Classifier Explanation

Leonhard Applis , Lisa Branz

Basic

LIME

SP-LIME

ANCHO

Ex I: Spamfil-

Ex II: Image Recogni-

Summar

Classifier Explanation Introduction to the Algorithms LIME and ANCHOR

Leonhard Applis , Lisa Branz

TH Nürnberg

21.1.2019

Classifier Explanation

Applis , Lisa Branz

Basics

LIM

SP-LIME

ANCHO

Ex I: Spamfil-

Ex II: Image Recognition

Summary

- Basics
- 2 LIME
- 3 SP-LIME
- 4 ANCHOR
- **5** Ex I: Spamfilter
- 6 Ex II: Image Recognition
- Summary

Trusting a Prediction

Explanation

Applis , Lisa Branz

Basics

LIMI

SP-LIME

NCHO

ANCHO.

Ex 1: Spamfilter

Ex II: Image Recogni-

Summar

What do Humans need to unterstand a Prediction

Example

Classifier Explanation

Applis , Lisa Branz

 $_{\mathrm{Basics}}$

LIMI

SP-LIME

ANGHO

T3 T

Spamfil ter

Ex II: Image Recogni-

Summar

Maybe just pick the Atheist Christian Example from the paper

Trusting a Model

Classifier Explanation

> Applis , Lisa Branz

Basics

LIME

SP-LIME

ANGHO

ANCHO

Spamfilter

Ex II: Image Recogni-

Summar

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

Basic

LIME

SP-

ANCHO

Ex I: Spamfil-

Ex II: Image Recogni-

Summar

Prooving a Model

Classifier Explanation

Applis , Lisa Branz

 $_{\mathrm{Basics}}$

LIIVIL

LIME

ANCHO

Ex I: Spamfil-

Ex II: Image

C.....

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

LIME

SP-LIME

ANCHO

Ex I: Spamfil

Ex II: Image Recogni

Summar

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

Classifier Explanation

Leonhar Applis Lisa Branz

Basic

LIME

SP-Lime

ANCHO

Ev. I

Ex I: Spamfilter

Ex II: Image Recognition

ummary

- 1 Basics
- 2 LIME
- 3 SP-LIME
- 4 ANCHOR
- 5 Ex I: Spamfilter
- 6 Ex II: Image Recognition
- Summary

Summary: What do we want

Classifier Explanation

> Applis , Lisa Branz

Basic

LIME

SP-LIME

NCHO

Ex 1: Spamfilter

Ex II: Image Recogni-

ummai

Human Readable Model Explanation For Every Classifier For Every Input

Definitions

Classifier Explanation

Applis , Lisa Branz

Basic

LIME

LIME

ANCHO

Ex I: Spamfil

Ex II: Image Recogni

Summar

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

Minimizing Fidelity \cdot Interpretability

Classifier Explanation

Leonhard Applis , Lisa Branz

Basic

LIME

SP-LIME

NCHOI

Ex I: Spamfil-

ter Ex II:

Image Recognition

Summar

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

Basic

LIME

SP-LIME

VCHO

Ex I:

Spamfil ter

Ex II: Image Recogni-

Summai

Here is the Pseudocode.

Visualisation

Classifier Explanation

Applis , Lisa Branz

Basic

LIME

SP-LIME

NCHO

Ex I: Spamfil-

Ex II: Image

Summai

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

Classifier Explanation

Leonhard Applis, Lisa Branz

Basic

LIME

SP-LIME

ANCHOI

NCHOP

Ex I: Spamfilter

Ex II: Image Recogni-

2........

maybe: Example

Classifier Explanation

Leonhar Applis , Lisa Branz

20010

LIMI

SP-LIME

ANCHO

Ex I: Spamfil

Ex II: Image Recogni

ummary

- Basics
- 2 LIME
- 3 SP-LIME
- 4 ANCHOR
- **5** Ex I: Spamfilter
- 6 Ex II: Image Recognition
- Summary

Problem with Sampling

Classifier Explanation

Leonhar Applis , Lisa Branz

Basic

LIMI

LIME

ANCHOR

Ex I: Spamfilter

Ex II: Image Recogni-

Summar

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

Submodular Pick The SPLIME Algorithm

Classifie: Explanation

Applis , Lisa Branz

Basic

LIMI

LIME

TOTTO

Ex. I:

Ex 1: Spamfil ter

Ex II: Image Recogni-

lummar

Here is the Pseudocode

SPLIME Example

Classifier Explanation

Applis , Lisa Branz

Basic

LIMI

LIME

NCHO

Ex I: Spamfil

Ex II: Image Recogni-

Summar

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

Classifier Explanation

Leonhar Applis , Lisa Branz

Sacio

LIME

SP-

ANCHOR

Ex I: Spamfil-

Spamfilter Ex II:

Ex 11: Image Recognition

ummary

- Basics
- 2 LIME
- 3 SP-LIME
- 4 ANCHOR
- 5 Ex I: Spamfilter
- 6 Ex II: Image Recognition
- Summary

Summai

I have not read this yet. I guess we can leave out definitions and problem, and run directly to pseudocode and comparison Draft:

- Algo
- Algo Visualisation (the Picture with red blue crosses etc.)
- Example
- Comparison Anchor vs. Lime

Classifier Explanation

Leonhar Applis , Lisa Branz

Pagia

LIMI

SP-LIME

ANCHO

Ex I: Spamfil-

ter Ex II: Image

ummary

- Basics
- 2 LIME
- 3 SP-LIME
- 4 ANCHOR
- **5** Ex I: Spamfilter
- 6 Ex II: Image Recognition
- Summary

Spamfilter Explaining RandomForests for Textclassification

Classifier Explanation

> Applis Lisa Branz

Basic

LIMI

SP-LIME

ANCHO

Ex I: Spamfil-

Ex II: Image

Summar

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

Classifier Explanation

Leonhar Applis , Lisa Branz

2 noi o

LIMI

SP-LIME

ANCHO

Ex. I:

Ex 1: Spamfilter

Ex II: Image Recognition

ummar

- Basics
- 2 LIME
- 3 SP-LIME
- 4 ANCHOR
- **6** Ex I: Spamfilter
- **6** Ex II: Image Recognition
- Summary

Traffic Sign Detection Explaining NeuralNetworks for Imageclassification

Classifier Explanation

> Applis , Lisa Branz

Basic

LIME

SP-LIME

ANCHO:

Ex I: Spamfil-

Ex II: Image Recogni-

ummar

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

Classifier Explanation

Leonhar Applis Lisa Branz

20010

LIMI

SP-Lime

NCHC

Ex I: Spamfil-

Ex II: Image Recognition

 $_{
m ummary}$

- 1 Basics
- 2 LIME
- 3 SP-LIME
- 4 ANCHOR
- **5** Ex I: Spamfilter
- 6 Ex II: Image Recognition
- Summary

summary

Classifier Explanation

Applis , Lisa

Basic

LIME

SP-

ANCHO

Ex I: Spamfil-

Ex II: Image Recogni-

Summar

sources

Classifier Explanation

> Applis , Lisa

Basic

LIMI

SP-LIME

ANCHO

Ex I: Spamfil-

Ex II: Image Recogni-

Summar