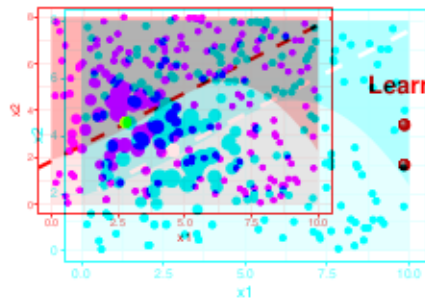


Interpretable Machine Learning



LIME Pitfalls



Learning goals

Learning goals

- Learn why LIME should be used with caution
- Possible pitfalls of LIME
- Learn why LIME should be used with caution
- Possible pitfalls of LIME



- LIME is one of the best known interpretable ML methods
- LIME is one of the best-known interpretable ML methods
 - ~ But several papers caution to be careful in practice
 - ~ But several papers caution to be careful in practice
- Problems can occur on different levels which are described subsequently:
- Problems can occur on different levels which are described subsequently:
 - Sampling procedure (extrapolation)
 - Definition of locality (sensitivity)
 - Scope of feature effects (local vs. global)
 - Faithfulness (trade-off with sparsity)
 - Surrogate model (hiding biases, robustness)
 - Definition of superpixels in case of image data (sensitivity)

PITFALL: SAMPLING



- **Pitfall:** Common sampling strategies for $\mathbf{z} \in \mathcal{Z}$ do not account for correlation between features
- **Implication:** Unlikely data points might be used to learn local explanation models
- **Implication:** Unlikely data points might be used to learn local explanation models

PITFALL: SAMPLING



- **Pitfall:** Common sampling strategies for $\mathbf{z} \in \mathcal{Z}$ do not account for correlation between features
- **Implication:** Unlikely data points might be used to learn local explanation models
- **Implication:** Unlikely data points might be used to learn local explanation models
- **Solution I:** Use a local sampler directly on \mathcal{X}
 - ↪ derivation is particularly difficult for high dimensional or mixed feature spaces
- **Solution II:** Use training data to fit surrogate model
 - ↪ only works well with enough data near \mathbf{x}

LIME PITFALL: LOCALITY

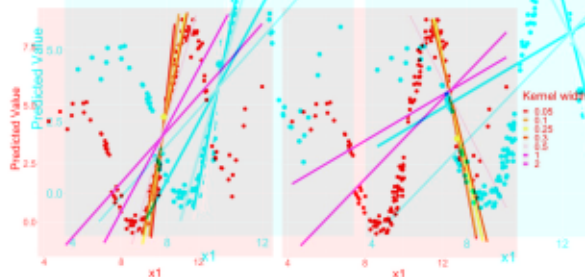
- Pitfall: Difficult to define locality (= how samples are weighted locally)
- Pitfall: Difficult to define locality (= how samples are weighted locally)
 - ~ Strongly affects local model, but there is no automatic procedure for choosing neighborhood
 - ~ Strongly affects local model, but there is no automatic procedure for choosing neighborhood
- Originally, an exponential kernel as proximity measure between \mathbf{x} and \mathbf{z} was proposed:
$$\phi_{\mathbf{x}}(\mathbf{z}) = \exp(-d(\mathbf{x}, \mathbf{z})^2 / \sigma^2)$$
 where d is a distance measure and σ is the kernel width
- Originally, an exponential kernel as proximity measure between \mathbf{x} and \mathbf{z} was proposed:
$$\phi_{\mathbf{x}}(\mathbf{z}) = \exp(-d(\mathbf{x}, \mathbf{z})^2 / \sigma^2)$$
 where d is a distance measure and σ is the kernel width



LIME PITFALL: LOCALITY

- Pitfall: Difficult to define locality (= how samples are weighted locally)
- Pitfall: Difficult to define locality (= how samples are weighted locally)
 - ~ Strongly affects local model, but there is no automatic procedure for choosing neighborhood
- Originally, an exponential kernel as proximity measure between \mathbf{x} and \mathbf{z} was proposed:

