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

Genji Ohara, Keigo Kimura, Mineichi Kudo

{genji-ohara, kimura5, mine}@ist.hokudai.ac.jp

Division of Computer Science and Information Technology
Graduate School of Information Sci. and Tech., Hokkaido University

1. Background

Interpretable Machine Learning

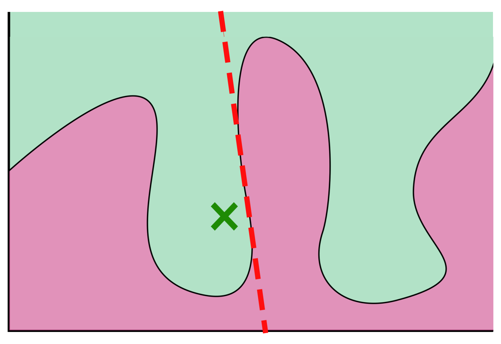
- simple models
  - linear models
  - decision trees
- → process is clear
- complex models
  - deep neural networks
  - ensemble models
- → process is unclear

Locally approximate complex models by simple models

2. Related Work

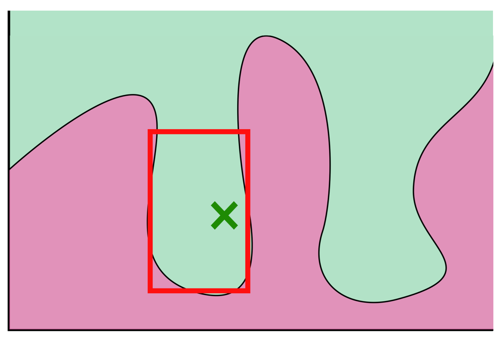
LIME (Local Interpretable Model-agnostic Explanations)[1]

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



Anchor[2]

1. Maximize the rectangular region as long as the model's outputs are consistent with high probability



LIME vs. Anchor

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

Figure: The focal point

{"not", "bad"} → Positive

Figure: Anchor's explanation

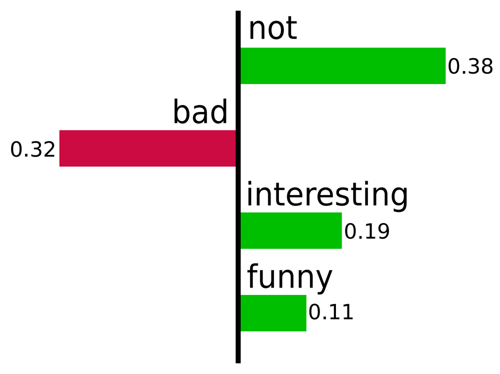


Figure: LIME's explanation

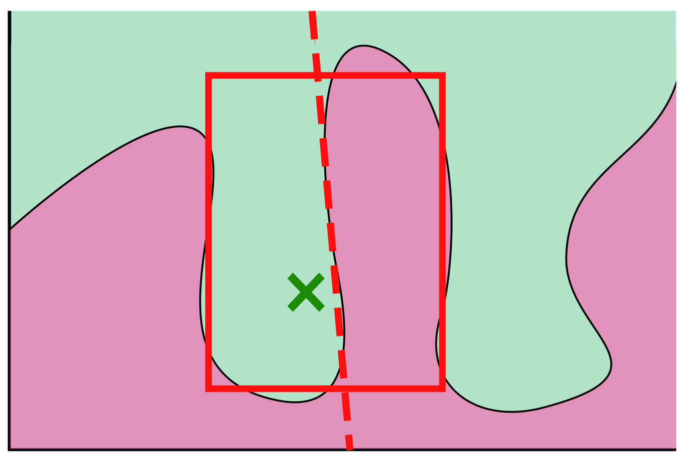
- LIME: unclear and not optimal scope
- Anchor: users get less insight

