Master’s Thesis Presentation

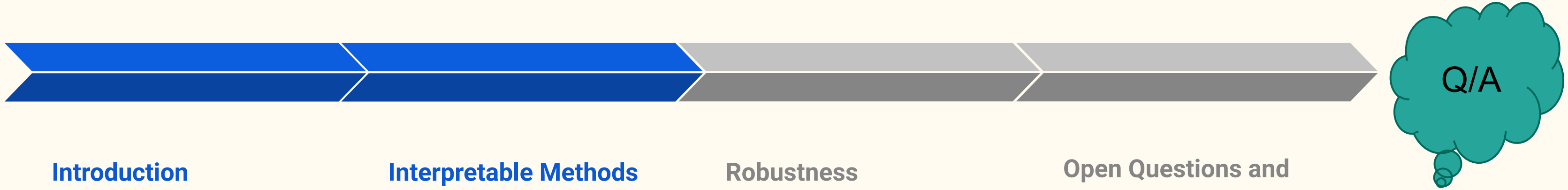
Model Agnostic Local Interpretable Methods And An Empirical Notion Of Robustness

By Rohit Singh

Chennai Mathematical Institute
(June 2020)

Internal Guide: Prof. Madhavan Mukund
External Guide: Prof. K.V. Subrahmanyam

Outline Of Presentation



Introduction

Background details, Motivation

Interpretable Methods

Specifically Local Interpretable Methods

Robustness

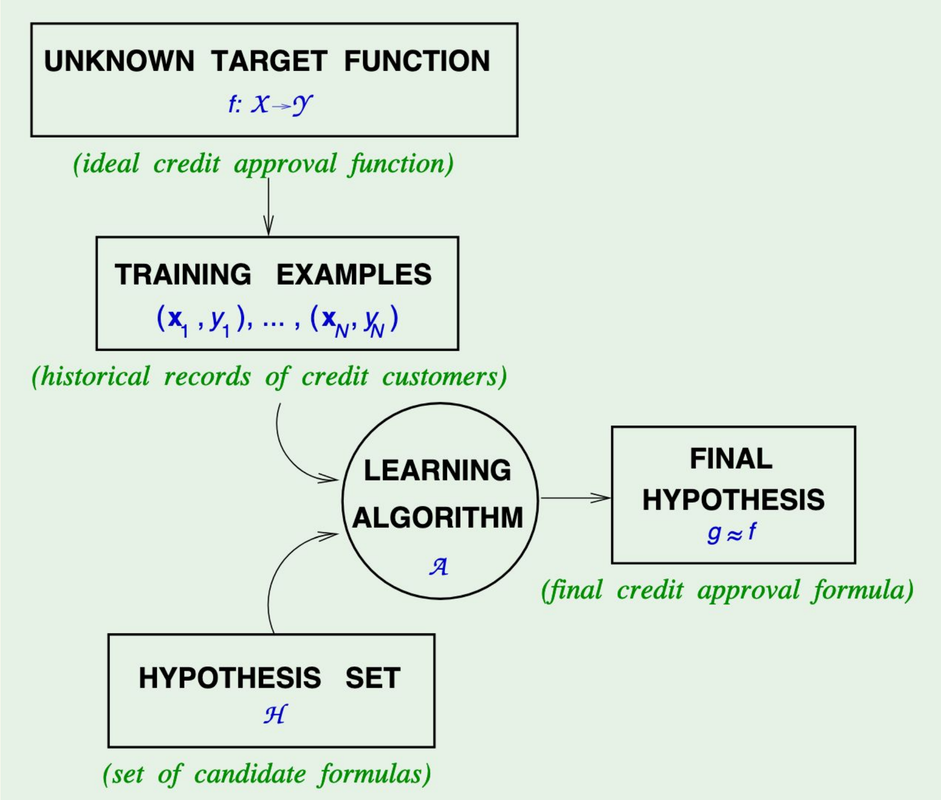
Motivation, Previous Works, Our Work, Experiments

Open Questions and Conclusion

A brief introduction to ML

<u>The essence of Machine Learning:</u>

- A pattern exists.
- We do not have a mathematical formula.
- We have sample data (examples) illustrating the pattern.


UNKNOWN TARGET FUNCTION
$f: \mathcal{X} \to \mathcal{Y}$
(ideal credit approval function)
TRAINING EXAMPLES
$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$
(historical records of credit customers)
LEARNING ALGORITHM
$\mathcal{A}$
FINAL HYPOTHESIS
$g \approx f$
(final credit approval formula)
HYPOTHESIS SET
$\mathcal{H}$
(set of candidate formulas)

Example:

Credit score of a person.

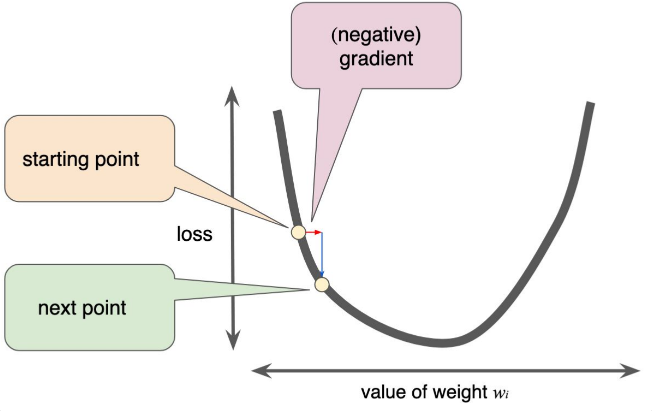
Hypothesis set,
Linear Functions.

