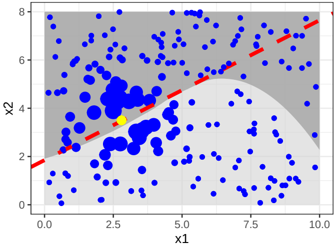
Interpretable Machine Learning

Local Explanations: LIME

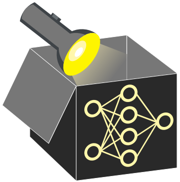
Pitfalls



Learning goals

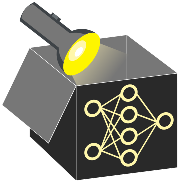
- Learn why LIME should be used with caution
- Possible pitfalls of LIME

LIME PITFALLS



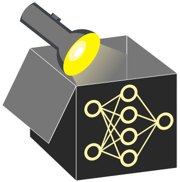
- LIME is one of the most widely used methods for local interpretability
  $\rightsquigarrow$ But several papers highlight important (practical) limitations
- Pitfalls arise at multiple levels, which will be discussed in detail:
  - **Sampling** – ignores feature dependencies, risks extrapolation
  - **Locality definition** – kernel width and dist. metrics affect sensitivity
  - **Local vs. global feats** – global signals may overshadow local ones
  - **Faithfulness** – trade-off between sparsity and local accuracy
  - **Hiding biases** – explanations can be manipulated to appear fair
  - **Robustness** – explanations vary for similar points
  - **Superpixels (images)** – instability due to segmentation method

PITFALL: SAMPLING



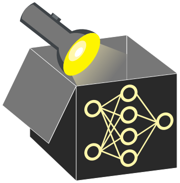
- **Pitfall**: Common sampling strategies for $\mathbf{z} \in \mathcal{Z}$ ignore feat dependencies
- **Implication:** Surrogate model may be trained on unrealistic points
  $\rightsquigarrow$ Undermines the fidelity and validity of the explanation

PITFALL: SAMPLING



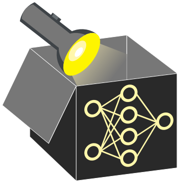
- **Pitfall**: Common sampling strategies for $\mathbf{z} \in \mathcal{Z}$ ignore feat dependencies
- **Implication:** Surrogate model may be trained on unrealistic points
  $\rightsquigarrow$ Undermines the fidelity and validity of the explanation
- **Solution I**: Sample locally from the true data manifold $\mathcal{X}$
  $\rightsquigarrow$ Challenging in high-dimensional or mixed-type data settings
- **Solution II**: Restrict sampling to training data near $\mathbf{x}$
  $\rightsquigarrow$ Requires enough training data points near $\mathbf{x}$

LIME PITFALL: LOCALITY



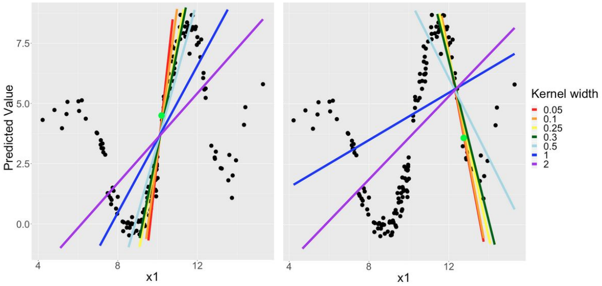
- **Pitfall**: Difficult to define locality (= how samples are weighted locally)
- **Implication:** Local model and explanation quality depend heavily on this weighting, but no principled way exists to choose it
- **Default:** Use exponential kernel as proximity measure between $\mathbf{x}$ and $\mathbf{z}$: $\phi_{\mathbf{x}}(\mathbf{z}) = exp(-d(\mathbf{x}, \mathbf{z})^2/\sigma^2)$ with distance measure d and kernel width σ

LIME PITFALL: LOCALITY



- **Pitfall**: Difficult to define locality (= how samples are weighted locally)
- **Implication:** Local model and explanation quality depend heavily on this weighting, but no principled way exists to choose it
- **Default:** Use exponential kernel as proximity measure between **x** and **z**: $\phi_{\mathbf{x}}(\mathbf{z}) = exp(-d(\mathbf{x}, \mathbf{z})^2/\sigma^2)$ with distance measure d and kernel width σ

Example: For 2 obs. (green points), fit local surr. models (lines) using only x_1