$\phi_{\mathbf{x}}(\mathbf{z}) = \exp(-d(\mathbf{x}, \mathbf{z})^2 / \sigma^2)$ where d is a distance measure and σ is the kernel width



- Surrogate models for 2 obs. (green points) for same model with one feature x_1
- First line refers to a linear surrogate model with different kernel width
- Each line refers to a linear surrogate model with different kernel width
- Right figure: larger kernel widths influence lines more
- Right figure: larger kernel widths influence lines more





- **Solution I:** Kernel width strongly interacts with locality:
- **Solution I:** Kernel width strongly interacts with locality:
 - Large kernel width leads to interaction with points further away (unwanted)
 - Small kernel width leads to small neighborhood
 - ~> risk of few data points
 - ~> potentially fitting more noise



- **Solution I:** Kernel width strongly interacts with locality:
 - **Solution I:** Kernel width strongly interacts with locality:
 - Large kernel width leads to interaction with points further away (unwanted)
 - Large kernel width leads to interaction with points further away (unwanted)
 - Small kernel width leads to small neighborhood
 - ~> risk of few data points
 - ~> risk of few data points
 - ~> potentially fitting more noise
 - ~> potentially fitting more noise
 - **Solution II:** Use Gower distance where no kernel width needs to be specified
- **Solution II:** Use Gower distance where no kernel width needs to be specified
 - **Problem:** data points far away receive weight > 0
 - **Problem:** data points far away receive weight > 0
 - ~> resulting explanations are rather global than local surrogates
 - ~> resulting explanations are rather global than local surrogates



- **Problem:**

By sampling obs. for the surrogate model from the whole input space, the influence of local features might be hidden in favor of features with global influence (even for small kernel width)



- **Problem:**

By sampling obs. for the surrogate model from the whole input space, the influence of local features might be hidden in favor of features with global influence (even for small kernel width)

- **Implication:**

- **Implication:**

- Some features influence the **global** shape of the black-box model
- Some features influence the **global** shape of the black-box model
- Some **local** features impact predictions only in smaller regions of \mathcal{X}
- Other **local** features impact predictions only in smaller regions of \mathcal{X}



- **Problem:**

By sampling obs. for the surrogate model from the whole input space, the influence of local features might be hidden in favor of features with global influence (even for small kernel width)

- **Implication:**

- **Implication:**

- Some features influence the **global** shape of the black-box model
- Some features influence the **global** shape of the black-box model
- Other **local** features impact predictions only in smaller regions of \mathcal{X}

- **Example: Decision trees**

- **Example: Decision trees**

⇒ Split features close to root have a more global influence than the ones close to leaves

Laugel et al. 2018

Binary classification model

Binary classification model

Right figure:

Right figure:

Black and grey crosses: training data

Green dot: Obs. to be explained

Background color: Classification of random forest

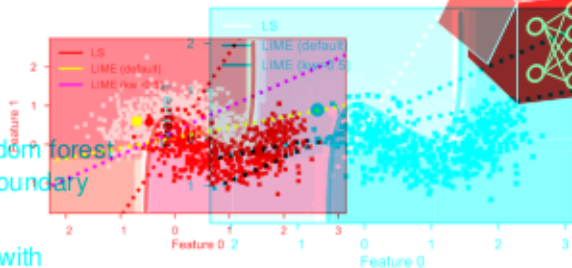
Dark grey curve: Classifier's decision boundary

Dotted lines: Local decision boundary

Observation: Decision boundaries of LIME with different kernels (blue and green lines) do not match

Dotted lines: Local decision boundary

Observation: Decision boundaries of LIME with different kernels (blue and green lines) do not match the direction of the local decision boundary (which appears steeper)



Half-moons dataset

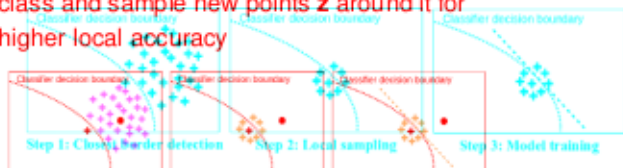
Half-moons dataset



SOLUTION

Laugel et al. 2018

- **Solution:** Find closest point to x from other class and
- **Solution:** Find closest point to x from other local accuracy class and sample new points z around it for higher local accuracy



