Local Interpretability of Machine Learning Models

joint with Przemysław Biecek

Mateusz Staniak

Wrocław, 18 IX 2018

Warsaw University of Technology

Introduction

Agenda

1. Interpretable Machine Learning (IML) / Explainable Artificial Intelligence (xAI): a new research area
2. Local Explanations of Machine Learning Models
3. LIME and live methodology
4. Break Down method
5. Examples & summary

IML: the what

1. Growing area of research
   - first work: PDP (Friedman), prediction decompositions (Robnik-Šikonja)
   - Breakthrough: LIME (2016)
2. Many faces:
   - building explainable methods
   - explaining *black box* models
   - knowledge extraction from complex models

IML: the why

DAVID LAZER AND RYAN KENNEDY OPINION 10.01.15 07:00 AM

WHAT WE CAN LEARN FROM THE EPIC FAILURE OF GOOGLE FLU TRENDS

Figure 1: Taken from
https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/

IML: the why

SCIENCE TECH HEALTH

IBM's Watson gave unsafe recommendations for treating cancer

Doctors fed it hypothetical scenarios, not real patient data

By Angela Chen | @chengela | Jul 26, 2018, 4:29pm EDT

SHARE

Figure 2: Taken from https://www.theverge.com/2018/7/26/17619382/ibms-watson-cancer-ai-healthcare-science

IML: areas of research

Problems & questions in IML

- How well does the model perform? (Model performance)
- Which variables are most important in the model? (Feature importance)
- What is the relationship between predictors and response? (Variable contribution / variable response)
- What factors drive a particular prediction? (local explanations)

Types of explanations

1. Intrinsic vs **post-hoc**
2. Model-specific vs **model-agnostic**
3. Global vs **local** (with respect to predictors or **observations**)

Local Explanation

1. Goal: discover which factors drive the prediction for a single instance
2. Motivation:
    - understanding the model
    - model validation
    - most interesting explanation from end-user point of view
3. Two main approaches:
    - Local approximations
    - Feature contributions (prediction decomposition)

LIME & live

LIME methodology: basics

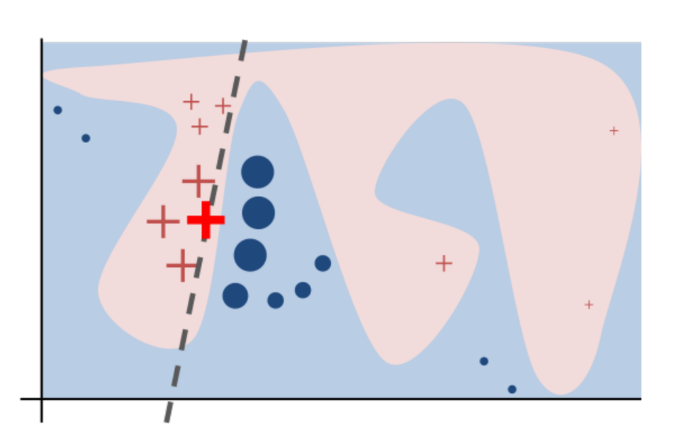

Figure 3: Intuition behind LIME methodology. Taken from [10]

LIME methodology: basics

- $x \in \mathbb{R}^d$ - instance being explained
- $x' \in \{0, 1\}^{d'}$ - interpretable representation
- $g \in G$ - a model that belongs to a class of interpretable models
- $\Omega(g)$ - measure of complexity of g (penalty term)
- $f(x)$ - explained model
- $\pi_x(z)$ - measure of closeness of z and x (kernel)
- $\mathcal{L}(f, g, \pi_x(z))$ - measure of unfaithfulness of local approximation

LIME explanation $\xi(x)$ is obtained by

$$\xi(x) = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_x(z)) + \Omega(g)$$

