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

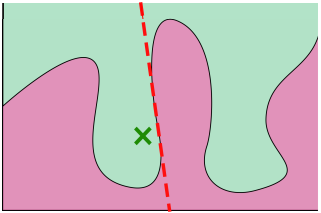
→ Decision process is unclear

Approximate Locally

Related Work: LIME & Anchor

LIME

- Approximate the model locally based on the perturbed samples

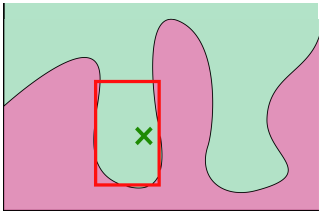


Visual illustration of LIME.

But how general is this explanation?

Anchor

- Maximize the rectangular region as long as the model's outputs are mostly consistent.



Visual illustration of Anchor.

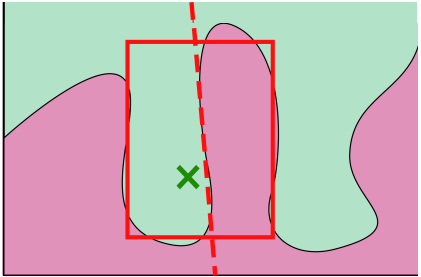
But how much influence does each feature have?

Proposed Method: R-LIME

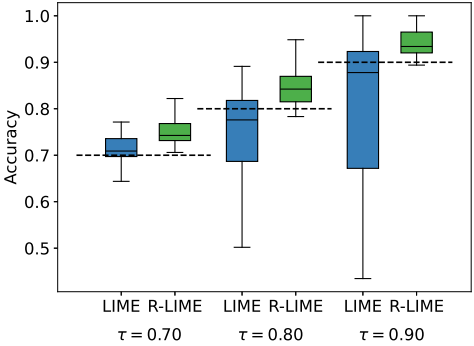
R-LIME (Ruled LIME) = LIME + Anchor

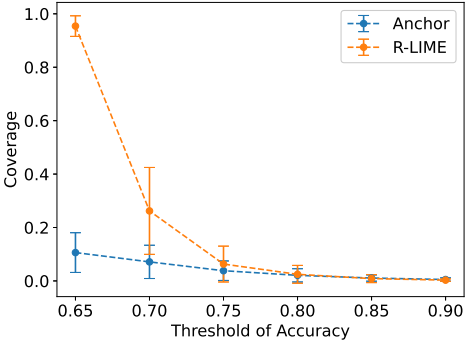
- Approximate in rectangular region
- Maximize the rectangular region as long as the approximation is accurate.
- Express the region as a conjunction of feature predicates

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



Experiments





- Much higher accuracy of R-LIME than LIME, especially for large τ
- Much higher coverage of R-LIME than Anchor, especially for small τ

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