Stellenbosch
UNIVERSITY
IYUNIVESITHI
UNIVERSITEIT

Machine Learning 441/741

Assignment 2: Nearest Neighbours & Decision Trees

Total: [170]
Deadline: 5 September 2025, 08:00

Instructions

When completing this assignment, follow the instructions given below:

- The assignment must be completed by each student individually.
- You may use any programming language to complete the assignment, and you may make use of any libraries, including machine learning libraries.
- You have to write a report, using the IEEE conference template (google!), in two-column, 10pt format. Please see section 5 for tips on report writing. Also see the writing rules – uploaded to your STEMlearn module – for a list of writing rules. Please consult these writing rules and apply them. Note that only the report will be evaluated, not your code. Therefore, a bad report will result in bad marks.
- Submit your report in pdf format. Note that no format other than pdf will be accepted. Make sure that your report has a reference to your git repository for your code. Code will not be evaluated, but may be scrutinized if found necessary. Make sure that you name your report file as `????????RWxxxassignment2.pdf`, where you replace the question marks with your student number, and `xxx` with the module code that you are registered for. Please note that I use a script to pull out all reports, and if you do not follow this file naming convention, your report **will not** be extracted for evaluation.
- Make sure that your name, surname and student number are clearly indicated in the front matter (title section) of your report. If there is no identification of the author of the report in the front matter of your report, your report **will not** be evaluated.
- Generative AI tools may not be used.
- Upload your report via STEMlearn before the deadline of **5 September 2025, 08:00**. Note that late assignments will not be accepted. After this deadline, I will extract all reports to start with evaluation of the reports that day.

1 Purpose of the Assignment

The main objectives of this assignment are to test your ability to develop classification models for the provided dataset, to identify data quality issues that need to be addressed for the classification models employed, to identify and implement only the necessary data transformations for the employed classification models, to compare the performance of the implemented models, and to write a report. This assignment focuses on only two machine learning classification models, namely k-nearest neighbours and classification trees.

With reference to the machine learnig models, you may use any libraries to implement these models. You may also use any libraries to perform the necessary pre-processing and statistical analysis of the performance of the models. If you do make use of libraries, please provide details of these libraries in the approriate section of your report.

With reference to the report, you have to write a report wherein you provide responses in clear narrative on the aspects enumerated below, under appropriate section headings. Note that if figures and tables are provided in the report, that figures and tables will not be looked at if you do not refer to them in the report and if you have not discussed them.

2 The Dataset: Forest Cover Type

You have been provided with a dataset for assignment 1, with the goal to identify data quality issues. This dataset was an edited version of the forest cover type dataset, with various data quality issues explicitly injected within that dataset. This section provides detail on the original forest cover type dataset, the data quality issues that were present in the assignment 1 version of the dataset, and then a short explanation of data quality issues that remain in the dataset provided for assignment 2.

Details on the Forest Cover Type Dataset

The dataset contains tree observations from four areas of the Roosevelt National Forest in Colorado. All observations are cartographic variables (no remote sensing) from 30 meter $\times$ 30 meter sections of forest. This dataset includes information on tree type, shadow coverage, distance to nearby landmarks, soil type, and local topography. The problem is to classify each observation into one of seven forest cover types, namely spruce/fir, lodgepole pine, ponderosa pine, cottonwood/willow, aspen, douglas-fir, or krummholz. For more detail on the original dataset and its descriptive features, please refer to the uploaded file `covtype.info`.

The Data Quality Issues

The dataset that you have received for assignment 1 was a doctored version of the forest cover dataset, with a number of data quality issues injected into this dataset. A list of these data quality issues is provided below, in no particular order:

- A1 values are order of magnitude larger than that of the other descriptive features.
- A3 and A2 are correlated, with A3 generated as A2$\times$(2.5 + rand()/100). Therefore, redundant feature.
- A4 has missing values.
- A5 has an incorrect negative value; most of the values for this feature are positive, so the negative value raises a concern.
- A7 is irrelevant – all values are randomly selected from a uniform distribution.
- A8 has large positive outliers.
- A10 has invalid character values for this numerical descriptive feature.
- A16 has cardinality of 1
- A17 has cardinality of 1
- A18 has potential irregular cardinality due to an invalid categorical value; most of the values are binary, with very few incorrect values set to 2 instead of binary.

