Anchors: High-Precision Model-Agnostic Explanations

Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin

Shorya Consul

Mónica Ribero

$$f : X \to Y$$

Black-box model

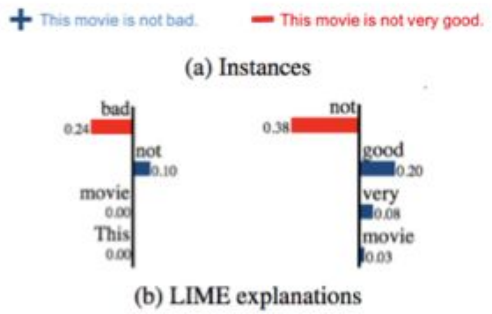
$$x \in X$$

An instance

Global Interpretability vs. Local Interpretability

- Allows the user to predict model's behaviour on any example (low *human accuracy*)
  - A set of rules
  - Simple model that imitates f
- Trades off flexibility, accuracy and/or efficiency
- Not suitable for image or text

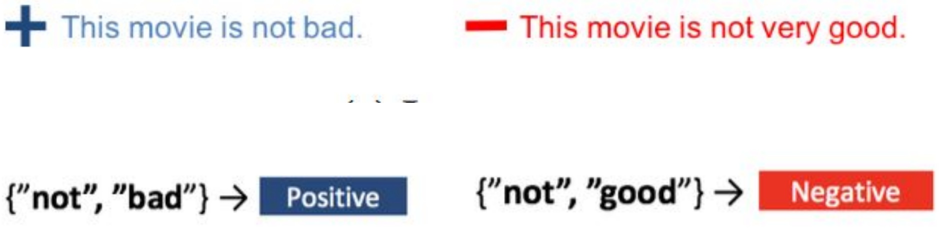
- Explain individual predictions
  - E.g. linear combination of input features
- Unclear *coverage*
  - *Region where explanation applies*



Anchors

- Local, model-agnostic explanations

- For instances where the anchor holds, prediction will be the same with high probability



Anchors

- Formally: Let A be a rule

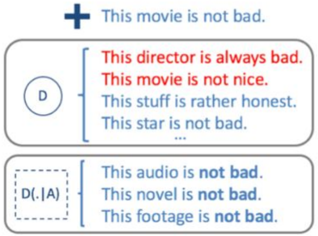
Interpretable explanation

$$A(x) := \begin{cases} 1 & \text{if all predicates hold for } x \\ 0 & \text{otherwise} \end{cases}$$

- A is an Anchor if

$$A(x) = 1, \quad \mathbb{E}_{\mathcal{D}(z|A)}[\mathbb{1}_{f(x)=f(z)}] \geq \tau$$

High precision



(a) $\mathcal{D}$ and $\mathcal{D}(.|A)$

Computing Anchors

- Computing precision is intractable for arbitrary f and D

$$P\left(\mathbb{E}_{\mathcal{D}(z|A)}[\mathbb{1}_{f(x)=f(z)}] \geq \tau \right) \geq 1-\delta$$

Can be estimated using the KL-LUCB algorithm[1]

- Anchors with higher coverage are preferred. Solve:

$$\max_{A \text{ s.t. } P(prec(A)\geq\tau)\geq 1-\delta} \mathbb{E}_{\mathcal{D}(z)}[A(z)]$$

Coverage:
Region where the explanation applies

[1]Kaufmann, E., and Kalyanakrishnan, S. 2013. Information complexity in bandit subset selection. In Proceedings of the Twenty-sixth Annual Conference on Learning Theory (COLT 2013), volume 30 of JMLR Workshop and Conference Proceedings, 228–251. JMLR.

Computing Anchors

Algorithm 1 Identifying the *Best* Candidate for Greedy

function GenerateCands($\mathcal{A}$, c)
  $\mathcal{A}_r = \emptyset$
  for all $A \in \mathcal{A}; a_i \in x, a_i \notin A$ **do**
    if $\text{cov}(A \wedge a_i) > c$ **then** {Only high-coverage}
      $\mathcal{A}_r \leftarrow \mathcal{A}_r \cup (A \wedge a_i)$ {Add as potential anchor}
  return $\mathcal{A}_r$ {Candidate anchors for next round}

function BestCand($\mathcal{A}, \mathcal{D}, \epsilon, \delta$)
  initialize $\text{prec}, \text{prec}_{ub}, \text{prec}_{lb}$ estimates $\forall A \in \mathcal{A}$
  $A \leftarrow \arg\max_A \text{prec}(A)$
  $A' \leftarrow \arg\max_{A' \neq A} \text{prec}_{ub}(A', \delta)$ {δ implicit below}
  while $\text{prec}_{ub}(A') - \text{prec}_{lb}(A) > \epsilon$ **do**
    sample $z \sim \mathcal{D}(z|A), z' \sim \mathcal{D}(z'|A')$ {Sample more}
    update $\text{prec}, \text{prec}_{ub}, \text{prec}_{lb}$ for A and A'
    $A \leftarrow \arg\max_A \text{prec}(A)$
    $A' \leftarrow \arg\max_{A' \neq A} \text{prec}_{ub}(A')$
  return A

Given an instance x *and level of precision* τ

Initialize empty, max coverage.

At each iteration

Extend by one additional predicate

Select A with max *estimated* precision

Break when level of precision is met.

