>>> Understanding VADER
>>> Using LIME to interpret black-box models

Name: Arturo Soberón
Date: May 18, 2022

>>> Table of Contents

>>> Black-Box Models

High-complexity Machine Learning models are capable of making really accurate predictions. Sadly, this gain in accuracy is often accompanied by a loss in interpretability.

A *black-box* model is a software that receives an input, returns an output and everything that happens in between is a mystery.

Local Interpretable Model-agnostic Explanations (LIME) makes many modifications to a given observation (x_0) and fits an interpretable model on these new data points to explain the decision making process of a black-box model.

>>> VADER

VADER is a Natural Language Processing (NLP) model that receives a text and, among other things, replies with a score that measures its negativity.[1]

$$x_o \rightarrow f(\cdot) \rightarrow y_0$$

As with most NLP models, VADER serves requests as a black-box model.

[1] Hutto et al., 2014

>>> Example

For example, the text:

> *Dear Dice,*
> *This game sucks. The specialists are trash. This trailer was a lie. I want my money back.*

receives a negative score of 0.112.

What drove this black-box model to make this prediction?

* Does the word *Dear* reduce the negative score?
* Do the words *sucks* and *trash* increase the negative score?

>>> LIME Procedure

LIME uses simple *surrogate* models to explain *individual* predictions of a black-box Machine Learning model.

* LIME does not train a simple global model

Given a data point x_0 and a predicted value $f(x_0)$, LIME makes many slight variations to x_0 to test what happens to the predictions made by the complex model.[2]

$$\big(x_0, f(x_0)\big) \rightarrow \{\big(x_0^1, f(x_0^1)\big), ..., \big(x_0^n, f(x_0^n)\big)\}$$

LIME then fits an interpretable model to the data such that

$$\min_g \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

[2] Ribeiro et al., 2016

>>> Pseudo-code

1. Select data point of interest x_0
2. Feed x_0 to $f(\cdot)$ to obtain y_0
3. Make multiple modifications around x_0
4. Get the predictions from the perturbed data set
5. Fit a simple model to this data set
6. Explain $f(x_0)$ by interpreting the local model

For NLP models, LIME takes out words from the original text x_0 to create the perturbed data set.

>>> Battlefield 2042

I mined a few thousand comments from Battlefield 2042's launch trailer. I chose to study Battlefield 2042 because it was a highly-anticipated yet poorly-received game.

> *One of the biggest games of the year on one of the most popular digital stores in the world on one of the biggest gaming platforms in the world [...] isn't able to keep up with Farming Simulator 22.*[3]

Due to the game's popularity and poor reception by the general public, the comment section is large and overwhelmingly negative, making it a good case study for this project.

---

[3] Kotaku, 2021

>>> Example

Given x_0:

How tf they managed to screw this up?

We create the following perturbed data set:

How	tf	they	managed	to	screw	this	up?	weight	score^0	score^1
1	0	1	1	1	1	1	0	0.750	0.167	0.167
1	1	1	1	1	0	1	1	0.875	0.167	0.000
0	1	1	1	1	1	1	1	0.875	0.167	0.149
1	0	1	1	1	1	1	1	0.875	0.167	0.149
1	0	1	1	1	1	0	1	0.750	0.167	0.167

>>> Example (continued)

By fitting

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \sum_{p=1}^{8} \beta_p x_{ip})^2 + 0.1 \sum_{p=1}^{8} |\beta_p|$$

we get the following explanation at x_0:

Word	Coef
How	0.43
tf	0.43
they	0.00
managed	0.11
to	0.00
screw	1.62
this	0.00
up?	0.00

>>> Next Steps

1. Tune λ
2. Formalize code for LIME procedure
    * Remove n-grams?
3. Generalize results for all 5K comments?
4. Written document
5. Web app?