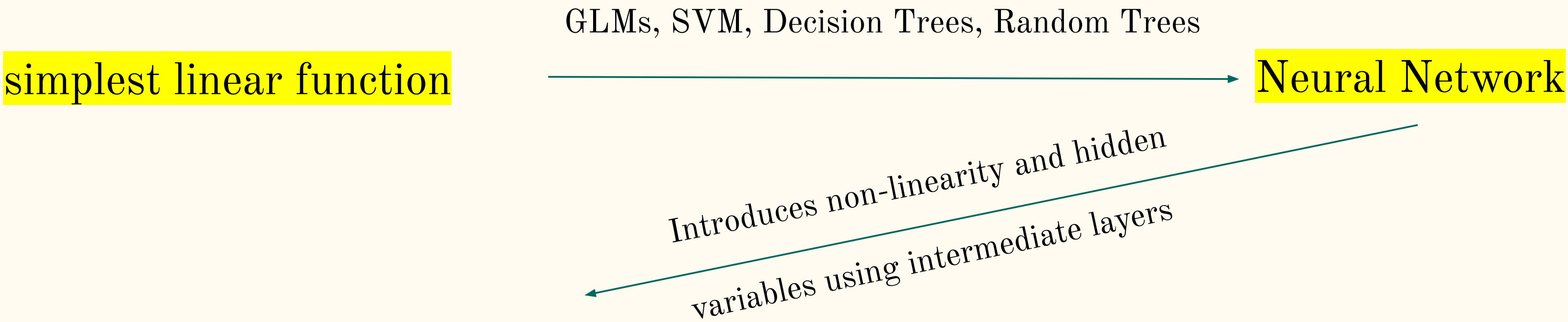
$$h(\mathbf{x}) = \mathbf{w}^{\mathsf{T}}\mathbf{x}$$

Loss Function

$$E_{\text{in}}(h) = \frac{1}{N}\sum_{n=1}^{N}\left(h(\mathbf{x}_n) - y_n\right)^2$$

Learning Algorithm,
Gradient Descent



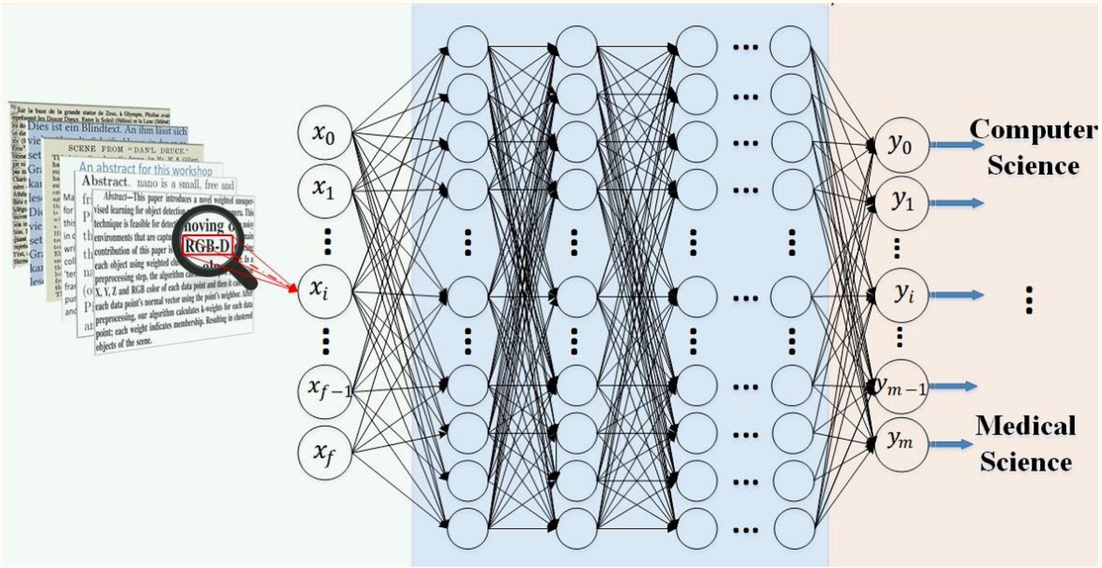


simplest linear function

GLMs, SVM, Decision Trees, Random Trees

Neural Network

Introduces non-linearity and hidden variables using intermediate layers



- **Simple linear function**
  - Can interpret weights.
- **Neural Network**
  - What do the weights mean?

Neural networks can learn very complex classification functions, but they behave like black boxes.

What is the meaning of Explanation/Interpretation in machine learning?

Artifacts either visual or textual that can help the user (a domain expert or layman) to understand the prediction of machine learning model.

Why should we care?

- In case of high-stakes decisions making
- A nice research problem

In case of high-stakes decision making

Case study 1 On May 7th 2016 at 3:40 p.m. on U.S. in Williston, Florida, 45-year-old man was killed when his Tesla Model S went under the trailer of a truck and the roof of his car was torn off by the impact.
Post-facto diagnosis suggested that the autopilot failed to detect white truck against bright sky, and radar system failed to detect the same.

Case study 2 ML is used to interpret medical scans. A qualified doctor should be able to interpret machine learning model output in terms of his expert knowledge of the disease. For instance, “This image indicates TB because of these black patches”.

A nice research problem

The objective of understanding the behaviour (or some part of it) of a complex and human level accurate mathematical model can be a motivating goal. In order to achieve this we need to explain ML models.

- Other task like loan approval can also use explanation to tell why a customer’s loan got denied.
- The above reasons gives us enough motivation to pursue the domain that deals with explaining the ML Models.

