IDA
2023

LEMON

Alternative Sampling for More Faithful Explanation through Local Surrogate Models

Dennis Collaris, Pratik Gajane, Joost Jorritsma,
Jarke J. van Wijk, Mykola Pechenizkiy

TU/e
EINDHOVEN
UNIVERSITY OF
TECHNOLOGY

Introduction

Why do we need explanations?

LEMON

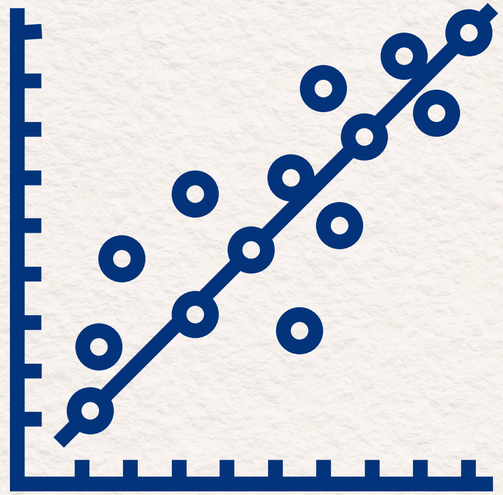
Background

How do we explain ML predictions?

BLACK BOX

≈

Linear regression



Neural network

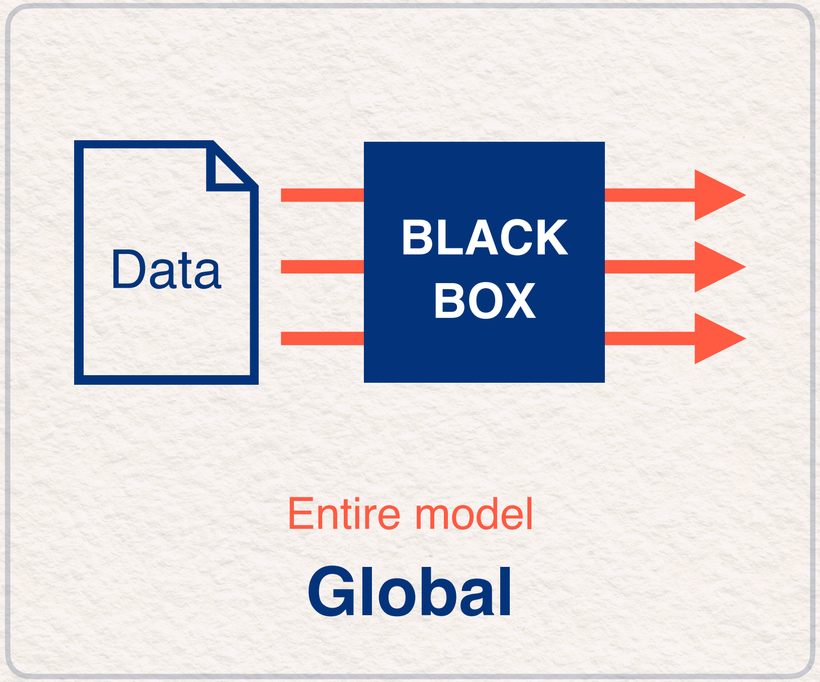
Support vector machines

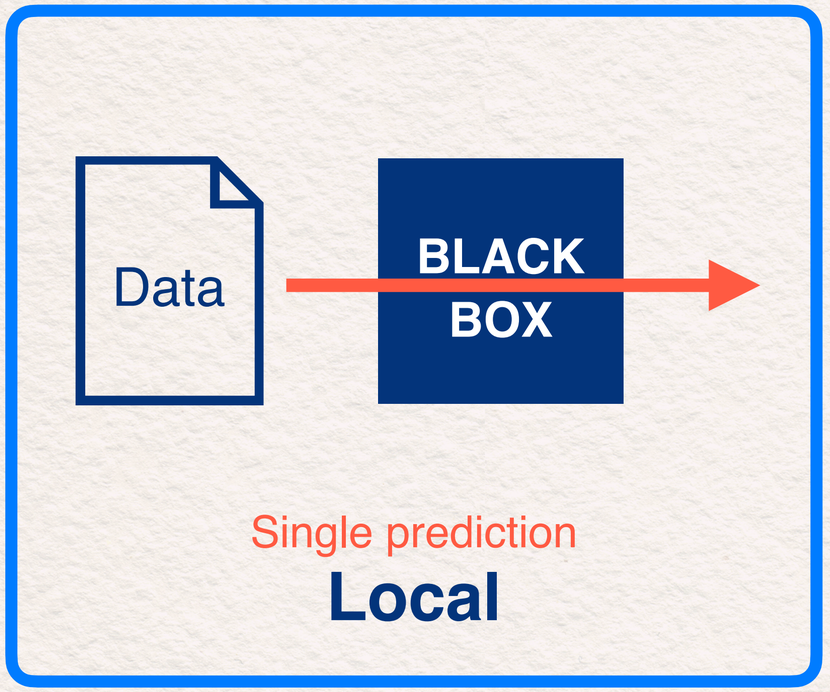
Random Forest

LEMON

Background

How do we explain ML predictions?

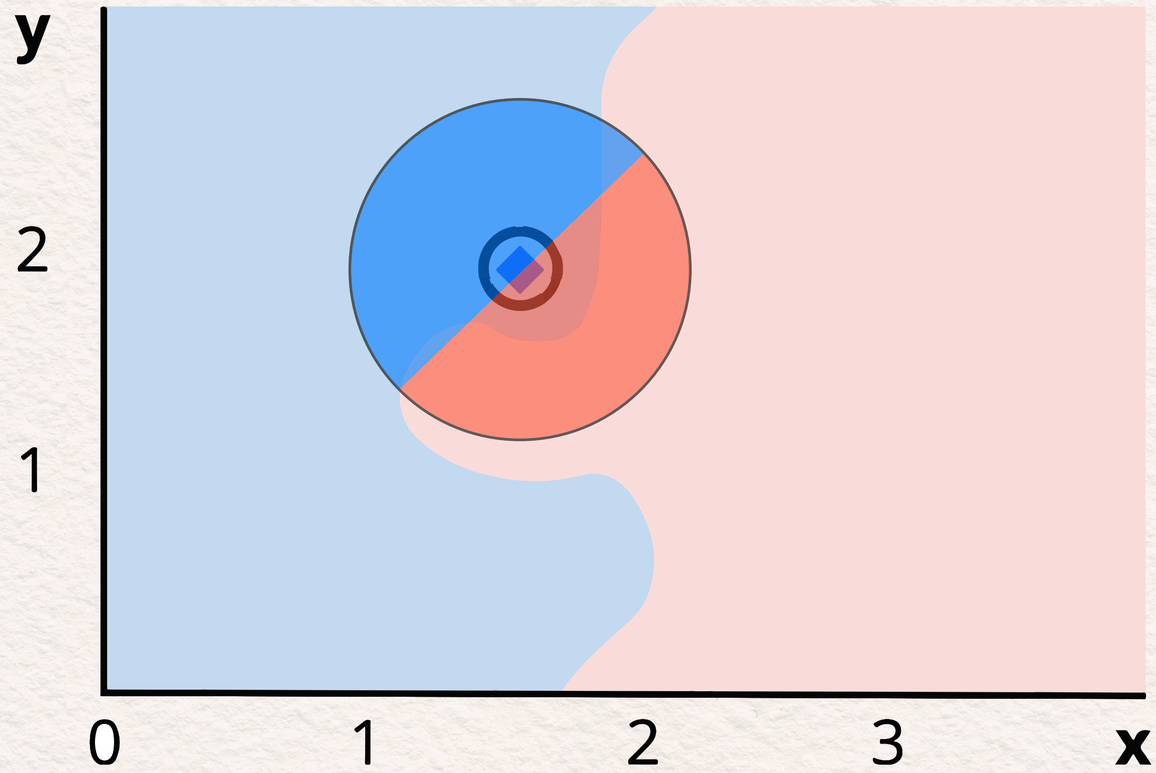




LEMON

Background

How do we explain ML predictions?



