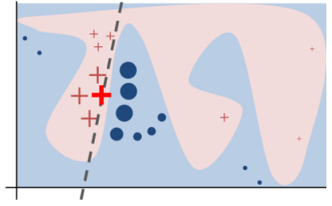
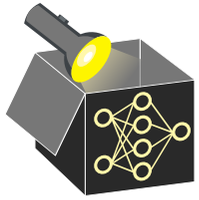
Interpretable Machine Learning

Introduction to Local Explanations



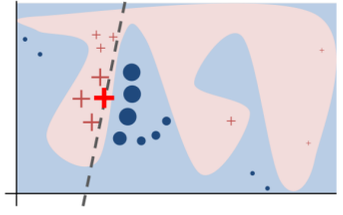
Learning goals

- Understand motivation for local explanations
- Develop an intuition for possible use-cases
- Know characteristics of local explanation methods



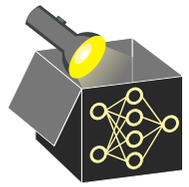
Interpretable Machine Learning

Local Explanations: LIME
Introduction to Local Explanations



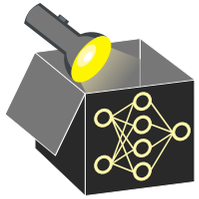
Learning goals

- Understand motivation for local explanations
- Develop an intuition for possible use-cases
- Know characteristics of local explanation methods



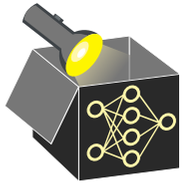
METHODOLOGICAL MOTIVATION

- Purpose of local explanations:
  - Insight into the driving factors for a **particular prediction/decision**
  - Understand ML model decisions in a **local neighborhood** of a given input (e.g., feature vector)

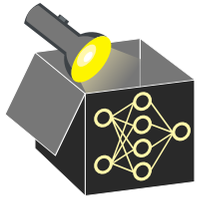


METHODOLOGICAL MOTIVATION

- Purpose of local explanations:
  - Insight into the driving factors for a **particular prediction/decision**
  - Understand ML model decisions in a **local neighborhood** of a given input (e.g., feature vector)



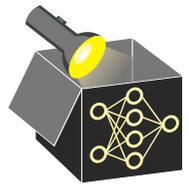
METHODOLOGICAL MOTIVATION



- Purpose of local explanations:
  - Insight into the driving factors for a **particular prediction/decision**
  - Understand ML model decisions in a **local neighborhood** of a given input (e.g., feature vector)

- Local Methods can address questions such as:
  - **Why** did the model decide to predict $\hat{y}$ for input **x**?
  - **How** does the model behave for observations similar to **x**?
  - **What if** some features of **x** had different values?
  - **Where** (in which regions in $\mathcal{X}$) does the model fail?

METHODOLOGICAL MOTIVATION

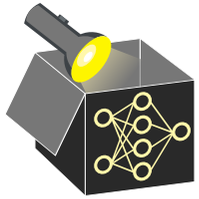


- Purpose of local explanations:
  - Insight into the driving factors for a **particular prediction/decision**
  - Understand ML model decisions in a **local neighborhood** of a given input (e.g., feature vector)

- Local Methods can address questions such as:
  - **Why** did the model decide to predict $\hat{y}$ for input **x**?
  - **How** does the model behave for observations similar to **x**?
  - **What if** some features of **x** had different values?
  - **Where** (in which regions in $\mathcal{X}$) does the model fail?

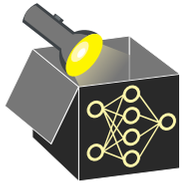
SOCIAL MOTIVATION

- Explanations for laypersons should be tailored to the **explainee**
  $\rightsquigarrow$ **case specific**, **human-intelligible**, and **faithful** to the explained mechanism



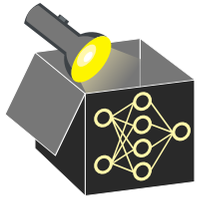
SOCIAL MOTIVATION

- Explanations for laypersons should be tailored to the **explainee**
  $\rightsquigarrow$ **case specific**, **human-intelligible**, **faithful** to explained mechanism



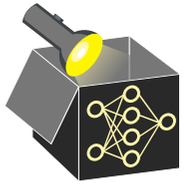
SOCIAL MOTIVATION

- Explanations for laypersons should be tailored to the **explainee**
  $\rightsquigarrow$ **case specific**, **human-intelligible**, and **faithful** to the explained mechanism
- If algorithms make decisions in **socially/safety critical domains**, end users have a justified interest in receiving explanations

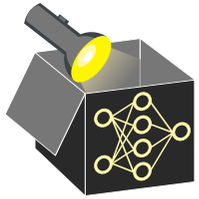


SOCIAL MOTIVATION

- Explanations for laypersons should be tailored to the **explainee**
  $\rightsquigarrow$ **case specific**, **human-intelligible**, **faithful** to explained mechanism
- If algorithms make decisions in **socially/safety critical domains**, end users have a justified interest in receiving explanations

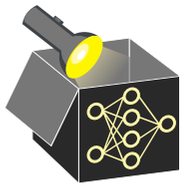


SOCIAL MOTIVATION



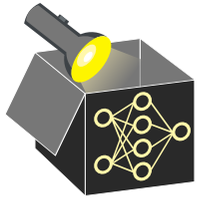
- Explanations for laypersons should be tailored to the **explainee**
  $\rightsquigarrow$ **case specific**, **human-intelligible**, and **faithful** to the explained mechanism
- If algorithms make decisions in **socially/safety critical domains**, end users have a justified interest in receiving explanations
- Local explanations cannot only increase **user trust**, but also help to detect **critical local biases** in algorithmic decision making

SOCIAL MOTIVATION



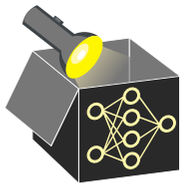
- Explanations for laypersons should be tailored to the **explainee**
  $\rightsquigarrow$ **case specific**, **human-intelligible**, **faithful** to explained mechanism
- If algorithms make decisions in **socially/safety critical domains**, end users have a justified interest in receiving explanations
- Local explanations cannot only increase **user trust**, but also help to detect **critical local biases** in algorithmic decision making

SOCIAL MOTIVATION



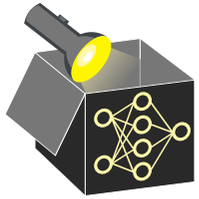
- Explanations for laypersons should be tailored to the **explainee**
  $\rightsquigarrow$ **case specific**, **human-intelligible**, and **faithful** to the explained mechanism
- If algorithms make decisions in **socially/safety critical domains**, end users have a justified interest in receiving explanations
- Local explanations cannot only increase **user trust**, but also help to detect **critical local biases** in algorithmic decision making
- European citizens have the legally binding **right to explanation** as given in the General Data Protection Regulation (GDPR) and the AI Act
  $\rightsquigarrow$ Instead of explaining the entire (complex) model (with potential market secrets), explanations in a case-by-case usage are more reasonable

SOCIAL MOTIVATION



- Explanations for laypersons should be tailored to the **explainee**
  $\rightsquigarrow$ **case specific**, **human-intelligible**, **faithful** to explained mechanism
- If algorithms make decisions in **socially/safety critical domains**, end users have a justified interest in receiving explanations
- Local explanations cannot only increase **user trust**, but also help to detect **critical local biases** in algorithmic decision making
- European citizens have the legally binding **right to explanation** as given in the General Data Protection Regulation (GDPR) and the AI Act
  $\rightsquigarrow$ Instead of explaining the entire (complex) model (with potential market secrets), explanations in a case-by-case usage are more reasonable

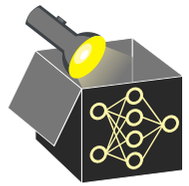
GDPR & AI ACT: THE RIGHT TO EXPLANATION



“The data subject should have the right not to be subject to a decision [...] based solely on automated processing [...], such as automatic refusal of an online credit application or e-recruiting practices without any human intervention.
[. . .]
In any case, such processing should be subject to suitable safeguards, which should include [...] the **right to obtain** [...] **an explanation of the decision reached after such assessment and to challenge the decision**.” ▸ Recital 71, GDPR, 2016

“Any affected person [...] shall have the right to obtain from the deployer clear and meaningful explanations of the role of the AI system in the decision-making procedure and the main elements of the decision taken.” ▸ Art. 86, AI Act, 2021

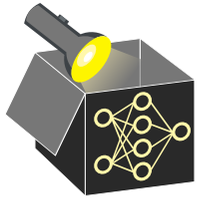
GDPR & AI ACT: THE RIGHT TO EXPLANATION



“The data subject should have the right not to be subject to a decision [...] based solely on automated processing [...], such as automatic refusal of an online credit application or e-recruiting practices without any human intervention.
[. . .]
In any case, such processing should be subject to suitable safeguards, which should include [...] the **right to obtain** [...] **an explanation of the decision reached after such assessment and to challenge the decision**.” ▸ GDPR 2016

“Any affected person [...] shall have the right to obtain from the deployer clear and meaningful explanations of the role of the AI system in the decision-making procedure and the main elements of the decision taken.”
▸ Act 2021

EXAMPLE: HUSKY OR WOLF?

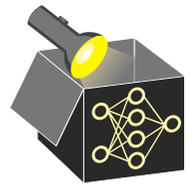


- We trained a model to predict if an image shows a wolf or a husky
- Below predictions on six test images are given
- Do you trust our predictor?

- Sometimes the ML model is wrong
- Can you guess the pattern the ML model learned to identify a wolf?

Source: [Sameer Singh 2018]

EXAMPLE: HUSKY OR WOLF?

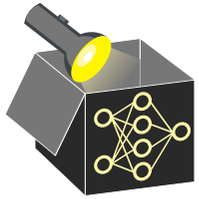


- We trained a model to predict if an image shows a wolf or a husky
- Below predictions on six test images are given
- Do you trust our predictor?

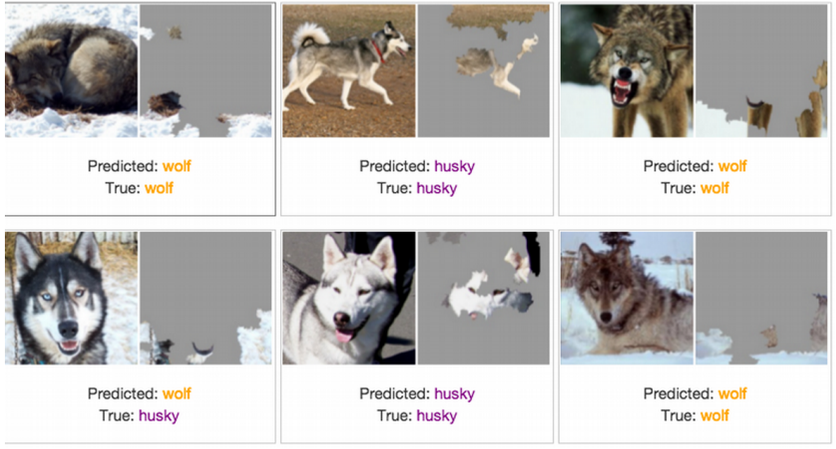
- Sometimes the ML model is wrong
- Can you guess the pattern the ML model learned to identify a wolf?

