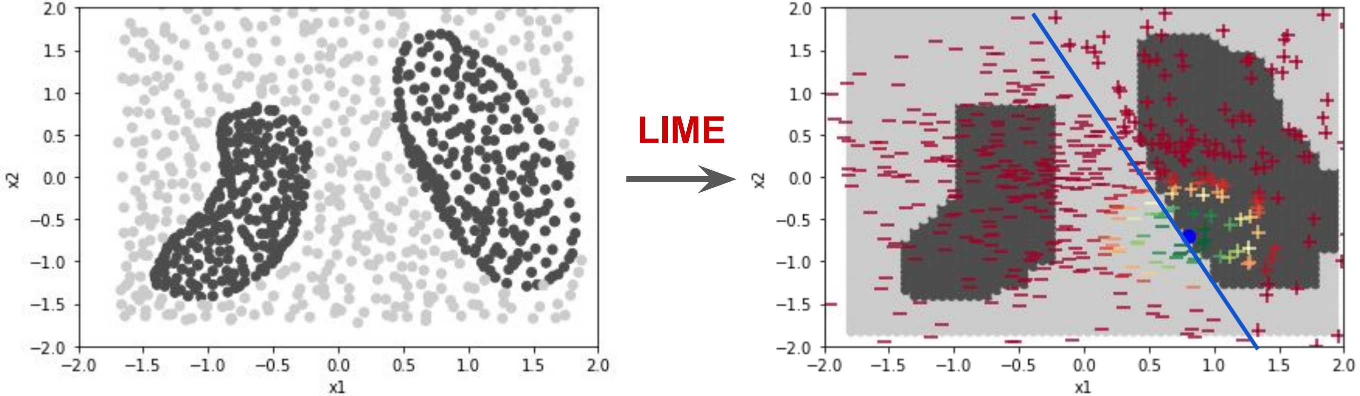
Interpretable Machine Learning with LIME



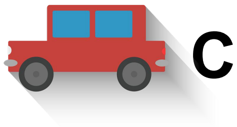
Cristian Arteaga, arteagac.github.io

The need for interpretability - examples

Health Care

- Blood test results
- Symptoms
- Health history of family

→ ML Model → Presence of pathology



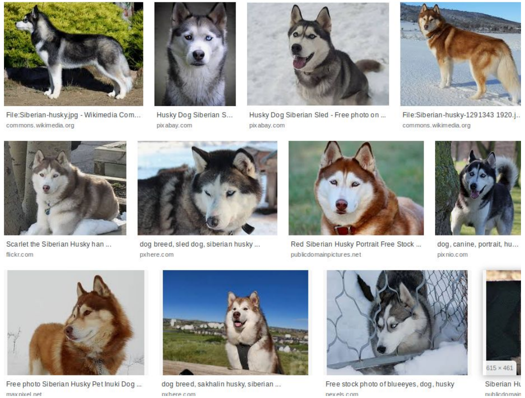
Car Insurance

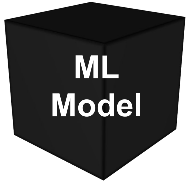
- Vehicle year and model
- Driving history
- Driver’s age

→ ML Model → Cost of insurance premium

The need for interpretability - examples

Computer Vision





Wolf or Husky?

http://innovation.uci.edu/2017/08/husky-or-wolf-using-a-black-box-learning-model-to-avoid-adoption-errors/

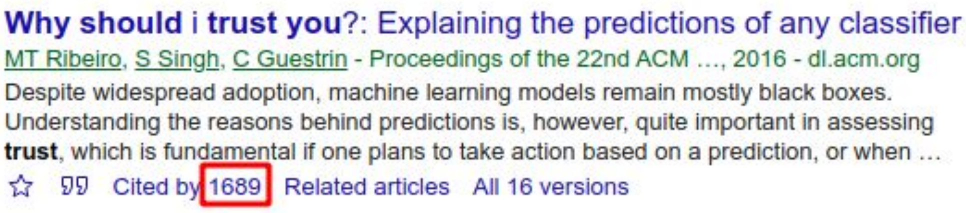
Some popular ML interpretability techniques