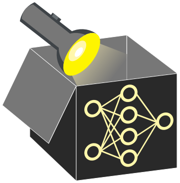


Line colors: different kernel widths used for proximity weighting

Right: larger kernel widths affect lines more

LIME PITFALL: LOCALITY

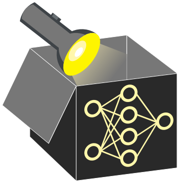
▸ "Kopper et al." 2019



- **Pitfall:** Choice of kernel width (σ) critically influences locality

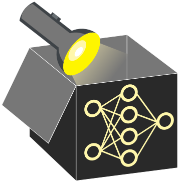
LIME PITFALL: LOCALITY

▸ "Kopper et al." 2019



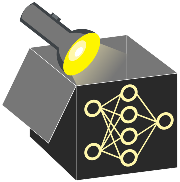
- **Pitfall:** Choice of kernel width (σ) critically influences locality
- **Implication of edge cases:**
  - *Large* $\sigma \rightarrow$ overemphasize distant points, hurting locality
  - *Small* $\sigma \rightarrow$ too few points may lead to unstable or noisy explanations

LIME PITFALL: LOCALITY ▸ "Kopper et al." 2019



- **Pitfall:** Choice of kernel width (σ) critically influences locality
- **Implication of edge cases:**
  - *Large* $\sigma \rightarrow$ overemphasize distant points, hurting locality
  - *Small* $\sigma \rightarrow$ too few points may lead to unstable or noisy explanations
- **Solution I:** Use Gower similarity directly as weights:
  $\pi(\mathbf{z}) = 1 - d_{\text{Gower}}(\mathbf{x}, \mathbf{z})$
  $\leadsto$ No kernel width required, but far points still receive (too high) weight
  $\leadsto$ Explanation may reflect more global than local structure
  $\leadsto$ Used in practical LIME implementations ▸ "lime package" n.d.

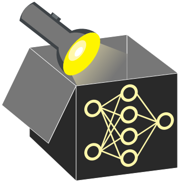
LIME PITFALL: LOCALITY ▸ "Kopper et al." 2019



- **Pitfall:** Choice of kernel width (σ) critically influences locality
- **Implication of edge cases:**
  - *Large* $\sigma \rightarrow$ overemphasize distant points, hurting locality
  - *Small* $\sigma \rightarrow$ too few points may lead to unstable or noisy explanations
- **Solution I:** Use Gower similarity directly as weights:
  $\pi(\mathbf{z}) = 1 - d_{\text{Gower}}(\mathbf{x}, \mathbf{z})$
  $\rightsquigarrow$ No kernel width required, but far points still receive (too high) weight
  $\rightsquigarrow$ Explanation may reflect more global than local structure
  $\rightsquigarrow$ Used in practical LIME implementations ▸ "lime package" n.d.
- **Solution II:** s-LIME adaptively selects σ to balance fidelity and stability
  ▸ "Gaudel et al." 2022

PITFALL: LOCAL VS. GLOBAL FEATURES

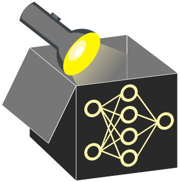
▸ "Laugel et al." 2018



- **Pitfall:** Sampling from entire input space may hide influence of locally relevant feat in favor of globally relevant ones, even for narrow kernels.
- **Feature types:**
  - *Global features* influence predictions broadly across whole input space $\mathcal{X}$
  - *Local features* affect predictions only in small subregions of $\mathcal{X}$

PITFALL: LOCAL VS. GLOBAL FEATURES

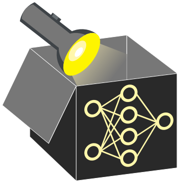
▸ **"Laugel et al." 2018**



- **Pitfall:** Sampling from entire input space may hide influence of locally relevant feat in favor of globally relevant ones, even for narrow kernels.
- **Feature types:**
  - *Global features* influence predictions broadly across whole input space $\mathcal{X}$
  - *Local features* affect predictions only in small subregions of $\mathcal{X}$
- **Implication:** LIME's surrogate may over-weight global features, producing explanations that miss critical local signals.

PITFALL: LOCAL VS. GLOBAL FEATURES

▸ "Laugel et al." 2018



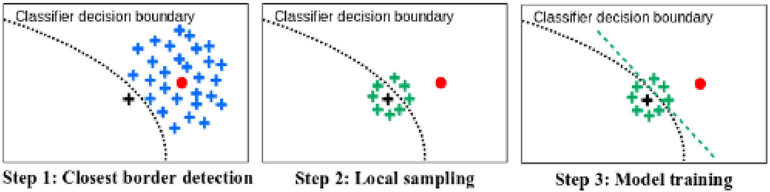
- **Pitfall:** Sampling from entire input space may hide influence of locally relevant feat in favor of globally relevant ones, even for narrow kernels.
- **Feature types:**
  - *Global features* influence predictions broadly across whole input space $\mathcal{X}$
  - *Local features* affect predictions only in small subregions of $\mathcal{X}$
- **Implication:** LIME's surrogate may over-weight global features, producing explanations that miss critical local signals.
- **Example**: Decision trees
  - Features near the root impact many instances $\rightarrow$ global
  - Features in lower nodes act locally

