Classifier Explanation

Introduction to the Algorithms LIME and ANCHOR

Leonhard Applis , Lisa Branz

TH Nürnberg

21.1.2019

Table of Contents

Me: Hey Siri, order me a Pizza

Siri: *(After a short break that nearly drains your whole battery)* Ok, I'm calling your mother...

Me: Wait! Why would you do this!?

Siri: This is the 5th time you ordered Pizza this week.

What do we want from our model?

1. Why did failed predictions fail?
2. Why did correct predictions succeed?
3. Why is my model uncertain about a prediction?

special importance:
setting a model *live*, where it's not *prelabeled*

Interpretations must be ...

- *human-readable*
- reproducable (same input + same model → same output)
- **model angostic**, meaning they can work with any (black-box) model

Difficulties:

- Models can be huge (millions of weights)
- Inputvectors can be huge (e.g. images)
- Some models are to complex by it's structure to be readable, (e.g. neural networks)

Example

Maybe just pick the Atheist Christian Example from the paper

Trusting a Model

Whats the next step after trusting a single prediction
Mention time-required for manually checking a lot of pictures

Example

Prooving a Model

Classifier Explanation is required to set anything up IRL, or everyone will hate you very bad

Improving a Model

How Classifier Explanation helps you to improve your models
Better Filtering Maybe better weighting of features Finding Links in Classification (Similiar Classes?)

Table of Contents

Summary: What do we want

Human Readable Model Explanation For Every Classifier For Every Input

Definitions

All the stuff from the Lime Paper with the fancy ∇ and λ and greeky cheeky mumbo jumbo

Minimizing Fidelity · Interpretability

The part talking about making a minimal complex model which is quite accurate

Here is the Pseudocode.

Visualisation

Put the funky red-blue image with the red-crosses from the paper here

maybe: Example

Table of Contents

Problem with Sampling

Explain that we have to little time to inspect everything Looking for a new way to pick samples

Here is the Pseudocode

SPLIME Example

I guess this needs more than 2 Pages, we should add an example

Table of Contents

Setup Problem, Show Code, Plot Examples, nice This could be left out from the presentation, and just be a live demo
Do both: LIME and ANCHOR and sample with SPLIME