- **PDP**: Partial Dependence Plots
- **LIME**: Local Interpretable Model-Agnostic Explanations
- **SHAP**: SHapley Additive exPlanations
- **CAM**: Class Activation Mapping

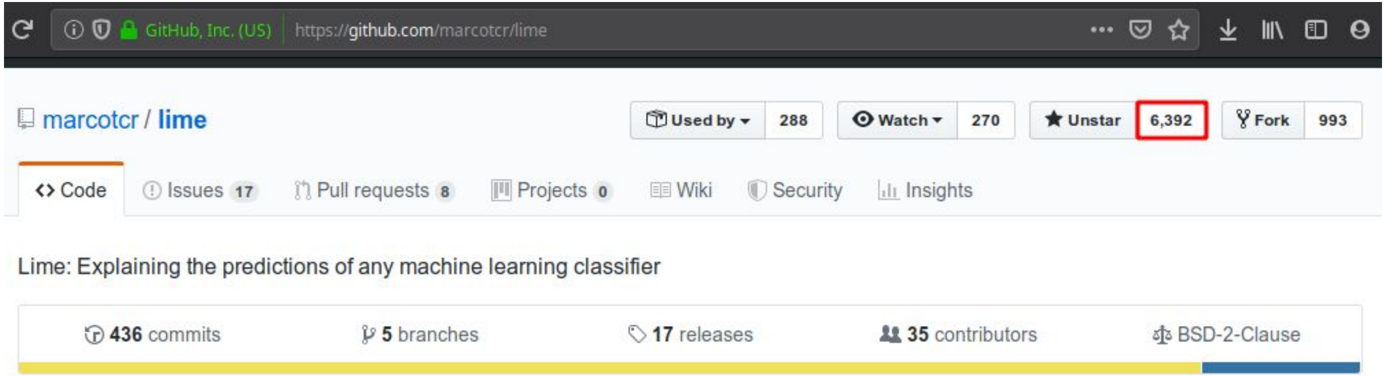
More details:
Molnar, Christoph. "Interpretable machine learning." *A Guide for Making Black Box Models Explainable* 7 (2018).

LIME

Paper: https://dl.acm.org/citation.cfm?Id=2939778



GitHub: https://github.com/marcotcr/lime



LIME: Local Interpretable Model-Agnostic Explanations

Local

Explanations are locally faithful instead of globally.

Interpretable

Humans are limited by an amount of information that can be processed and understood.

e.g. The weights of a neural network are not meaningful for a human.

Model-Agnostic

Any machine learning algorithm can be used as predictive model. Works with text, image and tabular data.

Explanations

Artifacts that provide an understanding between input to a ML model and the model's prediction.

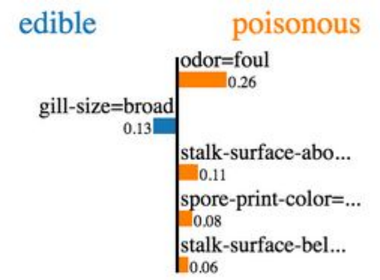
LIME for Tabular Data Classification

Mushroom Characteristics (Odor, size, etc.) → **ML Model** → Poisonous

Why?

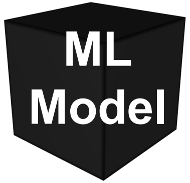
LIME Explanations

Because



LIME for Text Classification

From: conor@owlnet.rice.edu (Conor Frederick Prischmann)
Subject: Re: Genocide is Caused by Theism: Evidence?
Organization: Rice University
Lines: 23
In article [c60A0s.DvI@mailer.cc.fsu.edu] dekorte@dirac.scri.fsu.edu (Stephen L. Dekorte) writes:
I saw a 3 hour show on PBS the other day about the history of the

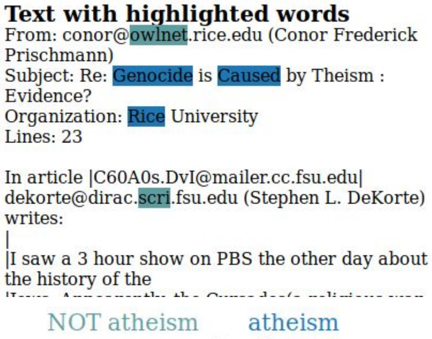


Atheism

Why?