LEMON

Issues

Where do current techniques fall short?

Problem

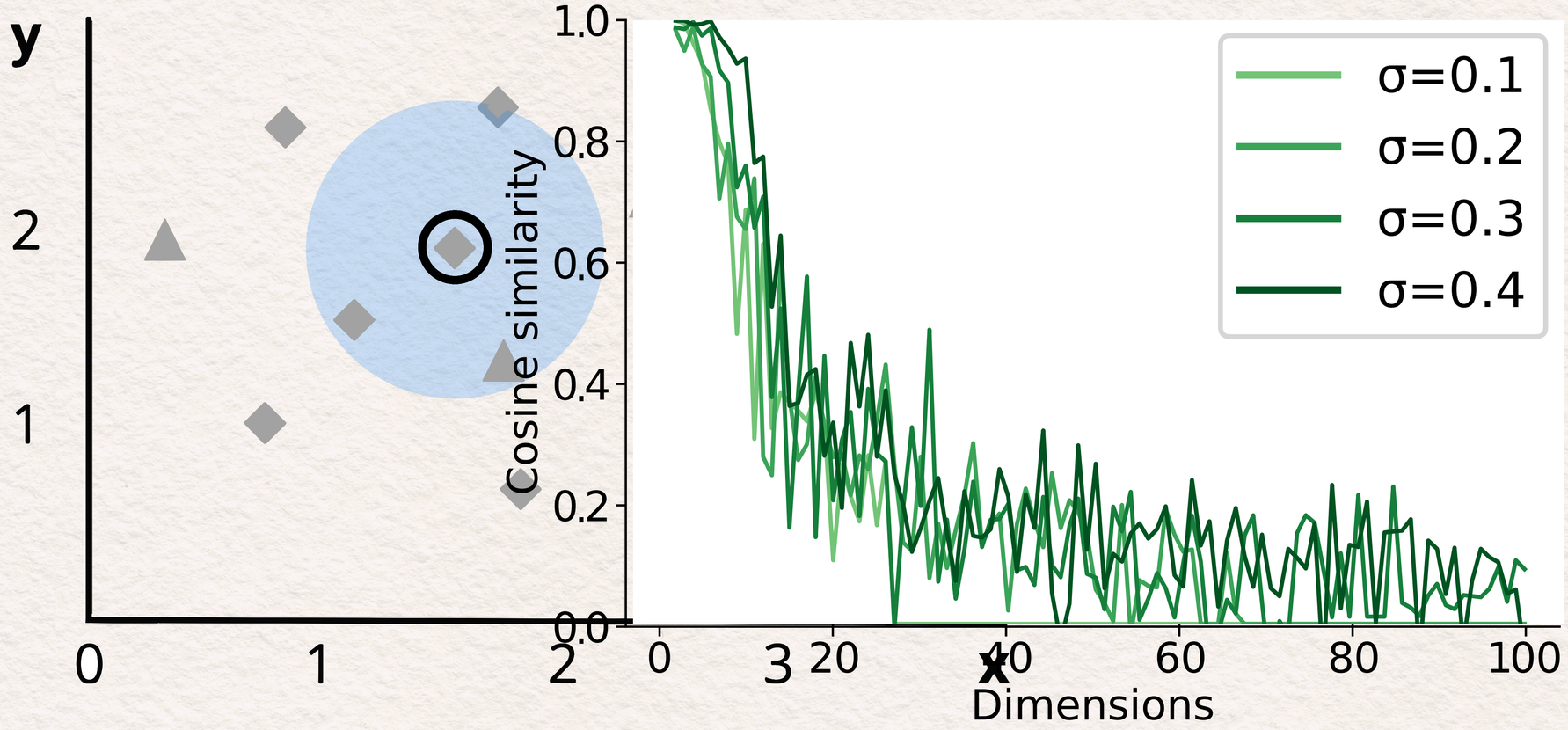
In high dimensions LIME explanations are less robust and less faithful.

[1] Alvarez-Melis, D., Jaakkola, T.S.: On the robustness of interpretability methods. Workshop on Human Interpretability in Machine Learning, 2018.

