Local Interpretability of Machine Learning Models

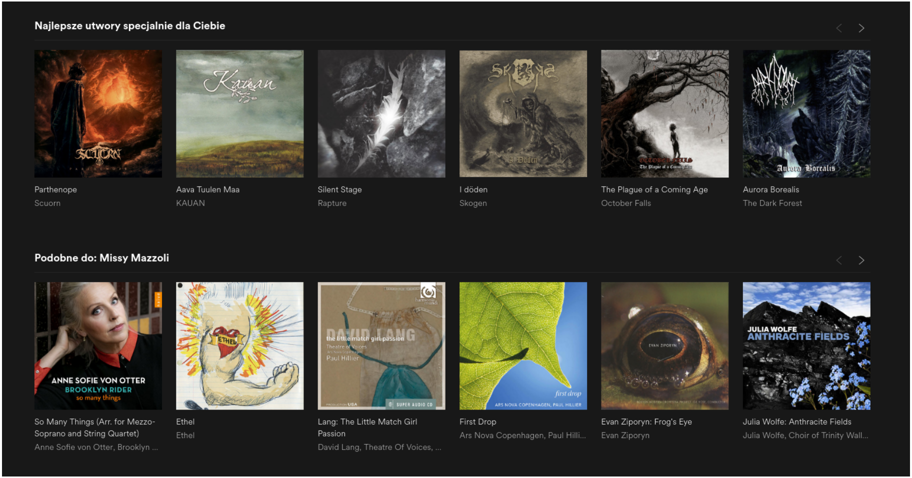
Mateusz Staniak

University of Wrocław

Berlin, 14.12.2017

Joint work with Przemysław Biecek (Technical University of Warsaw)

Example


Najlepsze utwory specjalnie dla Ciebie
Parthenope
Scuorn
Aava Tuulen Maa
KAUAN
Silent Stage
Rapture
I döden
Skogen
The Plague of a Coming Age
October Falls
Aurora Borealis
The Dark Forest
Podobne do: Missy Mazzoli
ANNE SOFIE VON OTTER
BROOKLYN RIDER
so many things
So Many Things (Arr. for Mezzo-Soprano and String Quartet)
Anne Sofie von Otter, Brooklyn ...
Ethel
Ethel
the little match girl passion
Paul Hillier
Lang: The Little Match Girl Passion
David Lang, Theatre Of Voices, ...
first drop
ARS NOVA COPENHAGEN, PAUL HILLIER
First Drop
Ars Nova Copenhagen, Paul Hilli...
EVAN ZIPORYN
Evan Ziporyn: Frog's Eye
Evan Ziporyn
JULIA WOLFE
ANTHRACITE FIELDS
Julia Wolfe: Anthracite Fields
Julia Wolfe, Choir of Trinity Wall...

Crucial questions

- Do we **understand** the model?
- Do we **trust** the model?

Possible approaches

Modeling

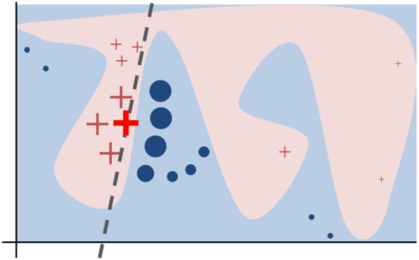
- Interpretable models only
- GAMs
- Surrogate models
- Model-specific explanations (`randomForestExplainer`, `xgboostExplainer`, ...)
- Model-agnostic explanations

Visualization

- Partial dependence plots
- Residual analysis
- Forest floor plots
- Other methods...

Main concepts I

- **interpretable representation**
- **local fidelity**
- **local exploration**



Formulation

- $x \in \mathbb{R}^d$ - 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
- $\mathcal{L}(f, g, \pi_x(z))$ - 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: 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.

live: Motivation & Explanation

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.
- Optional variable selection.
- Provide tools for model visualization.
- Focus on interpretable models easy to visualize.

Some good news

- The method finds the right local model.
- The method is pretty stable (similar results for different *fake* datasets).
- White box predictions are close to black box predictions at and around chosen instance.
- General framework: Shapley values.