Classification Of Explainable/Interpretable Methods

1. **Intrinsic vs Post-hoc Interpretability**

   In intrinsic interpretability we use models that are inherently interpretable like decision trees or put interpretable constraint on the model. In post-hoc interpretability is creating a new model to explain the original one.

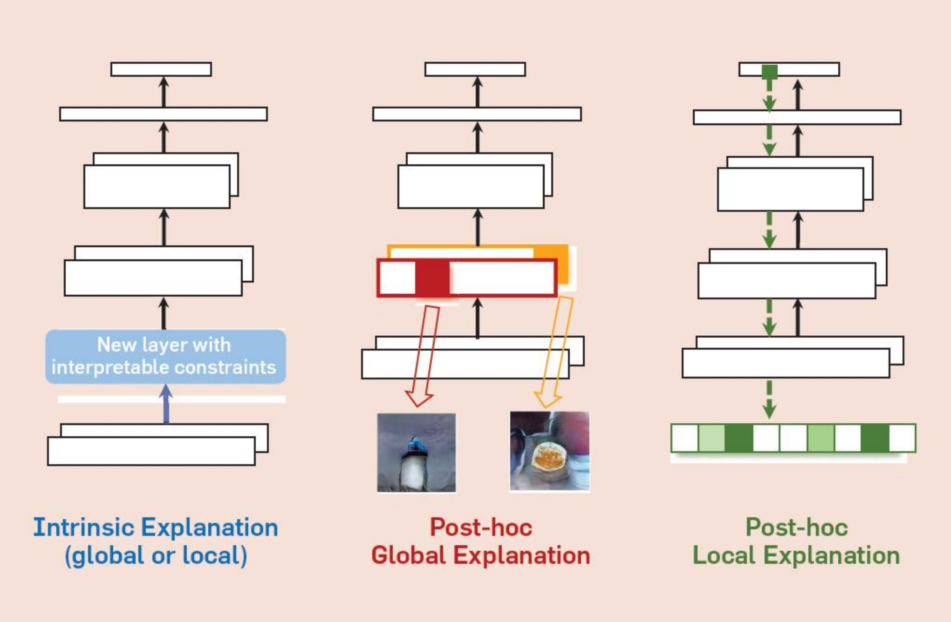
2. **Global vs Local Interpretability**

   Does the interpretation method explain an individual prediction or the entire model behavior?

3. Model agnostic vs Model based Interpretability

Model-specific interpretation tools are limited to specific model classes. Model-agnostic tools can be used on any machine learning model and are applied after the model has been trained (post hoc).

NOTE: These classifications aren't disjoint.


New layer with interpretable constraints
Intrinsic Explanation (global or local)
Post-hoc Global Explanation
Post-hoc Local Explanation

An illustration of three lines of interpretable machine learning techniques, taking DNN as an example.

Bottlenecks to post-hoc global interpretability and a way to get around using Local Interpretable Methods.

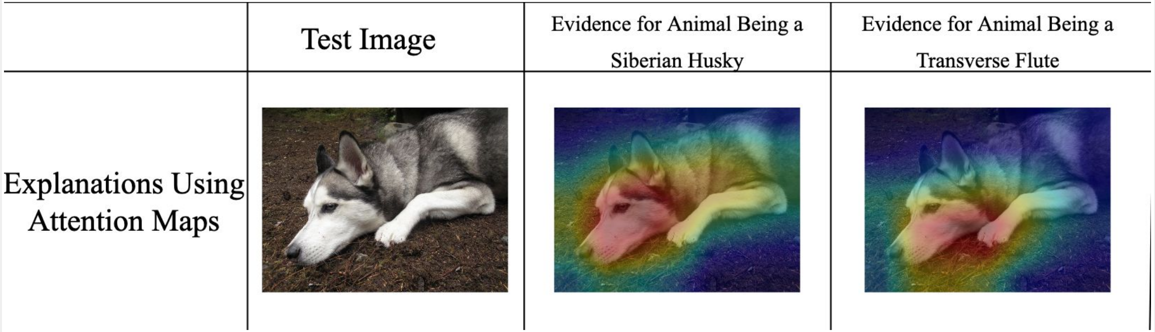
Here are some issues raised by Cynthia Rudin:

- Explainable ML methods provide explanations that are not faithful to what the original model computes.

  An explainable model that has a 90% agreement with the original model indeed explains the original model most of the time. However, an explanatory model that is correct 90% of the time is wrong 10% of the time. If a tenth of the explanations is incorrect, one cannot trust the explanations, and thus one cannot trust the original black box.

Bottlenecks to post-hoc global interpretability and a way to get around using Local Interpretable Methods.

- Explanations often do not make sense or do not provide enough detail to understand what the Black Box is doing.



