

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

Background

Interpretable Machine Learning

- Simple ML models (White-box)

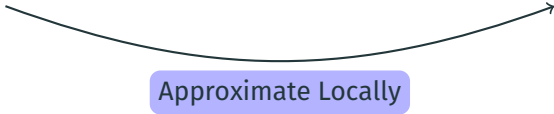
- Linear Models
- Decision Trees

→ Decision process is clear

- Complex ML models (Black-box)

- Deep Neural Networks
- Ensemble Models

→ Decision process is unclear

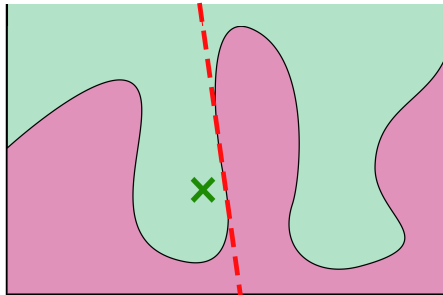


Approximate Locally

Related Work

Related Work 1 — LIME (Local Interpretable Model-agnostic Explanations)¹

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

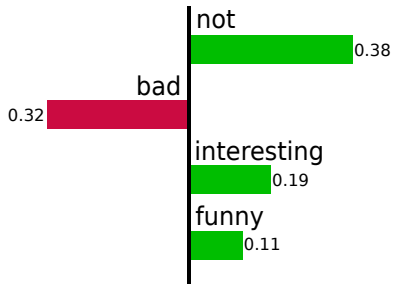


¹Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: Association for Computing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

Related Work / LIME / Example Output

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

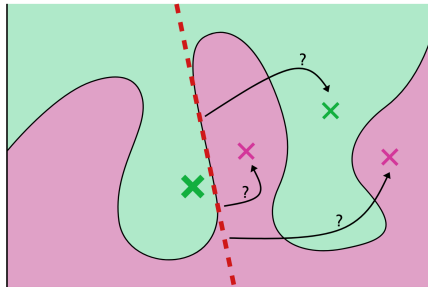
Example of the focal point. The sentiment prediction model predicted this sentence as "Positive".



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

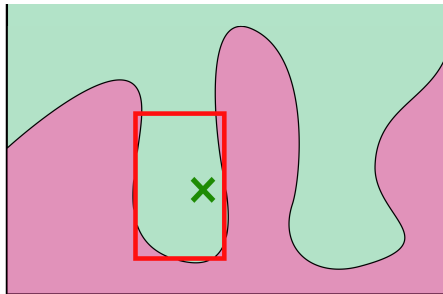
Unclear scope of explanation

- How general is the knowledge derived from the explanation?



Related Work 2 — **Anchor**

- Maximize the rectangular region as long as the model's outputs are consistent with high probability.
- Use the feature of the predicate to express the optimal rectangular region.
ex. Gender = 'Male' AND $20 \leq \text{Age} < 30$



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

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

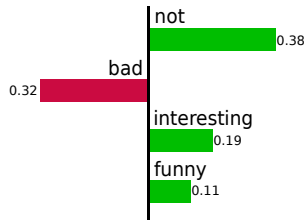
{"not", "bad"} → Positive

Example of Anchor explanation for the sentiment prediction model.

Related Work / Anchor / Drawbacks of Anchor

Users get less insight

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



{"not", "bad"} → Positive

Comparison of LIME and Anchor outputs for the sentiment prediction model

Related Work / Our Goals

	LIME	Anchor	Our Method
Feature Importance	✓	×	✓
Optimal Scope	×	✓	✓
Interpretable Scope	×	✓	✓

Achieve interpretability of both explanation and its scope

→ Users can apply the derived insight within reasonable range

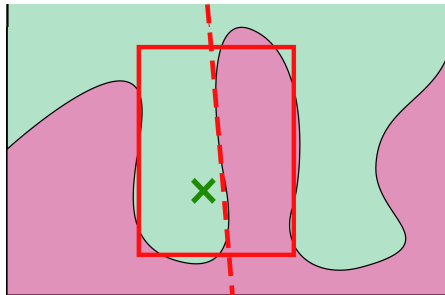
Proposed Method: R-LIME

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 $20 \leq \text{Age} < 30$



Proposed Method: R-LIME

m -dim input space (discretized)

$$\mathbb{D}^m$$

A black-box classifier

$$f : \mathbb{D}^m \rightarrow \{0, 1\}$$

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

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

Expected accuracy of
approx. model g in A

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

Probability that global
sample z is inside A

Our problem:

$$\tilde{A} = \arg \max_{A \text{ s.t. } P(\text{acc}(A) \geq \tau) \geq 1 - \delta, A(x)=1} \text{cov}(A)$$

Maximize coverage under the constraint of accuracy

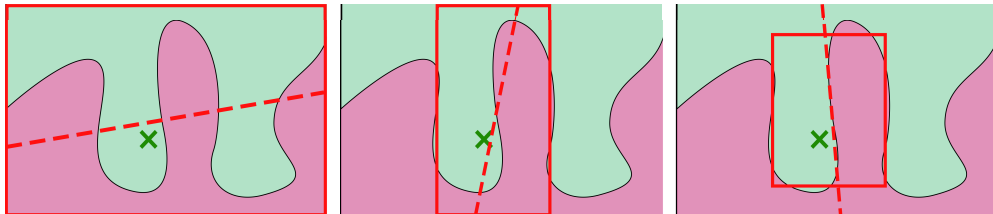
Proposed Method: R-LIME / Algorithm (Beam Search)

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.