Case study

- `live` package: `https://www.github.com/MI2DataLab/live`
  - `sample_locally` → `add_predictions`
  - `fit_explanation` → `plot_explanation`
- 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>
```

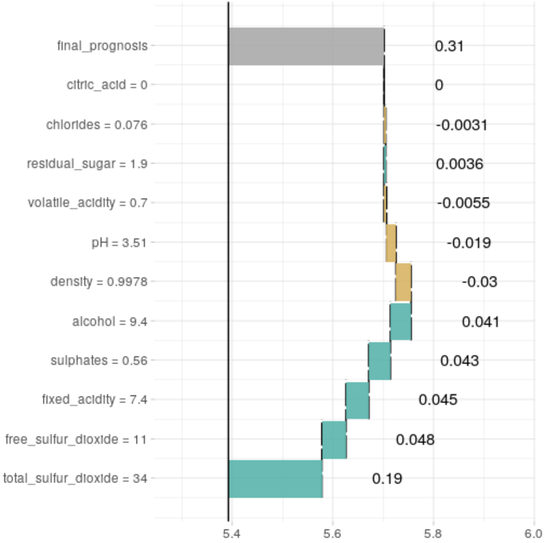
Case study

```
similar <- sample_locally(data = winequality_red,
                          explained_instance = winequality_red[5, ],
                          explained_var = "quality",
                          size = 100,
                          standardise = TRUE)
similar1 <- add_predictions(winequality_red, similar,
                            "regr.randomForest")
similar2 <- add_predictions(winequality_red, similar, "regr.svm")
trained <- fit_explanation(live_object = similar1,
                           white_box = "regr.lm",
                           selection = TRUE)
plot_explanation(trained,
                 regr_plot_type = "forestplot",
                 explained_instance = winequality_red[5, ])
```

Case study

Variable	Observed	Estimate	Lower	Upper
sulphates	0.56	0.08	0.06	0.09
citric_acid	0	0.06	0.05	0.07
chlorides	0.08	-0.04	-0.05	-0.03
density	1	-0.03	-0.04	-0.02

Case study


final_prognosis
citric_acid = 0
chlorides = 0.076
residual_sugar = 1.9
volatile_acidity = 0.7
pH = 3.51
density = 0.9978
alcohol = 9.4
sulphates = 0.56
fixed_acidity = 7.4
free_sulfur_dioxide = 11
total_sulfur_dioxide = 34
0.31
0
-0.0031
0.0036
-0.0055
-0.019
-0.03
0.041
0.043
0.045
0.048
0.19
5.4
5.6
5.8
6.0

Challenges

- LIME in high dimensional setting,
- optimal way of generating *fake* dataset,
- measures of fit,
- visualizing shrinkage methods...

Acknowledgements

Uniwersytet Wrocławski

MI

References I

- Bernd Bischl, Michel Lang, Lars Kotthoff, Julia Schiffner, Jakob Richter, Erich Studerus, Giuseppe Casalicchio, and Zachary M. Jones, *mlr: Machine learning in r*, Journal of Machine Learning Research **17** (2016), no. 170, 1–5.
- Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis, *Modeling wine preferences by data mining from physicochemical properties*, Decis. Support Syst. **47** (2009), no. 4, 547–553.
- Max Gordon and Thomas Lumley, *forestplot: Advanced forest plot using 'grid' graphics*, 2017, R package version 1.7.
- Brandon M. Greenwell, *pdp: An R Package for Constructing Partial Dependence Plots*, The R Journal **9** (2017), no. 1, 421–436.

References II

- S. Lundberg and S.-I. Lee, *A unified approach to interpreting model predictions*, ArXiv e-prints (2017).
- Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis, *Conditional variable importance for random forests*, BMC Bioinformatics **9** (2008), no. 307.
- M. Tulio Ribeiro, S. Singh, and C. Guestrin, *Model-Agnostic Interpretability of Machine Learning*, ArXiv e-prints (2016).
- Hadley Wickham, Dianne Cook, and Heike Hofmann, *Visualizing statistical models: Removing the blindfold*, Statistical Analysis and Data Mining: The ASA Data Science Journal **8** (2015), no. 4, 203–225.