LIME methodology: details

- Sampling observations before fitting local model is called **local exploration**
- The **interpretable representation** in binary space is created before sampling
- Loss $\mathcal{L}(f, g, \pi_x(z))$ describes **local fidelity** of the explanation - how well the simple model fits the complex model
- Interpretability of the resulting approximation is ensured via the penalty term $\Omega(g)$

LIME methodology: summary

Note that LIME

- performs discretization of features before fitting local model
- depends on many hyper-parameters (local model, distance function, discretization, sampling method etc)
- is a special case of a more general framework called Shapley values[7]

In summary, LIME addresses

- **understanding** issue by approximating the complex model with an interpretable model,
- **trust** issue using accompanying sp-LIME algorithm, which picks representative instances and their explanations.

LIME methodology: example

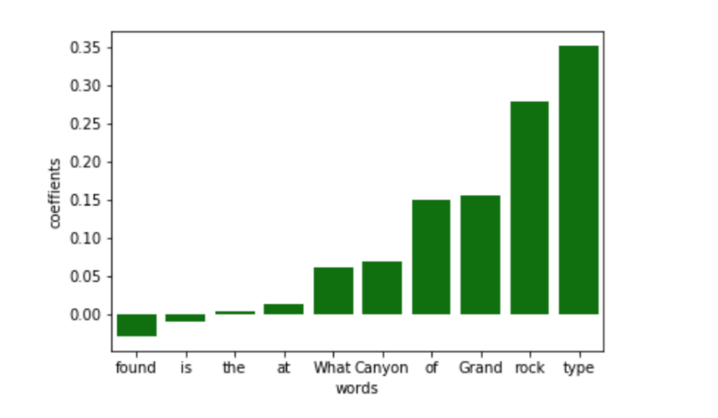

Figure 4: Example LIME application: words importance in QA systems. Taken from *How much should you ask? On the question structure in QA systems*, https://arxiv.org/pdf/1809.03734.pdf

live: introduction

Why?

- LIME for regression problems
- Model visualization in aid of LIME

How?

- Create dataset for local exploration by perturbing the explained instance
- Use original variables as interpretable inputs
- Provide optional variable selection
- Provide tools for model visualization
- Focus on interpretable models easy to visualize

live: work flow

1. Create dataset for local exploration: `sample_locally2` function
2. Add black box model predictions: `add_predictions2` function
3. Fit local explanation model to the prediction: `fit_explanation2` function
4. Visualize the result: `plot` function

live: more details

- Different methods of creating the new dataset are available, including by permuting each variable and by changing one feature per observations (the method does matter)
- We can control which variables are allowed to vary through fixed variables variable argument to sample locally (keeping date/factor/correlated variables unchanged)
- Black box model can be pre-trained or it can be trained using mlr, hyperparameters of both black box and explanation models can be set
- Variable selection is performed via LASSO regression

Break Down Plots

Break Down: basics

- Method of prediction decomposition
- Computes variable contributions
- Related to Shapley values
- Exact for linear models, greedy algorithm for any model
- Visualization method: waterfall plots AKA Break Down plots

Break Down for Linear Models

For linear regression, the difference between a particular prediction and an average prediction is given by

$$f(x^{new}) - \overline{f(x)} = (x_1^{new} - \bar{x}_1)\beta_1 + ... + (x_p^{new} - \bar{x}_p)\beta_p$$

Model-agnostic Break Down

All the following definition are taken from [9].

Definition (Relaxed model prediction)

Let $f^{IndSet}(x^{new})$ denote an expected model prediction for x^{new} relaxed on the set of indexes $IndSet \subset \{1, \ldots, p\}$.

$$f^{IndSet}(x^{new}) = E[f(x)|x_{IndSet} = x^{new}_{IndSet}]. \tag{1}$$

Thus $f^{IndSet}(x^{new})$ is an expected value for model response conditioned on variables from set IndSet in such a a way, that $\forall_{i \in IndSet} x_i = x_i^{new}$. Estimate:

$$\widehat{f^{IndSet}(x^{new})} = \frac{1}{n} \sum_{i=1}^{n} f(x^i_{-IndSet}, x^{new}_{IndSet}). \tag{2}$$