LEMON

Issues

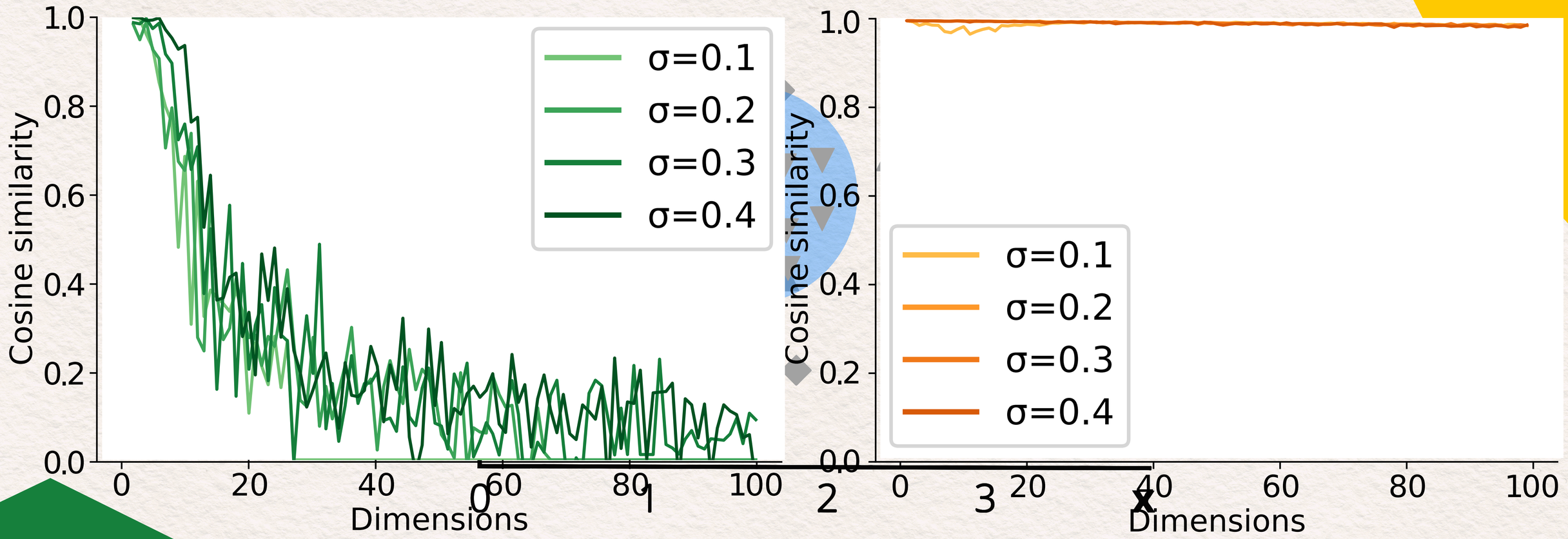
Where do current techniques fall short?