- A19 has a number of noisy instances (sorry, this one would have been difficult to detect).
- A20 has missing values.
- A21 has two issues. Firstly it is highly correlated with A20. Secondly, it has many missing values.
- A61 is a unique descriptive feature.
- A62 has a constant value.
- The target feature has some missing values.
- The target feater has a skew class distribution.
- Some of the observations have too many missing features.

Revised Forest Cover Type Dataset

The dataset uploaded to STEMlearn as `forestCover.csv` remains an adapted version of the original dataset. For this version of the dataset some of the data quality issues as listed above have been retained. The following data quality issues remain in the datset:

- There remain features with missing values, with missing values indicated by the character symbol '?'.
- The `Facet` feature is correlated with the `Aspect` feature.
- The `Inclination` feature contains only noisy values.
- There remain features with outliers.
- There remain features with numeric ranges that differ significantly from one another.
- There are numerical and categorical features.
- Feature `Water_Level` has cardinality of one.
- Feature `Observation_ID` has a unique value for each observation.
- The class distribution remains skew.

3 Tasks to Complete

Complete the assignment in the following steps:

1. Download the **forestCover.csv** dataset, which is a comma delimited dataset. You have to use this dataset and not the dataset available online. Note that the target feature is the last feature, and that there are a total of 58 descriptive features (this is a little more than the original dataset that you will find online).
2. Provide in your report a discussion of k-nearest neighbours and classification trees.
3. Follow this discussion with your expectations of the impact of the different remaining data quality issues for each of the two machine learning models.
4. For each of the machine learning approaches, discuss the data-preprocessing steps that you have implemented to optimally transform the dataset for that specific machine learning approach and to correct data quality issues. Note: do not do unnecessary data transformations. Carefully think about the data transformations needed for each of the machine learning algorithms, and apply only those. Provide justifications for each of these pre-processing steps. Should you decide not to address a data quality issue, justify this decision. Note that the bulk of the marks are for your justifications.
5. Develop the two predictive models and evaluate the performance of the two models on your pre-processed dataset. Make sure to construct optimal configurations of your chosen models both with respect to architecture and values for control parameters. Describe the process that you have followed to produce an optimal configuration for each model. For this purpose, carefully decide on the performance metrics that you will use. Conclude on which one of the two approaches is best for this problem, and support your conclusion with justifications. For the purposes of this assignment, make sure to report the performance based on a k-fold cross-validation. Decide on the number of folds with a justification.

4 Mark Rubric

The assignment will be assessed as follows:

Aspect		Mark
Title & Abstract		5
Introduction		10
Background		
	Machine learning algorith description	10
	Expectations wrt data quality issues	20
Implementation		10
Empirical process		
	Data pre-processing	20
	Control parameter tuning	10
	Performance metric	5
	Analysis process	10
Results & discussion		40
Conclusions		5
References		5
Linguistic quality		20
	Total	170

5 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. The purpose of the abstract is to convince the reader to continue reading your report.

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 be a summary of your entire report. Start by stating the context, moving towards the goals. Then elaborate on how these goals have been obtained, what you have done. Give a motivation for why this is done. Summarize the main observations of the study. You basically give a teaser to the reader, to convince the reader to continue reading the report. Give an outline of the remainder of the 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 of existing approaches. 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. Keep in mind that this is a background section, and does not contain any detail on what you have done, but only provides a summary of related background to understand what you have done.

3. Methodology

In this section you discuss how you have approached, implemented and solved your assignment problem. You provide pseudo code where necessary (only for new algorithms) 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, to clearly describe what you have done and how you have done it.

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. If statistical tests are used, these are discussed here. 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, using the empirical procedure above. 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

Start this section by stating again the goals of the report, what was done and how. Very general conclusions about the assignment that you have done are given. This section "answers" the questions and issues that you have raised and investigated. 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, and to provide any ideas for future work.

References

Provide all references that you have consulted.