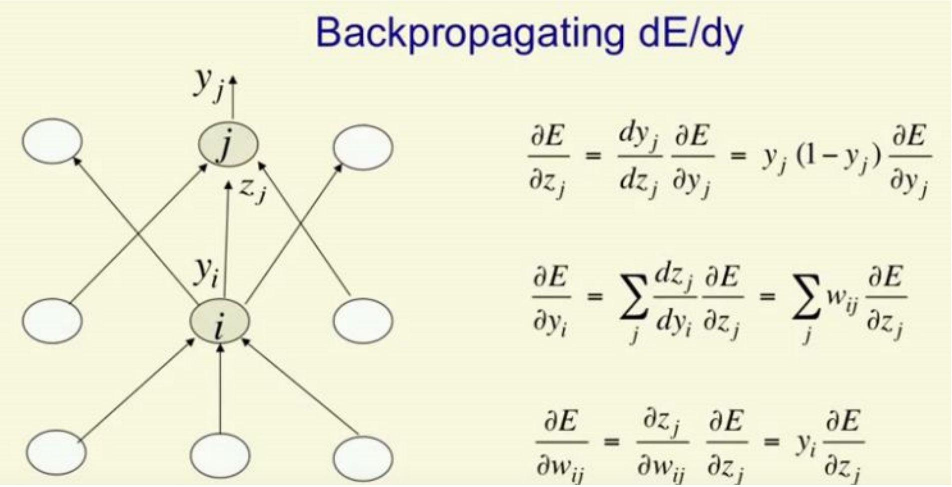
	Test Image	Evidence for Animal Being a Siberian Husky	Evidence for Animal Being a Transverse Flute
Explanations Using Attention Maps			

Local Interpretable Methods

- We try to provide explanations at the level of individual outcomes rather than for the model as a whole.
- It is useful to perceive local interpretable methods as the problem of assigning an Attribution value, to each input feature of a network.

Model Based Local Interpretable Methods

These are backpropagation based methods that compute the attributions for all input features in a single forward and backwards pass through the network, though,sometimes several of these steps are necessary, but the number does not depend on the number of input features.

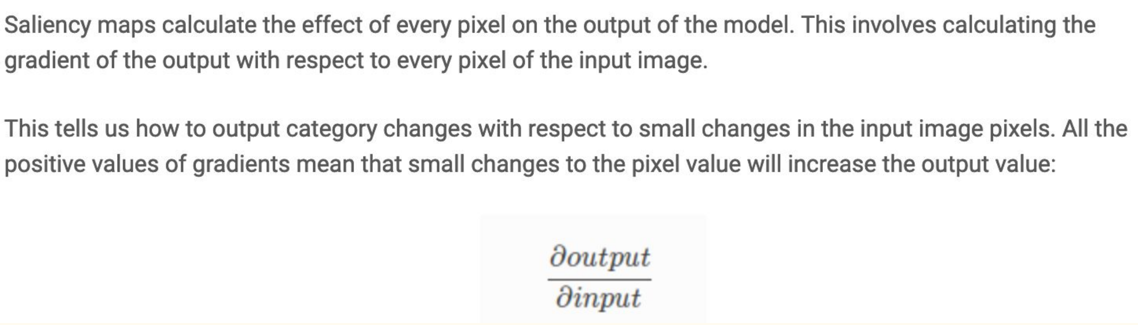


Backpropagating dE/dy

$$\frac{\partial E}{\partial z_j} = \frac{dy_j}{dz_j}\frac{\partial E}{\partial y_j} = y_j\,(1-y_j)\,\frac{\partial E}{\partial y_j}$$

$$\frac{\partial E}{\partial y_i} = \sum_j \frac{dz_j}{dy_i}\frac{\partial E}{\partial z_j} = \sum_j w_{ij}\frac{\partial E}{\partial z_j}$$

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial z_j}{\partial w_{ij}}\frac{\partial E}{\partial z_j} = y_i\frac{\partial E}{\partial z_j}$$