LIME

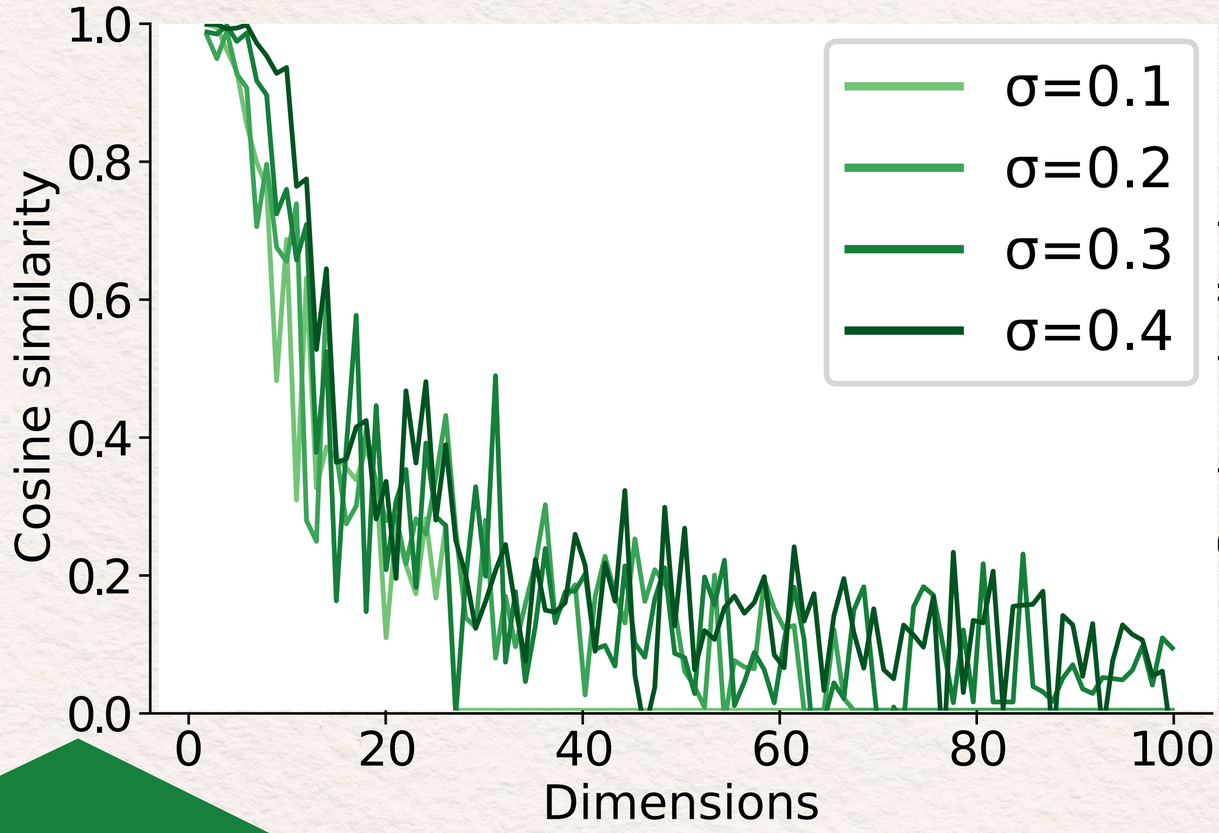
Solution

Introducing LEMON sampling

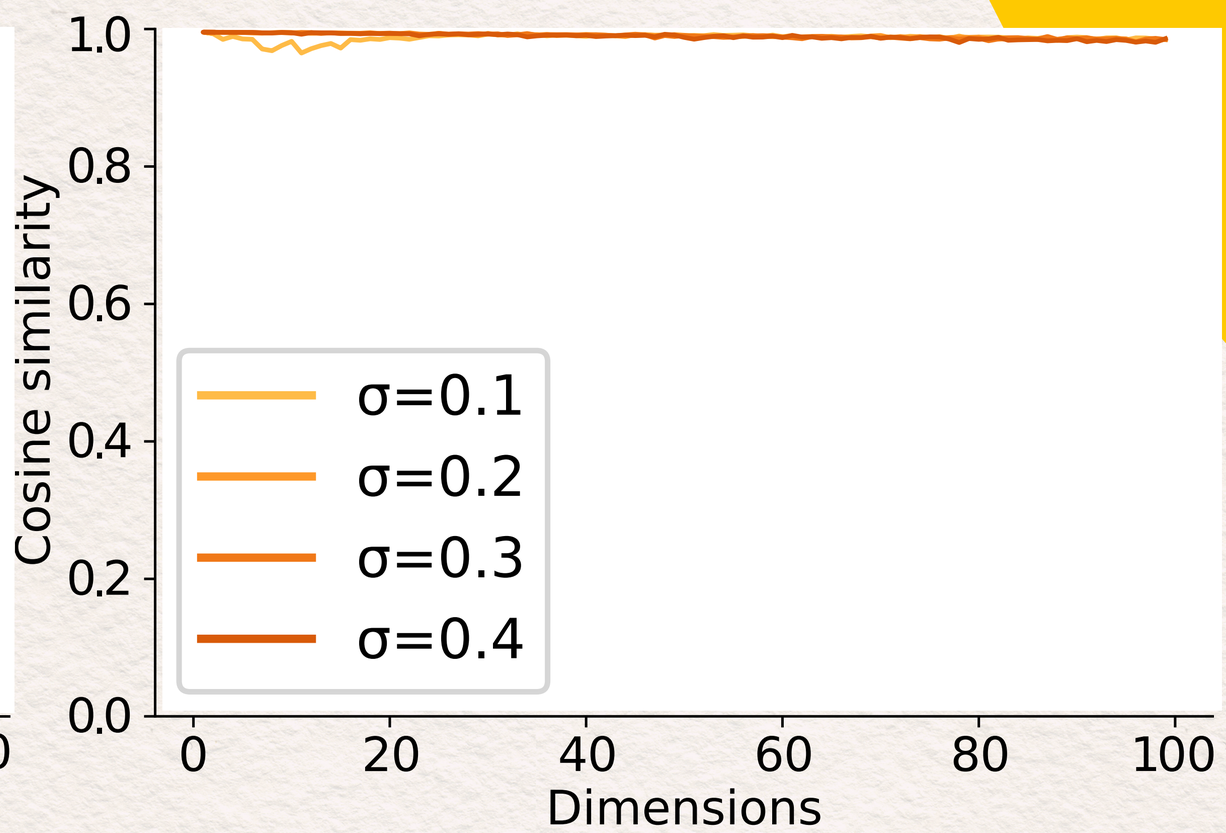


