Master Thesis Project Description

—

Inside the Black Box: How to Explain Individual Predictions of a Machine Learning Model

Marc Beillevaire
marcbei@kth.se

January 24, 2017

Abstract

This document describes the project that I will be conducting for my Master Thesis. This project will be carried on at Dataiku, a french data-science start-up in Paris. Into their datascience team, I will be researching on how to better understand predictions made by a machine learning algorithm, which is quite difficult in most cases today. This document provides both the project description and practical information on how it will be conducted.

1 Background and conditions

Dataiku is a software editor that develops Dataiku DSS (*Data Science Studio*), a collaborative data science platform that enables companies to efficiently process their data and build machine learning models. Dataiku also have an important datascience team that provide support in datascience, along with the software, for the clients.

During a six-month internship in their data science team, I will be researching on how to explain predictions of a machine learning model. Managing to have an insight on how a single prediction is made by models is a key element in many areas where machine learning algorithms are used. Research on this subject is starting to emerge as more and more use cases appear. But it is not really integrated into data processing softwares and thus still not used at a large scale by most companies.

The main goals are: exploring bibliography on the subject, testing several methods on various datasets and linking theses methods to Dataiku DSS to make it easy to explain predictions.

2 Research question

As stated before, the research will focus on how to produce understandable explanations related to a prediction made by a machine learning algorithm.

Most of these algorithms today still look like black boxes for most users: it is not easy - if not impossible - to understand how a given prediction is made by a Random Forest algorithm with several hundreds of trees. Several tools already exist to understand models but there are mostly global coefficients like the variable importance in a tree or a forest for instance.

Yet in many cases having an insight locally would be really helpful to users. An insurance company would like to know why a particular client has been detected as a fraudster, or a manufacturer doing predictive maintenance would like to know not only which device is likely to break soon but also how this device's lifetime could be increased practically. Moreover a European regulation has been adopted by the EU parliament and will force companies to provide *"explanations"* on every decision made by an algorithm that impact their client. Even if the application conditions has not been stated clearly up to now, this will require good explanations system to understand a particular decision. For instance, someone who has been refused a load could ask the bank for a specific and logical reason, even if this decision has been obtained by a computer.

A first look to this problem leads to two types of solutions to answer this question: either one would look at how a specific algorithm works, or do it in a model-agnostic way. In the first case, the goal is to retrieve the origins of how an output probability could be computed by looking inside the model. This is easy with linear models, such as Logistic Regression. They are in itself easily understandable. But for trees and tree ensemble methods, that would mean retrieving the contributions of each features from the decision trees.

3 Evaluation method

Most of the time, methods will be evaluated using practical tests on several dataset. Good dataset sources could be websites like UCI, or datascience challenges platforms such as Kaggle.

Some datasets would probably better suits some methods, but to be genral enough, testing datasets should be diverse: regression and classification datasets, text data and tabular data, ...

4 Background knowledge

I have attended Machine Learning courses at both my home university (Télécom ParisTech) and KTH, and this project is here to conclude my master in Machine Learning at KTH. I believe it fulfills all the requirements to be qualified as a good research project for a master thesis in this KTH master's program.

5 Supervisor at Dataiku and ressources

Pierre Gutierrez (pierre.gutierrez@dataiku.com), datascientist at Dataiku will supervise this project in the company. I will have weekly (at least) meetups with him to check how the project is going on.

Dataiku provides me with a computer and a *Dataiku Data Science Studio* license to test my results in the company's software.

6 Eligibility & study planning

My courses at KTH included all the mandatory courses in Machine learning and artificial intelligence. I've also passed the Introduction to the Philosophy of Science and Research Methodology course (DA2205).

As a foreign student in double degree I need to complete 90 ects from courses at KTH, of which I got 85.5. So I am missing 4.5 credits, and I am planning on taking another course this autumn to get all the 90 course credits to complete my master.