R-LIME: Rectangular Constraints and Optimization for Local Interpretable Model-agnostic Explanation Methods

Genji Ohara, Keigo Kimura and Mineichi Kudo

December 2024

Division of Computer Science and Information Technology
Graduate School of Information Sci. and Tech., Hokkaido University
Sapporo 060–0814, JAPAN

Agenda

Agenda

- Background
- Related Work
- Proposed Method: R-LIME
- Experiments
- Discussion
- Conclusion

Background

Background

Interpretable Machine Learning

- Complex ML models (Black Box)
  - Deep Neural Networks
  - Ensemble Models

→Decision process is unambiguous

- Simple ML models (White Box)
  - Linear Models
  - Decision Trees

→Decision process is ambiguous

Approximate Locally

Related Work

Related Work

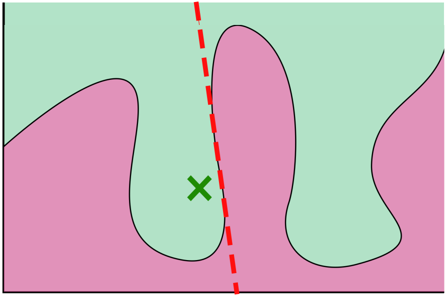
- LIME[1]
- Anchor[2]

[1] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. **""Why Should I Trust You?": Explaining the Predictions of Any Classifier".** In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* KDD '16. San Francisco, California, USA: Association for Computing Machinery, 2016, pp. 1135–1144. ISBN: 978-1-4503-4232-2.

[2] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. **"Anchors: High-Precision Model-Agnostic Explanations".** In: *Proceedings of the AAAI Conference on Artificial Intelligence* 32.1 (Apr. 2018), pp. 1527–1535.

Related Work 1 — **LIME (Local Interpretable Model-agnostic Explanations)**[3]

1. Generate perturbed instances around the given focal point
2. Learn a linear model on the perturbed instances

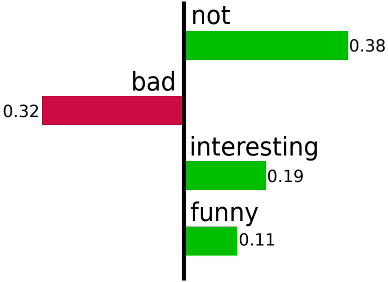


[3] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. **""Why Should I Trust You?": Explaining the Predictions of Any Classifier".** In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: Association for Computing Machinery, 2016, pp. 1135–1144. ISBN: 978-1-4503-4232-2.

Related Work / LIME

This book is not bad.
It is funny and interesting.

Example of the focal point. The sentiment prediction model predicted this sentence as "Positive".

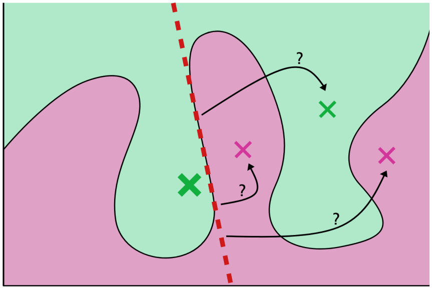


Example of LIME's explanation for the output by the sentiment prediction model.

Related Work 1 — Drawbacks of LIME

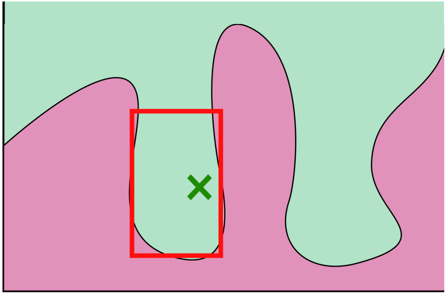
Scope of Explanation is Unknown

- How general is the knowledge derived from the explanation?



Related Work 2 — Anchor

- Search for the rectangular region in which the model's outputs for the focal point and other points are consistent with high probability.
- Use the feature of the predicate to express the optimal rectangular region.
  ex. Gender = 'Male' AND 20 ¡= Age ¡ 30



This book is not bad.
It is funny and interesting.