Model-agnostic Break Down

Definition (Distance to relaxed model prediction)

Let us define the distance between model prediction and relaxed model prediction for a set of indexes IndSet.

$$d(x^{new}, IndSet) := |f^{IndSet}(x^{new}) - f(x^{new})|. \tag{3}$$

Definition (Added feature contribution)

For j-th feature we define its contribution relative to a set of indexes IndSet (added contribution) as

$$contribution^{IndSet}(j) = f^{IndSet \cup \{j\}}(x^{new}) - f^{IndSet}(x^{new}). \tag{4}$$

It is the change in model prediction for x^{new} after relaxation on j.

Break Down: illustration

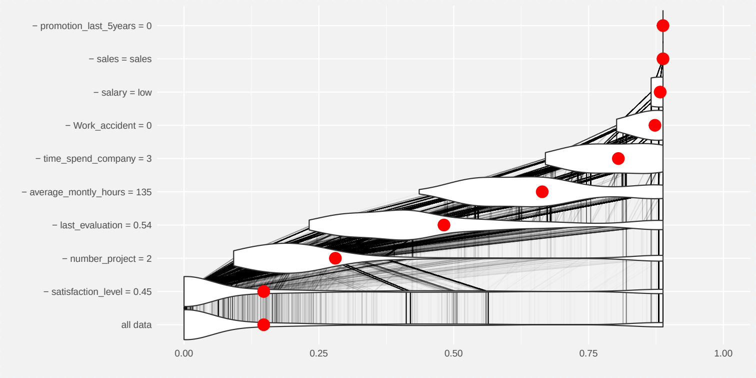

Figure 5: Illustration of Break Down method[9]

Example

live: R package

- R package `live` is available on CRAN
- Wine quality data.

```
# A tibble: 6 x 12
  fixed_acidity volatile_acidity citric_acid residual_sugar chlorides free_sulfur_dioxide total_sulfur_dioxide
          <dbl>            <dbl>       <dbl>          <dbl>     <dbl>               <dbl>                <dbl>
1           7.4             0.70        0.00            1.9     0.076                  11                   34
2           7.8             0.88        0.00            2.6     0.098                  25                   67
3           7.8             0.76        0.04            2.3     0.092                  15                   54
4          11.2             0.28        0.56            1.9     0.075                  17                   60
5           7.4             0.66        0.00            1.8     0.075                  13                   40
6           7.9             0.60        0.06            1.6     0.069                  15                   59
# ... with 5 more variables: density <dbl>, pH <dbl>, sulphates <dbl>, alcohol <dbl>, quality <int>
```

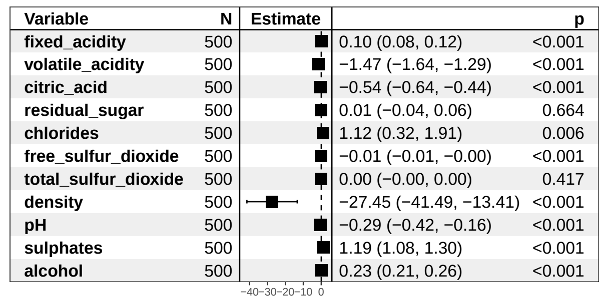

Figure 6: Forest plot[6] for the local linear model.

live: example

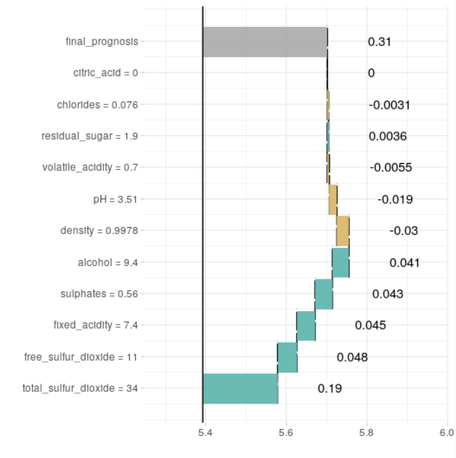

Figure 7: Waterfall plot for the local linear model.