Source: [Sameer Singh 2018]

EXAMPLE: HUSKY OR WOLF? USING LIME

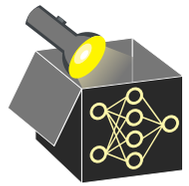


- Local explanations highlight the parts of an image which led to the prediction
  ⇝ our predictor is actually a snow detector

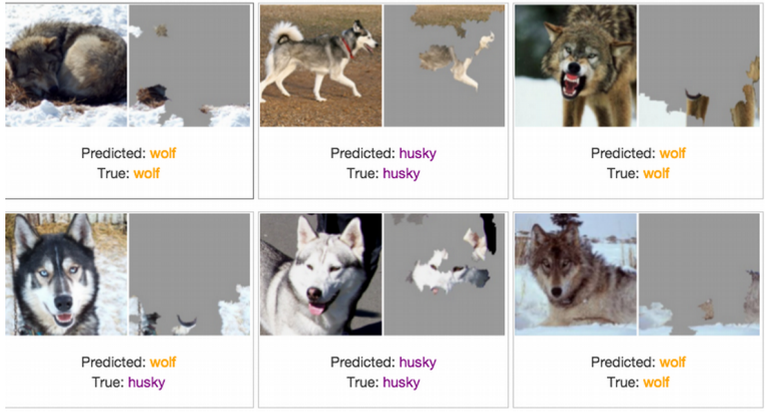


Source: [Sameer Singh 2018]

EXAMPLE: HUSKY OR WOLF? USING LIME

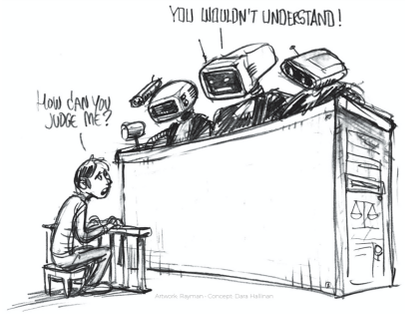


- Local explanations highlight parts of image which led to the prediction
  ⇝ our predictor is actually a snow detector

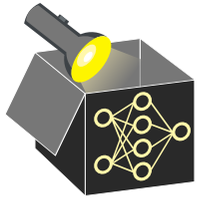


Source: [Sameer Singh 2018]

EXAMPLE: LOAN APPLICATION



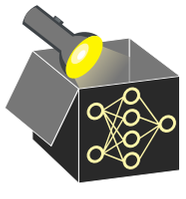
- Imagine: You apply for a loan at an online bank and are immediately rejected without reasons



Source: [https://www.elte.hu]

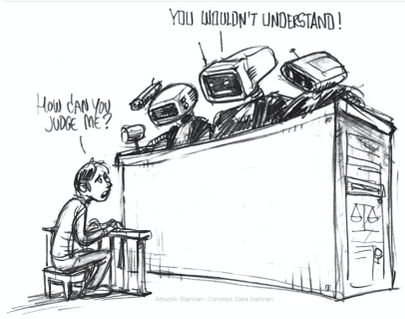
EXAMPLE: LOAN APPLICATION

- Imagine: You apply for a loan at an online bank and are immediately rejected without reasons



Source: [https://www.elte.hu]

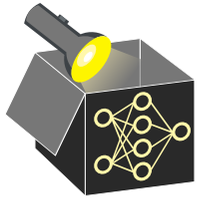
EXAMPLE: LOAN APPLICATION



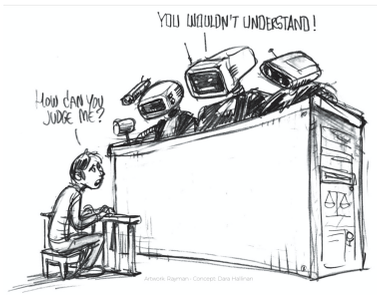
Source: [https://www.elte.hu]

- Imagine: You apply for a loan at an online bank and are immediately rejected without reasons
- Bank could e.g. provide a counterfactual explanation using local explanation methods:

  “If you were older than 21, your loan application would have been accepted."



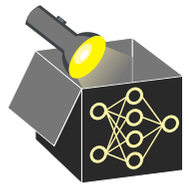
EXAMPLE: LOAN APPLICATION



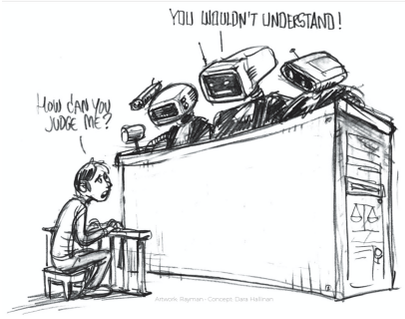
Source: [https://www.elte.hu]

- Imagine: You apply for a loan at an online bank and are immediately rejected without reasons
- Bank could e.g. provide a counterfactual explanation using local explanation methods:

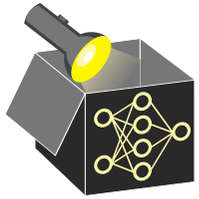
  “If you were older than 21, your loan application would have been accepted."



EXAMPLE: LOAN APPLICATION



Source: [https://www.elte.hu]



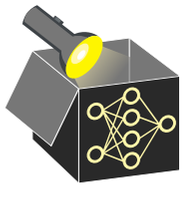
- Imagine: You apply for a loan at an online bank and are immediately rejected without reasons
- Bank could e.g. provide a counterfactual explanation using local explanation methods:

  > "If you were older than 21, your loan application would have been accepted."

$\rightsquigarrow$ helps to understand the decision and to take actions for recourse (if req.)

EXAMPLE: LOAN APPLICATION

Source: [https://www.elte.hu]



- Imagine: You apply for a loan at an online bank and are immediately rejected without reasons
- Bank could e.g. provide a counterfactual explanation using local explanation methods:

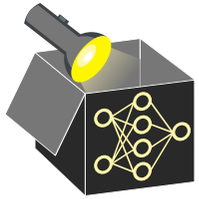
  > "If you were older than 21, your loan application would have been accepted."

$\rightsquigarrow$ helps to understand the decision and to take actions for recourse (if req.)

EXAMPLE: STOP OR RIGHT-OF-WAY?

- Imagine:
  - You work at a car company that develops image classifiers for autonomous driving
  - You show your model the following image (an adversarial example)

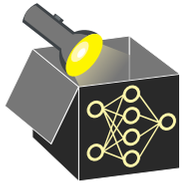
Source: [Eykholt et. al 2018]



EXAMPLE: STOP OR RIGHT-OF-WAY?

- Imagine:
  - You work at a car company that develops image classifiers for autonomous driving
  - You show your model the following image (an adversarial example)

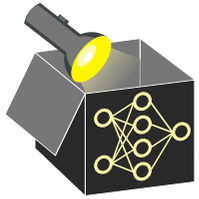
Source: [Eykholt et. al 2018]



EXAMPLE: STOP OR RIGHT-OF-WAY?

- Imagine:
  - You work at a car company that develops image classifiers for autonomous driving
  - You show your model the following image (an adversarial example)
  - Classifier is 99% sure it describes a right-of-way sign
- Would you entrust other people's lives into the hands of this software?

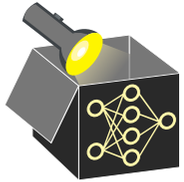
Source: [Eykholt et. al 2018]



EXAMPLE: STOP OR RIGHT-OF-WAY?

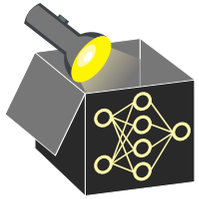
- Imagine:
  - You work at a car company that develops image classifiers for autonomous driving
  - You show your model the following image (an adversarial example)
  - Classifier is 99% sure it describes a right-of-way sign
- Would you entrust other people's lives into the hands of this software?

Source: [Eykholt et. al 2018]



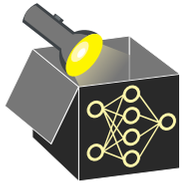
CHARACTERISTICS OF LOCAL EXPLANATIONS

- **Explanation scope:** Specific to one prediction, valid only in local environment



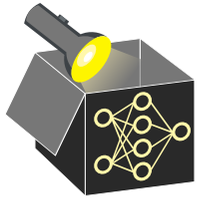
CHARACTERISTICS OF LOCAL EXPLANATIONS

- **Explanation scope:** Specific to one prediction, valid only in local environment



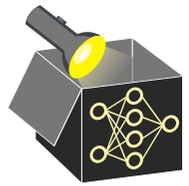
CHARACTERISTICS OF LOCAL EXPLANATIONS

- **Explanation scope:** Specific to one prediction, valid only in local environment
- **Applicable model classes:**
  - Model-agnostic (by design)
  - Model-specific variants (exploit internal structure for speed/accuracy)



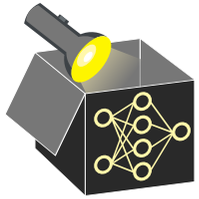
CHARACTERISTICS OF LOCAL EXPLANATIONS

- **Explanation scope:** Specific to one prediction, valid only in local environment
- **Applicable model classes:**
  - Model-agnostic (by design)
  - Model-specific variants (exploit internal structure for speed/accuracy)



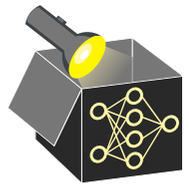
CHARACTERISTICS OF LOCAL EXPLANATIONS

- **Explanation scope:** Specific to one prediction, valid only in local environment
- **Applicable model classes:**
  - Model-agnostic (by design)
  - Model-specific variants (exploit internal structure for speed/accuracy)
- **Audience:** ML engineers, laypersons, and domain experts

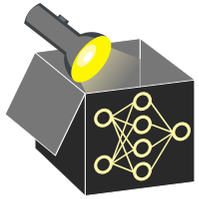


CHARACTERISTICS OF LOCAL EXPLANATIONS

- **Explanation scope:** Specific to one prediction, valid only in local environment
- **Applicable model classes:**
  - Model-agnostic (by design)
  - Model-specific variants (exploit internal structure for speed/accuracy)
- **Audience:** ML engineers, laypersons, and domain experts

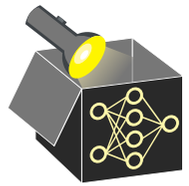


CHARACTERISTICS OF LOCAL EXPLANATIONS



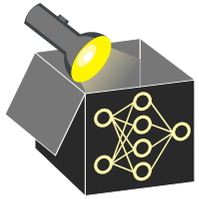
- **Explanation scope:** Specific to one prediction, valid only in local environment
- **Applicable model classes:**
  - Model-agnostic (by design)
  - Model-specific variants (exploit internal structure for speed/accuracy)
- **Audience:** ML engineers, laypersons, and domain experts
- **Supported data types:** Broad applicability across modalities (tabular, image, text, audio), but method-specific adaptations often required

CHARACTERISTICS OF LOCAL EXPLANATIONS



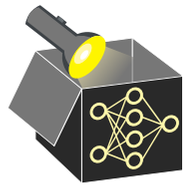
- **Explanation scope:** Specific to one prediction, valid only in local environment
- **Applicable model classes:**
  - Model-agnostic (by design)
  - Model-specific variants (exploit internal structure for speed/accuracy)
- **Audience:** ML engineers, laypersons, and domain experts
- **Supported data types:** Broad applicability across modalities (tabular, image, text, audio), but method-specific adaptations often required

CHARACTERISTICS OF LOCAL EXPLANATIONS



- **Explanation scope:** Specific to one prediction, valid only in local environment
- **Applicable model classes:**
  - Model-agnostic (by design)
  - Model-specific variants (exploit internal structure for speed/accuracy)
- **Audience:** ML engineers, laypersons, and domain experts
- **Supported data types:** Broad applicability across modalities (tabular, image, text, audio), but method-specific adaptations often required
- **Main method families**
  - Single ICE curves
  - Shapley / SHAP values
  - LIME / Anchors
  - Counterfactual explanations
  - Adversarial examples

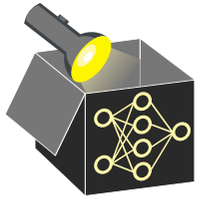
CHARACTERISTICS OF LOCAL EXPLANATIONS



- **Explanation scope:** Specific to one prediction, valid only in local environment
- **Applicable model classes:**
  - Model-agnostic (by design)
  - Model-specific variants (exploit internal structure for speed/accuracy)
- **Audience:** ML engineers, laypersons, and domain experts
- **Supported data types:** Broad applicability across modalities (tabular, image, text, audio), but method-specific adaptations often required
- **Main method families**
  - Single ICE curves
  - Shapley / SHAP values
  - LIME / Anchors
  - Counterfactual explanations
  - Adversarial examples

CREDIT DATASET

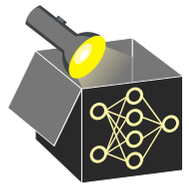
- We illustrate local explanation methods on the German credit data ▸ see Kaggle
- 522 observations, 9 features containing credit and customer information
- Binary target “risk” indicates if a customer has a ‘good’ or ‘bad’ credit risk
- We merged categories with few observations



name	type	range
age	numeric	[19, 75]
sex	factor	{male, female}
job	factor	{0, 1, 2, 3}
housing	factor	{free, own, rent}
saving.accounts	factor	{little, moderate, rich}
checking.accounts	factor	{little, moderate, rich}
credit.amount	numeric	[276, 18424]
duration	numeric	[6, 72]
purpose	numeric	{others, car, furniture, radio/TV}
risk	factor	{good, bad}

CREDIT DATASET

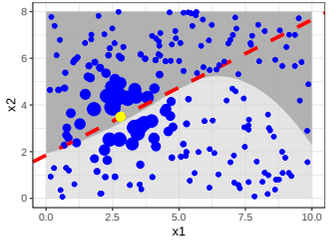
- We illustrate local explanation methods on the German credit data ▸ Kaggle n.d.
- 522 observations, 9 features containing credit and customer information
- Binary target “risk” indicates if a customer has a ‘good’ or ‘bad’ credit risk
- We merged categories with few observations



name	type	range
age	numeric	[19, 75]
sex	factor	{male, female}
job	factor	{0, 1, 2, 3}
housing	factor	{free, own, rent}
saving.accounts	factor	{little, moderate, rich}
checking.accounts	factor	{little, moderate, rich}
credit.amount	numeric	[276, 18424]
duration	numeric	[6, 72]
purpose	numeric	{others, car, furniture, radio/TV}
risk	factor	{good, bad}

