Some interesting title

Your Names Go Here + Group Number

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. .

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, using libraries such as scikit-learn [1] is highly encouraged).

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! You can deviate from this provided structure if it increases the readability of your submission.

2 Data Set and Problem Formulation

The data set (available on Brightspace) contains data about different astronomical observations and their classification into one of three types of objects. The following columns are available:

alpha: Right Ascension angle (at J2000 epoch)

delta: Declination angle (at J2000 epoch)

u: Ultraviolet filter in the photometric system

g: Green filter in the photometric system

r: Red filter in the photometric system

i: Near Infrared filter in the photometric system

z: Infrared filter in the photometric system

field_ID: Field number to identify each field

MJD: Modified Julian Date, used to indicate when a given piece of SDSS data was taken

redshift: Redshift value based on the increase in wavelength

plate: Plate ID, identifies each plate in SDSS

class: Class labels (target for your classification tasks)

In the remaining part of this section please add your description of the data set you are provided with and the learning tasks you will tackle in the next sections. You could look at 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]. You could make use of different ways to visualize the data or look at the correlations between different features.[2]

3 Experiments

This is the main section of your report. All methods, experiment descriptions and results should be included here. You have a lot of freedom in the type of experiments you choose to perform, but you should make sure to clearly explain (and motivate) the setup and describe the methods you use. Any results you obtain should be discussed clearly. Your experiments should contain at least the following:

- Pick one of the classifiers we discussed throughout the course, and run it on the prediction task (using cross-validation). Identify at least one key parameter of your chosen classifier, and show how this impact the cross-validation score. Explain why this is the case!
- Use any of the dimensionality reduction methods discussed in the course to reduce the dimensionality of the data set to 2 (use all features except the class you predicted, and any ID variables). Perform clustering on both the original and the dimension-reduced data. What differences do you find when changing the ordering of dimensionality reduction and clustering? Do the clusters found match with the type of object you were predicting before?
- Use any method discussed during the lectures to get a predictor which is as accurate as possible. Motivate why you choose this method, and identify why this manages to achieve these levels of accuracy.

You can include any number of additional experiments, e.g. parameter tuning, comparing different dimensionality reduction techniques, visualizing decision boundaries. . . , whatever experiments you find most interesting.

Within this report, make sure you briefly 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 should 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,. . .)!

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.

[1] Hint: Look at the variable types, and check if these are directly usable or might be better to transform (e.g. strings to numbers). You should remove attributes which are not relevant to the prediction task.

[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

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)

5 Submission

Submission of the assignment goes through Brightspace only, where you submit a pdf of you report, and a single python (or notebook) file containing the code used to run your experiments and create your plots. Your report should be a maximum of 8 pages long (excluding references).