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

Solve as multi-armed bandit problem



Experiments

Visually compare LIME and R-LIME on the real dataset

- Use recidivism dataset²
- Train black-box classifier (random forest)
- Compare the output explanations of LIME and R-LIME

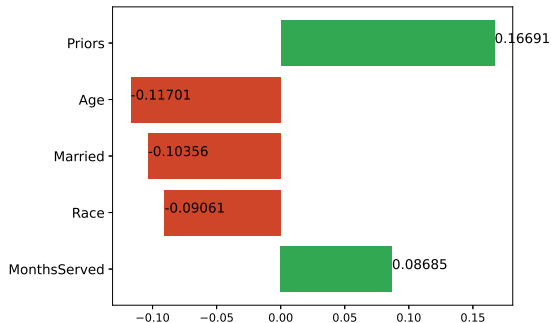
²Peter Schmidt and Ann D Witte. *Predicting Recidivism in North Carolina, 1978 and 1980*. Inter-university Consortium for Political and Social Research, 1988.

Experiments / Qualitative Evaluation

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)
------------	--------------------

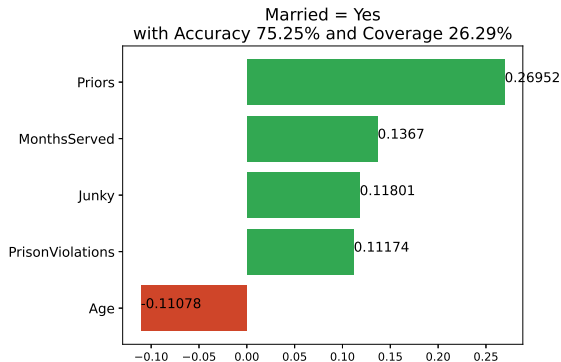
LIME: Only feature importance → Users cannot understand how general it is



Experiments / Qualitative Evaluation / Results — R-LIME ($\tau = 0.70$)

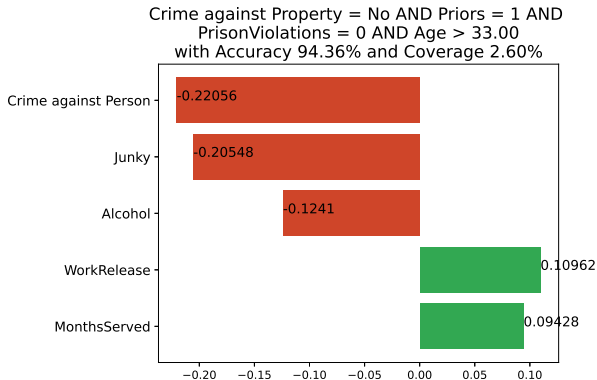
R-LIME: Not only feature importance but also its application scope

- “This can be only applied to married prisoners”



Experiments / Qualitative Evaluation / Results — R-LIME ($\tau = 0.90$)

Too low coverage \rightarrow Users understand its limited generality

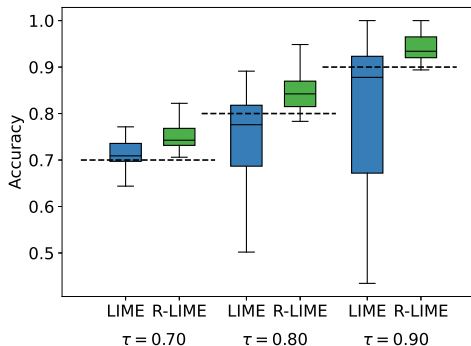


Compare local approximation accuracy of LIME and R-LIME

- Use recidivism dataset³
- Train black-box classifier (random forest)
- 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

³Peter Schmidt and Ann D Witte. *Predicting Recidivism in North Carolina, 1978 and 1980*. Inter-university Consortium for Political and Social Research, 1988.

Experiments / Quantitative Evaluation (LIME vs. R-LIME) / Results



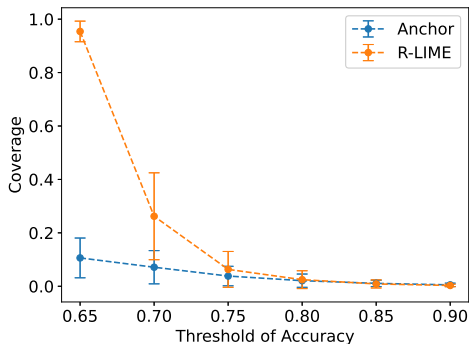
- R-LIME learns high-accuracy model adapted to the optimized region
- LIME may not effectively approximate depending on how the region selected

