

LIME

(Local Interpretable Model- Agnostic Explanations)

Christine Chen

What is LIME?

Local **I**nterpretable **M**odel-Agnostic **E**xplanations

- * Framework to determine model interpretability which can be applied to any model
- * another metric for your model beside accuracy
- * post-processing step after building your model

When might it be useful?

- * ethics of a business use cases (e.g. health)
- * assessing business risk when model is in production

Example of an application

text classification

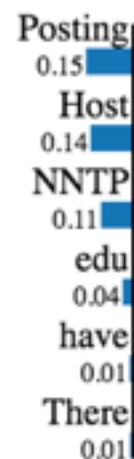
Training a random forest with 500 trees, we get a test set accuracy of 92.4%, which is surprisingly high.

Prediction probabilities



atheism

christian



Text with highlighted words

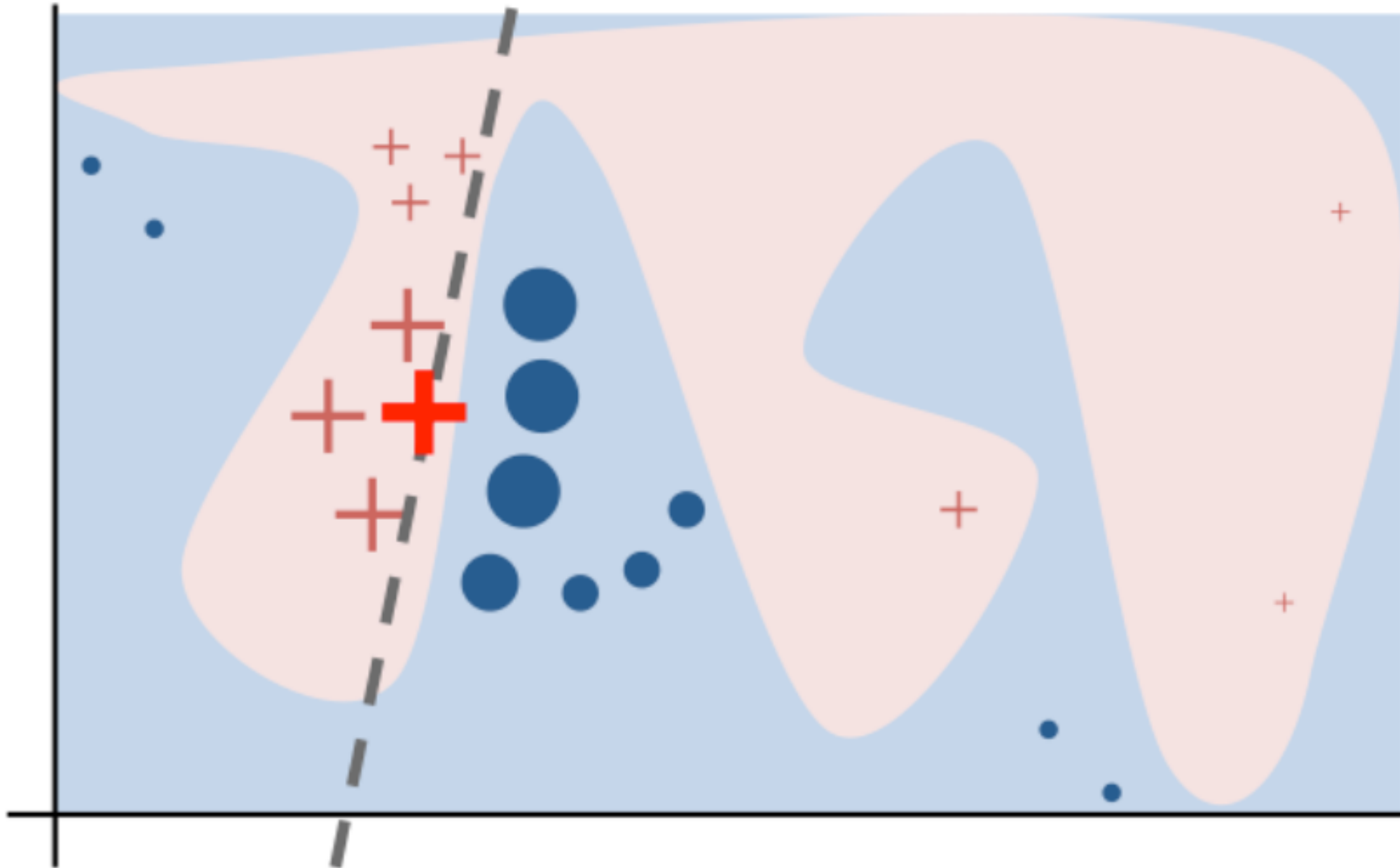
From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu

Hello Gang,

There have been some notes recently asking where to obtain the DARWIN fish.
This is the same question I have and I have not seen an answer on the net. If anyone has a contact please post on the net or email me.

Explanation is a very sparse linear model (with only 6 features).

How does it work?



Example of an application



(a) Original Image

(b) Explaining *Electric guitar*

(c) Explaining *Acoustic guitar*

(d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception network, highlighting positive pixels. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)