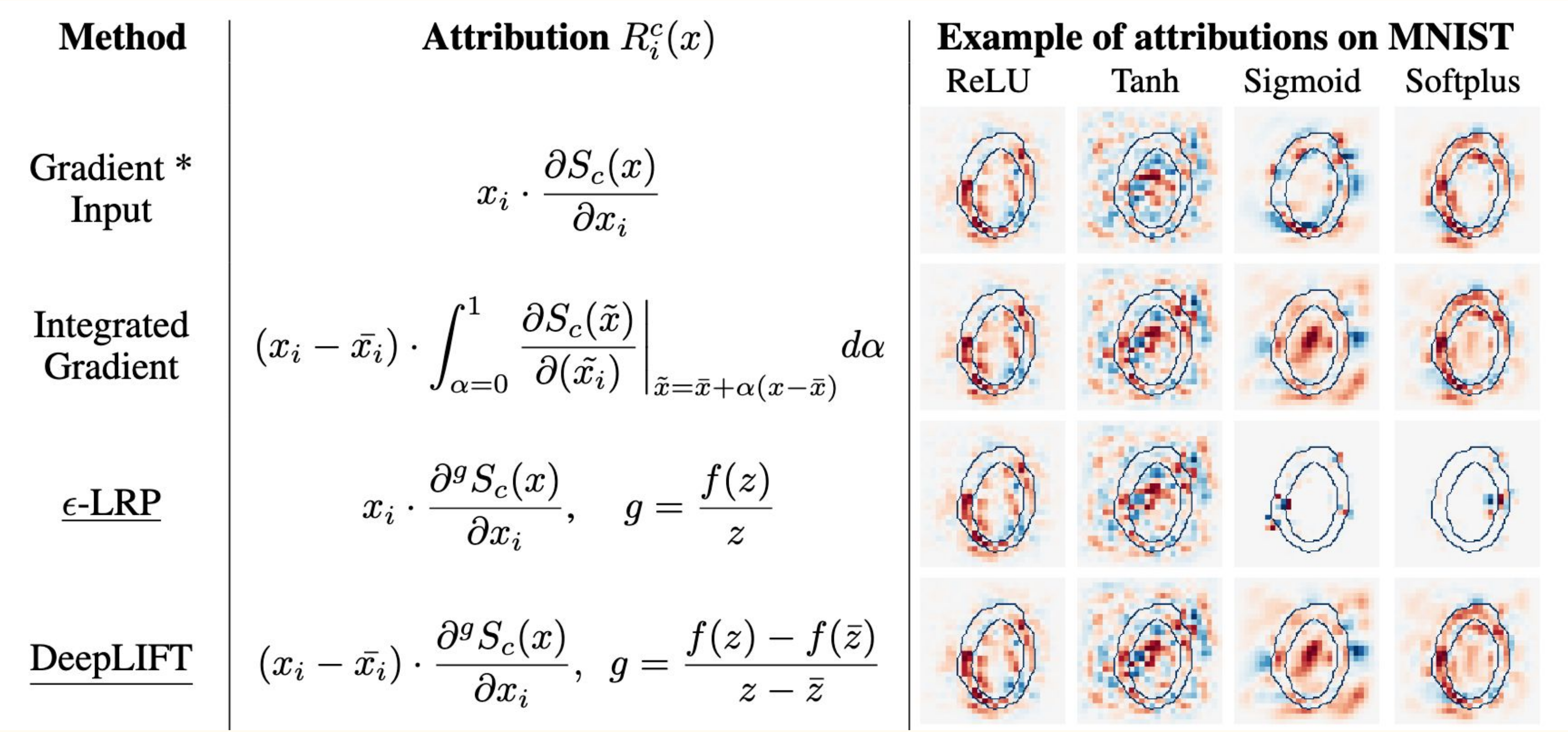


Saliency maps calculate the effect of every pixel on the output of the model. This involves calculating the gradient of the output with respect to every pixel of the input image.

This tells us how to output category changes with respect to small changes in the input image pixels. All the positive values of gradients mean that small changes to the pixel value will increase the output value:

$$\frac{\partial output}{\partial input}$$

Gradient Based Local Interpretable Methods

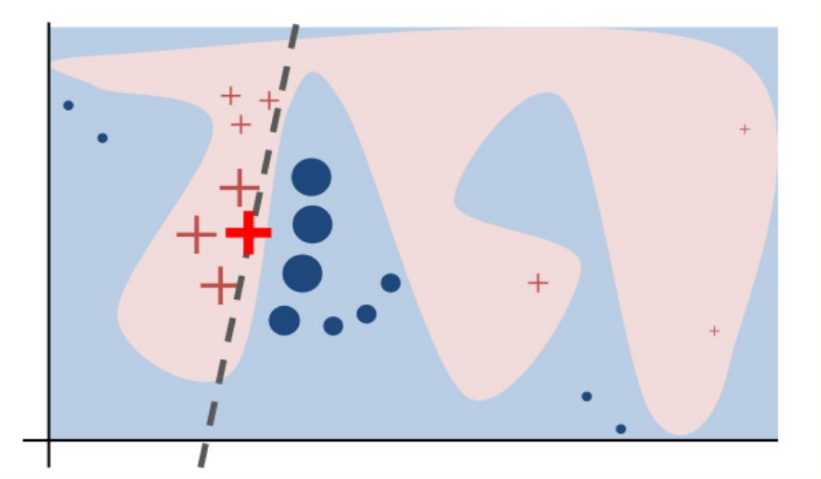


Method	Attribution $R_i^c(x)$	Example of attributions on MNIST			
		ReLU	Tanh	Sigmoid	Softplus
Gradient * Input	$x_i \cdot \frac{\partial S_c(x)}{\partial x_i}$				
Integrated Gradient	$(x_i - \bar{x_i}) \cdot \int_{\alpha=0}^{1} \frac{\partial S_c(\tilde{x})}{\partial (\tilde{x_i})}\Big\vert_{\tilde{x}=\bar{x}+\alpha(x-\bar{x})} d\alpha$				
ϵ-LRP	$x_i \cdot \frac{\partial^g S_c(x)}{\partial x_i}, \quad g = \frac{f(z)}{z}$				
DeepLIFT	$(x_i - \bar{x_i}) \cdot \frac{\partial^g S_c(x)}{\partial x_i}, \; g = \frac{f(z) - f(\bar{z})}{z - \bar{z}}$				

Model Agnostic, Perturbation Based Method (LIME)