LIME Explanations

Because



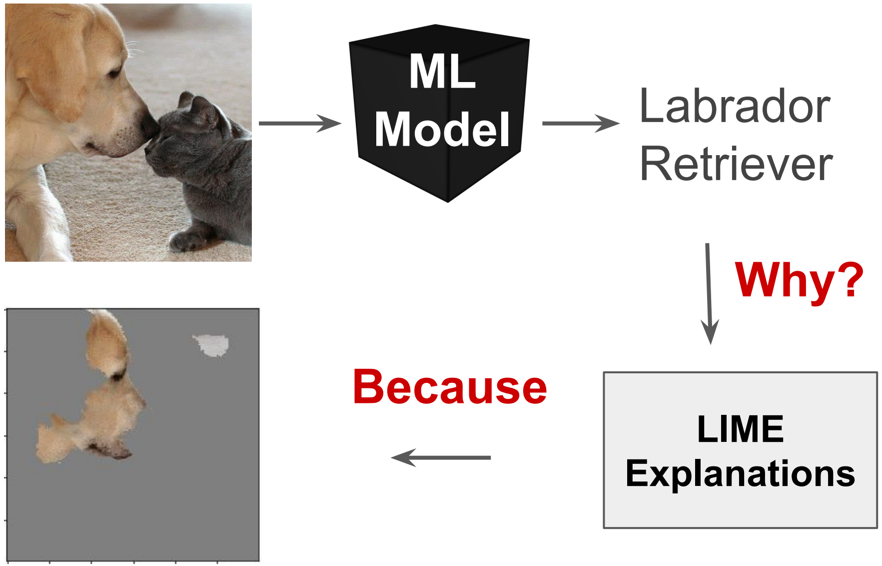
Text with highlighted words

From: conor@owlnet.rice.edu (Conor Frederick Prischmann)
Subject: Re: Genocide is Caused by Theism : Evidence?
Organization: Rice University
Lines: 23

In article |C60A0s.DvI@mailer.cc.fsu.edu| dekorte@dirac.scri.fsu.edu (Stephen L. DeKorte) writes:
|
|I saw a 3 hour show on PBS the other day about the history of the

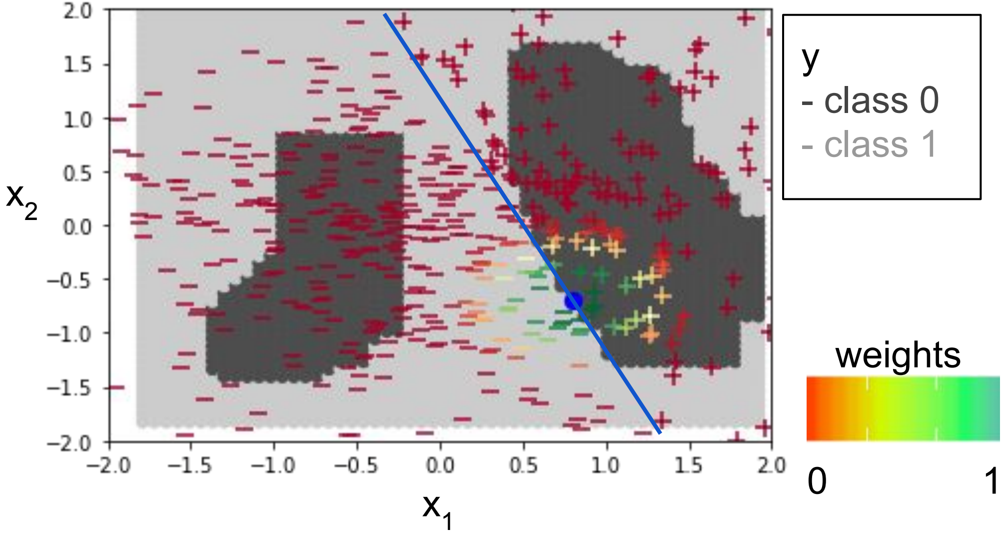
NOT atheism atheism

LIME for Image Classification



How LIME works?

Fit a **local** interpretable (simpler) model around the instance to be explained.