PITFALL: LOCAL VS. GLOBAL FEATURES

▸ "Laugel et al." 2018

- **Problem:** Sampling around observation to be explained **x** may miss decision boundary
- **Solution (LS: Local Surrogate Method):**
  1. Find closest point to **x** (red dot) from opposite class (black cross)
  2. Sample around that point to better capture boundary
  3. Train local surrogate using those samples

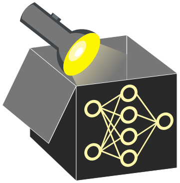
  $\rightsquigarrow$ better approximates the local direction of the decision boundary



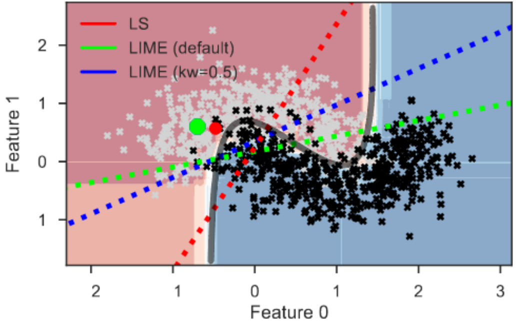
Example: **x** (red point), closest point from other class (black cross)

- LIME: What does the model do around this point?
- LS: How does the model change when crossing boundary near this point?

PITFALL: LOCAL VS. GLOBAL FEATS – EXAMPLE

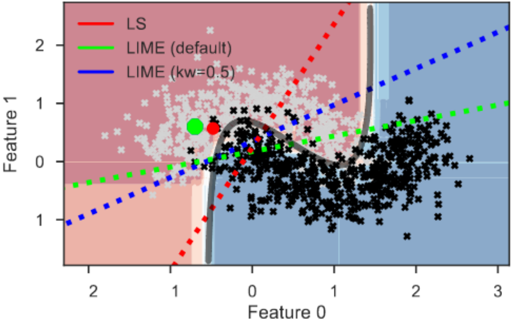


- Random forest (RF) classification on half-moons dataset
- **Background color:** Classification of RF (prediction surface)
- **Black/grey crosses:** training data
- **Green dot:** Obs. to be explained; **Red dot:** nearest opposite-class point
- **Grey curve:** RF's decision boundary; **Dotted lines:** LIME dec. bound.
- **Red line:** Local surrogate (LS) method ▸ "Laugel et al." 2018



PITFALL: LOCAL VS. GLOBAL FEATS – EXAMPLE

- Random forest (RF) classification on half-moons dataset
- **Background color:** Classification of RF (prediction surface)
- **Black/grey crosses:** training data
- **Green dot:** Obs. to be explained; **Red dot:** nearest opposite-class point
- **Grey curve:** RF's decision boundary; **Dotted lines:** LIME dec. bound.
- **Red line:** Local surrogate (LS) method ▸ "Laugel et al." 2018

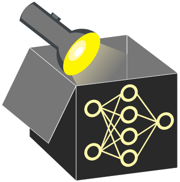


Feature 0 is global; class always flips when moving left (red) to right (blue)

Feature 1 is local; class flips only near boundary when moving up/down

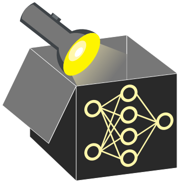
Observation: LIME decision boundaries (blue/green) fail to match the steep local bound. captured by LS (red)

PITFALL: FAITHFULNESS



- **Problem**: Trade-off between local fidelity vs. sparsity
- **Observation**:
  - Too simple model $\rightsquigarrow$ low fidelity $\rightsquigarrow$ unreliable explanations
  - Complex model $\rightsquigarrow$ high fidelity $\rightsquigarrow$ difficult to interpret surrogate

PITFALL: FAITHFULNESS



- **Problem**: Trade-off between local fidelity vs. sparsity
- **Observation**:
  - Too simple model $\rightsquigarrow$ low fidelity $\rightsquigarrow$ unreliable explanations
  - Complex model $\rightsquigarrow$ high fidelity $\rightsquigarrow$ difficult to interpret surrogate
- **Example: Credit data**
  - Random forest prediction for $\mathbf{x}$: $\hat{f}(\mathbf{x}) = \hat{\mathbb{P}}(y = \text{bad} \mid \mathbf{x}) = 0.143$
  - Sparse LM with 3 features (`age`, `checking.account`, `duration`):

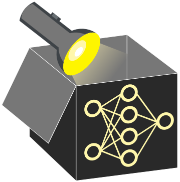
$$\hat{g}_{lm}(\mathbf{x}) = \hat{\theta}_0 + \hat{\theta}_1 x_{age} + \hat{\theta}_2 x_{checking.account} + \hat{\theta}_3 x_{duration} = 0.283$$