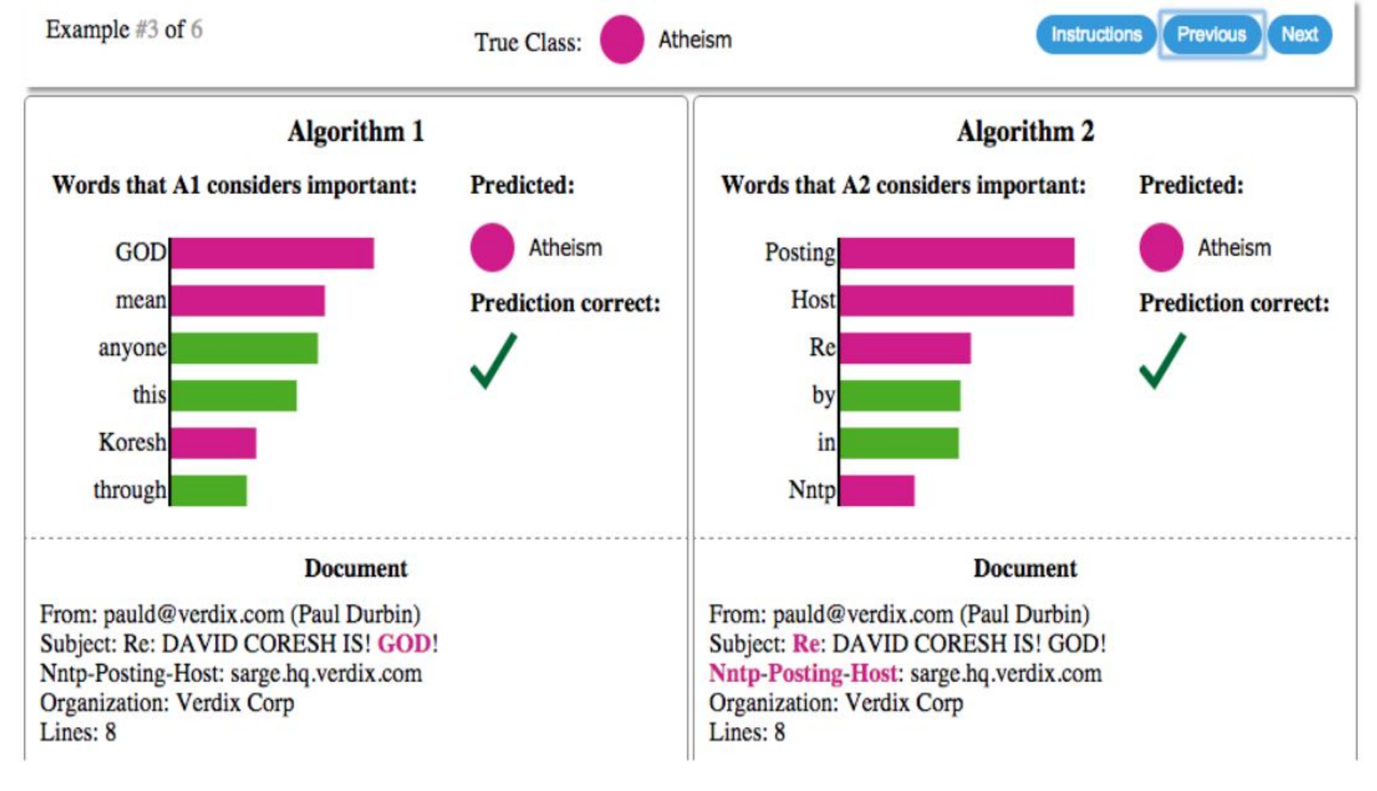
The recipe for LIME:

- Select your instance of interest for which you want to have an explanation of its black box prediction.
- Perturb your dataset and get the black box predictions for these new points.
- Weight the new samples according to their proximity to the instance of interest.
- Train a weighted, interpretable model on the dataset with the variations.
- Explain the prediction by interpreting the local model.



Toy example to present intuition for LIME.

LIME...



Example #3 of 6

True Class: ● Atheism

Algorithm 1

Words that A1 considers important:

Predicted: ● Atheism

Prediction correct: ✓

Document

From: pauld@verdix.com (Paul Durbin)
Subject: Re: DAVID CORESH IS! GOD!
Nntp-Posting-Host: sarge.hq.verdix.com
Organization: Verdix Corp
Lines: 8

Algorithm 2

Words that A2 considers important:

Predicted: ● Atheism

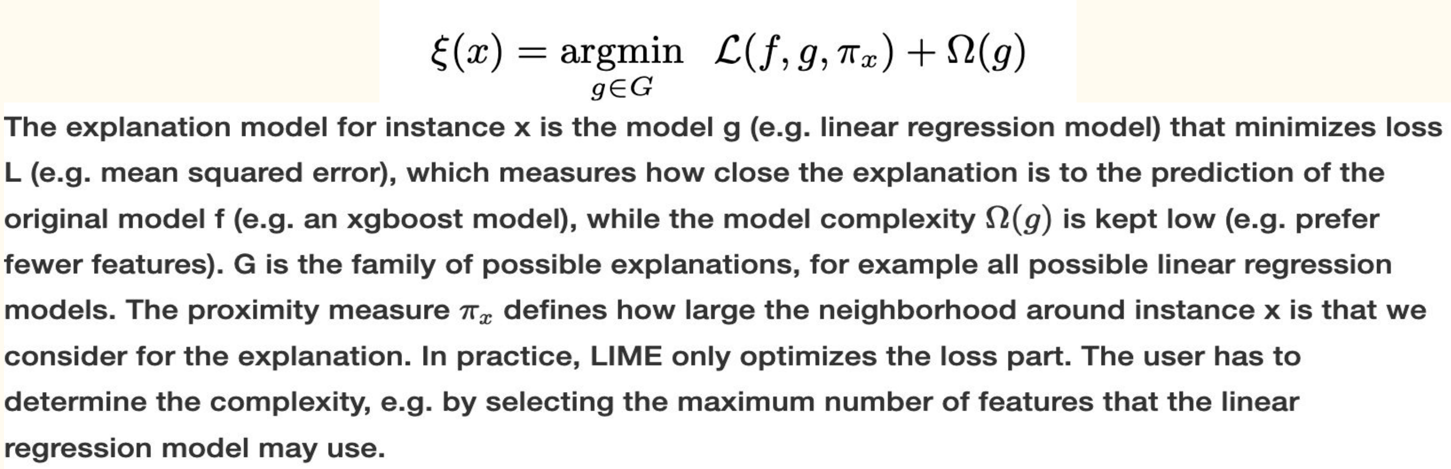
Prediction correct: ✓

Document

From: pauld@verdix.com (Paul Durbin)
Subject: Re: DAVID CORESH IS! GOD!
Nntp-Posting-Host: sarge.hq.verdix.com
Organization: Verdix Corp
Lines: 8

Explaining individual predictions of competing classifiers trying to determine if a document is about “Christianity” or “Atheism”.