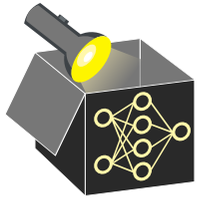
Interpretable Machine Learning

Local Interpretable Model-agnostic Explanations (LIME)



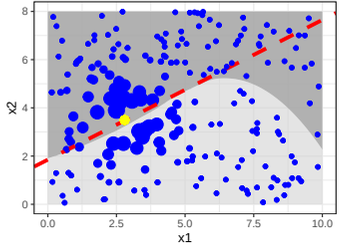
Learning goals

- Understand motivation for LIME
- Develop a mathematical intuition



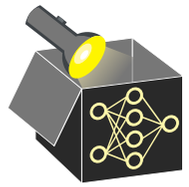
Interpretable Machine Learning

Local Explanations: LIME
Local Interpretable Model-agnostic Explanations (LIME)

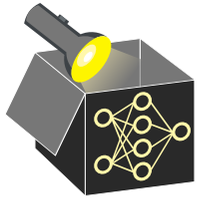


Learning goals

- Understand motivation for LIME
- Develop a mathematical intuition



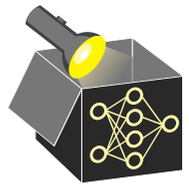
LIME



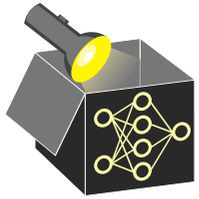
- **Locality assumption:** $\hat{f}$ behaves similarly simple in small neighborhood of $\mathbf{x}$
  $\rightsquigarrow$ Approximate $\hat{f}$ near $\mathbf{x}$ using an interpretable surrogate model $\hat{g}$
- **Interpretation strategy:** Use $\hat{g}$'s simple internal structure to explain $\hat{f}(\mathbf{x})$ locally
  $\rightsquigarrow$ **Common surrogates:** Sparse linear models, shallow decision trees
- **Applicability:** Model-agnostic; supports tabular, image, and text data
- **In practice:** Generate samples near $\mathbf{x}$, predict with $\hat{f}$, and fit $\hat{g}$ to these samples using $\hat{f}$'s outputs as targets, weighting samples by their proximity/closeness to $\mathbf{x}$

LIME

- **Locality assumption:**
  $\hat{f}$ behaves similarly simple in small neighborhood of $\mathbf{x}$
  $\rightsquigarrow$ Approximate $\hat{f}$ near $\mathbf{x}$ using an interpretable surrogate model $\hat{g}$
- **Interpretation strategy:**
  Use $\hat{g}$'s simple internal structure to explain $\hat{f}(\mathbf{x})$ locally
  $\rightsquigarrow$ **Common surrogates:** Sparse linear models, shallow decision trees
- **Applicability:** Model-agnostic; supports tabular, image, and text data
- **In practice:** Generate samples near $\mathbf{x}$, predict with $\hat{f}$, and fit $\hat{g}$ to these samples using $\hat{f}$'s outputs as targets, weighting samples by their proximity/closeness to $\mathbf{x}$



LIME: CHARACTERISTICS



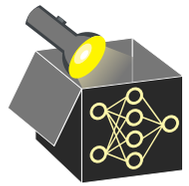
Definition: LIME provides a local explanation for a black-box model $\hat{f}$ in form of a surrogate model $\hat{g} \in \mathcal{G}$, where $\mathcal{G}$ is a class of interpretable models

Surrogate model $\hat{g}$ should satisfy two characteristics:

1. **Interpretable**: Provide human-understandable insights into the relationship between input features and prediction (e.g. via coefficients, model structure)
2. **Local fidelity** / **faithfulness**: $\hat{g}$ closely approximates $\hat{f}$ in the vicinity of the input **x** being explained

Goal: Find $\hat{g}$ with **minimal complexity and maximal local fidelity**

LIME: CHARACTERISTICS



Definition: LIME provides a local explanation for a black-box model $\hat{f}$ in form of a surrogate model $\hat{g} \in \mathcal{G}$, where $\mathcal{G}$ is a class of interpretable models

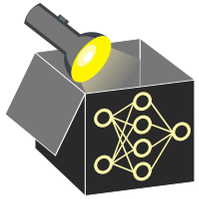
Surrogate model $\hat{g}$ should satisfy two characteristics:

1. **Interpretable**: Provide human-understandable insights into the relationship between input features and prediction (e.g. via coefficients, model structure)
2. **Local fidelity** / **faithfulness**: $\hat{g}$ closely approximates $\hat{f}$ in the vicinity of the input **x** being explained

Goal: Find $\hat{g}$ with **minimal complexity and maximal local fidelity**

MODEL COMPLEXITY

We can measure the complexity of $\hat{g} \in \mathcal{G}$ using a complexity measure $J : \mathcal{G} \to \mathbb{R}_0$



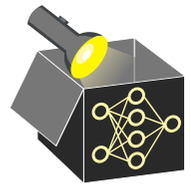
Example: (Sparse) Linear Models

- Let $\mathcal{G} = \left\{g : \mathcal{X} \to \mathbb{R} \mid g(\mathbf{x}) = s(\boldsymbol{\theta}^\top \mathbf{x})\right\}$ be the class of linear models
- $s(\cdot)$ is identity (linear model) or logistic sigmoid function (logistic regression)

$\rightsquigarrow$ $J(g) = \sum_{j=1}^{p} \mathcal{I}_{\{\theta_j \neq 0\}}$: Count number of non-zero coefficients (via L_0-norm of $\boldsymbol{\theta}$)

MODEL COMPLEXITY

We can measure complexity of $\hat{g} \in \mathcal{G}$ using a complexity measure $J : \mathcal{G} \to \mathbb{R}_0$



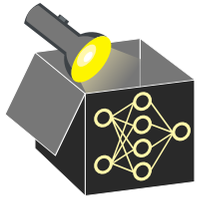
Example: (Sparse) Linear Models

- Let $\mathcal{G} = \left\{g : \mathcal{X} \to \mathbb{R} \mid g(\mathbf{x}) = s(\boldsymbol{\theta}^\top \mathbf{x})\right\}$ be the class of linear models
- $s(\cdot)$ is identity (linear model) or logistic sigmoid function (log. reg.)

$\rightsquigarrow$ $J(g) = \sum_{j=1}^{p} \mathcal{I}_{\{\theta_j \neq 0\}}$: Count number of non-zero coeffs (via L_0-norm of $\boldsymbol{\theta}$)

MODEL COMPLEXITY

We can measure the complexity of $\hat{g} \in \mathcal{G}$ using a complexity measure $J : \mathcal{G} \to \mathbb{R}_0$



Example: (Sparse) Linear Models

- Let $\mathcal{G} = \left\{ g : \mathcal{X} \to \mathbb{R} \mid g(\mathbf{x}) = s(\boldsymbol{\theta}^\top \mathbf{x}) \right\}$ be the class of linear models
- $s(\cdot)$ is identity (linear model) or logistic sigmoid function (logistic regression)

$\rightsquigarrow$ $J(g) = \sum_{j=1}^{p} \mathcal{I}_{\{\theta_j \neq 0\}}$: Count number of non-zero coefficients (via L_0-norm of $\boldsymbol{\theta}$)

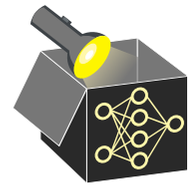
Example: Decision Trees

- Let $\mathcal{G} = \left\{ g : \mathcal{X} \to \mathbb{R} \mid g(\mathbf{x}) = \sum_{m=1}^{M} c_m \mathcal{I}_{\{\mathbf{x} \in Q_m\}} \right\}$ be the class of trees
- Q_m are disjoint axis parallel regions (leaves) and $c_m \in \mathbb{R}$ constant predictions

$\rightsquigarrow$ $J(g) = M$: Count number of terminal/leaf nodes

MODEL COMPLEXITY

We can measure complexity of $\hat{g} \in \mathcal{G}$ using a complexity measure $J : \mathcal{G} \to \mathbb{R}_0$



Example: (Sparse) Linear Models

- Let $\mathcal{G} = \left\{ g : \mathcal{X} \to \mathbb{R} \mid g(\mathbf{x}) = s(\boldsymbol{\theta}^\top \mathbf{x}) \right\}$ be the class of linear models
- $s(\cdot)$ is identity (linear model) or logistic sigmoid function (log. reg.)

$\rightsquigarrow$ $J(g) = \sum_{j=1}^{p} \mathcal{I}_{\{\theta_j \neq 0\}}$: Count number of non-zero coeffs (via L_0-norm of $\boldsymbol{\theta}$)

Example: Decision Trees

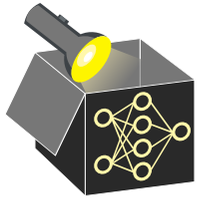
- Let $\mathcal{G} = \left\{ g : \mathcal{X} \to \mathbb{R} \mid g(\mathbf{x}) = \sum_{m=1}^{M} c_m \mathcal{I}_{\{\mathbf{x} \in Q_m\}} \right\}$ be the class of trees
- Q_m are disjoint axis parallel regions (leaves); $c_m \in \mathbb{R}$ constant predictions

$\rightsquigarrow$ $J(g) = M$: Count number of terminal/leaf nodes

LOCAL FIDELITY OF SURROGATE MODELS

- Surrogate $\hat{g}$ is **locally faithful** to a black-box model $\hat{f}$ around an input $\mathbf{x}$ if

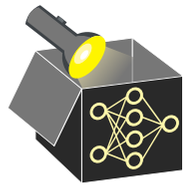
  $\hat{g}(\mathbf{z}) \approx \hat{f}(\mathbf{z})$ for synthetic samples $\mathbf{z} \in \mathcal{Z} \subseteq \mathbb{R}^p$ generated around $\mathbf{x}$



LOCAL FIDELITY OF SURROGATE MODELS

- Surrogate $\hat{g}$ is **locally faithful** to a black-box model $\hat{f}$ around an input $\mathbf{x}$ if

  $\hat{g}(\mathbf{z}) \approx \hat{f}(\mathbf{z})$ for synthetic samples $\mathbf{z} \in \mathcal{Z} \subseteq \mathbb{R}^p$ generated around $\mathbf{x}$

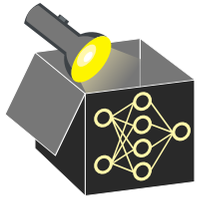


LOCAL FIDELITY OF SURROGATE MODELS

- Surrogate $\hat{g}$ is **locally faithful** to a black-box model $\hat{f}$ around an input $\mathbf{x}$ if

  $$\hat{g}(\mathbf{z}) \approx \hat{f}(\mathbf{z}) \quad \text{for synthetic samples } \mathbf{z} \in \mathcal{Z} \subseteq \mathbb{R}^p \text{ generated around } \mathbf{x}$$

- **Optimization principle:** The closer $\mathbf{z}$ is to $\mathbf{x}$, the more $\hat{g}(\mathbf{z})$ should match $\hat{f}(\mathbf{z})$

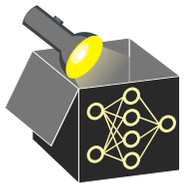


LOCAL FIDELITY OF SURROGATE MODELS

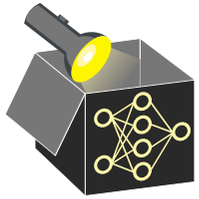
- Surrogate $\hat{g}$ is **locally faithful** to a black-box model $\hat{f}$ around an input $\mathbf{x}$ if

  $$\hat{g}(\mathbf{z}) \approx \hat{f}(\mathbf{z}) \quad \text{for synthetic samples } \mathbf{z} \in \mathcal{Z} \subseteq \mathbb{R}^p \text{ generated around } \mathbf{x}$$

- **Optimization principle:** Closer $\mathbf{z}$ is to $\mathbf{x}$, the more $\hat{g}(\mathbf{z})$ should match $\hat{f}(\mathbf{z})$



LOCAL FIDELITY OF SURROGATE MODELS



- Surrogate $\hat{g}$ is **locally faithful** to a black-box model $\hat{f}$ around an input $\mathbf{x}$ if

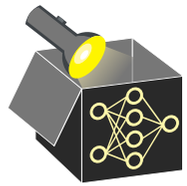
  $$\hat{g}(\mathbf{z}) \approx \hat{f}(\mathbf{z}) \quad \text{for synthetic samples } \mathbf{z} \in \mathcal{Z} \subseteq \mathbb{R}^p \text{ generated around } \mathbf{x}$$