Assumes shorter anchors have higher coverage

Beam-Search

- Shortcomings of greedy approach:
  - Greedy approach can only maintain a single rule at a time - suboptimal choice irreversible
  - Greedy approach not concerned with coverage, returns shortest anchor

- Beam-search
  - Maintain a set of candidate rules (addresses shortcoming one)
  - Pick anchor with highest coverage (addresses shortcoming two)
  - Do not store any rule with coverage < best anchor so far (efficient pruning of search space)
  - More likely to return anchor with higher coverage than greedy approach

Beam Search

Algorithm 2 Outline of the Beam Search

function BeamSearch($f, x, \mathcal{D}, \tau$)
  hyperparameters B, ϵ, δ
  $A^* \leftarrow$ **null**, $\mathcal{A}_0 \leftarrow \emptyset$ {Set of candidate rules}
  loop
    $\mathcal{A}_t \leftarrow \text{GenerateCands}(\mathcal{A}_{t-1}, \text{cov}(A^*))$
    $\mathcal{A}_t \leftarrow \text{B-BestCand}(\mathcal{A}_t, \mathcal{D}, B, \delta, \epsilon)$ {LUCB}
    if $\mathcal{A}_t = \emptyset$ **then break loop**
    for all $A \in \mathcal{A}_t$ s.t. $\text{prec}_{lb}(A, \delta) > \tau$ **do**
      if $\text{cov}(A) > \text{cov}(A^*)$ **then** $A^* \leftarrow A$
  return A^*

Initialization

Generate candidates with higher coverage and take B best

Update solution if higher coverage and high enough precision

Hyperparameters (tolerance (ϵ), width (δ) or maximum number of samples can be tuned to reasonable values so that the algorithm generates anchors quickly.

Anchor - Examples

(a) Original image (b) Anchor for "beagle" (c) Images where Inception predicts $P(\text{beagle}) > 90\%$

What animal is featured in this picture ?	**dog**
What floor is featured in this picture?	dog
What toenail is paired in this flowchart ?	dog
What animal is shown on this depiction ?	dog

(d) **VQA:** Anchor (bold) and samples from $\mathcal{D}(z|A)$

Where is the **dog**?	on the floor
What color is the **wall?**	white
When was this picture taken?	during the day
Why is he lifting his paw?	to play

(e) **VQA:** More example anchors (in bold)

Figure 3: Anchor Explanations for Image Classification and Visual Question Answering (VQA)

Experiments

Apply LIME explanations without considering distance to test instance

Use explanations if application of linear explanation yield probability higher than threshold; threshold set to give same average precision as anchor approach (*"cheating"*)

		Precision		Coverage	
		anchor	**lime-n**	**anchor**	**lime-t**
adult	logistic	95.6	81.0	10.7	21.6
	gbt	96.2	81.0	9.7	20.2
	nn	95.6	79.6	7.6	17.3
rcdv	logistic	95.8	76.6	6.8	17.3
	gbt	94.8	71.7	4.8	2.6
	nn	93.4	65.7	1.1	1.5
lending	logistic	99.7	80.2	28.6	12.2
	gbt	99.3	79.9	28.4	9.1
	nn	96.7	77.0	16.6	5.4

Table 4: Average precision and coverage with **simulated users** on 3 tabular datasets and 3 classifiers. *lime-n* indicates direct application of LIME to unseen instances, while *lime-t* indicates a threshold was tuned using an oracle to achieve the same precision as the anchor approach. The anchor approach is able to maintain very high precision, while a naive use of linear explanations leads to varying degrees of precision.

- Explanations from validation set, metrics obtained on test set
- 3 models used - logistic regression, 400 gradient boosted trees, MLP with 2 x 50 units
- Submodular pick used to pick subset of anchors

Anchor approach gives higher average precision

Experiments

Method	Precision				Coverage (perceived)				Time/pred (seconds)			
	adult	rcdv	vqa1	vqa2	adult	rcdv	vqa1	vqa2	adult	rcdv	vqa1	vqa2
No expls	54.8	83.1	61.5	68.4	79.6	63.5	39.8	30.8	29.8 ±14	35.7±26	18.7±20	13.9±20
LIME(1)	68.3	98.1	57.5	76.3	89.2	55.4	71.5	54.2	28.5±10	24.6±6	8.6±3	11.1±8
Anchor(1)	100.0	97.8	93.0	98.9	43.1	24.6	31.9	27.3	13.0±4	14.4±5	5.4±2	3.7±1
LIME(2)	89.9	72.9	-	-	78.5	63.1	-	-	37.8±20	24.4±7	-	-
Anchor(2)	87.4	95.8	-	-	62.3	45.4	-	-	10.5±3	19.2±10	-	-

Table 5: **Results of the User Study.** Underline: significant w.r.t. anchors in the same dataset and same number of explanations. Results show that users consistently achieve high precision with anchors, as opposed to baselines, with less effort (time).

Anchors lead to much higher average precision

Measured by number of instances users made predictions (very confident)

Users found anchors easier to use. Also were better able to judge when they could make accurate prediction

Limitations

- Predictions near boundary of decision function may have very complex anchors
  - Require very specific sufficient conditions, so low coverage
  - Does not generalize well to other instances
- Anchors predict different outcomes on same test instance - *"conflicting"*
  - Unlikely due to high-probability precision guarantee
  - Submodular pick favors anchors with low overlap
- Complex output spaces
  - Experiments focussed on explaining functions of output, not full output
  - Example: Multi-label classifier
    - Explanation for each label might be overwhelming with lots of labels
    - Rules may be too complex if set of labels is considered as an entity

Future work

- Realistic perturbation distribution
  - Local perturbation distribution expressive enough to reveal model’s behavior
  - Resulting components/rules must be interpretable
  - Still an active line of research