Machine Learning 441/741

Assignment 1: Data Quality Issues, Nearest Neighbours, Decision Trees

Total: [204]
Deadline: 2 September 2024, 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 for tips on report writing. Also see the writing rules uploaded to your SUNlearn module for a list of writing guidelines. Please consult these writing guidelines 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 `????????assignment1.pdf`, where you replace the question marks with your student number. 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 SUNlearn before the deadline of **2 September 2024, 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.

Classification of Dry Beans

The main objective of this assignment is to test your ability to identify data quality issues in a large dataset, and to decide on the best approaches to handle these data quality issues with reference to the two classification algorithms that will be applied to classify dry beans type. For the purposes of this assignment you will develop two different classification predictive models, i.e. a k-nearest neighbour algorithm and a classification tree.

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 these 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.

Complete the assignment in the following steps:

1. Download the `DryBeanDataSet.xlsx` dataset. The dataset contains 13611 instances, 20 descriptive features, and the class feature `Class` in column U.
2. Without changing anything in the provided dataset, provide a table wherein you characterize all of the features of the dataset. You need to think of characteristics that are important to facilitate the process of understanding the data and the identification of data quality issues. (20)
3. You now have to very carefully explore the dataset to identify data quality issues. For this part of your report, only identify the data quality issues and provide justifications for these issues. Note that the bulk of the marks will go for your justifications. (40)
4. Provide in your report a discussion of the k-nearest neighbour algorithm and classification tree algorithm that you have implemented (or used). (20)
5. 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. Again, the bulk of the marks are for your justifications. (100)
6. 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. (60)

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.