- **Optimization principle:** The closer $\mathbf{z}$ is to $\mathbf{x}$, the more $\hat{g}(\mathbf{z})$ should match $\hat{f}(\mathbf{z})$
- To operationalize this optimization, we need:
  1. **A proximity (similarity) measure** $\phi_{\mathbf{x}}(\mathbf{z})$ between $\mathbf{z}$ and $\mathbf{x}$, e.g.:

     $\phi_{\mathbf{x}}(\mathbf{z}) = \exp(-d(\mathbf{x}, \mathbf{z})^2/\sigma^2)$ (exponential kernel), where

     - d is a distance metric (e.g., Euclidean or Gower for mixed types)
     - σ is the kernel width that controls locality

LOCAL FIDELITY OF SURROGATE MODELS



- Surrogate $\hat{g}$ is **locally faithful** to a black-box model $\hat{f}$ around an input $\mathbf{x}$ if

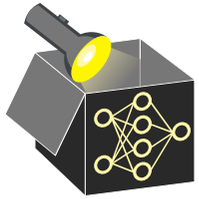
  $$\hat{g}(\mathbf{z}) \approx \hat{f}(\mathbf{z}) \quad \text{for synthetic samples } \mathbf{z} \in \mathcal{Z} \subseteq \mathbb{R}^p \text{ generated around } \mathbf{x}$$

- **Optimization principle:** Closer $\mathbf{z}$ is to $\mathbf{x}$, the more $\hat{g}(\mathbf{z})$ should match $\hat{f}(\mathbf{z})$
- To operationalize this optimization, we need:
  1. **A proximity (similarity) measure** $\phi_{\mathbf{x}}(\mathbf{z})$ between $\mathbf{z}$ and $\mathbf{x}$, e.g.:

     $\phi_{\mathbf{x}}(\mathbf{z}) = \exp(-d(\mathbf{x}, \mathbf{z})^2/\sigma^2)$ (exponential kernel), where

     - d: distance metric (e.g., Euclidean or Gower for mixed types)
     - σ is the kernel width that controls locality

LOCAL FIDELITY OF SURROGATE MODELS



- Surrogate $\hat{g}$ is **locally faithful** to a black-box model $\hat{f}$ around an input $\mathbf{x}$ if

$$\hat{g}(\mathbf{z}) \approx \hat{f}(\mathbf{z}) \quad \text{for synthetic samples } \mathbf{z} \in \mathcal{Z} \subseteq \mathbb{R}^p \text{ generated around } \mathbf{x}$$

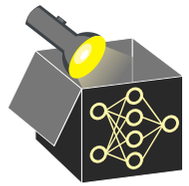
- **Optimization principle:** The closer $\mathbf{z}$ is to $\mathbf{x}$, the more $\hat{g}(\mathbf{z})$ should match $\hat{f}(\mathbf{z})$
- To operationalize this optimization, we need:
  1. **A proximity (similarity) measure** $\phi_{\mathbf{x}}(\mathbf{z})$ between $\mathbf{z}$ and $\mathbf{x}$, e.g.:

     $\phi_{\mathbf{x}}(\mathbf{z}) = \exp(-d(\mathbf{x}, \mathbf{z})^2/\sigma^2)$ (exponential kernel), where

     - d is a distance metric (e.g., Euclidean or Gower for mixed types)
     - σ is the kernel width that controls locality
  2. **A loss function** $L(\hat{f}(\mathbf{z}), \hat{g}(\mathbf{z}))$, e.g. the L_2 loss/squared error:

$$L(\hat{f}(\mathbf{z}), \hat{g}(\mathbf{z})) = \left(\hat{g}(\mathbf{z}) - \hat{f}(\mathbf{z})\right)^2$$

LOCAL FIDELITY OF SURROGATE MODELS



- Surrogate $\hat{g}$ is **locally faithful** to a black-box model $\hat{f}$ around an input $\mathbf{x}$ if

$$\hat{g}(\mathbf{z}) \approx \hat{f}(\mathbf{z}) \quad \text{for synthetic samples } \mathbf{z} \in \mathcal{Z} \subseteq \mathbb{R}^p \text{ generated around } \mathbf{x}$$

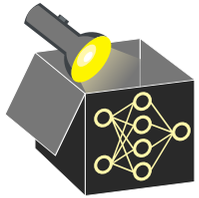
- **Optimization principle:** Closer $\mathbf{z}$ is to $\mathbf{x}$, the more $\hat{g}(\mathbf{z})$ should match $\hat{f}(\mathbf{z})$
- To operationalize this optimization, we need:
  1. **A proximity (similarity) measure** $\phi_{\mathbf{x}}(\mathbf{z})$ between $\mathbf{z}$ and $\mathbf{x}$, e.g.:

     $\phi_{\mathbf{x}}(\mathbf{z}) = \exp(-d(\mathbf{x}, \mathbf{z})^2/\sigma^2)$ (exponential kernel), where

     - d: distance metric (e.g., Euclidean or Gower for mixed types)
     - σ is the kernel width that controls locality
  2. **A loss function** $L(\hat{f}(\mathbf{z}), \hat{g}(\mathbf{z}))$, e.g. the L_2 loss/squared error:

$$L(\hat{f}(\mathbf{z}), \hat{g}(\mathbf{z})) = \left(\hat{g}(\mathbf{z}) - \hat{f}(\mathbf{z})\right)^2$$

LOCAL FIDELITY OF SURROGATE MODELS



- Surrogate $\hat{g}$ is **locally faithful** to a black-box model $\hat{f}$ around an input $\mathbf{x}$ if

$$\hat{g}(\mathbf{z}) \approx \hat{f}(\mathbf{z}) \quad \text{for synthetic samples } \mathbf{z} \in \mathcal{Z} \subseteq \mathbb{R}^p \text{ generated around } \mathbf{x}$$

- **Optimization principle:** The closer $\mathbf{z}$ is to $\mathbf{x}$, the more $\hat{g}(\mathbf{z})$ should match $\hat{f}(\mathbf{z})$
- To operationalize this optimization, we need:
  1. **A proximity (similarity) measure** $\phi_{\mathbf{x}}(\mathbf{z})$ between $\mathbf{z}$ and $\mathbf{x}$, e.g.:

     $\phi_{\mathbf{x}}(\mathbf{z}) = \exp(-d(\mathbf{x}, \mathbf{z})^2/\sigma^2)$ (exponential kernel), where

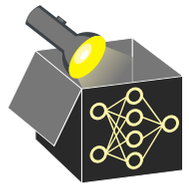
     - d is a distance metric (e.g., Euclidean or Gower for mixed types)
     - σ is the kernel width that controls locality
  2. **A loss function** $L(\hat{f}(\mathbf{z}), \hat{g}(\mathbf{z}))$, e.g. the L_2 loss/squared error:

$$L(\hat{f}(\mathbf{z}), \hat{g}(\mathbf{z})) = \left(\hat{g}(\mathbf{z}) - \hat{f}(\mathbf{z})\right)^2$$

- The overall **local fidelity objective** is measured by a weighted loss:

$$L(\hat{f}, \hat{g}, \phi_{\mathbf{x}}) = \sum_{\mathbf{z} \in \mathcal{Z}} \phi_{\mathbf{x}}(\mathbf{z}) \cdot L(\hat{f}(\mathbf{z}), \hat{g}(\mathbf{z}))$$

LOCAL FIDELITY OF SURROGATE MODELS



- Surrogate $\hat{g}$ is **locally faithful** to a black-box model $\hat{f}$ around an input $\mathbf{x}$ if

$$\hat{g}(\mathbf{z}) \approx \hat{f}(\mathbf{z}) \quad \text{for synthetic samples } \mathbf{z} \in \mathcal{Z} \subseteq \mathbb{R}^p \text{ generated around } \mathbf{x}$$

- **Optimization principle:** Closer $\mathbf{z}$ is to $\mathbf{x}$, the more $\hat{g}(\mathbf{z})$ should match $\hat{f}(\mathbf{z})$
- To operationalize this optimization, we need:
  1. **A proximity (similarity) measure** $\phi_{\mathbf{x}}(\mathbf{z})$ between $\mathbf{z}$ and $\mathbf{x}$, e.g.:

     $\phi_{\mathbf{x}}(\mathbf{z}) = \exp(-d(\mathbf{x}, \mathbf{z})^2/\sigma^2)$ (exponential kernel), where

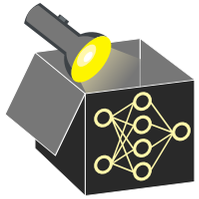
     - d: distance metric (e.g., Euclidean or Gower for mixed types)
     - σ is the kernel width that controls locality
  2. **A loss function** $L(\hat{f}(\mathbf{z}), \hat{g}(\mathbf{z}))$, e.g. the L_2 loss/squared error:

$$L(\hat{f}(\mathbf{z}), \hat{g}(\mathbf{z})) = \left(\hat{g}(\mathbf{z}) - \hat{f}(\mathbf{z})\right)^2$$

- The overall **local fidelity objective** is measured by a weighted loss:

$$L(\hat{f}, \hat{g}, \phi_{\mathbf{x}}) = \sum_{\mathbf{z} \in \mathcal{Z}} \phi_{\mathbf{x}}(\mathbf{z}) \cdot L(\hat{f}(\mathbf{z}), \hat{g}(\mathbf{z}))$$

LIME OPTIMIZATION TASK

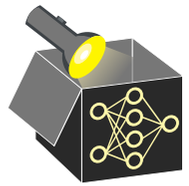


- Optimization problem of LIME:

$$\arg\min_{\hat{g} \in \mathcal{G}} L(\hat{f}, \hat{g}, \phi_{\mathbf{x}}) + J(\hat{g})$$

- **In practice** LIME uses a two-stage approach:
  1. User specifies complexity $J(\hat{g})$ beforehand (e.g., LASSO with k features)
  2. Optimize $L(\hat{f}, \hat{g}, \phi_{\mathbf{x}})$ (model fidelity) for fixed complexity
- **Goal:** Build a **model-agnostic** explainer
  - $\rightsquigarrow$ Optimize $L(\hat{f}, \hat{g}, \phi_{\mathbf{x}})$ without making any assumptions on the form of $\hat{f}$
  - $\rightsquigarrow$ Surrogate $\hat{g}$ approximates $\hat{f}$ locally through sampling and fitting

LIME OPTIMIZATION TASK



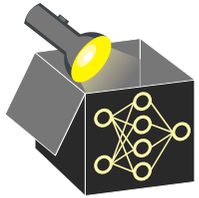
- Optimization problem of LIME:

$$\arg\min_{\hat{g} \in \mathcal{G}} L(\hat{f}, \hat{g}, \phi_{\mathbf{x}}) + J(\hat{g})$$

- **In practice** LIME uses a two-stage approach:
  1. User sets complexity $J(\hat{g})$ beforehand (e.g., LASSO with k features)
  2. Optimize $L(\hat{f}, \hat{g}, \phi_{\mathbf{x}})$ (model fidelity) for fixed complexity
- **Goal:** Build a **model-agnostic** explainer
  - $\rightsquigarrow$ Optimize $L(\hat{f}, \hat{g}, \phi_{\mathbf{x}})$ without making assumptions on the form of $\hat{f}$
  - $\rightsquigarrow$ Surrogate $\hat{g}$ approximates $\hat{f}$ locally through sampling and fitting

LIME ALGORITHM: OUTLINE

Ribeiro. 2016

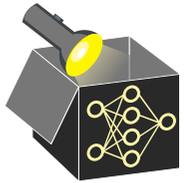


Input:

- Pre-trained black-box model $\hat{f}$
- Observation $\mathbf{x}$ whose prediction $\hat{f}(\mathbf{x})$ we want to explain
- Interpretable model class $\mathcal{G}$ for local surrogate (to limit complexity)

LIME ALGORITHM: OUTLINE

RIBEIRO_2016

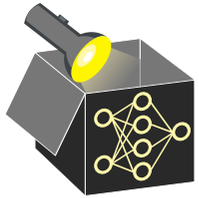


Input:

- Pre-trained black-box model $\hat{f}$
- Observation $\mathbf{x}$ whose prediction $\hat{f}(\mathbf{x})$ we want to explain
- Interpretable model class $\mathcal{G}$ for local surrogate (to limit complexity)

LIME ALGORITHM: OUTLINE

▸ Ribeiro. 2016



Input:

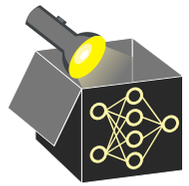
- Pre-trained black-box model $\hat{f}$
- Observation $\mathbf{x}$ whose prediction $\hat{f}(\mathbf{x})$ we want to explain
- Interpretable model class $\mathcal{G}$ for local surrogate (to limit complexity)

Algorithm:

1. Independently sample new points $\mathbf{z} \in \mathcal{Z}$
2. Retrieve predictions $\hat{f}(\mathbf{z})$ for obtained points $\mathbf{z}$
3. Weight $\mathbf{z} \in \mathcal{Z}$ by their proximity $\phi_{\mathbf{x}}(\mathbf{z})$ to quantify closeness to $\mathbf{x}$
4. Train interpretable surrogate model $\hat{g}$ on data points $\mathbf{z} \in \mathcal{Z}$ using weights $\phi_{\mathbf{x}}(\mathbf{z})$
   $\rightsquigarrow$ Predictions $\hat{f}(\mathbf{z})$ are used as target of this model
5. Return $\hat{g}$ as the local explanation for $\hat{f}(\mathbf{x})$

LIME ALGORITHM: OUTLINE

▸ RIBEIRO_2016



Input:

- Pre-trained black-box model $\hat{f}$
- Observation $\mathbf{x}$ whose prediction $\hat{f}(\mathbf{x})$ we want to explain
- Interpretable model class $\mathcal{G}$ for local surrogate (to limit complexity)

Algorithm:

1. Independently sample new points $\mathbf{z} \in \mathcal{Z}$
2. Retrieve predictions $\hat{f}(\mathbf{z})$ for obtained points $\mathbf{z}$
3. Weight $\mathbf{z} \in \mathcal{Z}$ by their proximity $\phi_{\mathbf{x}}(\mathbf{z})$ to quantify closeness to $\mathbf{x}$
4. Train interpretable surrogate model $\hat{g}$ on data $\mathbf{z} \in \mathcal{Z}$ using weights $\phi_{\mathbf{x}}(\mathbf{z})$
   $\rightsquigarrow$ Predictions $\hat{f}(\mathbf{z})$ are used as target of this model
5. Return $\hat{g}$ as the local explanation for $\hat{f}(\mathbf{x})$

LIME ALGORITHM: EXAMPLE

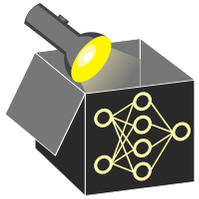
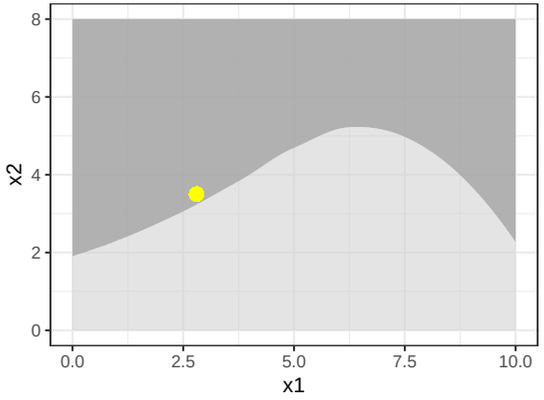


Illustration of LIME based on a classification task:

- Light/dark gray background: prediction surface of a classifier
- Yellow point: **x** to be explained
- $\mathcal{G}$: class of logistic regression models



LIME ALGORITHM: EXAMPLE

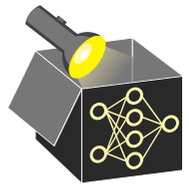
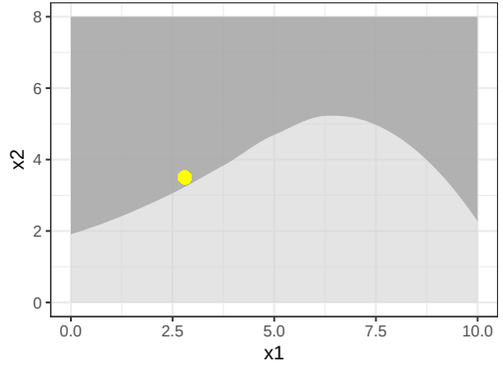
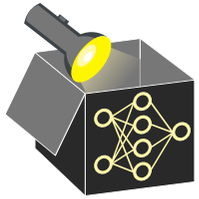


Illustration of LIME based on a classification task:

- Light/dark gray background: prediction surface of a classifier
- Yellow point: **x** to be explained
- $\mathcal{G}$: class of logistic regression models



LIME ALGORITHM: EXAMPLE (STEP 1+2: SAMPLING)



Strategies for sampling:

- Uniformly sample new points from the feasible feature range
- Use the training data set with or without perturbations
- Draw samples from the estimated univariate distribution of each feature
- Create an equidistant grid over the supported feature range

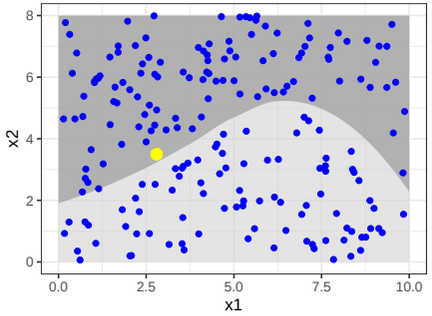


Figure: Uniformly sampled

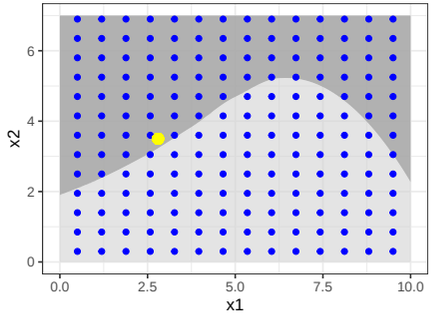
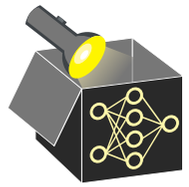


Figure: Equidistant grid

LIME ALGO.: EXAMPLE (STEP 1+2: SAMPLING)



Strategies for sampling:

- Uniformly sample new points from the feasible feature range
- Use the training data set with or without perturbations
- Draw samples from the estimated univariate distribution of each feature
- Create an equidistant grid over the supported feature range

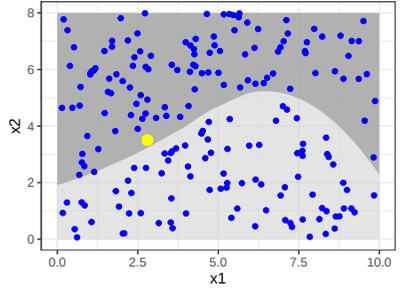


Figure: Uniformly sampled

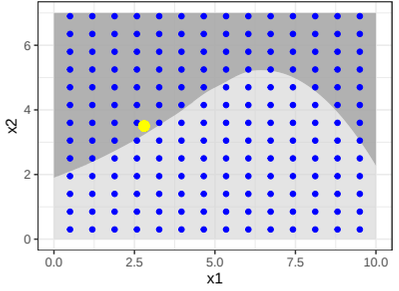
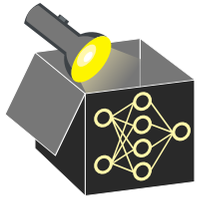


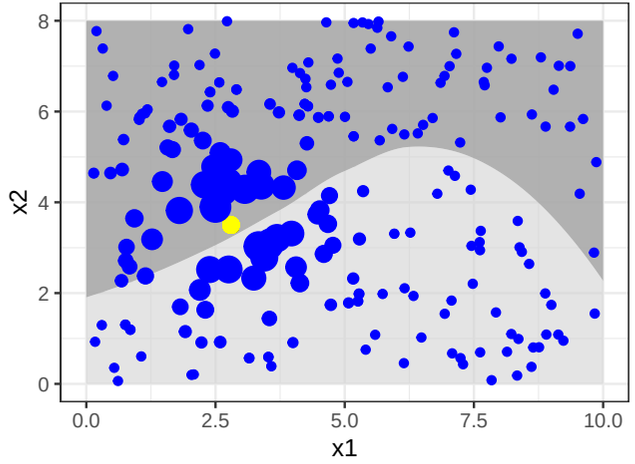
Figure: Equidistant grid

LIME ALGORITHM: EXAMPLE (STEP 3: PROXIMITY)



In this example, we use the exponential kernel defined on the Euclidean distance d

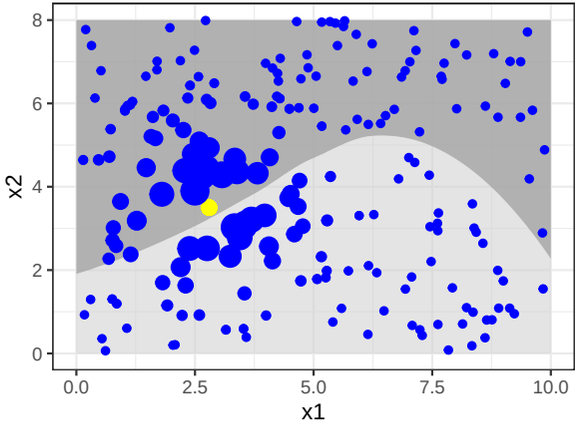
$$\phi_{\mathbf{x}}(\mathbf{z}) = exp(-d(\mathbf{x}, \mathbf{z})^2/\sigma^2).$$



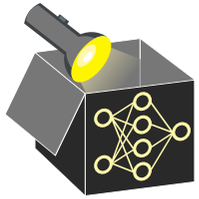
LIME ALGO.: EXAMPLE (STEP 3: PROXIMITY)

In this example, we use the exponential kernel defined on the Euclidean distance d

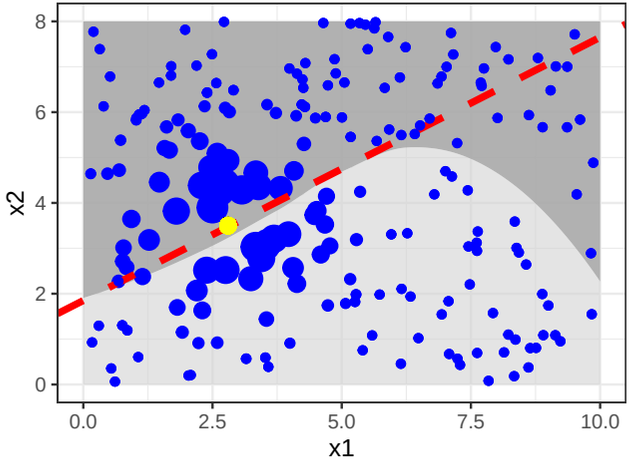
$$\phi_{\mathbf{x}}(\mathbf{z}) = exp(-d(\mathbf{x}, \mathbf{z})^2/\sigma^2).$$



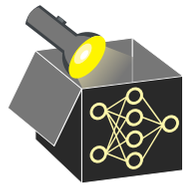
LIME ALGORITHM: EXAMPLE (STEP 4: SURROGATE)



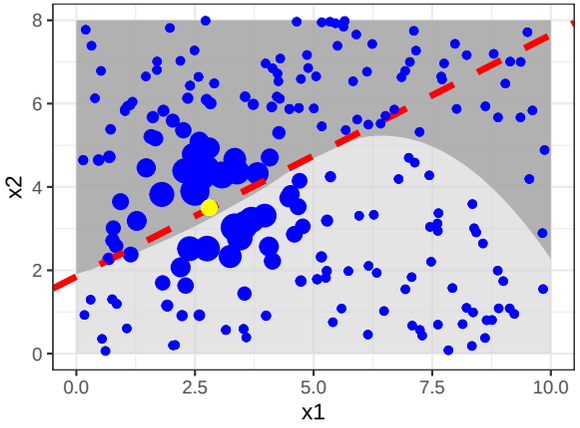
In this example, we fit a **logistic regression** model
$\rightsquigarrow L(\hat{f}(\mathbf{z}), \hat{g}(\mathbf{z}))$ is the Bernoulli loss



LIME ALGO.: EXAMPLE (STEP 4: SURROGATE)

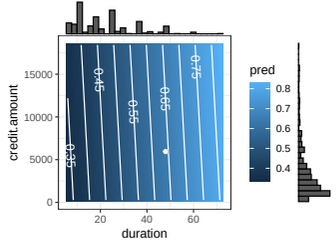


In this example, we fit a **logistic regression** model
$\rightsquigarrow L(\hat{f}(\mathbf{z}), \hat{g}(\mathbf{z}))$ is the Bernoulli loss



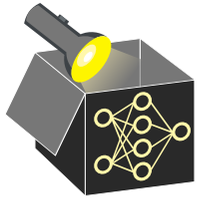
Interpretable Machine Learning

LIME Examples



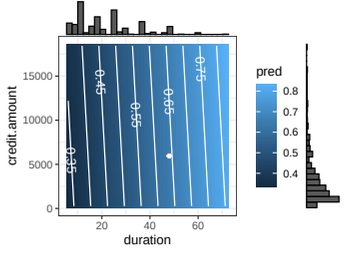
Learning goals

- See real-world data examples
- See application to image and text data



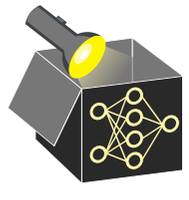
Interpretable Machine Learning

Local Explanations: LIME
LIME Examples

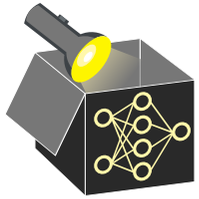


Learning goals

- See real-world data examples
- See application to image and text data



LIME EXAMPLE: CREDIT SCORING (TABULAR DATA)

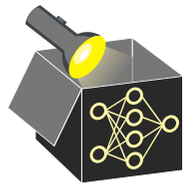


- **Black-box model $\hat{f}_{bad}$:** SVM with RBF kernel (predicts prob. of bad credit risk)
- **Instance to explain x:** First row in the dataset, with $\hat{f}_{bad}(\mathbf{x}) = 0.658$

duration	sex	credit.amount	purpose	housing	age	saving	checking	...
48	female	5951	radio/TV	own	22	little	moderate	...

