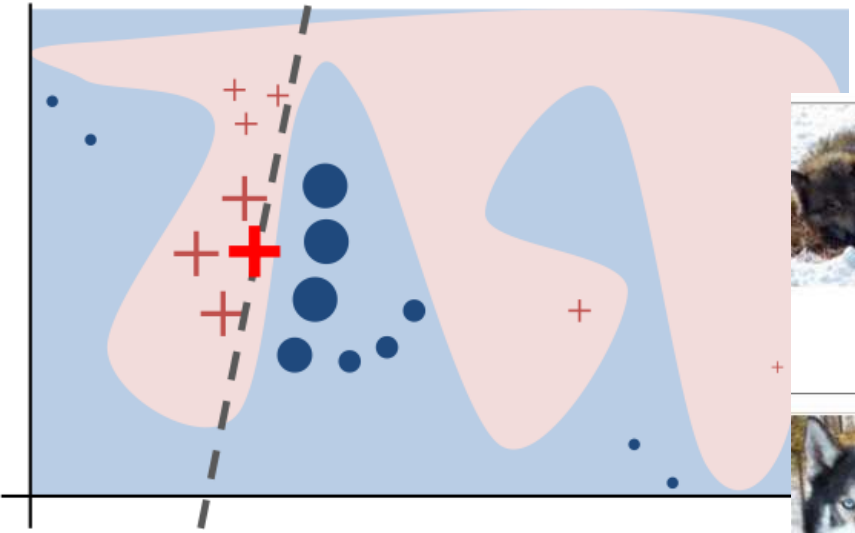
Anchors

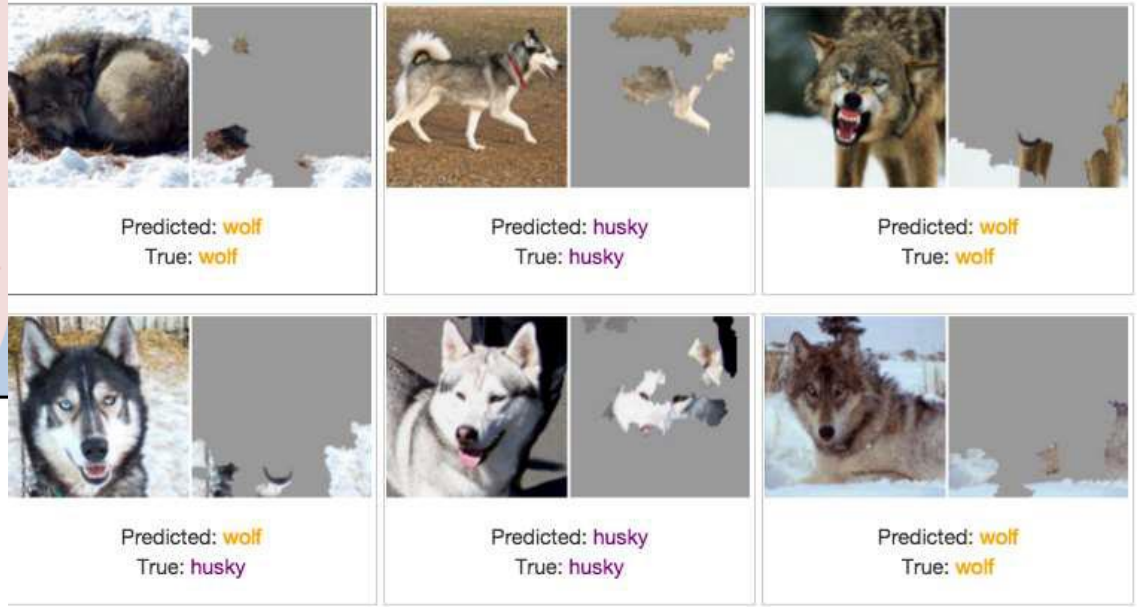
Alicja Gosiewska

Wydział Matematyki i Nauk Informacyjnych
Politechniki Warszawskiej

LIME

Local Interpretable Model-Agnostic Explanations





This movie is not bad.

This movie is not very good.