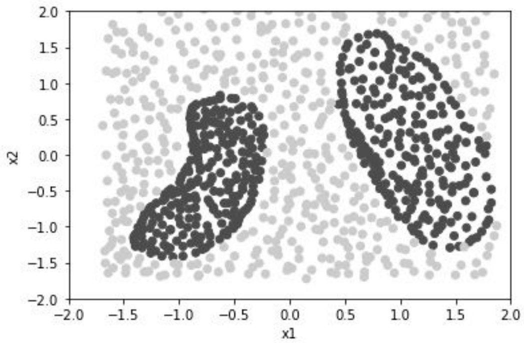


How LIME works?

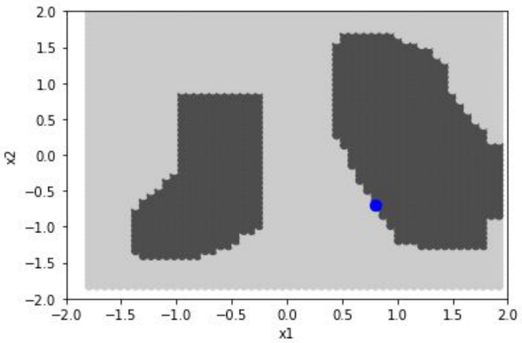
Data:

X = $\{x_1, x_2\}$ **y** = labels (gray or black)

Data Generation Process



ML (Random Forest) Prediction

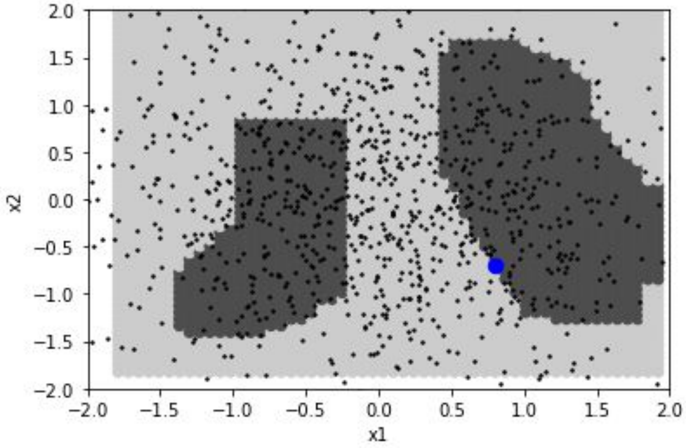


Task: Explain classifier's prediction on **one instance** (blue dot).

Step 1

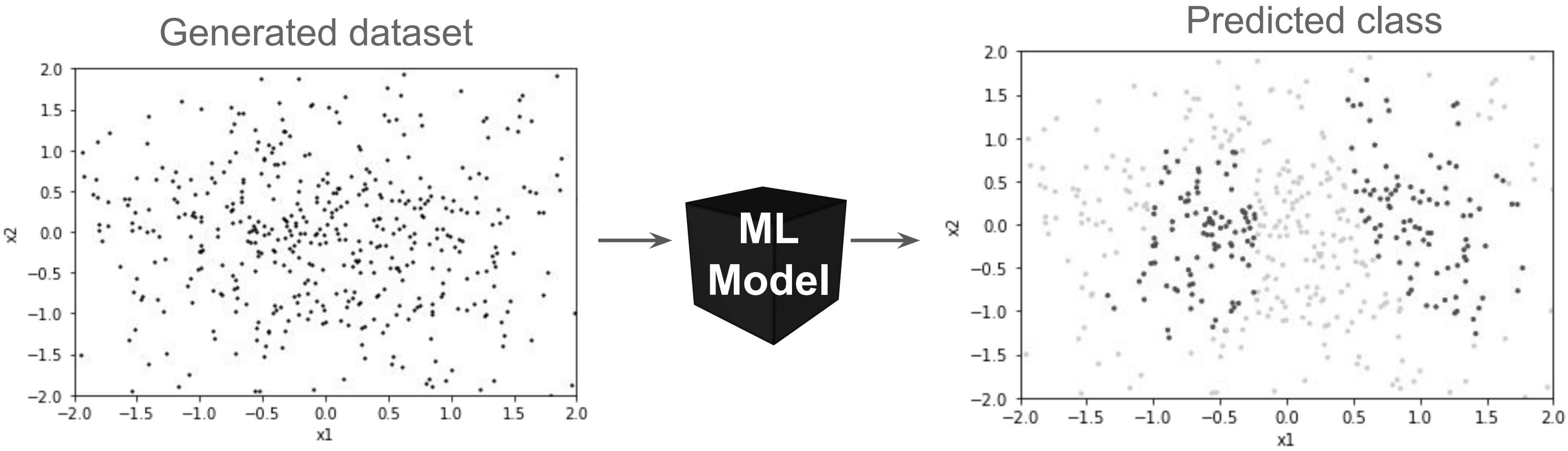
Generate a new dataset by sampling around the instance to be explained.