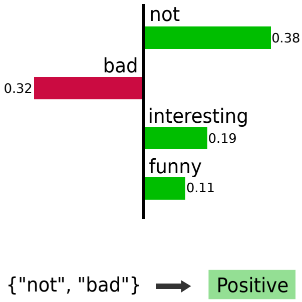
Example of the focal point. The sentiment prediction model predicted this sentence as “Positive”.

{"not", "bad"} → Positive

Example of Anchor explanation for the sentiment prediction model.

Users get less insight

- How much influence does each feature have on the prediction?



Comparison of LIME and Anchor outputs for the sentiment prediction model

Related Work / Our Goals

	LIME	Anchor	Proposed Method
Feature Importance	✓	×	✓
Optimal Region	×	✓	✓
Interpretable Region	×	✓	✓

Juggle Interpretability of explanation and its region

→ Users can utilize knowledge derived from explanation within reasonable range

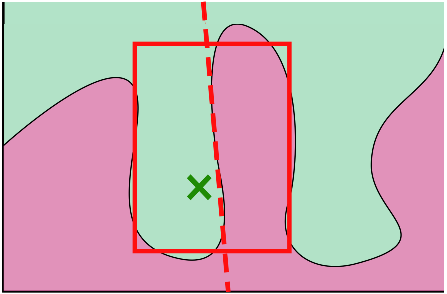
Proposed Method: R-LIME

Proposed Method: R-LIME

R-LIME (Ruled LIME) = LIME + Anchor

- Approximate in rectangular region
- Express the region as a conjunction of feature predicates

  ex. Gender = ‘Male’ AND 20 ¡= Age ¡ 30



Proposed Method: R-LIME / Setting

m-dim input space (discretized)	$\mathbb{D}^m$
A black-box classifier	$f : \mathbb{D}^m \to \{0, 1\}$
A focal point	$x \in \mathbb{D}^m$
Distribution on input space	$\mathcal{D}$
Set of all possible approx. model	G

Rule: a conjunction of predicates

$$A(z) = a_{i_1}(z) \wedge a_{i_2}(z) \wedge \cdots \wedge a_{i_k}(z), \quad a_i(z) = \mathbb{1}_{z_i = x_i}$$

Proposed Method: R-LIME / Setting

Expected accuracy of approx. model g in A

Accuracy of rule A: $\mathrm{acc}(A) = \max_{g \in G} \mathbb{E}_{z \sim \mathcal{D}(z|A)}[\mathbb{1}_{f(z)=g(z)}]$

Coverage of rule A: $\mathrm{cov}(A) = \mathbb{E}_{z \sim \mathcal{D}(z)}[A(z)]$

Probability that global sample z is inside A

Our problem: $\tilde{A} = \mathop{\arg\max}_{A \text{ } s.t. \text{ } P(\mathrm{acc}(A) \geq \tau) \geq 1-\delta, A(x)=1} \mathrm{cov}(A)$

Maximize coverage under the constraint of accuracy

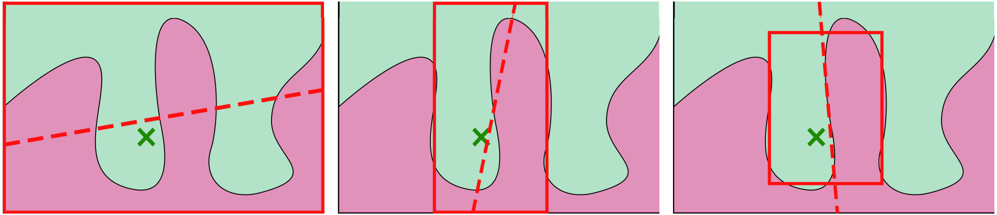
Proposed Method: R-LIME / Algorithm (Beam Search)

Add a new predicate to each of the rules in $\mathcal{A}_{i-1}$

Solve as multi-armed bandit problem

Repeat the following steps:

