

Classifier Explanation

Introduction to the Algorithms LIME and ANCHOR

Leonhard Applis , Lisa Branz

TH Nürnberg

21.1.2019

Table of Contents

Classifier
Expla-
nation

Leonhard
Applis ,
Lisa
Branz

Basics

LIME

SP-
LIME

ANCHOR

Ex I:
Spamfil-
ter

Ex II:
Image
Recogni-
tion

Summary

1 Basics

2 LIME

3 SP-LIME

4 ANCHOR

5 Ex I: Spamfilter

6 Ex II: Image Recognition

7 Summary

Trusting a Prediction

Classifier
Expla-
nation

Leonhard
Applis ,
Lisa
Branz

Basics

LIME

SP-
LIME

ANCHOR

Ex I:
Spamfil-
ter

Ex II:
Image
Recogni-
tion

Summary

What do Humans need to understand a Prediction

Example

Classifier
Expla-
nation

Leonhard
Applis ,
Lisa
Branz

Basics

LIME

SP-
LIME

ANCHOR

Ex I:
Spamfil-
ter

Ex II:
Image
Recogni-
tion

Summary

Maybe just pick the Atheist Christian Example from the paper

Trusting a Model

Classifier
Expla-
nation

Leonhard
Applis ,
Lisa
Branz

Basics

LIME

SP-
LIME

ANCHOR

Ex I:
Spamfil-
ter

Ex II:
Image
Recogni-
tion

Summary

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

Example

Classifier
Explan-
ation

Leonhard
Applis ,
Lisa
Branz

Basics

LIME

SP-
LIME

ANCHOR

Ex I:
Spamfil-
ter

Ex II:
Image
Recogni-
tion

Summary

Prooving a Model

Classifier
Expla-
nation

Leonhard
Applis ,
Lisa
Branz

Basics

LIME

SP-
LIME

ANCHOR

Ex I:
Spamfil-
ter

Ex II:
Image
Recogni-
tion

Summary

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

Improving a Model

Classifier
Expla-
nation

Leonhard
Applis ,
Lisa
Branz

Basics

LIME

SP-
LIME

ANCHOR

Ex I:
Spamfil-
ter

Ex II:
Image
Recogni-
tion

Summary

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
Expla-
nation

Leonhard
Applis ,
Lisa
Branz

Basics

LIME

SP-
LIME

ANCHOR

Ex I:
Spamfil-
ter

Ex II:
Image
Recogni-
tion

Summary

1 Basics

2 LIME

3 SP-LIME

4 ANCHOR

5 Ex I: Spamfilter

6 Ex II: Image Recognition

7 Summary

Summary: What do we want

Classifier
Expla-
nation

Leonhard
Applis ,
Lisa
Branz

Basics

LIME

SP-
LIME

ANCHOR

Ex I:
Spamfil-
ter

Ex II:
Image
Recogni-
tion

Summary

Human Readable Model Explanation For Every Classifier For Every Input

Definitions

Classifier
Explan-
ation

Leonhard
Applis ,
Lisa
Branz

Basics

LIME

SP-
LIME

ANCHOR

Ex I:
Spamfil-
ter

Ex II:
Image
Recogni-
tion

Summary

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

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

Local Interpretable Model-Agnostic Explanations

The LIME-Algorithm

Classifier
Expla-
nation

Leonhard
Applis ,
Lisa
Branz

Basics

LIME

SP-
LIME

ANCHOR

Ex I:
Spamfil-
ter

Ex II:
Image
Recogni-
tion

Summary

Here is the Pseudocode.

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

maybe: Example

Table of Contents

Classifier
Expla-
nation

Leonhard
Applis ,
Lisa
Branz

Basics

LIME

SP-
LIME

ANCHOR

Ex I:
Spamfil-
ter

Ex II:
Image
Recogni-
tion

Summary

1 Basics

2 LIME

3 SP-LIME

4 ANCHOR

5 Ex I: Spamfilter

6 Ex II: Image Recognition

7 Summary

Problem with Sampling

Classifier
Expla-
nation

Leonhard
Applis ,
Lisa
Branz

Basics

LIME

SP-
LIME

ANCHOR

Ex I:
Spamfil-
ter

Ex II:
Image
Recogni-
tion

Summary

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

Submodular Pick

The SPLIME Algorithm

Classifier
Expla-
nation

Leonhard
Applis ,
Lisa
Branz

Basics

LIME

SP-
LIME

ANCHOR

Ex I:
Spamfil-
ter

Ex II:
Image
Recogni-
tion

Summary

Here is the Pseudocode

SPLIME Example

Classifier
Expla-
nation

Leonhard
Applis ,
Lisa
Branz

Basics

LIME

SP-
LIME

ANCHOR

Ex I:
Spamfil-
ter

Ex II:
Image
Recogni-
tion

Summary

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

Table of Contents

Classifier Expla- nation

Leonhard
Applis ,
Lisa
Branz

Basics

LIME

SP-
LIME

ANCHOR

Ex I:
Spamfil-
ter

Ex II:
Image
Recogni-
tion

Summary

1 Basics

2 LIME

3 SP-LIME

4 ANCHOR

5 Ex I: Spamfilter

6 Ex II: Image Recognition

7 Summary

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

Table of Contents

Classifier Expla- nation

Leonhard
Applis ,
Lisa
Branz

Basics

LIME

SP-
LIME

ANCHOR

Ex I:
Spamfil-
ter

Ex II:
Image
Recogni-
tion

Summary

1 Basics

2 LIME

3 SP-LIME

4 ANCHOR

5 Ex I: Spamfilter

6 Ex II: Image Recognition

7 Summary

Spamfilter

Explaining RandomForest for Textclassification

Classifier
Explan-
ation

Leonhard
Applis ,
Lisa
Branz

Basics

LIME

SP-
LIME

ANCHOR

Ex I:
Spamfil-
ter

Ex II:
Image
Recogni-
tion

Summary

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

Table of Contents

Classifier
Expla-
nation

Leonhard
Applis ,
Lisa
Branz

Basics

LIME

SP-
LIME

ANCHOR

Ex I:
Spamfil-
ter

Ex II:
Image
Recogni-
tion

Summary

1 Basics

2 LIME

3 SP-LIME

4 ANCHOR

5 Ex I: Spamfilter

6 Ex II: Image Recognition

7 Summary

Traffic Sign Detection

Explaining NeuralNetworks for Imageclassification

Classifier
Expla-
nation

Leonhard
Applis ,
Lisa
Branz

Basics

LIME

SP-
LIME

ANCHOR

Ex I:
Spamfil-
ter

Ex II:
Image
Recogni-
tion

Summary

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

Table of Contents

Classifier Expla- nation

Leonhard
Applis ,
Lisa
Branz

Basics

LIME

SP-
LIME

ANCHOR

Ex I:
Spamfil-
ter

Ex II:
Image
Recogni-
tion

Summary

1 Basics

2 LIME

3 SP-LIME

4 ANCHOR

5 Ex I: Spamfilter

6 Ex II: Image Recognition

7 Summary

summary

Classifier
Expla-
nation

Leonhard
Applis ,
Lisa
Branz

Basics

LIME

SP-
LIME

ANCHOR

Ex I:
Spamfil-
ter

Ex II:
Image
Recogni-
tion

Summary

Classifier
Explaination

Leonhard
Applis ,
Lisa
Branz

Basics

LIME

SP-
LIME

ANCHOR

Ex I:
Spamfilter

Ex II:
Image
Recognition

Summary