Solution

Introducing LEMON sampling



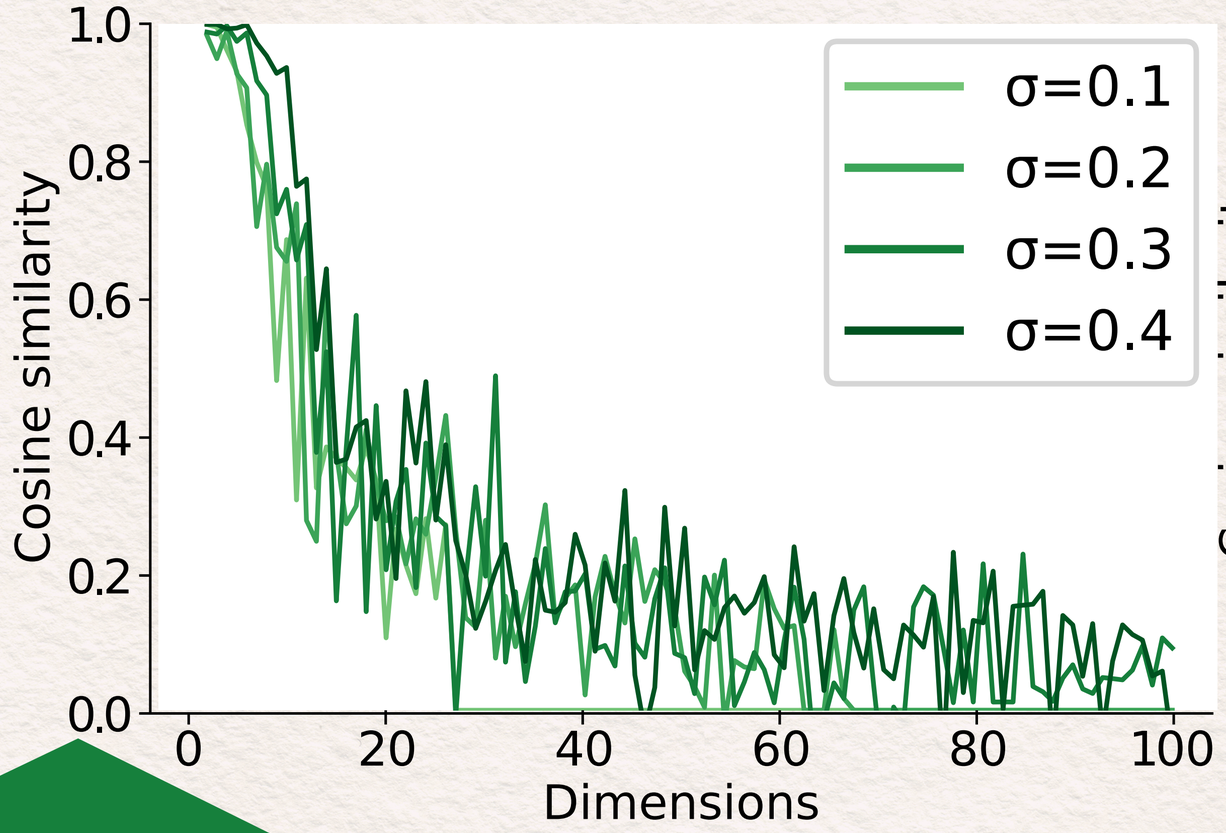
LIME



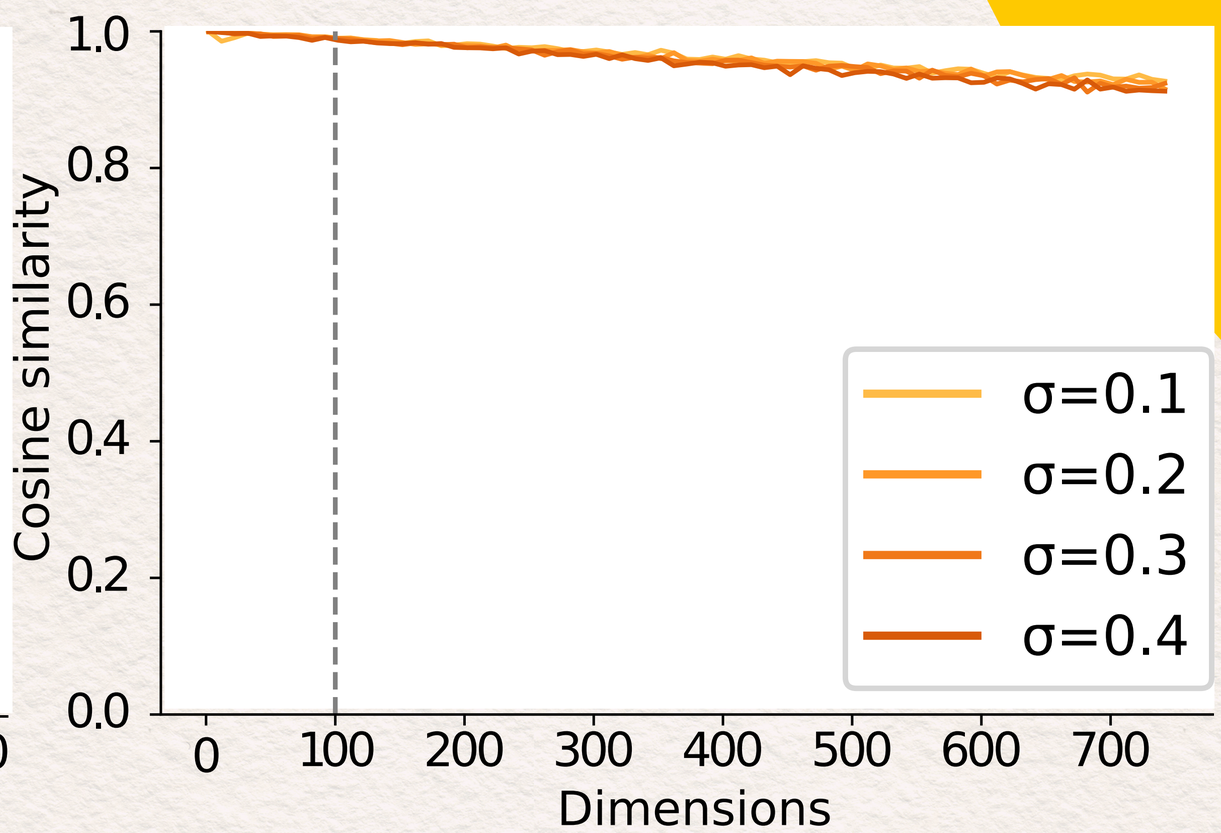
LEMON

Solution

Introducing LEMON sampling



LIME



LEMON

Evaluation

How does it perform in the real world?

