Classifier Explanation

> Applis , Lisa Branz

Basics

SP-

LIME

Ex I: Spamfilter

Classifier Explanation Introduction to the Algorithms LIME and ANCHOR

Leonhard Applis , Lisa Branz

TH Nürnberg

21.1.2019

Table of Contents

Classifier Explanation

Applis , Lisa Branz

Basics

Basics

111/11/22

2 LIME

3 SP-LIME

4 Ex I: Spamfilter

Leonhar Applis Lisa Branz

Basics

SP-

LIME

Ex I: Spamf 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?

- Why did failed predictions fail?
- Why did correct predictions succeed?
- 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 \rightarrow same output)
- model angostic, meaning they can work with any (black-box) model

Difficulties:

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

Example

Classifier Explanation

Applis , Lisa Branz

Basic

LIM

SP-LIME

Ex I: Spamfil Maybe just pick the Atheist Christian Example from the paper

Trusting a Model

Classifier Explanation

Applis , Lisa Branz

Basics

LIME

SP-LIME

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

Example

Classifier Explanation

Leonhard Applis , Lisa Branz

Basic

LIME

SP-LIME

Ex I: Spamfilter

Prooving a Model

Expla-

Basics

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

Improving a Model

Classifier Explanation

> Applis , Lisa Branz

Basics

SP-

Ex I: Spamf 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

Classifier Explanation

Applis , Lisa Branz

Basics

LIME

2 LIME

IME x I:

3 SP-LIME

4 Ex I: Spamfilter

Summary: What do we want

Classifier Explanation

Applis , Lisa Branz

Basic

LIME

SP-LIME

Ex I: Spamfil Human Readable Model Explanation For Every Classifier For Every Input

Definitions

Classifier Explanation

Applis , Lisa Branz

Basic

LIME

SP-LIME

Ex I: Spamfi All the stuff from the Lime Paper with the fancy ∇ and λ and greeky cheeky mumbo jumbo

Minimizing Fidelity \cdot Interpretability

Classifier Explanation

Leonhar Applis , Lisa Branz

 $_{
m Basic}$

LIME

SP-LIME

Ex I: acc

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

Local Interpretable Model-Agnostic Explanations $_{\rm The\ LIME-Algorithm}$

Classifie: Explanation

> Applis , Lisa Branz

Sasic

LIME

SP-LIME

Ex I: Spamfil ter Here is the Pseudocode.

Visualisation

Classifier Explanation

Applis , Lisa Branz

Basic

LIME

SP-LIME

Ex I: Spamfil Put the funky red-blue image with the red-crosses from the paper here

Classifier Explanation

Applis,
Lisa
Branz

Basic

LIME

SP-LIME

Ex I: Spamfilter maybe: Example

Table of Contents

Classifier Explanation

Applis , Lisa Branz

Basic

LIM

SP-LIMI

Ex I: SpamfilBasics

2 LIME

3 SP-LIME

4 Ex I: Spamfilter

Problem with Sampling

Classifier Explanation

Applis , Lisa Branz

Basic

an.

SP-LIME

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

Submodular Pick The SPLIME Algorithm

Classifier Explanation

Applis , Lisa Branz

Basic

LIMI

SP-LIME

Ex I: Spamfil Here is the Pseudocode

SPLIME Example

Classifier Explanation

Applis , Lisa Branz

Basic

LIMI

SP-LIME

Ex I: Spamfil I guess this needs more than 2 Pages, we should add an example $\,$

Table of Contents

Classifier Explanation

Applis , Lisa Branz

Basic

LIM

SP-LIME

Ex I: SpamfilBasics

2 LIME

3 SP-LIME

4 Ex I: Spamfilter

Spamfilter Explaining RandomForests for Textclassification

Classifier Explanation

Applis Lisa Branz

Basic LIME

SP-LIME

Ex I: SpamfilSetup 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