Explainable Machine Learning

Overview

Machine Learning has seen an unprecedented development over the last decade and offers a great promise of solving various predictive problems. However, the most accurate models are often black-box, i.e. they don't provide an explanation **why** they made a particular decision. This uncertainty is undesirable not only from the technical (model improvements) or legal point of view (e.g. GDPR). It can also introduce bias in a high-stakes decision that operate at scale (loan & credit card applications, promotions and recruitment, sentence parole). In order to gain/maintain **trust** in model predictions it's crucial to have access to model's interpretable explanations.

Prerequisites

This course is designed for practitioners who want to get a better understanding of their ML models and get a better knowledge of popular explainable ML frameworks. Attendees should:

- write R code at an intermediate level
- have a basic knowledge of binary classification and regression models
- familiarity with ML model training, testing and evaluation workflow

Outline

In this tutorial, we are going to show the pitfalls of trusting models based on their accuracy alone. Then, we will introduce two model-agnostic frameworks for explainable ML: LIME (Local Interpretable Model-agnostic Interpretations) and DALEX (Descriptive Machine Learning Explanations). We will show how to apply these methods to standard classification and regression problems and how this may change how much we trust original model predictions.

After the course, you will be able to:

1) Approach with caution pure accuracy measures.
2) Use a number of tools to explore and understand global drivers behind model predictions (DALEX).

3) Understand local drivers behind correctly- and –incorrectly predicted instances (LIME).
4) Compare a range of Explainable ML packages and frameworks (e.g. xgboostExplainer, iml, SHAP, Anchors) and choose the one that is most relevant to your predictive problem.

Facilitators

Kasia Kulma, Mango Solutions

Kasia Kulma is a Senior Data Scientist at Mango Solutions and holds a PhD in evolutionary biology from Uppsala University, Sweden. She has experience in building recommender systems, customer segmentations, web applications and running NLP projects. She is the author of the blog **r-tastic.co.uk** and is a mentor and organiser in R-Ladies London. She is an R-enthusiast interested in data (science) ethics and evidence-based medicine.

Hannah Frick, Mango Solutions

Hannah Frick is a Data Scientist at Mango Solutions and holds a PhD in statistics from Universitaet Innsbruck, Austria. She has authored and maintains several R packages on CRAN. She's a co-founder of the R-Ladies Global organisation and part of the leadership team. You can follow her on Twitter @hfcfrick.