For tabular data, sampling around the the mean and std. dev. is recommended.



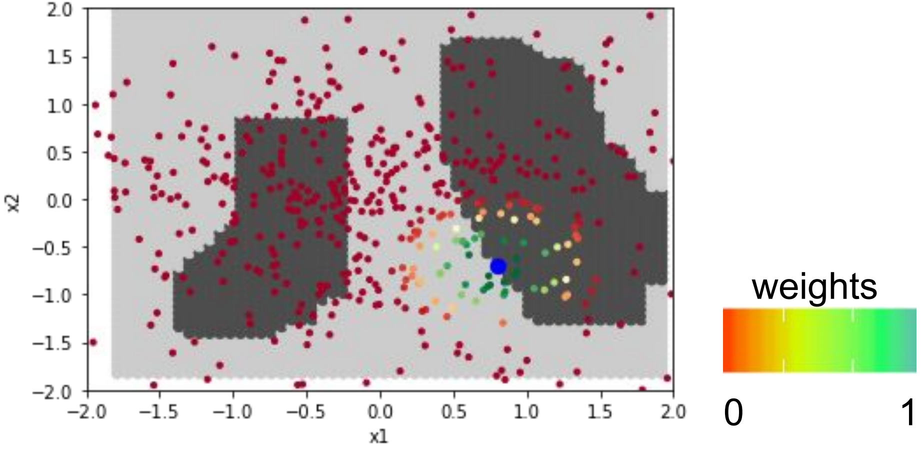
Step 2

Use the machine learning model to predict the classes of the new generated dataset.



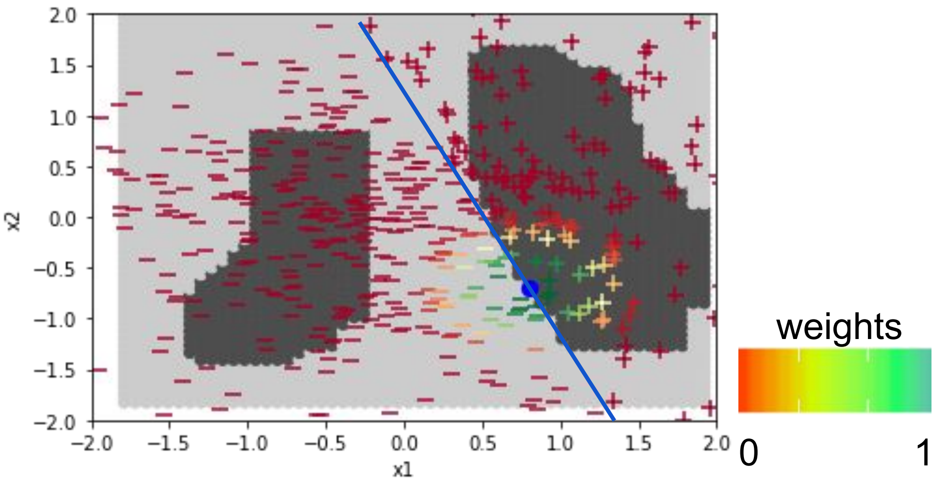
Step 3

Use a kernel function to **weight** the importance of each instance of the new dataset on the locality of the instance to be explained.



Step 4

Fit a weighted linear model (blue line) using the **new dataset**, the **predicted classes** for the new dataset and the **weights**. This linear model can be used to explain the prediction



Python code:

https://arteagac.github.io/blog