Step 1: Closest border detection Step 2: Local sampling Step 3: Model training

Example: x (red point), closest point from other class (black cross)
(black cross)



SOLUTION

Laugel et al. 2018

- **Solution:** Find closest point to x from other class and
- **Solution:** Find closest point to x from other local accuracy class and sample new points z around it for higher local accuracy

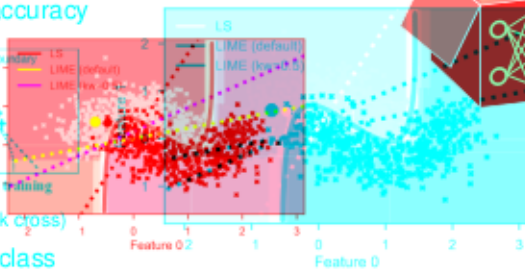


Example: x (red point), closest point from other class (black cross)

- **Red dot (right figure):** Closest point from other class (black cross)
- **Red line:** Local surrogate (LS) method
- **Red dot (right figure):** Closest point from other class
- **Red line:** Local surrogate (LS) method

Laugel et al. 2018

~> better approximates the local direction of the decision boundary



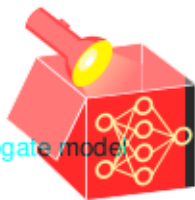
Half-moons dataset

Half-moons dataset

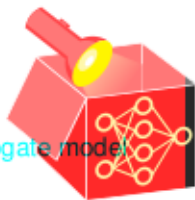


PITFALL: FAITHFULNESS

- **Problem:** Trade-off between local fidelity vs. sparsity
- **Observation I:** Low fidelity \rightsquigarrow unreliable explanations
- **Observation II:** High fidelity requires complex models \rightsquigarrow difficult to interpret surrogate model



PITFALL: FAITHFULNESS



- **Problem:** Trade-off between local fidelity vs. sparsity
- **Observation I:** Low fidelity \rightarrow unreliable explanations
- **Observation II:** High fidelity requires complex models \rightarrow difficult to interpret surrogate model
- **Example: Credit data**
- **Example: Credit data**
 - Original prediction by random forest for one data point \mathbf{x} :
 - Original prediction by random forest for one data point \mathbf{x} :

$$\hat{f}(\mathbf{x}) = \hat{P}(y = 1 | \mathbf{x}) = 0.143$$

$$\hat{f}(\mathbf{x}) = \hat{P}(y = 1 | \mathbf{x}) = 0.143$$

- Linear model with only three selected features (age, checking.account, duration):

- Linear model with only three selected features (age, checking.account, duration):

$$g_{lm}(\mathbf{x}) = \hat{\beta}_0 + \hat{\beta}_1 x_{age} + \hat{\beta}_2 x_{checking.account} + \hat{\beta}_3 x_{duration} = 0.283$$

- Generalized additive model (with all 9 features) is more complex:

- Generalized additive model (with all 9 features) is more complex:

$$g_{gam}(\mathbf{x}) = \hat{\theta}_0 + f_{age}(x_{age}) + f_{checking.account}(x_{checking.account}) + f_{duration}(x_{duration}) + \dots = 0.148$$

$$g_{gam}(\mathbf{x}) = \hat{\theta}_0 + f_{age}(x_{age}) + f_{checking.account}(x_{checking.account}) + f_{duration}(x_{duration}) + \dots = 0.148$$

PITFALL: HIDING BIASES Slack et al. 2020

- **Problem:** Developer could manipulate their model to hide biases
- **Observation:** LIME can sample out-of-distribution points (extrapolation)





● **Problem:** Developer could manipulate their model to hide biases

● **Observation:** LIME can sample out-of-distribution points (extrapolation)

● **Attack with adversarial model:**

● **Attack with adversarial model:**

1 classifier to discriminate between in-distribution and out-of-distribution data points

2 for in-distribution points, use the original (biased) model

3 for out-of-distribution points, use the original (biased) model

3 for out-of-distribution points produced for local explanation, use an unbiased model