1. $\mathcal{A}_i \leftarrow$ A set of candidate rules
2. $\mathcal{A}_i \leftarrow B$ rules with highest accuracy
3. Search for the rule with highest coverage in $\mathcal{A}_i$
   If it is found, return it



Experiments

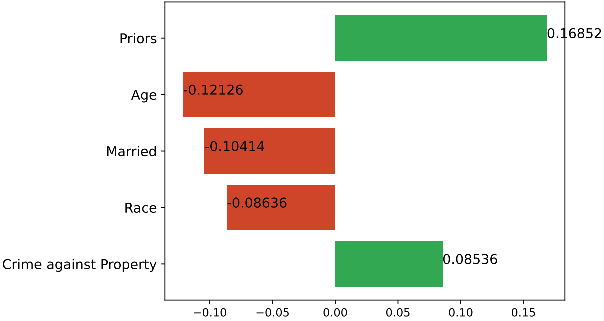
Visually compare LIME and R-LIME on the real dataset

- Use recidivism dataset[4]
- Train black-box classifier (random forest)
- Compare the output explanations of LIME and R-LIME

[4] Peter Schmidt and Ann D Witte. ***Predicting Recidivism in North Carolina, 1978 and 1980.*** Inter-university Consortium for Political and Social Research, 1988.

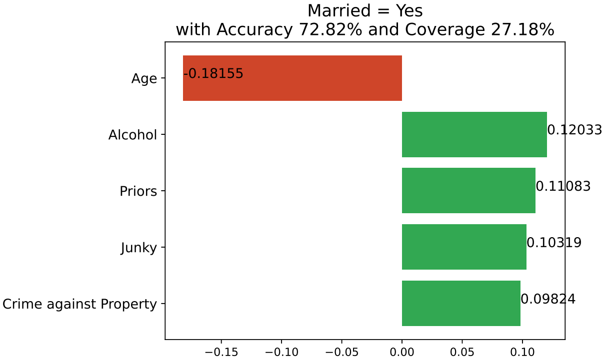
Race	Black (0)
Alcohol	No (0)
Junky	No (0)
Supervised Release	Yes (1)
Married	Yes (1)
Felony	No (0)
WorkRelease	Yes (1)
Crime against Property	No (0)
Crime against Person	No (0)
Gender	Male (1)
Priors	1
YearsSchool	8.00 ¡ YearsSchool ¡= 10.00 (1)
PrisonViolations	0
Age	Age ¿ 33.00 (3)
MonthsServed	4.00 ¡ MonthsServed ¡= 9.00 (1)
Recidivism	No more crimes (0)

LIME provides the importance of each feature to the prediction of the random forest

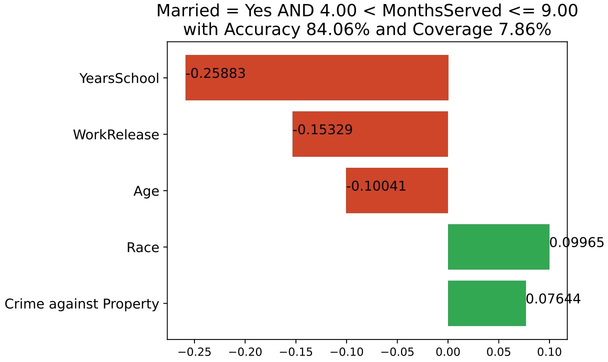


R-LIME provides not only the feature importance but also its application scope.

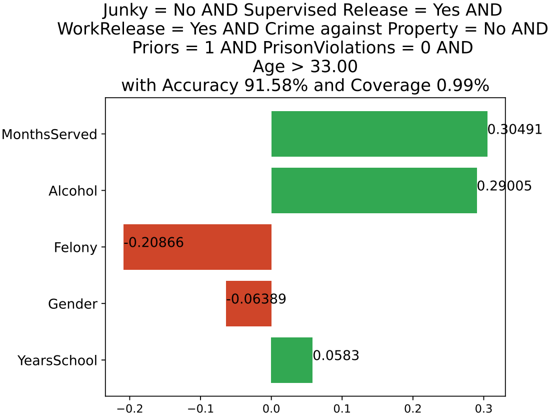
It can be only applied to married prisoners!



