**Anchors: High-Precision Model-Agnostic Explanations paper by Ribeiro et al., 2018**

**Central contribution of the paper**

This paper introduces a local[[1]](#footnote-1), model agnostic[[2]](#footnote-2) and post-hoc[[3]](#footnote-3) interpretability method, related to the research area, explainable AI. This research intends to address downfalls of Its’ predecessor research, LIME[[4]](#footnote-4) by Ribeiro et al,2016 which models the local behavior around some point in a linear way to generate explanations. Instead of fitting surrogate models as in LIME, this method utilizes rule-based conditions called anchors. The anchors work in a way, when the conditions hold true in the input, the predictions are same most of the time regardless of feature values; by design it is intended to attain higher precision. As conditions are well defined in a discrete manner and easy to comprehend, method has extremely clear coverage; means that the explanation can be applied on unseen instances in the region; this addresses unclear coverage in LIME.

Authors explains exploration based multi-armed bandit (MAB; KL-LUCB37[[5]](#footnote-5)) problem in reinforcement learning as an approach to, construct best anchor. The paper further explains anchor generation as a bottom-up construction which initialize with an empty rule, and later extend to search over a space of potential anchors. Best anchor in each MAB search is identified by highest estimated precision and extends the rest with the selected best anchor; this continues until it exceeds the predefined precision threshold. This greedy approach tries to find the shortest anchor which guarantees higher coverage with more interpretability. However as greedy algorithm is only able to maintain a single rule at a time and not directly concerned with the coverage of the anchors, the method has been extended to perform a beam-search.

**Related literature**

As mentioned before this research was published by same authors of LIME and expected to address its’ shortcomings. It mainly addresses unclear coverage and performs well when original model predictions are non-linear or complex, where LIME tends to fail as it only fits linear surrogate models.

LORE[[6]](#footnote-6) proposed by Guidotti et. al,2018 takes a similar approach to interpret models. There are two main differences. LORE generates perturbation samples using a genetic algorithm, in two sets, using two fitness functions. This perturbation sampling approach expects to overcome, the problem of not being data-agnostic nature of model-agnostic models. Then, instead of MAB, LORE trains a decision tree using the two sets of samples to extract a logic rule for interpretation.

QLIME[[7]](#footnote-7) by Bramhall et. al,2020 expands on the linear limitations of LIME by fitting non-linear relationships using a quadratic approximation. However, this method also suffers from unclear coverage. When expanding local region, higher complexities need to be captured and accordingly the order of the explainer need to be increased into higher ordered polynomials.

**Reason for selecting particular publication**

I have deployed LIME before to interpret a ML model which was designed to predict credit default probabilities of a lending portfolio. Before implementing the ML model, it was expected to inspect how the lending decision changes for different applicants, based on their local behavior. I studied the concept of local, model agnostic interpretations and as well as the theory behind the LIME. It helped me to grasp the follow up research on Anchors, which addresses downfalls of LIME. And I recently studied MAB problem, in reinforcement learning, and was curious to know how it has been applied to construct Anchors.

**Strengths and Weaknesses of the publication**

Strengths:

*\*As Anchor was introduced as an improved approach for LIME, most of its’ strengths can be pointed over LIME. Please refer attached Jupyter notebooks where I performed both Anchors and LIME for same datasets with same examples for tabular and text data.*

1. Easy to interpret the rules against the outcome, rather than explaining the effect of the relative importance of variables in LIME
2. Performs well when the model predictions are non-linear (LIME only fits linear models) or closed to decision boundary
3. Highly efficient with underlying MAB algorithm.

|  |  |  |  |
| --- | --- | --- | --- |
| **Data type** | **Model used** | **Time taken** | |
| **Anchors** | **LIME** |
| Tabular | Random forest | 438 ms | 468 ms |
| Text | Naive Bayes | 1.41 s | 527 ms |

Table : time taken for my examples (with fewer data); efficiency might be apparent with many data

1. Guarantees a higher precision as best anchor in each MAB search is identified by highest estimated precision.

|  |  |  |
| --- | --- | --- |
| **Data type** | **Model used** | **Precision** |
| **Anchors** |
| Tabular | Random forest | 95% |
| Text | Naive Bayes | 100% |

Table : precision of Anchors for my examples

1. Extremely clear coverage; relevance of an explanation to other instances is determined based on whether the rule is covering the new instance or not.

Experiment codes:

Weaknesses:

1. Requires all features to be discrete and they pre-process any continuous features into quartile bins by default, which can result in loss of information.
2. These model-agnostic methods are not data-agnostic. The perturbation samples must be estimated from a joint distribution of the source data.
3. Hyperparameters like the precision threshold or the beam width need to be configured properly to yield meaningful results.
4. Overly complex output spaces may lead to non-intuitive explanations.

**Relevance of this publication for neurocat, or more precisely for the AI quality tool aidkit**

One of key focus areas of AI quality tool aidkit is, understanding the risk associated with black-box AI models. Before accessing the risk, the model and its local/global behaviors need to be understood first by the team and end users. It is not much simple for end-users, to understand the other widely used model agnostic approaches for interpretability like LIME, SHARP; where Anchors come in handy as set of IF THEN conditions. Neurocat may serve SLA (service-level agreement) customers where model is already in production and original developers are not around to explain the model. In this scenario, Anchor would be an ideal method for understanding the model predictions at a glance with no detailed explanations, with minimum time and effort. Other than simplicity, in terms of comprehensibility Anchors outperforms existing model agnostic approaches with higher precision, coverage and efficiency as well.

1. replicate original model’s behavior in the vicinity of a selected instance [↑](#footnote-ref-1)
2. treat the original model as a black box, hence re-usable for any model [↑](#footnote-ref-2)
3. interpretation using reverse engineering [↑](#footnote-ref-3)
4. **L**ocal **I**nterpretable **M**odel-agnostic **E**xplanations [↑](#footnote-ref-4)
5. algorithm works by constructing confidence regions based on KL divergence; each step, it selects best mean and highest upper bound [↑](#footnote-ref-5)
6. **Lo**cal **R**ule-based **E**xplanations [↑](#footnote-ref-6)
7. **Q**uadratic **L**ocal **I**nterpretable **M**odel-Agnostic

   **E**xplanation Approach [↑](#footnote-ref-7)