Understanding black-box model predictions with LIME

András Bérczi
2019.05.29.

About me

Statistics + data analysis + R = <3

Data Scientist @ Emarsys

You can find me on Linkedin and on Github: andrasberczi

LIME

"Why should I trust you?" Explaining the Predictions of Any Classifier

A technique created by Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin (University of Washington) in 2016

LIME R package

R port of the Python package

Developed by Thomas Lin Pedersen

Why am I interested in LIME?

More complex models

Email image and text analysis

Interesting topic :)

Why I want to talk about LIME?

Spread the word!

Try it out and share your experience!

Explain prediction of a ‘black-box’ model

Why is it good to have an explanation?

- Check if the model really behaves as you expect it
  - Gives prediction for the ‘right’ reason
- More trust in complex model if it is explainable
- GDPR

Locally

Interpretable

Model-agnostic

Explanations

What should be expected of an explanation?

Local fidelity

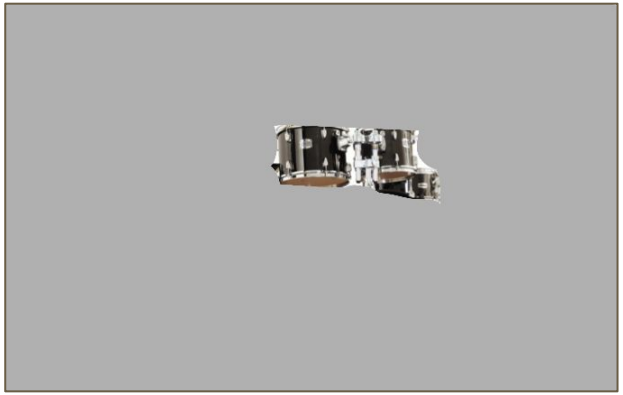
Interpretability

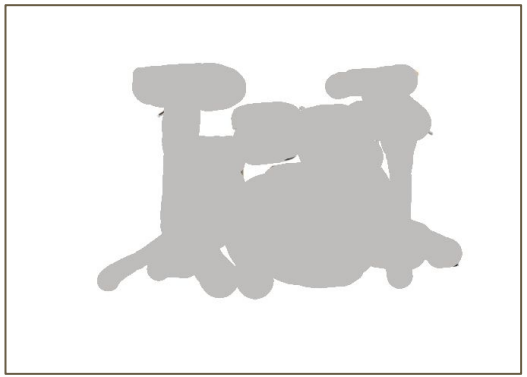
Model-agnostic

Select prediction(s) to be explained

Perturb features

Perturbe features





Predict with black-box model

P(drum) = 0.95

P(drum) = 0.01

Similarity scores

VS

Fit linear regression on perturbed instances

select m best features

weight results by similarity scores

Explain prediction locally

Label: drum, membranophone, tympan
Probability: 0.96
Explanation Fit: 0.50

Explain prediction locally

Why is it important?

Images classified as Husky





Images classified as Wolf





Why is it important?

Why is it important?

How to create an explanation in R

...A SHORT DEMO...

Runtime

Good for tabular + text ~ fraction of second per prediction

Slow for images ~ 10 minutes per image

Ending thoughts

It’s not all perfect (parameters, instability of explanations),

BUT

- good direction: models should be interpretable
- gives insight to black-box model

Try it out and tell us about it! :)

Thank you!
Any questions?

Links

https://arxiv.org/pdf/1602.04938.pdf

https://github.com/thomasp85/lime

https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime

https://christophm.github.io/interpretable-ml-book/lime.html

https://www.data-imaginist.com/2017/announcing-lime/