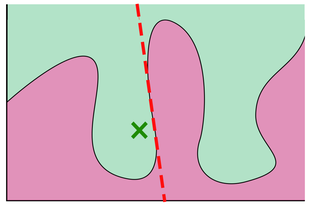
R-LIME: Rectangular Constraints and Optimization for Local Interpretable Model-agnostic Explanation Methods

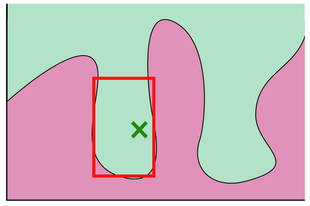
LIME (Local Interpretable Model-agnostic Explanations)

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



Anchor

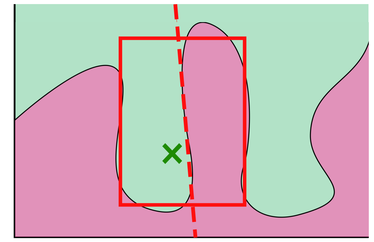
1. Maximize the rectangular region as long as the model's outputs are consistent with high probability

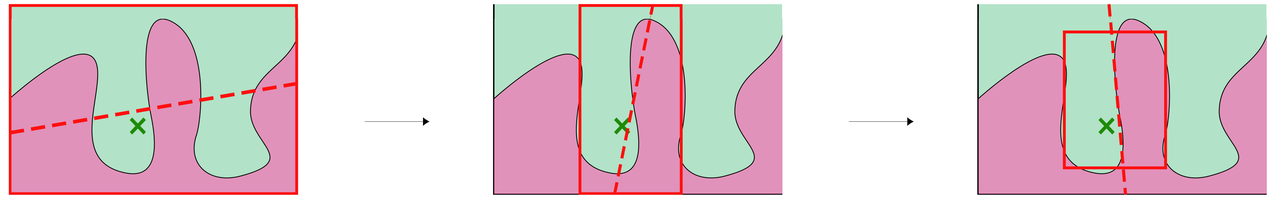


Our Method: R-LIME (Ruled LIME)

R-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





LIME vs. Anchor vs. R-LIME

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

Figure: The focal point

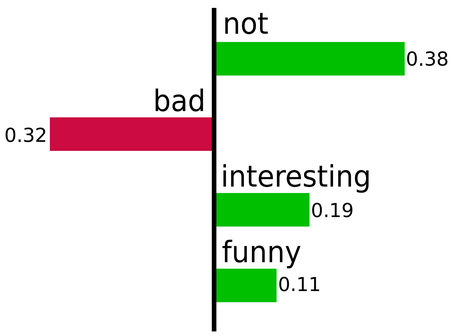


Figure: LIME's explanation

{"not", "bad"} → Positive

Figure: Anchor's explanation

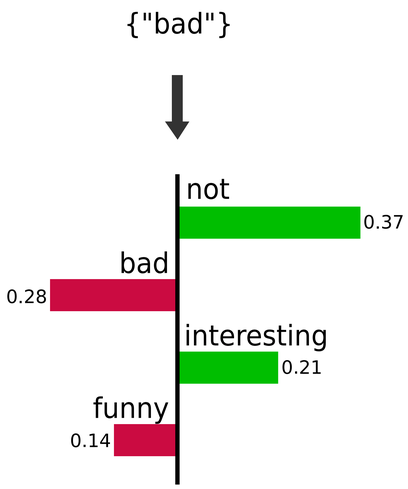


Figure: R-LIME's explanation

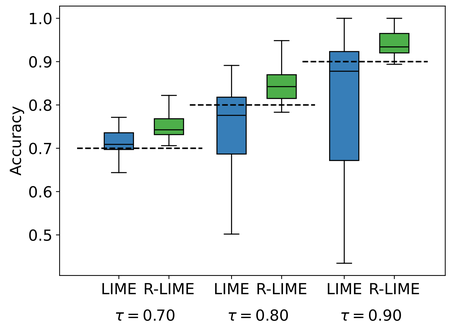


Figure: LIME vs. R-LIME (in accuracy)

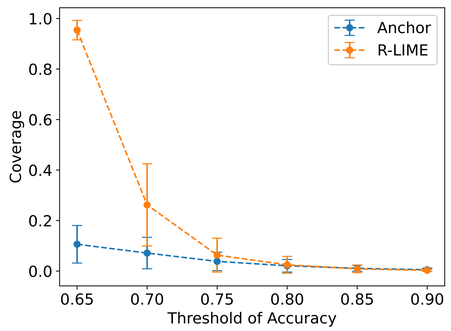


Figure: LIME vs. R-LIME (in coverage)

	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