LIME…



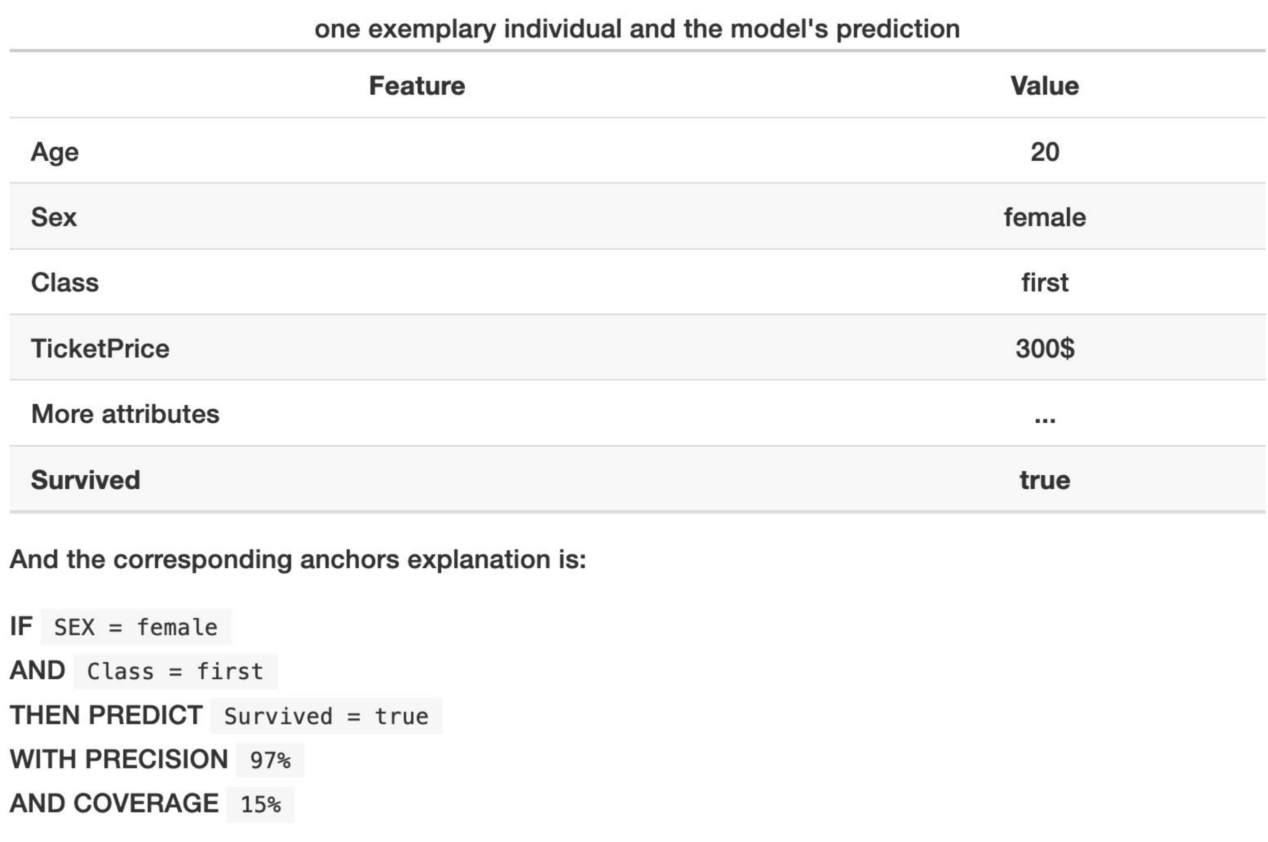
$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \ \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

The explanation model for instance x is the model g (e.g. linear regression model) that minimizes loss L (e.g. mean squared error), which measures how close the explanation is to the prediction of the original model f (e.g. an xgboost model), while the model complexity $\Omega(g)$ is kept low (e.g. prefer fewer features). G is the family of possible explanations, for example all possible linear regression models. The proximity measure π_x defines how large the neighborhood around instance x is that we consider for the explanation. In practice, LIME only optimizes the loss part. The user has to determine the complexity, e.g. by selecting the maximum number of features that the linear regression model may use.

LIMITATIONS:

- Requires a perturbation data set.
- Too many parameters to optimize in order to get good explanation.

Perturbation Based (Anchor)