Dataset
Wine

Model
Naive

$$RMSE(\hat{\mathbf{y}}^r, \hat{\mathbf{y}}^s) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\hat{\mathbf{y}}_i^r - \hat{\mathbf{y}}_i^s)^2}\,.$$

Dataset
Breast cancer

Model
Neural

LEMON

Evaluation

How do we evaluate?

1. Sample 50000 evaluation data points around the point

2. Label evaluation points with the:
    1. reference model;
    2. LIME surrogate; and
    3. LEMON surrogate

3. Calculate RMSE between reference and surrogate predictions

$$RMSE(\hat{\mathbf{y}}^r, \hat{\mathbf{y}}^s) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\hat{\mathbf{y}}_i^r - \hat{\mathbf{y}}_i^s)^2}\,.$$

LEMON

Evaluation

How does it perform in the real world?

			LIME	LEMON
Dataset **Breast cancer**	Model **Random**	$\sigma = \frac{3}{4}\sqrt{(n)}$	**0.367**	**0.065**
Dataset **Diabetes**	Model **Neural**	$\sigma = 1$	**0.067**	**0.063**

LEMON

Future work

What's next?

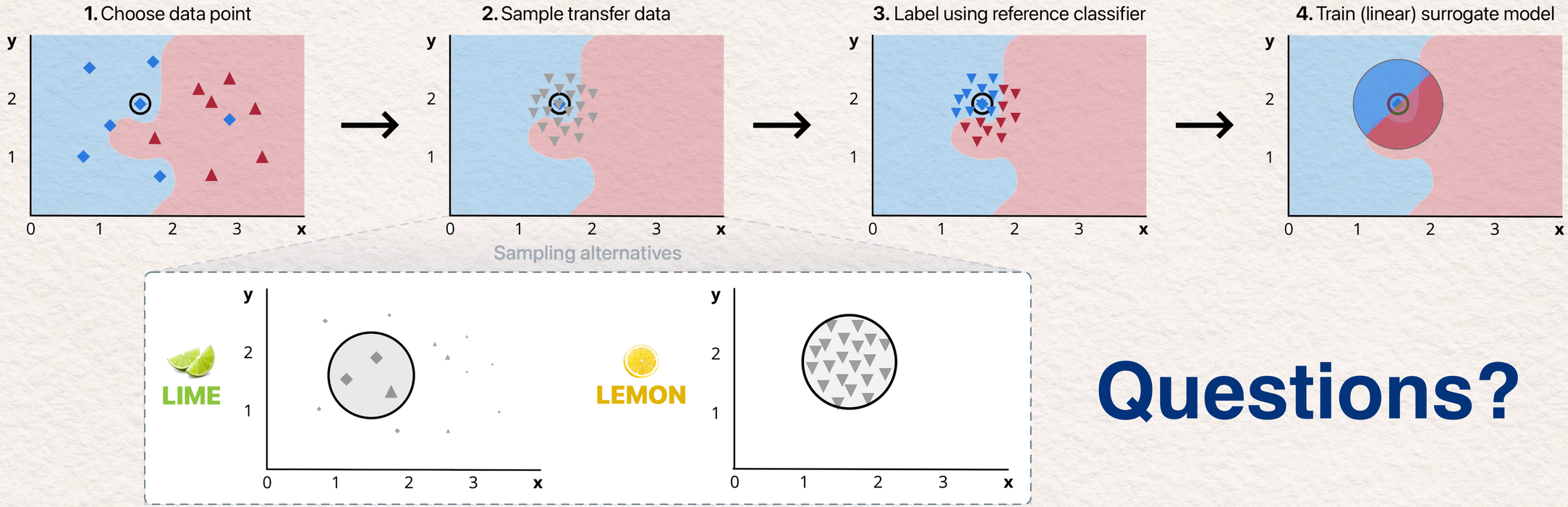
1. Supporting observation-based sampling
2. Investigating different kernel shapes
3. Explanations for multiple instances
4. Different faithfulness metrics

LEMON

LEMON

Alternative Sampling for More Faithful Explanation through Local Surrogate Models

More Info



Questions?