- **Surrogate model:** LASSO, restricted to 5 non-zero features (via regularization)
- **Training data for surrogate:** Samples **z**, weighted by Gower distance to **x**

EXAMPLE: CREDIT SCORING (TABULAR DATA)

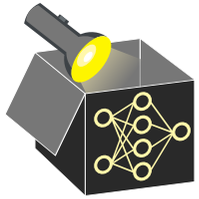


- **Black-box model $\hat{f}_{bad}$:** SVM with RBF kernel (predicts probability of bad credit risk)
- **Instance to explain x:** First row in the dataset, with $\hat{f}_{bad}(\mathbf{x}) = 0.658$

duration	sex	credit.amount	purpose	housing	age	saving	checking	...
48	female	5951	radio/TV	own	22	little	moderate	...

- **Surrogate model:** LASSO, restricted to 5 non-0 feats (via regularization)
- **Training data for surrogate:** Samples **z**, weighted by Gower dist. to **x**

LIME EXAMPLE: CREDIT SCORING (TABULAR DATA)



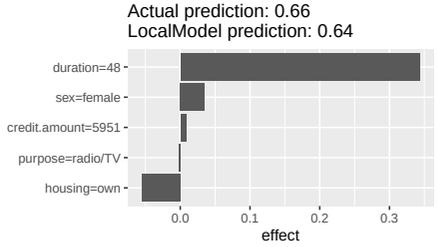
- **Black-box model $\hat{f}_{bad}$:** SVM with RBF kernel (predicts prob. of bad credit risk)
- **Instance to explain x:** First row in the dataset, with $\hat{f}_{bad}(\mathbf{x}) = 0.658$

duration	sex	credit.amount	purpose	housing	age	saving	checking	...
48	female	5951	radio/TV	own	22	little	moderate	...

- **Surrogate model:** LASSO, restricted to 5 non-zero features (via regularization)
- **Training data for surrogate:** Samples **z**, weighted by Gower distance to **x**
- **Prediction:**
  $\hat{g}(\mathbf{x}) = 0.640$ vs. $\hat{f}_{bad}(\mathbf{x}) = 0.658$

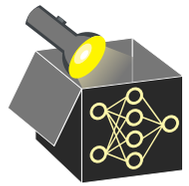
⇝ $\hat{g}$ provides good local approximation of $\hat{f}_{bad}$, but omits several features

⇝ Small mismatch reflects trade-off: **interpretability vs. fidelity**



Interpretation: Prediction is mainly driven by loan duration, with small positive effect from sex and credit.amount, and negative contributions from housing and purpose.

EXAMPLE: CREDIT SCORING (TABULAR DATA)



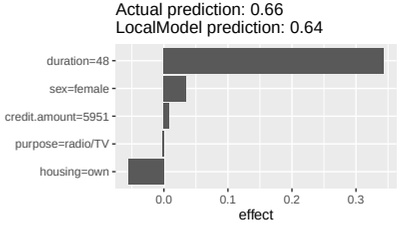
- **Black-box model $\hat{f}_{bad}$:** SVM with RBF kernel (predicts probability of bad credit risk)
- **Instance to explain x:** First row in the dataset, with $\hat{f}_{bad}(\mathbf{x}) = 0.658$

duration	sex	credit.amount	purpose	housing	age	saving	checking	...
48	female	5951	radio/TV	own	22	little	moderate	...

- **Surrogate model:** LASSO, restricted to 5 non-0 feats (via regularization)
- **Training data for surrogate:** Samples **z**, weighted by Gower dist. to **x**
- **Prediction:**
  $\hat{g}(\mathbf{x}) = 0.640$ vs. $\hat{f}_{bad}(\mathbf{x}) = 0.658$

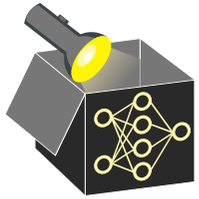
⇝ $\hat{g}$ provides good local approx. of $\hat{f}_{bad}$, but omits several features

⇝ Small mismatch reflects trade-off: **interpretability vs. fidelity**

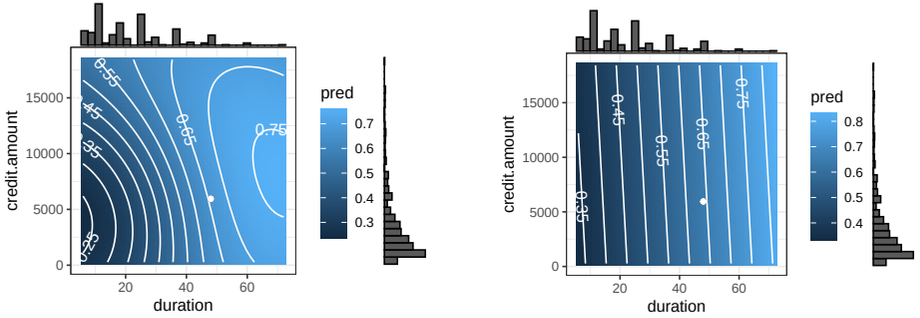


Interpretation: Prediction is mainly driven by loan duration, with small positive effect from sex and credit.amount, and negative contributions from housing and purpose.

EXAMPLE ON CREDIT DATASET (CONT'D)

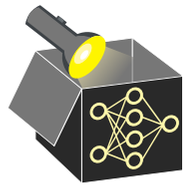


- 2D ICE plots (prediction surface plots) for `duration` and `credit.amount`
- Illustration how $\hat{g}$ linearly approximates the nonlinear decision surface of $\hat{f}_{bad}$

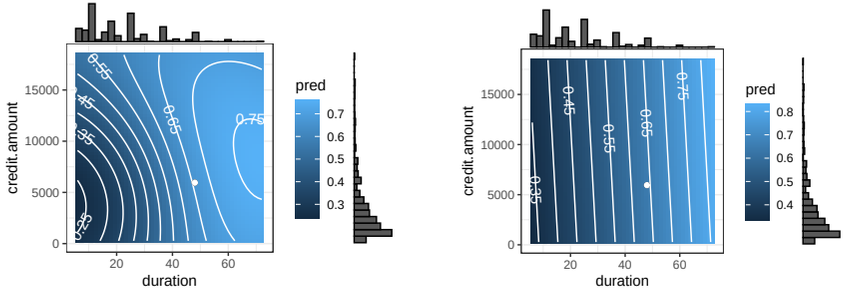


- **Left:** 2D ICE plot of $\hat{f}_{bad}$ showing decision surface
- **Right:** Linear approximation by surrogate model $\hat{g}$.
  ⇝ White dot indicates input **x** to be explained
  ⇝ Histograms show marginal distribution of features in training data

EXAMPLE ON CREDIT DATASET (CONT'D)



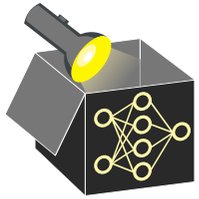
- 2D ICE plots (pred. surface plots) for `duration` and `credit.amount`
- Illustration how $\hat{g}$ linearly approximates nonlinear decision surface of $\hat{f}_{bad}$



- **Left:** 2D ICE plot of $\hat{f}_{bad}$ showing decision surface
- **Right:** Linear approximation by surrogate model $\hat{g}$.
  ⇝ White dot indicates input **x** to be explained
  ⇝ Histograms show marginal distribution of features in training data

LIME FOR TEXT DATA

Shen, Ian, (2019)



LIME can also be applied to text data:

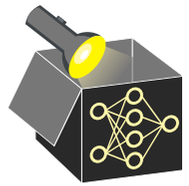
- Raw text representations:
  - Binary vector indicating the presence or absence of a word
  - A vector of word counts
- Examples for *"This text is the first text."* and *"Finally, this is the last one."*:

this	text	is	the	first	finally	last	one
1	2	1	1	1	0	0	0
1	0	1	1	0	1	1	1

- **Sampling**: Randomly set the entry of individual words to 0; equal to removing all occurrences of this word in the text.
- **Proximity**: Exponential kernel with cosine distance.
  - Neglects words that do not occur in both texts
  - Measures the distance irrespective of the text size

LIME FOR TEXT DATA

IAN_2019



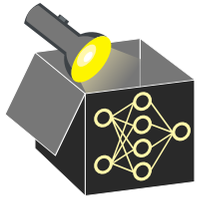
LIME can also be applied to text data:

- Raw text representations:
  - Binary vector indicating the presence or absence of a word
  - A vector of word counts
- Examples for *"This text is the first text."* and *"Finally, this is the last one."*:

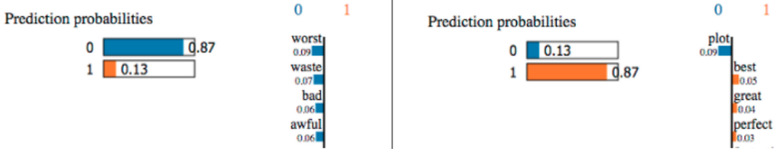
this	text	is	the	first	finally	last	one
1	2	1	1	1	0	0	0
1	0	1	1	0	1	1	1

- **Sampling**: Randomly set the entry of individual words to 0; equal to removing all occurrences of this word in the text.
- **Proximity**: Exponential kernel with cosine distance.
  - Neglects words that do not occur in both texts
  - Measures the distance irrespective of the text size

LIME FOR TEXT DATA (CONT'D) ▸ Shen, Ian, (2019)

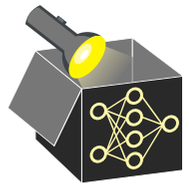


- Random forest classifier labeling movie reviews from IMDB
  - 0: negative
  - 1: positive
- Surrogate model is a sparse linear model



Words like "worst" or "waste" indicate negative review while words like "best" or "great" indicate positive review

LIME FOR TEXT DATA (CONT'D) ▸ IAN_2019



- Random forest classifier labeling movie reviews from IMDB
  - 0: negative
  - 1: positive
- Surrogate model is a sparse linear model

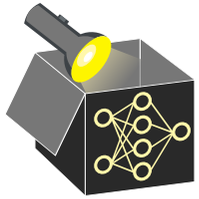
Words like "worst" or "waste" indicate negative review while words like "best" or "great" indicate positive review

LIME FOR IMAGE DATA

LIME also works for image data:

- **Idea**: Each obs. is represented by a binary vector indicating the presence or absence of superpixels
  ▸ Achanta et al. 2012
- Superpixels are interconnected pixels with similar colors (absence of a single pixel might not have a (strong) effect on the prediction)
- **Warning**: Size of superpixels needs to be determined before the segmentation takes place
- **Sampling**: Randomly switching some of the super pixels “off", i.e., by coloring some superpixels uniformly

Example for superpixels of different sizes

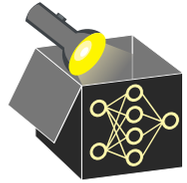


LIME FOR IMAGE DATA

LIME also works for image data:

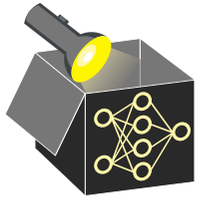
- **Idea**: Each obs. is represented by a binary vector indicating the presence or absence of superpixels
  ▸ Achanta 2012
- Superpixels are interconnected pixels with similar colors (absence of a single pixel might not have a (strong) effect on the prediction)
- **Warning**: Size of superpixels needs to be determined before the segmentation takes place
- **Sampling**: Randomly switching some of the super pixels “off", i.e., by coloring some superpixels uniformly

Example for superpixels of different sizes



Interpretable Machine Learning – 5 / 6

LIME FOR IMAGE DATA (CONT'D) ▸ Ribeiro. 2016

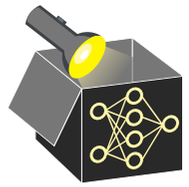


- Explaining prediction of pre-trained inception neural network classifier
- **Sampling**: Graying out all superpixels besides 10 superpixels
- **Surrogate**: Locally weighted sparse linear models
- **Proximity**: Exponential kernel with euclidean distance