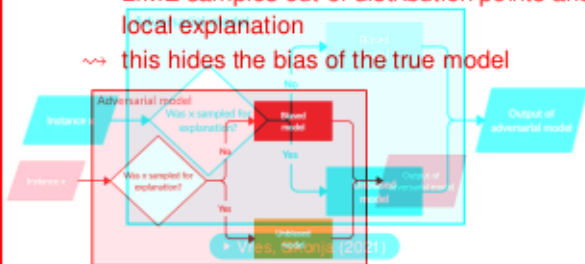
~ LIME samples out-of-distribution points and uses the unbiased model for local explanation

~ this hides the bias of the true model

~ LIME samples out-of-distribution points and uses the unbiased model for

local explanation

~ this hides the bias of the true model



► Vree, Sikojna (2021)



- **Problem:** Developer could manipulate their model to hide biases
- **Observation:** LIME can sample out-of-distribution points (extrapolation)
- **Attack with adversarial model:**

Attack with adversarial model:

- 1 classifier to discriminate between in-distribution and out-of-distribution data points
 - 2 for in-distribution points, use the original (biased) model
 - 3 for out-of-distribution points produced for local explanation, use an unbiased model
- ~ LIME samples out-of-distribution points and uses the unbiased model for local explanation
- ~ this hides the bias of the true model
- ~ LIME samples out-of-distribution points and uses the unbiased model for local explanation

Example: Not using 'gender' to approve a loan

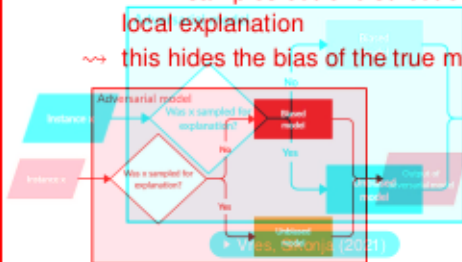
- biased model trained on features correlated with 'gender' (e.g. duration of parental leave)
- ~ used to make biased / unfair predictions

Example: Not using 'gender' to approve a loan

- biased model trained on features correlated with 'gender' (e.g. duration of parental leave)
- ~ used to make biased / unfair predictions
- unbiased model trained on features uncorrelated with 'gender' (e.g. duration of parental leave)
- ~ used to produce explanations based on unbiased predictions to hide bias

local explanation

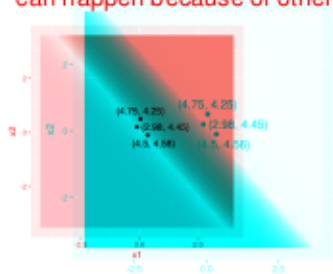
~ this hides the bias of the true model



Vree, Sikonja (2021)

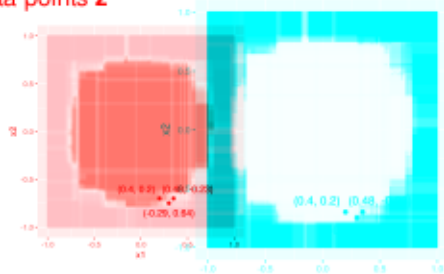


- **Problem:** Instability of explanations
 - **Problem:** Instability of explanations
 - **Observation:** Explanations of two very close points could vary greatly
 - **Observation:** Explanations of two very close points could vary greatly
- can happen because of other sampled data points z
- can happen because of other sampled data points z



Linear prediction task* (logistic regression).

Linear surrogate returns similar coefficients for similar points.



Circular prediction task (random forest).

Linear surrogate returns different coefficients for similar points.

- **Problem** Instability because of specification of superpixels for image data
- **Observation** Multiple specification of superpixels exist, influencing both the shape and size



- **Problem:** Instability because of specification of superpixels for image data
- **Observation:** Multiple specification of superpixels exist, influencing both the shape and size
- **Implication:** The specification of superpixel has a large influence on the explanations
- **Attack:** Change superpixels as part of an adversarial attack
- **Attack:** Change superpixels as part of an adversarial attack \rightsquigarrow changed explanation