LIME explanations

Anchor explanations

{"not", "bad"} → Positive

{"not", "good"} → Negative

$f : X \to Y$ - black box model

$x \in X$ - an instance to be explained

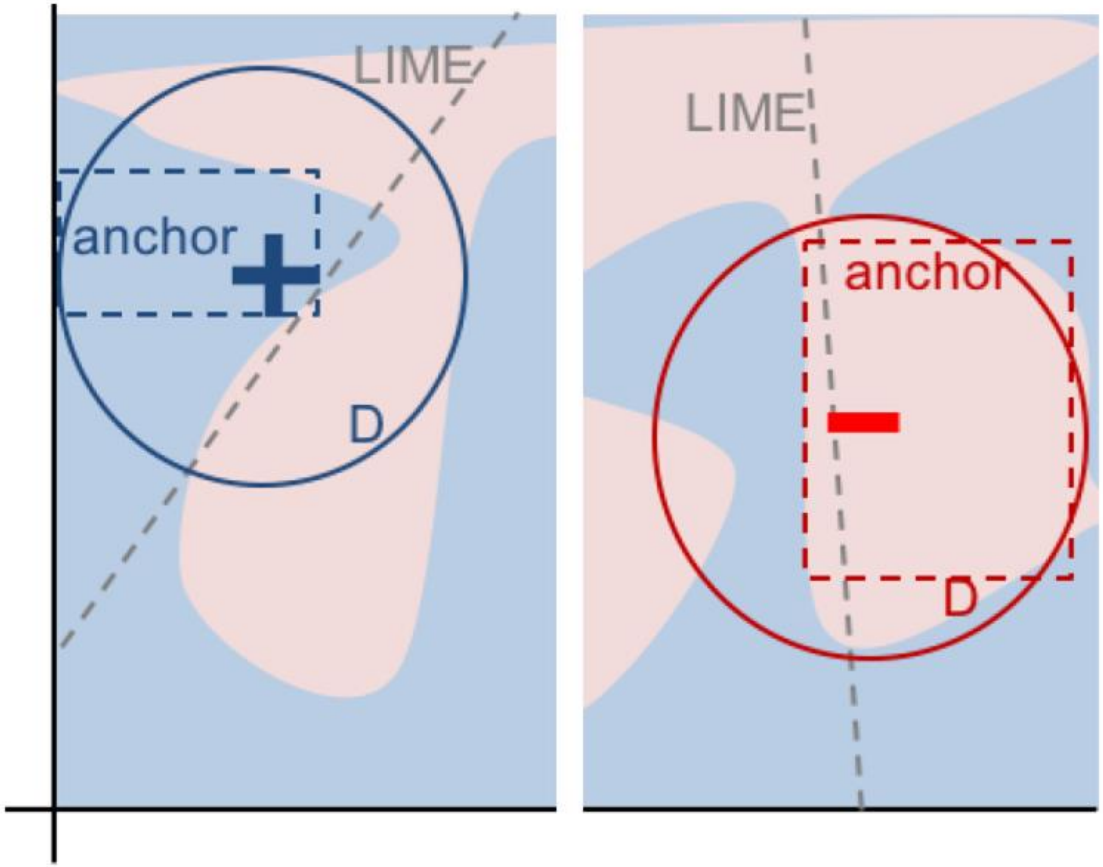
$\mathcal{D}_x$ - perturbation distribution

A - a rule (set of predicates)

$\mathcal{D}_x(\cdot|A)$ - conditional distribution when the rule A applies

A is an anchor if:

$$\mathbb{E}_{\mathcal{D}(z|A)}\left[\mathbb{1}_{f(x)=f(z)}\right] \geq \tau$$


LIME
anchor
D
LIME
anchor
D

	If	Predict
adult	No capital gain or loss, never married	$\leq$ 50K
	Country is US, married, work hours $>$ 45	$>$ 50K

An anchor A is a set of feature predicates that achieves $prec(A) \geqslant \tau$, where

$$\text{prec}(A) = \mathbb{E}_{\mathcal{D}(z|A)} \left[\mathbb{1}_{f(x)=f(z)} \right]$$

Probabilistic definition:

$$P\left(\text{prec}(A) \geq \tau\right) \geq 1 - \delta$$

Coverage of an anchor:

$$\text{cov}(A) = \mathbb{E}_{\mathcal{D}(z)}[A(z)]$$

Search for an anchor is the following combinatorial optimization problem:

$$\max_{A \text{ s.t. } P(\text{prec}(A) \geq \tau) \geq 1-\delta} \text{cov}(A)$$

MI

Algorithm 1 Identifying the *Best* Candidate for Greedy

function GenerateCands($\mathcal{A}, c$)
  $\mathcal{A}_r = \emptyset$
  for all $A \in \mathcal{A}; a_i \in x, a_i \notin A$ **do**
    if $\text{cov}(A \wedge a_i) > c$ **then** {Only high-coverage}
      $\mathcal{A}_r \leftarrow \mathcal{A}_r \cup (A \wedge a_i)$ {Add as potential anchor}
  return $\mathcal{A}_r$ {Candidate anchors for next round}

function BestCand($\mathcal{A}, \mathcal{D}, \epsilon, \delta$)
  initialize $\text{prec}, \text{prec}_{ub}, \text{prec}_{lb}$ estimates $\forall A \in \mathcal{A}$
  $A \leftarrow \arg\max_A \text{prec}(A)$
  $A' \leftarrow \arg\max_{A' \neq A} \text{prec}_{ub}(A', \delta)$ {δ implicit below}
  while $\text{prec}_{ub}(A') - \text{prec}_{lb}(A) > \epsilon$ **do**
    sample $z \sim \mathcal{D}(z|A), z' \sim \mathcal{D}(z'|A')$ {Sample more}
    update $\text{prec}, \text{prec}_{ub}, \text{prec}_{lb}$ for A and A'
    $A \leftarrow \arg\max_A \text{prec}(A)$
    $A' \leftarrow \arg\max_{A' \neq A} \text{prec}_{ub}(A')$
  return A

MI

Method	Precision		Coverage		Time/pred	
	adult	rcdv	adult	rcdv	adult	rcdv
No expls	54.8	83.1	79.6	63.5	29.8 ±14	35.7±26
LIME(1)	68.3	98.1	89.2	55.4	28.5±10	24.6±6
Anchor(1)	100.0	97.8	43.1	24.6	13.0±4	14.4±5

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin.
Anchors: High-Precision Model-Agnostic Explanations

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin.
Nothing Else Matters: Model-Agnostic Explanations
By Identifying Prediction Invariance

The anchor_exp python package