3. Proposed Method

R-LIME (Ruled LIME) = LIME + Anchor

- Approximate in rectangular region
- Maximize the region as long as approximation accuracy is higher than the given threshold
- Express the region as a conjunction of feature predicates

ex. Gender = 'Male' AND $20 \leq$ Age < 30



Settings

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	$\mathcal{D}$
all possible approx. model	G

Definitions

rule: a conjunction of predicates

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

accuracy: expected accuracy of approx. g in A

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

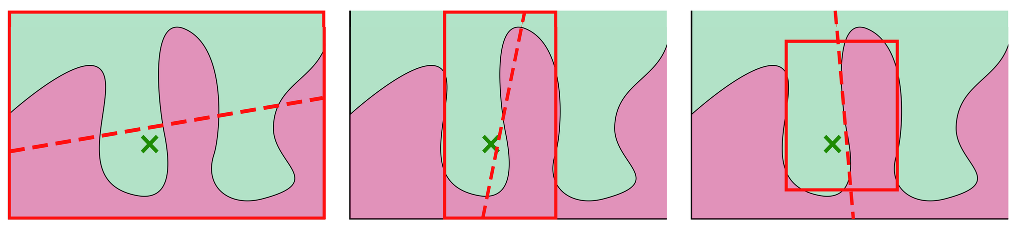
coverage: probability that sample z is inside A

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

our problem:

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

Maximize coverage under constraint of accuracy



Algorithm (beam search)

$$\mathcal{A}_{t-1} = \{A_1, \ldots, A_B\}$$

Generate a set of candidate rules
- add a new predicate to each rule

$\mathcal{A}_{t-1} = \{a_1\}$
$\to \mathcal{A}_t = \{a_1 \wedge a_2, a_1 \wedge a_3, a_1 \wedge a_4, \ldots\}$

$$\bar{\mathcal{A}}_t = \{A_1 \wedge a_1, A_1 \wedge a_2, \ldots, A_B \wedge a_m\}$$

Search rules with highest accuracy
- solve as best arm identification in multi-armed bandit problem using KL-LUCB algorithm[3]

$$\mathcal{A}_t = \{A'_1, A'_2, \ldots, A'_B\}$$

Search a rule with highest coverage under constraint of accuracy
- sample and update bounds acc_u, acc_l unless $\mathrm{acc}(A)_u \leq \tau$ or $\tau \leq \mathrm{acc}(A)_l$

$$A^* = \mathop{\arg\max}_{A \in \mathcal{A}_t \ s.t. \ P(\mathrm{acc}(A) \geq \tau) \geq 1-\delta} \mathrm{cov}(A)$$

if $A*$ is null → $t \leftarrow t + 1$; continue;

if $A*$ is not null

return A^*

4. Experiments

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)

Figure: Focal point

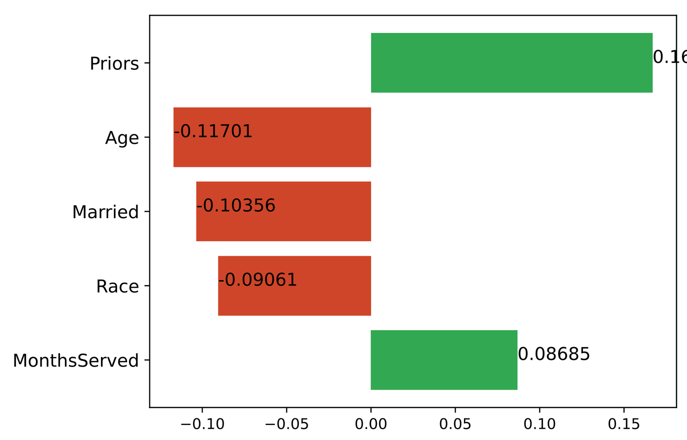


Figure: LIME's explanation

Age > 33.00 AND Priors = 1
with Accuracy 75.66% and Coverage 6.35%

Figure: Anchor's explanation ($\tau = 0.70$)