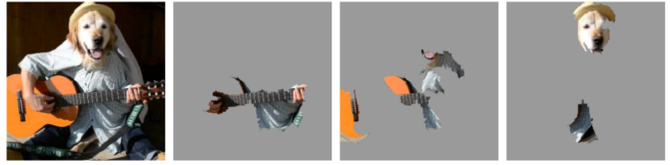
(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Top 3 classes predicted

LIME FOR IMAGE DATA (CONT'D) ▸ RIBEIRO_2016



- Explaining prediction of pre-trained inception neural network classifier
- **Sampling**: Graying out all superpixels besides 10 superpixels
- **Surrogate**: Locally weighted sparse linear models
- **Proximity**: Exponential kernel with euclidean distance



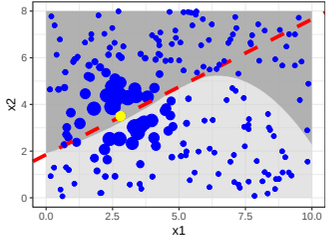
(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Top 3 classes predicted

Interpretable Machine Learning – 6 / 6

Interpretable Machine Learning

LIME Pitfalls

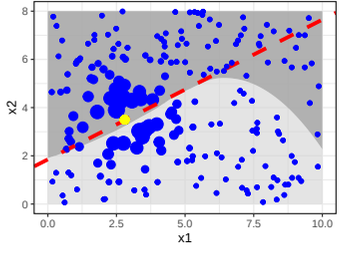


Learning goals

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

Interpretable Machine Learning

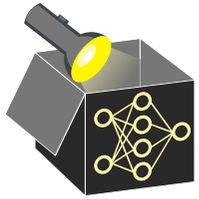
Local Explanations: LIME
LIME Pitfalls



Learning goals

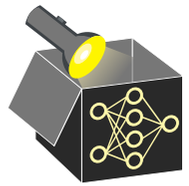
- Learn why LIME should be used with caution
- Possible pitfalls of LIME

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 distance metrics affect sensitivity
  - **Local vs. global features** – 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

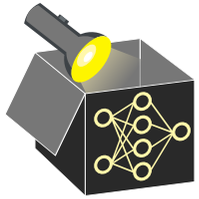
LIME PITFALLS



- 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

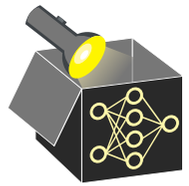
PITFALL: SAMPLING

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

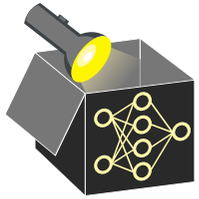


PITFALL: SAMPLING

- **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

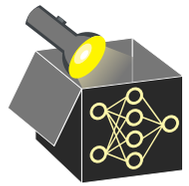


PITFALL: SAMPLING



- **Pitfall**: Common sampling strategies for $\mathbf{z} \in \mathcal{Z}$ ignore feature 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 $\mathbf{x}$

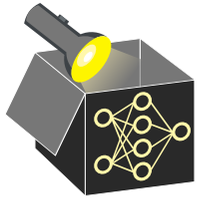
PITFALL: SAMPLING



- **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 $\mathbf{x}$

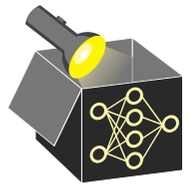
LIME PITFALL: LOCALITY

- **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

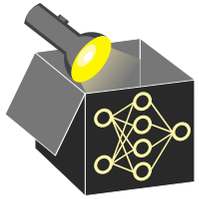


LIME PITFALL: LOCALITY

- **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 $\mathbf{x}$ 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

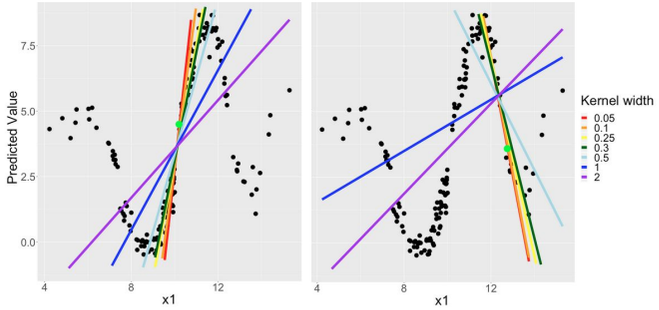


LIME PITFALL: LOCALITY



- **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

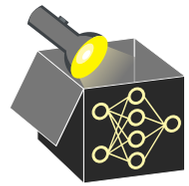
Example: For 2 obs. (green points), fit local surrogate models (lines) using only x_1



Line colors: different kernel widths used for proximity weighting

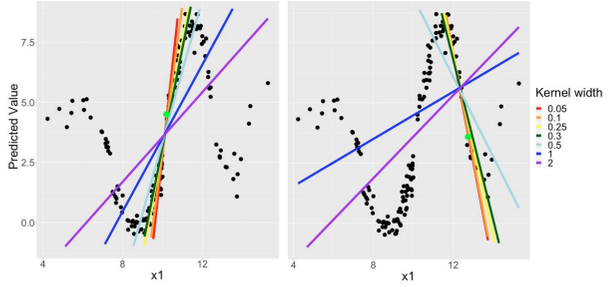
Right: larger kernel widths affect lines more

LIME PITFALL: LOCALITY



- **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

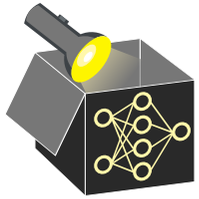


Line colors: different kernel widths used for proximity weighting

Right: larger kernel widths affect lines more

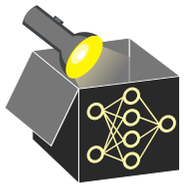
LIME PITFALL: LOCALITY ▸ Kopper et al. 2019

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



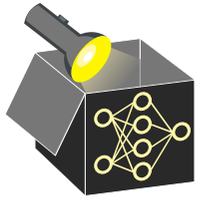
LIME PITFALL: LOCALITY ▸ KOPPER_2019

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



LIME PITFALL: LOCALITY

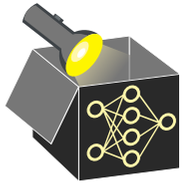
Kopper et al. 2019



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

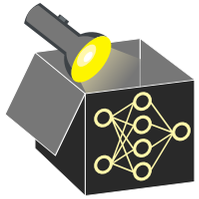
LIME PITFALL: LOCALITY

KOPPER_2019



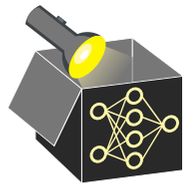
- **Pitfall:** Choice of kernel width (σ) critically influences locality
- **Implication of edge cases:**
  - *Large* $\sigma \rightarrow$ overemphasize distant points, hurting locality
  - *Small* $\sigma \rightarrow$ too few points may lead to unstable or noisy explanations

LIME PITFALL: LOCALITY ▸ Kopper et al. 2019



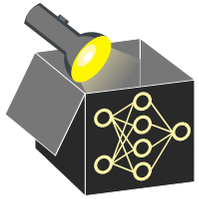
- **Pitfall:** Choice of kernel width (σ) critically influences locality
- **Implication of edge cases:**
  - *Large* $\sigma \rightarrow$ overemphasize distant points, hurting locality
  - *Small* $\sigma \rightarrow$ risk of too few points, leading to unstable or noisy explanations
- **Solution I:**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

LIME PITFALL: LOCALITY ▸ KOPPER_2019



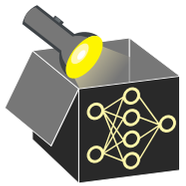
- **Pitfall:** Choice of kernel width (σ) critically influences locality
- **Implication of edge cases:**
  - *Large* $\sigma \rightarrow$ overemphasize distant points, hurting locality
  - *Small* $\sigma \rightarrow$ too few points may lead to unstable or noisy explanations
- **Solution I:** Use Gower 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$_{p}ac$ n.d.

LIME PITFALL: LOCALITY ▸ Kopper et al. 2019



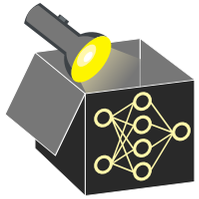
- **Pitfall:** Choice of kernel width (σ) critically influences locality
- **Implication of edge cases:**
  - *Large* $\sigma \rightarrow$ overemphasize distant points, hurting locality
  - *Small* $\sigma \rightarrow$ risk of too few points, leading to unstable or noisy explanations
- **Solution I:**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

LIME PITFALL: LOCALITY ▸ KOPPER_2019



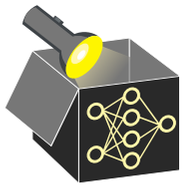
- **Pitfall:** Choice of kernel width (σ) critically influences locality
- **Implication of edge cases:**
  - *Large* $\sigma \rightarrow$ overemphasize distant points, hurting locality
  - *Small* $\sigma \rightarrow$ too few points may lead to unstable or noisy explanations
- **Solution I:** Use Gower 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$_{pac}$ n.d.
- **Solution II:** s-LIME adaptively selects σ to balance fidelity and stability
  ▸ Gaudel 2022

PITFALL: LOCAL VS. GLOBAL FEATURES ▸ Laugel et al. 2018



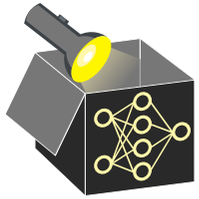
- **Pitfall:** Sampling from entire input space may hide influence of locally relevant features in favor of globally relevant ones, even for narrow kernels.
- **Feature types:**
  - *Global features* influence predictions broadly across whole imput space $\mathcal{X}$
  - *Local features* affect predictions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

PITFALL: LOCAL VS. GLOBAL FEATURES ▸ LAUGEL_2018



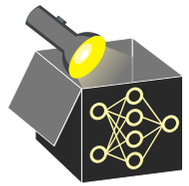
- **Pitfall:** Sampling from entire input space may hide influence of locally relevant feat in favor of globally relevant ones, even for narrow kernels.
- **Feature types:**
  - *Global features* influence predictions broadly across whole imput space $\mathcal{X}$
  - *Local features* affect predictions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

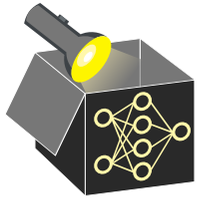
PITFALL: LOCAL VS. GLOBAL FEATURES ▸ LAUGEL_2018



- **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 imput 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

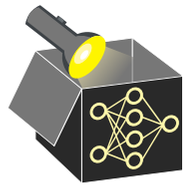
▸ Laugel et al. 2018



- **Pitfall:** Sampling from entire input space may hide influence of locally relevant features in favor of globally relevant ones, even for narrow kernels.
- **Feature types:**
  - *Global features* influence predictions broadly across whole imput 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

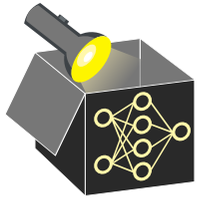
▸ LAUGEL_2018



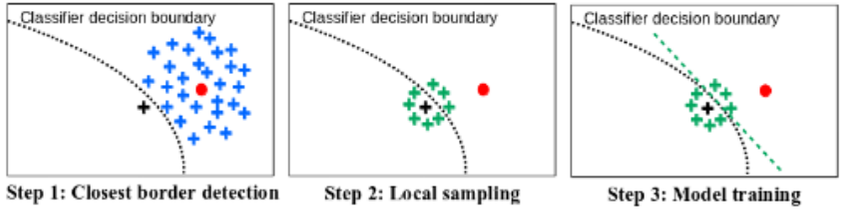
- **Pitfall:** Sampling from entire input space may hide influence of locally relevant feat in favor of globally relevant ones, even for narrow kernels.
- **Feature types:**
  - *Global features* influence predictions broadly across whole imput space $\mathcal{X}$
  - *Local features* affect predictions 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 obs. to be explained **x** may miss decision boundary
- **Solution (LS: Local Surrogate Method):**
  - Step 1: Find closest point to **x** (red dot) from opposite class (black cross)
  - Step 2: Sample around that point to better capture boundary
  - Step 3: Train local surrogate using those samples
    $\rightsquigarrow$ better approximates the local direction of the decision boundary

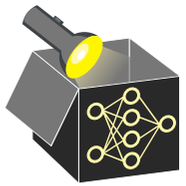


Example: x (red point), closest point from other class (black cross)

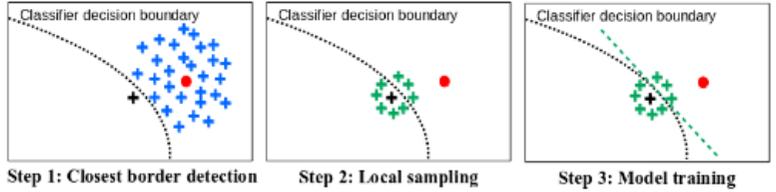
- LIME: What does the model do around this point?
- LS: How does the model change when crossing the boundary near this point?