Constraints of served months are added to the explanation

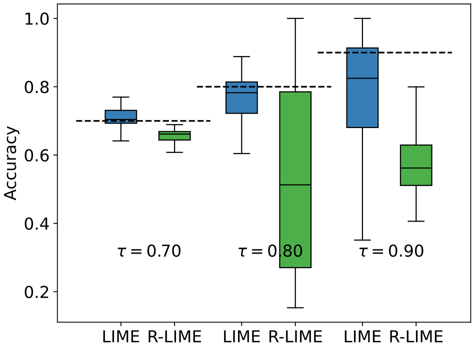


Too low coverage $\rightarrow$ Limited generality



Compare local approximation accuracy of LIME and R-LIME

- Train random forest model on recidivism dataset
- Repeat the following steps against 100 instances
  - Generate explanations of LIME and R-LIME
  - Sample 10000 instances within the region of the R-LIME explanation
  - Calculate the local approximation accuracy of LIME and R-LIME



- R-LIME learns high-accuracy model adapted to the oprimized region
- LIME may not effectively approximize depending on how the region selected

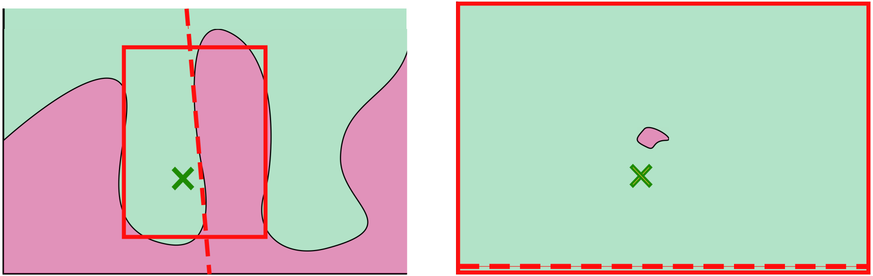
Discussion

Discussion

Topics of Discussion

- Behavior for imbalanced label distribution
- Changes in reword distribution in best arm identification
- Parameter selection

When the ratio of minority labels is less than $1 - \tau$



R-LIME covers the entire input space and always outputs the majority label

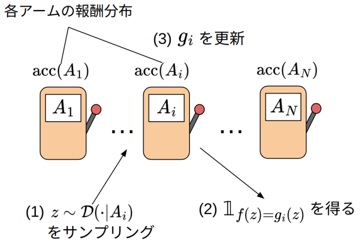
- Modify the loss function
  - weighted logistic loss
  - Focal Loss[5]
- Constraint on imbalanced label distribution
  - add the following constraint

$$\left(\mathbb{E}_{z\sim\mathcal{D}(z|A)}[\mathbb{1}_{f(z)=1}] - \frac{1}{2}\right)^2 < \mu$$

[5] Tsung Yi Lin et al. **"Focal Loss for Dense Object Detection".** In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.2 (2020), pp. 318–327.

Best arm identification using KL-LUCB algorithm[6]

- the original assumes constant reword distribution
- but in R-LIME, it changes with every update of the model after sampling



[6] Emilie Kaufmann and Shivaram Kalyanakrishnan. **"Information Complexity in Bandit Subset Selection".** In: *Proceedings of the 26th Annual Conference on Learning Theory.* Ed. by Shai Shalev-Shwartz and Ingo Steinwart. Vol. 30. Proceedings of Machine Learning Research. Princeton, NJ, USA: PMLR, Dec. 2013, pp. 228–251.

	Estimated acc.	True acc.	Deviation
Average	$.811$	$.829$	$.012$
Standard Deviation	$.018$	$.023$	$.017$

Comparison of true accuracy and estimated accuracy by R-LIME

Considering confidence level $1 - \delta = 0.95$, the deviation was relatively small.

Conclusion

Conclusion

	LIME	Anchor	**R-LIME**
Feature Importance	✓	×	✓
Optimal Scope	×	✓	✓
Interpretable Scope	×	✓	✓

Our methods achieves interpretability of both explanation and its scope!

