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-**??**.

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-C 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

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

1. Transform the problem of predicting the total number of bikes rented into a classification problem (around 5 classes). Turn the classification version of this problem into multiple binary classification problems. Use both one-vs-all and one-vs-one methods and solve this using logistic regression. Make an in-depth comparison of these two methods of splitting the problem into binary classification.
2. Use a perceptron method to predict the weather situation based on other some other features in the data set. Use feature subset selection to determine which features to use. Use the one-vs-one methods to turn it into binary classification. Does this algorithm find a perfect separating hyperplane? Why / why not? How does the scaling factor in the perceptron update rule ('learning rate') impact the behaviour of the algorithm and the train and test error (on one of the binary classification subtasks)?
3. Identify whether or not the data set contains any outliers. Motivate your answer by providing an example of an outlier detection method applied to the data.
4. Use an SVM regressor to predict the number of bike rentals by subscribers. Identify how using different kernels impacts the performance. Make sure to use cross-validation for this experiment, and motivate your choices.
5. Use PCA to reduce the dimensionality of the data set to 2 (use all features except the number of rented bikes). Identify what happens when some features are left out of the data set before applying this transformation. What does this tell you about these attributes?
6. Cluster the data using hierarchical clustering (use all features, except number of rented bikes). Do any intuitive clusters emerge? Select some limited number for the amount of clusters, and project this to the 2D PCA-space created in the last exercise. Also apply clustering to the PCA-transformed data. How do these two methods of combining PCA and clustering differ? What could this tell you about the data?
7. Use a random forest method to predict the number of bike rentals. Tune the parameters of this model with any method you like (e.g. grid search or random search). Make sure to use a train, test and validation set, and compare this to the out-of-bag score given by the forest itself. Compare the structure of the individual trees to the decision tree you got in part 1.
8. 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.

C Submission + Peer review

The assignment includes students individually carrying out the reviews of assignment reports from other groups. The peer review system will be opened on November 8, and instructions for submission will be provided at that time. Note tat at the final deadline for part 2 (which is December 6th, 16:00), the content of part 1 will be graded as well, including based on how you incorporate the received feedback.