FACULTY
OF INFORMATION
TECHNOLOGY
CTU IN PRAGUE

ASSIGNMENT OF BACHELOR'S THESIS

Title:	Benchmarking of algorithms for machine learning
Student:	Tom Svoboda
Supervisor:	Ing. Viktor Černý
Study Programme:	Informatics
Study Branch:	Web and Software Engineering
Department:	Department of Software Engineering
Validity:	Until the end of winter semester 2021/22

Instructions

In an unnamed company selling SaaS products, we want to offer the best service possible to our customers. To add additional value to them we want to predict their needs based on their input data. For this purpose, we can use machine learning algorithms, which create a prediction model for each customer. We assume various algorithms and their configurations can have different success rates for each customer type. The goal of this thesis is to create a tool, that can automatically evaluate the quality of created prediction models.

- Create a methodology, which evaluates the quality of prediction for each model against expected results.
- Apply this methodology in a tool, which automatizes the evaluation of these models.
- The tool will provide an output as feedback for developers of machine learning algorithms in a way, that will improve the quality of said models.

References

Will be provided by the supervisor.

Ing. Michal Valenta, Ph.D.
Head of Department

doc. RNDr. Ing. Marcel Jiřina, Ph.D.
Dean

Prague February 24, 2020

FACULTY
OF INFORMATION
TECHNOLOGY
CTU IN PRAGUE

Bachelor's thesis

Benchmarking of algorithms for machine learning

Tom Svoboda

Department of Software Engineering
Supervisor: Ing. Viktor Černý

August 25, 2020

Declaration

I hereby declare that the presented thesis is my own work and that I have cited all sources of information in accordance with the Guideline for adhering to ethical principles when elaborating an academic final thesis.

I acknowledge that my thesis is subject to the rights and obligations stipulated by the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular that the Czech Technical University in Prague has the right to conclude a license agreement on the utilization of this thesis as a school work under the provisions of Article 60 (1) of the Act.

In Prague on August 25, 2020 . . . . . .

Czech Technical University in Prague
Faculty of Information Technology

Citation of this thesis

Svoboda, Tom. *Benchmarking of algorithms for machine learning.* Bachelor's thesis. Czech Technical University in Prague, Faculty of Information Technology, 2020.

Abstrakt

Cílem práce je vytvořit metodiku pro hodnocení modelů strojového učení. Následně použít tuto metodiku v nástroji, který automatizuje hodnocení modelů a dává zpětnou vazbu jejich vývojářům.

Klíčová slova hodnocení modelů strojového učení, porovnání prediktivního modelování

Abstract

The goal of this work is to create methodology to evaluate machine learning models. Then use this methodology in a tool, to automate the evaluation of models and provides feedback to their developers.

Keywords machine learning model evaluation, benchmark predictive modeling

Contents

List of Figures

Introduction

Motivation

In an unnamed company selling Software as a Service (SaaS) products, we want to offer the best service possible to our customers. To add additional value to them we want to predict their needs based on their input data. For this purpose, we can use machine learning algorithms, which create a prediction model for each customer. We assume various algorithms and their configurations can have different success rates for each customer type.

Motivation of this thesis is to help with development of such algorithms and provide a tool for selecting the best one for each customer.

The goal is not to compare machine learning algorithms in general, but to compare their usability on selected problem.

This work is focused only on supervised machine learning.[1]

Aim of the Thesis

The goal of this thesis is to create a tool that can automatically evaluate the quality of machine learning algorithm for prepared data sets.

The initial goal is to create a methodology, which evaluates the quality of prediction for each model against expected results. The consequent goal is to apply this methodology in a tool, which automatizes the evaluation of these models. The final goal is for the tool to provide an output as feedback for developers of machine learning algorithms in a way that will improve the quality of said models.

[1] This choice is explained in chapter 2.2

Chapter 1

Machine Learning

1.1 What is machine learning?

Machine learning is set of methods that can automatically detect patterns in data, and then to use the uncovered patterns to predict future data or to perform other kinds of decision making under uncertainty.[1]

1.1.1 Types of machine learning:

1. supervised learning (predictive)
2. unsupervised learning (descriptive)
3. reinforced learning

The goal of **supervised learning** is to learn how to map inputs into outputs given a labeled set of input-output pairs.

The goal of **unsupervised learning** is to look for previously undetected patterns in a data set with a minimum of human supervision.