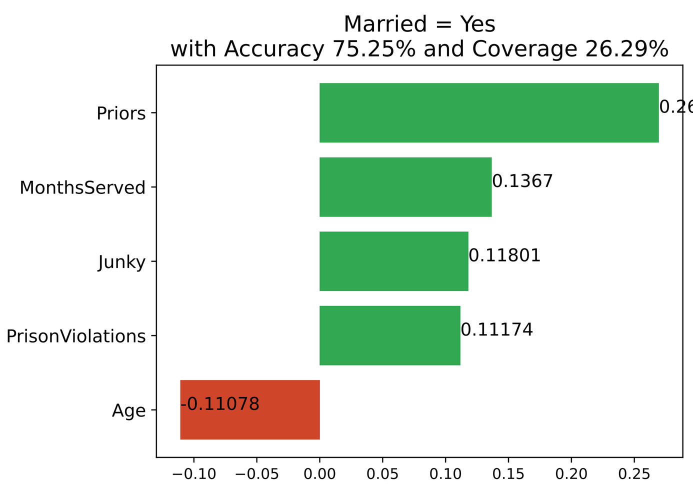


Figure: R-LIME's explanation ($\tau = 0.70$)

- More interpretable and optimal than LIME
- More descriptive than Anchor

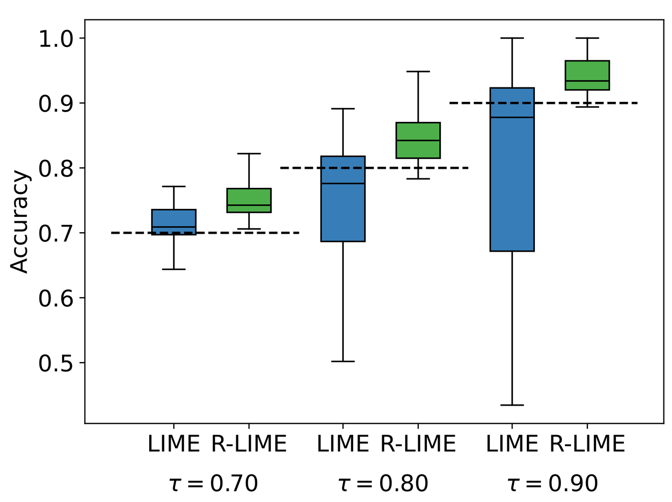


Figure: LIME vs. R-LIME (in accuracy)

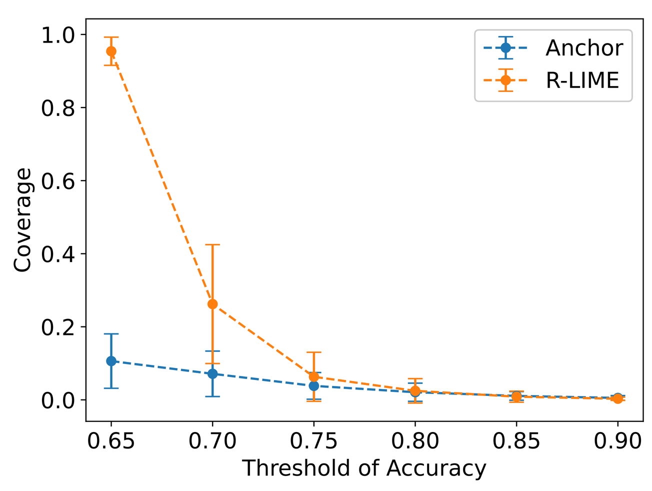


Figure: LIME vs. R-LIME (in coverage)

- More accurate than LIME
  - R-LIME adapts to optimized region flexibly
- More general than Anchor
  - R-LIME captures decision boundary more plecisely

5. Conclusion

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

- Achieved interpretability of both explanation and its scope!

Also:
- More accurate than LIME
- More general than Anchor

References

[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.

[3] 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.

27th International Conference on Pattern Recognition, December 01–05, 2024, Kolkata, India