the number of descriptive features and the bag sizes that they are constructed from. Investigate the influence of the number of members in the ensemble, the number of randomly selected features, and the size of the bags.

For the purposes of this assignment, only consider binary classification problems. Use at least three datasets, making sure that the datasets differ in complexity and that they are appropriate for illustrating the need for non-linear logistic regression models.

Your report should include the necessary background, detail on how the linear logistic regression ensemble and the non-linear logistic regression ensemble was implemented, detail on the empirical process followed and then finally your results and a discussion of the results.

Option 3: Isolation Forests

The main purpose of this assignment is to explore the impact of the control parameters of an isolation forest on the ability of the isolation forest to identify anomalies. For the purposes of this assignment, you have to identify all the control parameters of an isolation forest. In your report, discuss your opinion on the effect of different values for each of the control parameters. Then, find at least three anomaly datasets that you will use for your analysis. Define the empirical process that you will use to evaluate the sensitivity of the isolation forest to these control parameters. See if you can derive correlations among the various control parameters and if you can provide guidelines on deciding values for each of the control parameters.

Your reports should include the necessary background, a discussion of the control parameters. Provide detail of the implementation of your isolation forest. Provide details on the empirical process, present and discuss the results.

Option 4: An Ensemble-based Breast Cancer Predictive Model

For this assignment you will evaluate the performance of a homogeneous ensemble to that of a heterogeneous ensemble, using the breast cancer dataset described in the section below. Note that a homogeneous ensemble contains ensemble members that are all of the same machine learning algorithm. A heterogeneous ensemble contains ensemble members of different machine learning algorithms. The purpose of this assignment is to determine which of the homogeneous or the heterogeneous ensemble performs best for this classification problem. You may select any homogeneous ensemble. Carefully consider the algorithms that you will include in the heterogeneous ensemble.

Note that you have to investigate the best values for the control parameters of the ensembles. Also decide on an appropriate voting approach. Describe these control parameters and the process that you have followed to find best values.

Your report should compare the two approaches to determine which approach performs best. For this purpose, you have to decide on appropriate performance measures, which you have to describe in your report. In your report, offer an explanation for why the identified approach performed best.

The Breast Cancer Dataset

Both topics will use the same dataset, `breastCancer.csv`, a breast tumors dataset. The target feature is `diagnosis`, which indicates if a tumour is malignant (M) or benign (B). The dataset contains the following descriptive features: `id`, `diagnosis`, `radius_mean`, `texture_mean`, `perimeter_mean`, `area_mean`, `smoothness_mean`, `compactness_mean`, `concavity_mean`, `concave points_mean`, `symmetry_mean`, `fractal_dimension_mean`, `radius_se`, `texture_se`, `perimeter_se`, `area_se`, `smoothness_se`, `compactness_se`, `concavity_se`, `concave points_se`, `symmetry_se`, `fractal_dimension_se`, `radius_worst`, `texture_worst`, `perimeter_worst`, `area_worst`, `smoothness_worst`, `compactness_worst`, `concavity_worst`, `concave points_worst`, `symmetry_worst`, `fractal_dimension_worst`, `gender` and `Bratio`.

Study this dataset so that you fully understand the data types, and any potential data quality issues. In an appropriate section of your report, discuss the data quality issues, and any data transformation and pre-processing that you have done in relation to your chosen topic.

Technical Report writing

The following is a general guideline of how to structure your report.

Title Section

Provide your report with a title, and as author provide your initials, surname and student number. Also provide an email address.

Abstract

Provide a very concise summary of what this report provides. Provide some context, the goals, how these were achieved, and the main observation. The abstract should be short. No more than 300 words.

1. Introduction

The introduction sets the stage for the remainder of your report. You usually have very general statements here. The introduction prepares the reader for what to expect from reading your report. In general, the introduction should either contain or be a summary of your ENTIRE report.

2. Background

A very high level discussion on the problem domain and the algorithms and/or approaches that you have used. Do not be too specific on the algorithms and approaches. This section is typically where the “base cases” of concepts that appear throughout the remainder of your report are discussed. It is also an ideal place to refer a reader to other sources containing relevant information on the topic but which is outside the scope of your assignment. It is the perfect place for pseudo code. Remember to discuss very generally. After reading this section the marker should be able to determine whether or not you know what you’re talking about.

3. Implementation

In this section you discuss how you approached, implemented and solved your assignment choice. You provide pseudo code where necessary and discussions of the solutions that you have implemented. This is also the section where your discussion specializes on the concepts mentioned in the background section. Be very specific in your discussions in this section.

4. Empirical Procedure

Here you describe the empirical procedure followed to apply your algorithms to obtain answers to the goals/hypothesis of the study. You elaborate on the performance measures used and provide the benchmark problems used. Provide all control parameter values with a motivation for why you have used these, and state the number of independent runs. After reading this section (in addition to the background) the reader should be able to duplicate your experiments to obtain similar results to those obtained by you.

5. Research Results:

This is the section where you report your results obtained from running the experiments as discussed in the implementation section. You have to give, at least, averages and standard deviations for the experiments/simulations. Thoroughly discuss the results that you have obtained and provide clear arguments in support of your results and observations from these results. Answer questions like “are these results to be expected?”, “why did these results occur?” and “would different circumstances lead to different results?”.

6. Conclusion(s):

Very general conclusions about the assignment that you have done. This section “answers” the questions and issues that you’ve raised and investigated. This section is, in general, a summary of what you have done, what the results were and finally what you concluded from these results. This is the final section in your document so be sure that all the issues raised up until now are answered here. This is also the perfect section to discuss what you have learnt in doing this assignment.

References

Provide all references that you have consulted.

Mark Rubric

Your report will be evaluated as follows:

Aspect	Mark
Abstract	5
Introduction	10
Background	25
Implementation	10
Empirical process	20
Results & discussion	50
Conclusions	5
References	5
Linguistic quality	20
Total	150