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

$$\hat{g}_{gam}(\mathbf{x}) = \hat{\theta}_0 + f_1(x_{age}) + f_2(x_{checking.account}) + f_3(x_{duration}) + \cdots = 0.148$$

PITFALL: HIDING BIASES

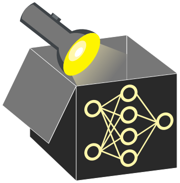
▸ "Slack et al." 2020



- **Problem:** LIME samples out-of-distribution (OOD) points, making it exploitable
- **Risk:** Developers can adversarially hide bias in the original model

PITFALL: HIDING BIASES

▸ "Slack et al." 2020



- **Problem:** LIME samples out-of-distribution (OOD) points, making it exploitable
- **Risk:** Developers can adversarially hide bias in the original model
- **Attack** with adversarial model:
  1. Train a detector to distinguish in-distribution vs. OOD points
  2. Use **biased model** for in-distribution inputs (i.e., true predictions)
  3. Use **unbiased model** for OOD samples to get LIME explanations

  $\rightsquigarrow$ LIME explanations rely on unbiased model
  $\Rightarrow$ hides bias in original model

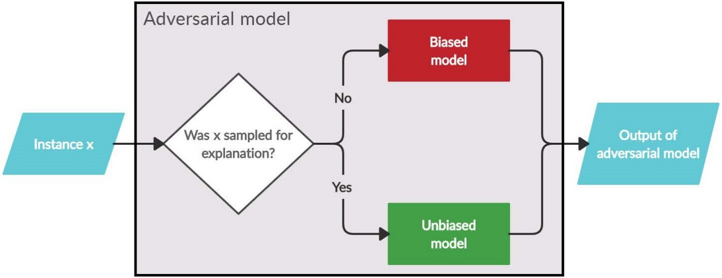
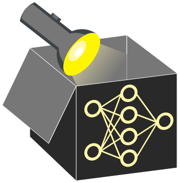


Image Source: ▸ "Vres, Sikonja" 2021

PITFALL: HIDING BIASES

▸ "Slack et al." 2020



Key insight: LIME can be fooled if explanations rely on model behavior outside the true data manifold.

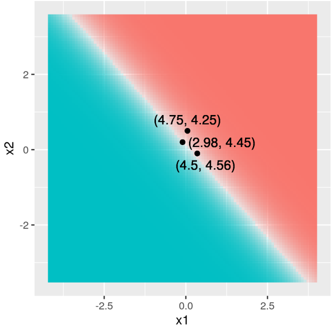
Example: Credit approval

- Biased model uses feats correlated with gender (parental leave duration)
  $\rightsquigarrow$ used to make biased/unfair predictions
- Unbiased model uses only features unrelated to gender for fairness
  $\rightsquigarrow$ used to produce explanations based on unbiased predictions in order to hide bias
- LIME's extrapolated samples trigger the unbiased model
  $\Rightarrow$ explanation appears fair, but original predictions are biased

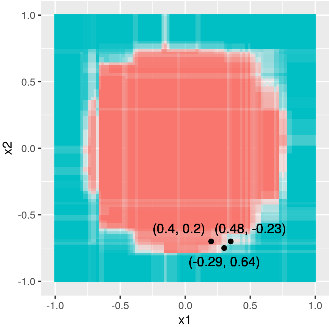
PITFALL: ROBUSTNESS

▸ "Alvarez-Melis, D., & Jaakkola, T." 2018

- **Problem**: Instability of LIME explanations
- **Observation**: Explanations of two very close points could vary greatly
  - ⇝ Variability driven by the stochastic sampling of **z** for each explanation
- **Example:**



Linear task (logistic regression). LIME returns similar coefficients for similar points.



Nonlinear task (random forest). LIME returns different coefficients for similar points.

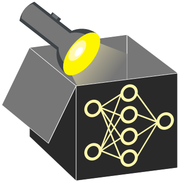
PITFALL: DEFINITION OF SUPERPIXELS

▸ "Achanta et al." 2012

- **Problem**: LIME relies on superpixels (but their definition differ) for image data
- **Observation**: Definition of superpixel differ, influencing their size, shape, and alignment

PITFALL: DEFINITION OF SUPERPIXELS

▸ "Achanta et al." 2012



- **Problem**: LIME relies on superpixels (but their definition differ) for image data
- **Observation**: Definition of superpixel differ, influencing their size, shape, and alignment
- **Implication**: Specification of superpixel has a large influence on LIME explanations
- **Attack**: Change superpixels as part of an adversarial attack $\rightsquigarrow$ changed explanation