Break Down: R package

Basic library: `breakDown`

`broken(model, instance, ...)`

As a part of DALEX package:

`single_prediction(dalex_explainer, instance)` or `prediction_breakdown(dalex_explainer, instance)`

Break Down: example

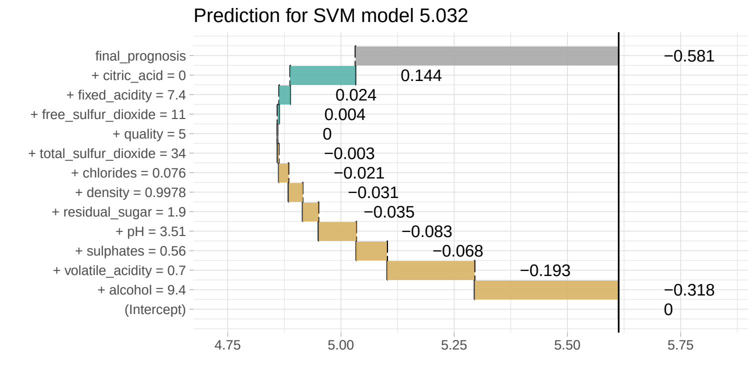

Figure 8: Waterfall plot for wine data - model-agnostic Break Down.

Summary

live & Break Down: challenges

- LIME & Break Down in high dimensional setting
- optimal way of generating *fake* dataset
- fit diagnostics for local and complex models
- detecting local interactions

IML: more

Software

- DALEX: R package
- iml: R package
- skater: Python library

Documentation

- DALEX docs: https://pbiecek.github.io/DALEX_docs/ (P. Biecek)
- Interpretable Machine Learning Book: https://christophm.github.io/interpretable-ml-book (Ch. Molnar)

Example open problems

- Interaction detection
- Model-specific explanations (deep learning...)
- Biased training

MI2DataLab

http://mi2.mini.pw.edu.pl/

https://github.com/mi2datalab

Acknowledgement

Uniwersytet Wrocławski

D. Basaj, B. Rychalska, P. Biecek, and A. Wroblewska.
How much should you ask? On the question structure in QA systems.
ArXiv e-prints, Sept. 2018.

P. Biecek.
DALEX: explainers for complex predictive models.
ArXiv e-prints, June 2018.

B. Bischl, M. Lang, L. Kotthoff, J. Schiffner, J. Richter, E. Studerus, G. Casalicchio, and Z. M. Jones.
mlr: Machine learning in r.
Journal of Machine Learning Research, 17(170):1–5, 2016.

G. Casalicchio, C. Molnar, and B. Bischl.
Visualizing the Feature Importance for Black Box Models.
ArXiv e-prints, Apr. 2018.

P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis.
Modeling wine preferences by data mining from physicochemical properties.
Decis. Support Syst., 47(4):547–553, Nov. 2009.

N. Kennedy.
forestmodel: Forest Plots from Regression Models, 2018.
R package version 0.5.0.

S. Lundberg and S.-I. Lee.
A unified approach to interpreting model predictions.
ArXiv e-prints, May 2017.

C. Molnar.
Interpretable Machine Learning.
https://christophm.github.io/interpretable-ml-book/, 2018.
`https://christophm.github.io/interpretable-ml-book/`.

M. Staniak and P. Biecek.
Explanations of model predictions with live and breakDown packages.
ArXiv e-prints, Apr. 2018.

M. Tulio Ribeiro, S. Singh, and C. Guestrin.
Model-Agnostic Interpretability of Machine Learning.
ArXiv e-prints, June 2016.