Assignment report, group number

Your Names Go Here

Leiden Institute of Advanced Computer Science, The Netherlands

Abstract. This document contains the format for the report required for submission of the practical assignment for the course Introduction to Machine Learning. The tasks for this assignment are provided in Appendices A-D.

1 Introduction

This document serves as a *description of the practical assignment* for the course Introduction to Machine Learning. For this assignment, you are provided with a data set which you should analyze using some of the algorithms discussed during the lectures or this course. The assignment report should be written as a *scientific paper* and submitted together with the code (in Python, mainly using the scikit-learn library [1]).

To help you structure your report, we provide you with a *brief report outline* in this document. Please complete the following sections with your own results, explanations and conclusions. This includes the abstract and this introduction!

Appendices A-D contain the *specification of the tasks of the assignment*. Do not include them in your report.

2 Data Set

The data set (available on Brightspace) contains data about bike rentals in a large European city. The main learning task for this data set is predicting the amount of bikes rented (by subscribers to the service and by non-subscribers) based on the other features in the data set, but we will also define some additional tasks during this assignment.

In the remaining part of this section please add your description of the data set you are provided with.

2.1 Problem formulation

Please add problem description here.

3 Experiments

This is the main section of your report. All methods, experiment descriptions and results should be included here.

4 Conclusion and future work

Conclude your most important findings, and what you can learn from them. Identify some points on which can be improved in future, or areas where other algorithms might be useful.

References

1. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)

A Content of part 1 of the assignment, deadline of 18.10.2021

1. Identify what variables are present in the data set, how they are distributed, what type of variables they are. Apply some pre-processing if this is needed to make the data usable[1]. Make use of different ways to visualize the data, and look at the correlations between different features[2]. (This should be part of Section 2 of your report.).
2. Formulate the problem of predicting the number of bikes rented based on the other features present. Use the terminology that has been used in the lectures. (This should be part of the 'problem formulation' part (Section 2.1 of your report.)
3. Split the data into two sets: train and test. Train a linear regression method to predict the total number of bikes rented based on the data in the training set, and verify the performance of this regressor on the test set. Identify how its performance varies based on how large the training set is (vizualise this, for example using matplotlib or similar packages). For this linear regressor, try to experiment with different transformations of the target, as this often has a large impact on the R-squared metric. Explain why this is the case! This part, and all of the following tasks, should be part of the experiments section (Section 3).
4. Create a decision tree regressor to predict the number of bikes rented by subscribers. For this algorithm, identify what parameter settings you can modify, and explain what these parameters control. Select the one which has the most impact on the test performance, and create a plot showing how this parameter impacts both train and test error, and identify the ideal setting based on this plot. Then, apply these same parameter settings to predicting the number of bikes rented by non-subscribers. Are these settings optimal in this case as well? Clearly motivate your answers.

For your report, make sure you explain the working principles of the methods you use and reason why they lead to the found results. Use relevant visualizations and explain what is being shown (every figure needs to have a caption, and be referenced in the text). The reasoning and discussion about the methods used is key in showing that you understand the concepts, and is thus the most important part in deciding your assignment grade. Since this is a scientific report, make sure to cite all references you use (papers, books,...)!

[1] Hint: Look at the variable types. Any strings should be transformed to numeric, and simple categorical variables might be better suited to be turned into binary features (look into one-hot-encoding),.... You might also want to exclude the 'date' feature.

[2] Hint: For some inspiration on the kind of plots you can create, you can look at the practicums, or go to https://seaborn.pydata.org/examples/index.html

A.1 Submission of assignment part 1

Each group should submit:

1. The python code used to generate all data and figures used in this report. This should be structured clearly, so it can be easily run by reviewers. Pay attention to your coding style and use enough comments (in English). You should use a jupyter notebook for this, as this allows you to give a clear structure to your code. Your code should be one file only![3]
2. Pdf file of the report typeset in LaTeX[4], following the format outlined in this document.

Submission for this part of assignment is mandatory and is to be made via Brightspace. You will not receive a grade for this, but you will be given feedback via Brightspace. This submission serves as a basis for part 2 of the assignment.

B Content of part 2 of the assignment, deadline of 06.12.2021

The second part of the assignment will be made available at a later date (middle of October).

C Peer review

The assignment includes students individually carrying out the reviews of assignment reports from other groups.

D Grading

The first part of the assignment will not be graded, but you will receive some feedback on it to improve the final submission. At the final deadline for part 2, the content of part 1 will be graded as well, including based on how you incorporate the received feedback.

[3] Hint: You can use https://git.liacs.nl to host and share code with your teammates. You can log in using your ULCN username and password.
[4] Hint: You can use free version of https://overleaf.com to edit your report as a group