Classifier Explanation

Introduction to the Algorithms LIME and ANCHOR

Leonhard Applis , Lisa Branz

TH Nürnberg

21.1.2019

Table of Contents

Basics

LIME

SP-LIME

ANCHOR

Ex I: Spamfilter

Ex II: Image Recognition

Summary

1 Basics

2 LIME

3 SP-LIME

4 ANCHOR

5 Ex I: Spamfilter

6 Ex II: Image Recognition

7 Summary

Trusting a Prediction

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

Leonhard Applis , Lisa Branz

Basics

LIME

SP-LIME

ANCHOR

Ex I: Spamfilter

Ex II: Image Recognition

Summary

Maybe just pick the Atheist Christian Example from the paper

Trusting a Model

Classifier Explanation
Leonhard Applis , Lisa Branz
Basics
LIME
SP-LIME
ANCHOR
Ex I: Spamfilter
Ex II: Image Recognition
Summary

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

Basics

LIME

SP-LIME

ANCHOR

Ex I: Spamfilter

Ex II: Image Recognition

Summary

Prooving a Model

Classifier Explanation

Leonhard Applis , Lisa Branz

Basics

LIME

SP-LIME

ANCHOR

Ex I: Spamfilter

Ex II: Image Recognition

Summary

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

Improving a Model

Classifier Explanation
Leonhard Applis , Lisa Branz
Basics
LIME
SP-LIME
ANCHOR
Ex I: Spamfilter
Ex II: Image Recognition
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

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 Explanation

Leonhard Applis , Lisa Branz

Basics

LIME

SP-LIME

ANCHOR

Ex I: Spamfilter

Ex II: Image Recognition

Summary

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

Classifier Explanation
Leonhard Applis , Lisa Branz
Basics
LIME
SP-LIME
ANCHOR
Ex I: Spamfilter
Ex II: Image Recognition
Summary

Here is the Pseudocode.

Visualisation

Classifier Explanation

Leonhard Applis , Lisa Branz

Basics

LIME

SP-LIME

ANCHOR

Ex I: Spamfilter

Ex II: Image Recognition

Summary

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

maybe: Example

Table of Contents

Basics

LIME

SP-LIME

ANCHOR

Ex I: Spamfilter

Ex II: Image Recognition

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 Explanation

Leonhard Applis , Lisa Branz

Basics

LIME

SP-LIME

ANCHOR

Ex I: Spamfilter

Ex II: Image Recognition

Summary

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

Classifier Explanation
Leonhard Applis , Lisa Branz
Basics
LIME
SP-LIME
ANCHOR
Ex I: Spamfilter
Ex II: Image Recognition
Summary

Here is the Pseudocode

SPLIME Example

Classifier Explanation

Leonhard Applis , Lisa Branz

Basics

LIME

SP-LIME

ANCHOR

Ex I: Spamfilter

Ex II: Image Recognition

Summary

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

Table of Contents

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

1 Basics

2 LIME

3 SP-LIME

4 ANCHOR

5 Ex I: Spamfilter

6 Ex II: Image Recognition

7 Summary

Classifier Explanation
Leonhard Applis , Lisa Branz
Basics
LIME
SP-LIME
ANCHOR
Ex I: Spamfilter
Ex II: Image Recognition
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

1 Basics

2 LIME

3 SP-LIME

4 ANCHOR

5 Ex I: Spamfilter

6 Ex II: Image Recognition

7 Summary

Classifier Explanation
Leonhard Applis , Lisa Branz
Basics
LIME
SP-LIME
ANCHOR
Ex I: Spamfilter
Ex II: Image Recognition
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

Basics

LIME

SP-LIME

ANCHOR

Ex I: Spamfilter

Ex II: Image Recognition

Summary

1 Basics

2 LIME

3 SP-LIME

4 ANCHOR

5 Ex I: Spamfilter

6 Ex II: Image Recognition

7 Summary

summary

Classifier Explanation

Leonhard Applis , Lisa Branz

Basics

LIME

SP-LIME

ANCHOR

Ex I: Spamfilter

Ex II: Image Recognition

Summary

sources

Classifier Explanation

Leonhard Applis , Lisa Branz

Basics

LIME

SP-LIME

ANCHOR

Ex I: Spamfilter

Ex II: Image Recognition

Summary