Reinforced learning applies to problems where an agent is interacting in an environment trying to accomplish some task. Reinforced learning teaches the agent through positive or negative feedback derived from its action.[1, 2]

1.2 Predicting with supervised learning

1.2.1 Types of problems

There are two types of problems supervised learning solves. First is regression - predicting continuous scalar value for input data (predicting the height of a person from their sex and age). Second is classification - to classify input data (predicting illness from the patient's symptoms).[3]

1.2.2 Process of developing prediction model

We have clearly defined a prediction problem we want to tackle. Now we want to create a model to predict the outcomes. At first, we need to obtain data from which the learning algorithm will learn. The data needs to be paired with expected outcomes. If the outcomes are not available in the data set, they need to be added manually. Now we choose or develop the learning algorithm that is suitable for the problem and data-set. When ready, we can run the algorithm on the prepared data-set with expected outcomes. The output of this algorithm is a trained prediction model. To make predictions, we can feed the prediction model with additional data-points, and it will output the prediction of outcome.

1.3 How to assess accuracy of prediction model?

Multiple statistical metrics can be used: accuracy, sensitivity, specificity, Matthews correlation coefficient

1.4 Existing tools

There are existing automated machine learning development tools (known as AutoML) eg.: H2O.ai, Azure Machine Learning.

They provide the necessary functionality and infrastructure to apply methods described in this document.

But problem is, they don't provide easy way, how to automate creating multiple models from set of data with same algorithm[2]. Also, as they can solve a wide range of problems, more upfront learning is required before one can use them. The last downside is they are proprietary licensed software.

[2]on May 14, 2020 Microsoft released Many Models Solution Accelerator, which implements this functionality for Azure Machine Learning

Chapter 2

Analysis and design

2.1 Terminology

Learning algorithm - algorithm that creates predictive model from input data Predictive model - parametric model, that makes prediction for data with same structure as data it was learned on.

2.2 Choosing area of machine learning to focus on

This work is focused only on supervised learning.

2.2.1 There are two parameters considered:

- it has to tackle problems in the area of SaaS, which this work will be supporting
- it is plausible to automate

The first requirement draws out reinforced learning. Both supervised and supervised learning can solve many problems in the selected area. In the case of supervised learning, we can think of business predictions such as customer churn, or as possible product features such as automating user actions. For unsupervised learning - there is automatic content processing such as cluster analysis or finding outliers in customer behavior.

The second requirement is satisfied by supervised learning. It is easier to evaluate then descriptive learning. The goal of descriptive learning is to find interesting patterns. Evaluating how much the finding is interesting depends on each particular problem. That makes it harder to generalize the evaluation process. The methodology of evaluating prediction learning is explained in chapter 3.3.

2.3 Methodology of evaluating learning algorithm

This methodology aims to find an algorithm that produces the best results for the problem it wants to solve.

The results are prediction models created for each customer[3].

The evaluation of results depends on the problem definition. As explained in chapter 1.3, several metrics can be useful for different problems. So first step is to select (or define) the metric according to the need.

As machine learning algorithms rely mostly on input data we need to treat it as constant input if we want to compare the algorithms. Now the only variable is the learning algorithm itself.

2.3.1 Process of evaluating algorithm:

1. split the input data to training and testing
2. run machine learning algorithm that produces a model
3. compute metric with testing data used on created model

For comparison of algorithms, the training and testing data must stay constant when comparing algorithms. Otherwise, the selection could affect results. When making predictions for customers based on their historical data, it is best to split the data by time (eg.: last month would be testing data, earlier data are training data).

2.3.2 Process of evaluating algorithm for customers:

1. split the input data by customer
2. evaluate learning algorithm for each customer using only their data
3. gather computed metrics

With computed metrics for customers data scientists can decide on next steps. They can select eligible customers for this algorithm. This selection process could be automated by defining a threshold that identifies success in selected metric.

[3]the process is specified in chapter 1.2.2

Conclusion

Defining the methodology of evaluating machine learning algorithms depends on the problem definition it tries to tackle. Because of that, it was necessary to limit the scope of this work to those areas of machine learning, that are less dependable on the problem definition and exact methods can be used. Within that scope, the methodology explained in this work provides satisfactory results on tested problems.

The tool that applies that methodology wasn't finished.

Future work could extend methodology and the tool to support unsupervised learning to enable a broader range of problems to be solved more efficiently. The tool could provide guidelines for selecting the best evaluation metric.

Bibliography

[1] James, G.; Witten, D.; et al. *An Introduction to Statistical Learning: with Applications in R*. Springer Texts in Statistics, Springer New York, 2013, ISBN 9781461471387.

[2] Shalev-Shwartz, S.; Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*. Understanding Machine Learning: From Theory to Algorithms, Cambridge University Press, 2014, ISBN 9781107057135. Available from: `https://www.cse.huji.ac.il/~shais/UnderstandingMachineLearning/`

[3] Murphy, K. *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning series, MIT Press, 2012, ISBN 9780262018029.