Classifier Explanation

Applis , Lisa Branz

Basics

LIME

SP-LIME

ANCHO

Ex 1: Spamfilter

Ex II: Image Recogni-

Summary

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: Spamfilter

Ex II: Image Recognition

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

Trusting a Prediction

Classifier Explanation

> Applis Lisa Branz

Basics

SP-

ANCHO

Ex I: Spamfil-

Ex II: Image Recogn tion

Summar

What do Humans need to unterstand a Prediction?

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.

Example

Classifier Explanation

Applis , Lisa Branz

Basics

LIMI

SP-LIME

ANCHO

AIVCIIO.

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

LIMIL

LIME

ANCHO

Ex I: Spamfil

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

Applis, Lisa

Basic

LIME

SP-

ANCHO

Ex I: Spamfil-

Ex II: Image Recogni-

C.mmn.

Prooving a Model

Classifier Explanation

Applis , Lisa Branz

Basics

LIMI

SP-LIME

ANCHOE

Ex I: Spamfil-

ter

Image Recognition

Summar

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

Improving a Model

Classifier Explanation

> Applis , Lisa Branz

 $_{\mathrm{Basics}}$

LIMI

SP-LIME

ANGHO

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

Raeic

LIME

SP-

ANCHO

Ex I:

Spamfilter

Ex II: Image Recognition

- 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

2 naia

LIME

SP-LIME

NCHO

Ex 1: Spamfilter

Ex II: Image Recogni-

Summa

Human Readable Model Explanation For Every Classifier For Every Input

Definitions

Classifier Explanation

Applis , Lisa Branz

Basic

LIME

SP-LIME

ANCHOI

Ex I: Spamfil

Ex II: Image

Summar

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

Minimizing Fidelity \cdot Interpretability

Classifier Explanation

Applis , Lisa Branz

Basic

LIME

SP-LIME

ANCHOE

Ex I:

Spamfilter

Ex II: Image Recognition

Summai

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

D-- T-

Spamfilter

Ex II: Image Recogni-

Here is the Pseudocode. $\,$

Visualisation

Classifier Explanation

Applis , Lisa Branz

Basic

LIME

SP-LIME

NCHO

Spamfilter

Ex II: Image Recogni-

Summar

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

Classifier Explanation

Applis, Lisa

Basic

LIME

SP-LIME

NCHOI

Spamfilter

Ex II: Image Recogni-

Summary

maybe: Example

Classifier Explanation

Leonhar Applis , Lisa Branz

2 naia

LIMI

SP-LIMI

ANCHO

Ex I:

Spamfilter

Ex II: Image Recognition

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

Problem with Sampling

Classifier Explanation

Applis , Lisa Branz

Basic

LIM

LIME

ANCHOR

Ex I: Spamfil ter

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

VICTIO:

Ex 1: Spamfil ter

Ex II: Image Recogni-

lummar

Here is the Pseudocode

SPLIME Example

Classifier Explanation

Applis , Lisa Branz

Basic

LIMI

LIME

NCHOR

Ex I: Spamfil-

Ex II: Image Recogni-

Summar

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

Classifier Explanation

Applis , Lisa Branz

Sacios

LIME

SP-

ANCHOR

Ex I:

Ex 1: Spamfilter

Ex II: Image Recognition

- 1 Basics
- 2 LIME
- 3 SP-LIME
- 4 ANCHOR
- **6** 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

D - -:-

LIMI

SP-LIME

ANCHO

Ex I: Spamfil-

Ex II: Image Recogni-

- 1 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

...

LIME

ANCHO

Ex I: Spamfil-

Ex II: Image Recogni-

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 I: Spamfilter

Ex II: Image Recognition

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

Traffic Sign Detection Explaining NeuralNetworks for Imageclassification

Classifier Explanation

> Applis , Lisa Branz

Basic

....

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

Sacio

LIMI

SP-Lime

ANCHO

Ev I

Ex I: Spamfilter

Ex II: Image Recognition

- 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: Spamfilter

Ex II: Image Recogni-

Summar

sources

Classifier Explanation

> Applis , Lisa

Basic

LIME

SP-

ANCHO

Ex I: Spamfil-

Ex II: Image Recogni-

Summar