Compare coverage of Anchor and R-LIME

- Use recidivism dataset⁴
- Train black-box classifier (random forest)
- Repeat the following steps under the thresholds $\tau \in \{0.65, 0.70, \dots, 0.90\}$
 - Generate explanations for 704 instances in test set
 - Calculate the coverage of Anchor and R-LIME

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

Experiments / Quantitative Evaluation (Anchor vs. R-LIME) / Results

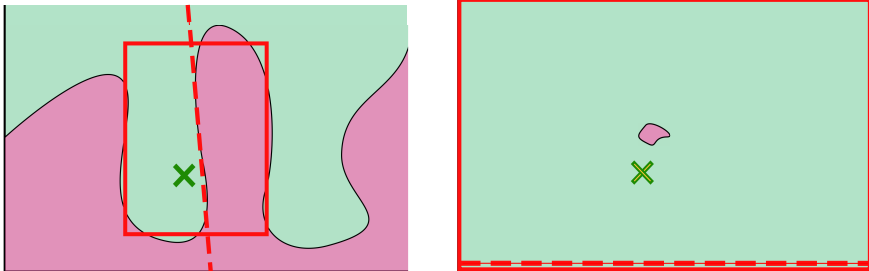


- Much higher coverage of R-LIME than Anchor, especially for small τ
 - R-LIME dynamically captures complex decision boundary.
 - Anchor uses only the intervals discretized in advance.

Discussion

Discussion / Behavior for Imbalanced Label Distribution

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



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

Discussion / Behavior for Imbalanced Label Distribution / Possible solutions

- Modify the loss function
 - weighted logistic loss
 - Focal Loss⁵
- Constraint on imbalanced label distribution
 - add the following constraint

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

⁵Tsung 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 algorithm⁶

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

⁶Emilie 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!

Also:

- More accurate than LIME
- More general than Anchor

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: $A^* \leftarrow \text{null}$, $\mathcal{A}_0 \leftarrow \emptyset$, $t \leftarrow 0$

\triangleright Initialize the set of candidate rules \mathcal{A}_0 to \emptyset

2: **while** $A^* = \text{null}$ **do**

3: $t \leftarrow t + 1$

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

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

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

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

Algorithm 2 Generating new candidate rules

```
1: function GENERATECANDS( $\mathcal{A}, x$ )
2:   if  $\mathcal{A} = \emptyset$  then return  $\{true\}$ 
3:    $\bar{\mathcal{A}} \leftarrow \emptyset$ 
4:   for all  $A \in \mathcal{A}$  do
5:     for all  $a \in (T(x) \setminus A)$  do
6:        $\bar{\mathcal{A}} \leftarrow \bar{\mathcal{A}} \cup (A \wedge a)$ 
7:   return  $\bar{\mathcal{A}}$ 
```

▷ An initial empty rule always returns *true*

▷ Get a new rule by adding a new predicate a to A

Algorithm 3 Searching rules with highest accuracy (KL-LUCB [4])

```
1: function B-BESTCANDS( $\bar{\mathcal{A}}, \mathcal{D}, B, \epsilon, \delta$ )
2:   initialize  $\text{acc}, \text{acc}_u, \text{acc}_l$  for  $\forall A \in \bar{\mathcal{A}}$ 
3:    $\mathcal{A} \leftarrow \text{B-PROVISIONALLYBESTCANDS}(\bar{\mathcal{A}})$ 
4:    $A \leftarrow \arg \min_{A \in \mathcal{A}} \text{acc}_l(A, \delta)$ 
5:    $A' \leftarrow \arg \max_{A' \notin (\bar{\mathcal{A}} \setminus \mathcal{A})} \text{acc}_u(A', \delta)$ 
6:   while  $\text{acc}_u(A', \delta) - \text{acc}_l(A, \delta) > \epsilon$  do
7:     sample  $z \sim \mathcal{D}(z|A), z' \sim \mathcal{D}(z'|A')$ 
8:     update  $\text{acc}, \text{acc}_u, \text{acc}_l$  for  $A$  and  $A'$ 
9:      $\mathcal{A} \leftarrow \text{B-PROVISIONALLYBESTCANDS}(\bar{\mathcal{A}})$ 
10:     $A \leftarrow \arg \min_{A \in \mathcal{A}} \text{acc}_l(A, \delta)$ 
11:     $A' \leftarrow \arg \max_{A' \notin (\bar{\mathcal{A}} \setminus \mathcal{A})} \text{acc}_u(A', \delta)$ 
12:  return  $\mathcal{A}$ 
```

$\triangleright B$ rules with highest accuracy
 \triangleright The rule with the smallest lower bound
 \triangleright The rule with the largest upper bound

Algorithm 4 Searching a rule with highest coverage under constraint

```
1: function LARGESTCAND( $\mathcal{A}, \tau, \delta$ )  
2:    $A^* \leftarrow \text{null}$   $\triangleright$  If no rule satisfies the constraint, return null  
3:   for all  $A \in \mathcal{A}$  s.t.  $\text{acc}_l(A, \delta) > \tau$  do  
4:      $\lfloor$  if  $\text{cov}(A) > \text{cov}(A^*)$  then  $A^* \leftarrow A$   
5:    $\lfloor$  return  $A^*$ 
```