Appendix

Algorithm 1 R-LIME

Input: Black-box model f, Target instance x, Distribution $\mathcal{D}$, Threshold τ, Beam width B, Tolerance ϵ, Confidence level $1-\delta$

Output: Rule A^* satisfying Eq. (1)

1: $A^* \leftarrow$ **null**, $\mathcal{A}_0 \leftarrow \emptyset,\ t \leftarrow 0$ ▷ *Initialize the set of candidate rules $\mathcal{A}_0$ to $\emptyset$*

2: **while** $A^* =$ **null do**

3: $\quad t \leftarrow t+1$

4: $\quad \bar{\mathcal{A}}_t \leftarrow$ GenerateCands($\mathcal{A}_{t-1}$)

5: $\quad \mathcal{A}_t \leftarrow$ B-BestCands($\bar{\mathcal{A}}_t, \mathcal{D}, B, \epsilon, \delta$)

6: $\quad A^* \leftarrow$ LargestCand($\mathcal{A}_t, \tau, \delta$)

$$\tilde{A} = \underset{A\ s.t.\ P(\mathrm{acc}(A)\geq\tau)\geq 1-\delta, A(x)=1}{\arg\max} \mathrm{cov}(A) \tag{1}$$

Algorithm 2 Generating new candidate rules

function GenerateCands($\mathcal{A}, x$)
    if $\mathcal{A} = \emptyset$ **then return** $\{true\}$ ▷ *An initial empty rule always returns* $true$
    $\bar{\mathcal{A}} \leftarrow \emptyset$
    for all $A \in \mathcal{A}$ **do**
        for all $a \in (T(x) \setminus A)$ **do**
            $\bar{\mathcal{A}} \leftarrow \bar{\mathcal{A}} \cup (A \wedge a)$ ▷ *Get a new rule by adding a new predicate* a *to* A
    return $\bar{\mathcal{A}}$

Algorithm 3 Searching rules with highest accuracy (KL-LUCB [5])

function B-BestCands($\bar{\mathcal{A}}, \mathcal{D}, B, \epsilon, \delta$)
    initialize $\mathrm{acc}, \mathrm{acc}_u, \mathrm{acc}_l$ for $\forall A \in \bar{\mathcal{A}}$
    $\mathcal{A} \leftarrow$ B-ProvisionallyBestCands($\bar{\mathcal{A}}$) ▷ *B rules with highest accuracy*
    $A \leftarrow \arg\min_{A \in \mathcal{A}} \mathrm{acc}_l(A, \delta)$ ▷ *The rule with the smallest lower bound*
    $A' \leftarrow \arg\max_{A' \notin (\bar{\mathcal{A}} \setminus \mathcal{A})} \mathrm{acc}_u(A', \delta)$ ▷ *The rule with the largest upper bound*
    while $\mathrm{acc}_u(A', \delta) - \mathrm{acc}_l(A, \delta) > \epsilon$ **do**
        sample $z \sim \mathcal{D}(z|A), z' \sim \mathcal{D}(z'|A')$
        update $\mathrm{acc}, \mathrm{acc}_u, \mathrm{acc}_l$ for A and A'
        $\mathcal{A} \leftarrow$ B-ProvisionallyBestCands($\bar{\mathcal{A}}$)
        $A \leftarrow \arg\min_{A \in \mathcal{A}} \mathrm{acc}_l(A, \delta)$
        $A' \leftarrow \arg\max_{A' \notin (\bar{\mathcal{A}} \setminus \mathcal{A})} \mathrm{acc}_u(A', \delta)$
    return $\mathcal{A}$

Algorithm 4 Searching a rule with highest coverage under constraint

```
function LARGESTCAND(𝒜, τ, δ)
    A* ← null                                  ▷ If no rule satisfies the constraint, return null
    for all A ∈ 𝒜 s.t. acc_l(A, δ) > τ do
        if cov(A) > cov(A*) then A* ← A
    return A*
```