PITFALL: LOCAL VS. GLOBAL FEATURES

▸ LAUGEL_2018



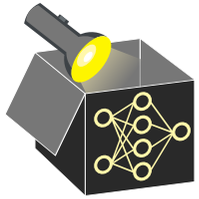
- **Problem:** Sampling around observation to be explained **x** may miss decision boundary
- **Solution (LS: Local Surrogate Method):**
  1. Find closest point to **x** (red dot) from opposite class (black cross)
  2. Sample around that point to better capture boundary
  3. Train local surrogate using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



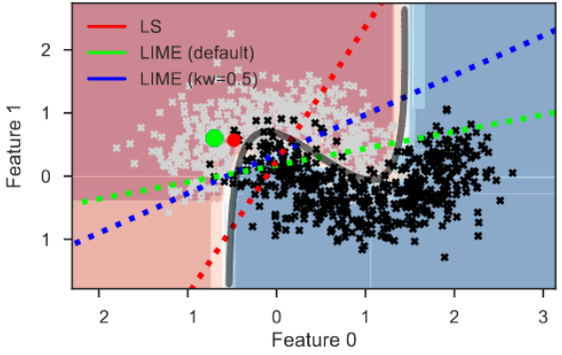
Example: x (red point), closest point from other class (black cross)

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

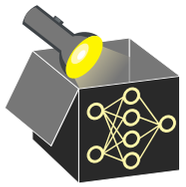
PITFALL: LOCAL VS. GLOBAL FEATURES – EXAMPLE



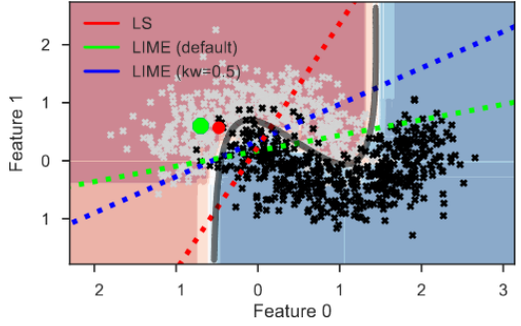
- Random forest (RF) classification on half-moons dataset
- **Background color:** Classification of RF (prediction surface)
- **Black/grey crosses:** training data
- **Green dot:** Obs. to be explained; **Red dot:** nearest point from opposite class
- **Grey curve:** RF's decision boundary; **Dotted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



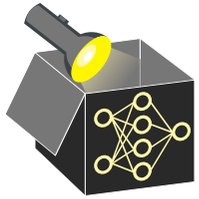
PITFALL: LOCAL VS. GLOBAL FEATS – EXAMPLE



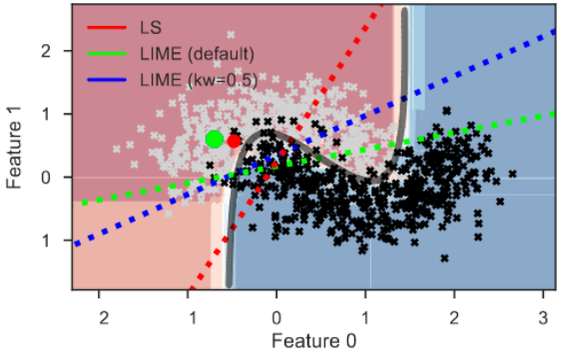
- Random forest (RF) classification on half-moons dataset
- **Background color:** Classification of RF (prediction surface)
- **Black/grey crosses:** training data
- **Green dot:** Obs. to be explained; **Red dot:** nearest opposite-class point
- **Grey curve:** RF's decision boundary; **Dotted lines:**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



PITFALL: LOCAL VS. GLOBAL FEATURES – 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 point from opposite class
- **Grey curve:** RF's decision boundary; **Dotted lines:** LIME decision boundaries
- **Red line:** Local surrogate (LS) method ▸ Laugel et al. 2018

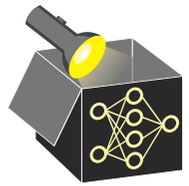


Feature 0 is global; class always flips when moving left (red) to right (blue)

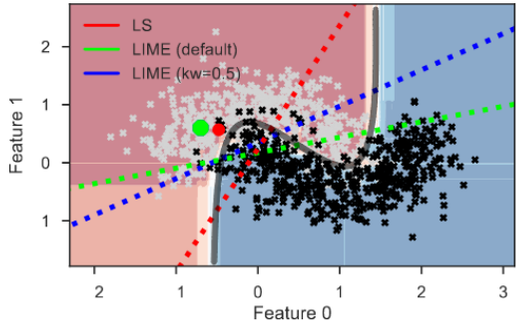
Feature 1 is local; class flips only near boundary when moving up/down

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

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 2018



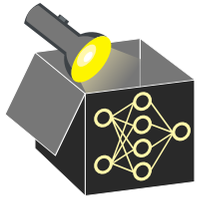
Feature 0 is global; class always flips when moving left (red) to right (blue)

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)

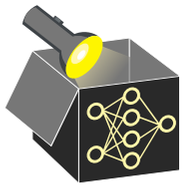
PITFALL: FAITHFULNESS

- **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

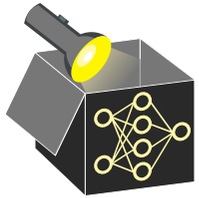


PITFALL: FAITHFULNESS

- **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



PITFALL: FAITHFULNESS



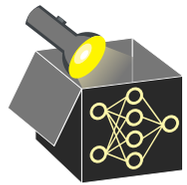
- **Problem**: Trade-off between local fidelity vs. sparsity
- **Observation**:
  - Too simple model $\rightsquigarrow$ low fidelity $\rightsquigarrow$ unreliable explanations
  - Complex model $\rightsquigarrow$ high fidelity $\rightsquigarrow$ difficult to interpret surrogate
- **Example: Credit data**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

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

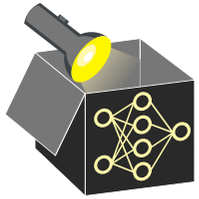
  - Generalized additive model (with all 9 features) is more complex:

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

Interpretable Machine Learning – 8 / 12

PITFALL: HIDING BIASES

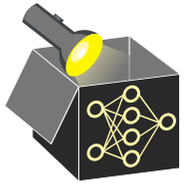
Slack et al. 2020



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

PITFALL: HIDING BIASES

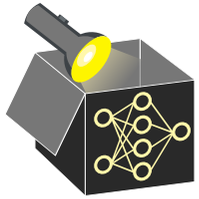
SLACK_2020



- **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 produce LIME explanations

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

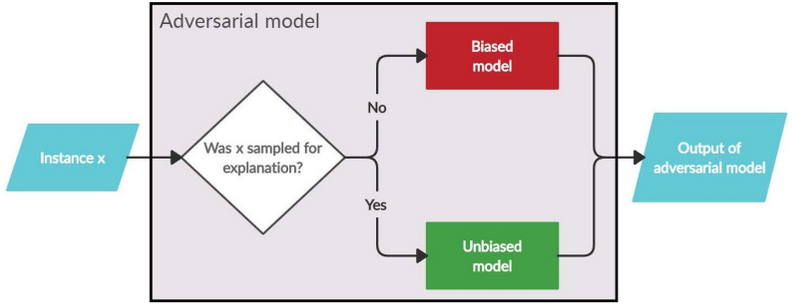
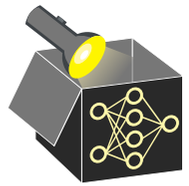


Image Source: Vres, Sikonja (2021)

PITFALL: HIDING BIASES

SLACK_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

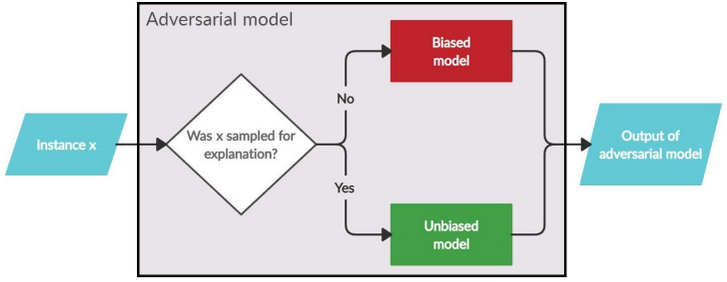
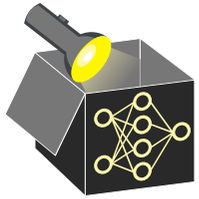


Image Source: Sikonja 2021

PITFALL: HIDING BIASES ▸ Slack et al. 2020

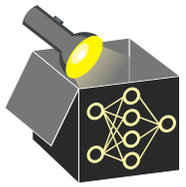


Key insight: LIME can be fooled if explanations rely on model behavior outside the true data manifold.

Example: Credit approval

- Biased model uses features 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 to hide bias
- LIME's extrapolated samples trigger the unbiased model
  $\Rightarrow$ explanation appears fair, but original predictions are biased

PITFALL: HIDING BIASES ▸ SLACK_2020



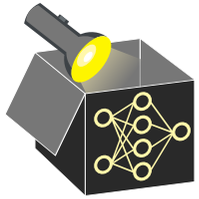
Key insight: LIME can be fooled if explanations rely on model behavior outside the true data manifold.

Example: Credit approval

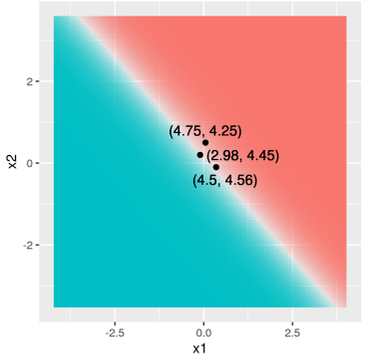
- Biased model uses feats correlated with gender (parental leave duration)
  $\rightsquigarrow$ used to make biased/unfair predictions
- Unbiased model uses only features unrelated to gender for fairness
  $\rightsquigarrow$ used to produce explanations based on unbiased predictions in order to hide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

PITFALL: ROBUSTNESS

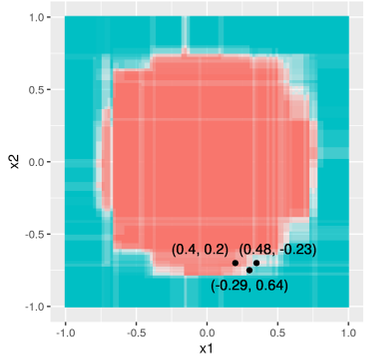
Alvarez-Melis, D., & Jaakkola, T. 2018



- **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:**



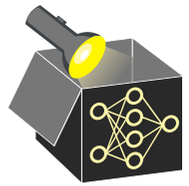
Linear task (logistic regression).
LIME returns similar coefficients for similar points.



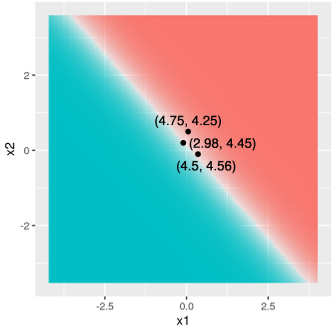
Nonlinear task (random forest).
LIME returns different coefficients for similar points.

PITFALL: ROBUSTNESS

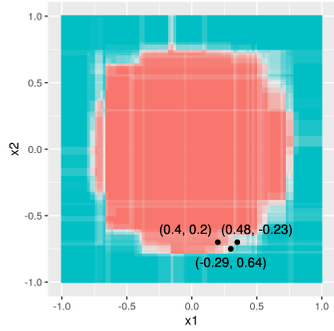
JAAKKOLA_2018



- **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:**



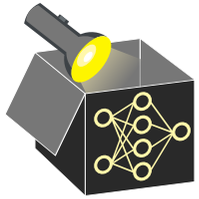
Linear task (logistic regression).
LIME returns similar coefficients for similar points.



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

PITFALL: DEFINITION OF SUPERPIXELS

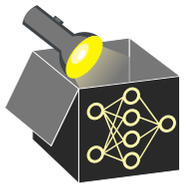
Achanta et al. 2012



- **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

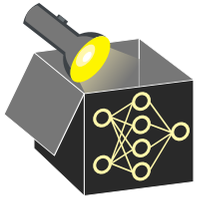
ACHANTA_2012



- **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

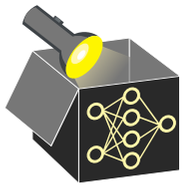
▸ Achanta et al. 2012



- **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

PITFALL: DEFINITION OF SUPERPIXELS

▸ ACHANTA_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