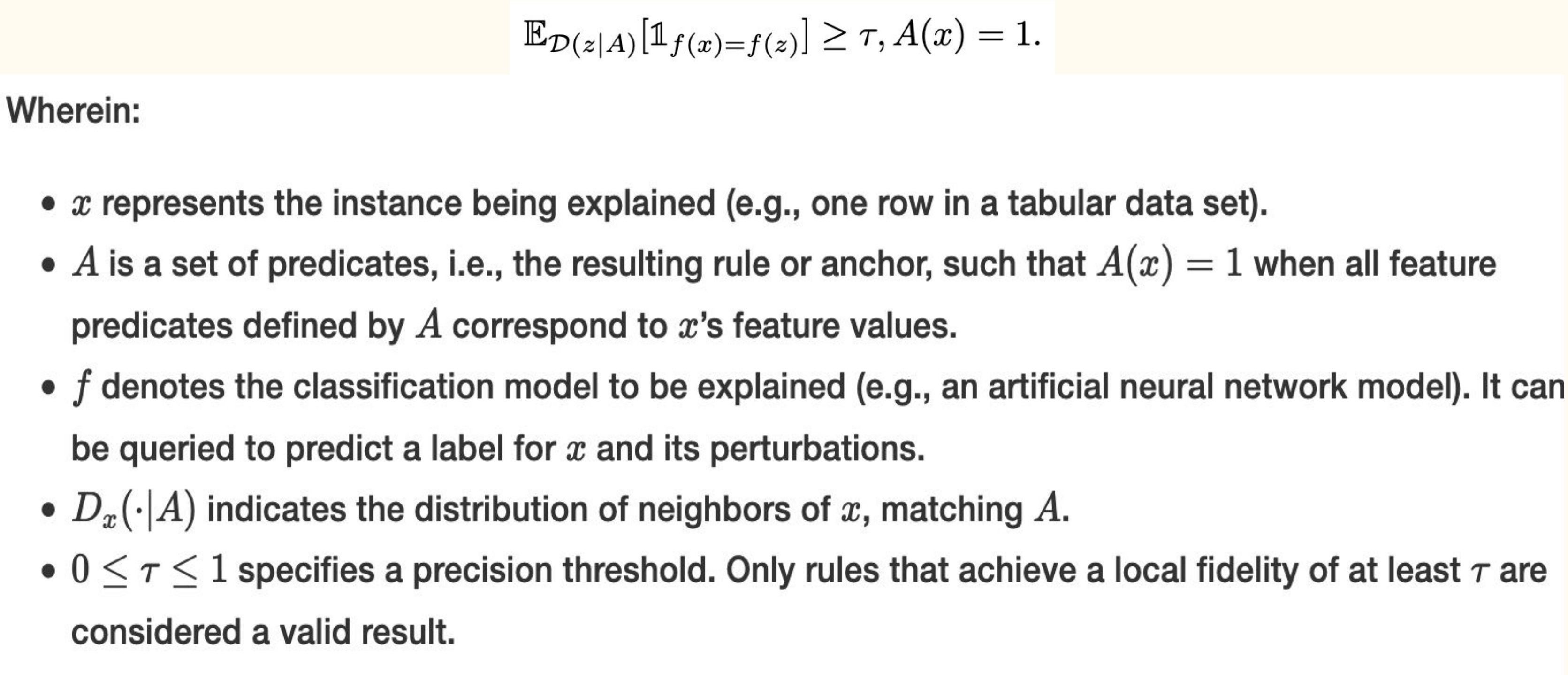
one exemplary individual and the model's prediction

Feature	Value
Age	20
Sex	female
Class	first
TicketPrice	300$
More attributes	...
Survived	**true**

And the corresponding anchors explanation is:

IF `SEX = female`
AND `Class = first`
THEN PREDICT `Survived = true`
WITH PRECISION `97%`
AND COVERAGE `15%`

Anchor…



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

Wherein:

- x represents the instance being explained (e.g., one row in a tabular data set).
- A is a set of predicates, i.e., the resulting rule or anchor, such that $A(x) = 1$ when all feature predicates defined by A correspond to x's feature values.
- f denotes the classification model to be explained (e.g., an artificial neural network model). It can be queried to predict a label for x and its perturbations.
- $D_x(\cdot|A)$ indicates the distribution of neighbors of x, matching A.
- $0 \leq \tau \leq 1$ specifies a precision threshold. Only rules that achieve a local fidelity of at least τ are considered a valid result.

Anchor...

The previous equation can be framed as an optimisation problem

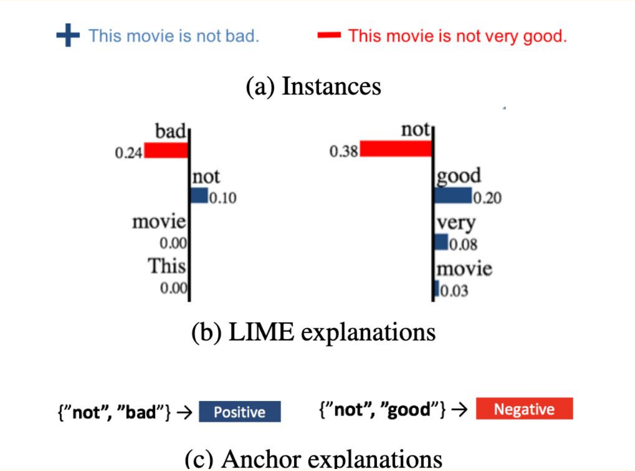
$$P(prec(A) \geq \tau) \geq 1 - \delta \quad \text{with} \quad prec(A) = \mathbb{E}_{\mathcal{D}_x(z|A)}[1_{f(x)=f(z)}]$$

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

$$\max_{A \text{ s.t. } P(prec(A)\geq\tau)\geq 1-\delta} cov(A)$$

LIMITATIONS:

- Similar to LIME, there are too many parameters to tweak.
- The existing methods to solve the above optimisation problem are slow specially for Image Data.

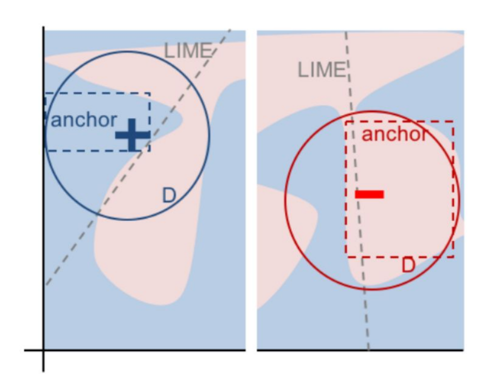


(a) Instances

(b) LIME explanations

(c) Anchor explanations

Sentiment predictions LSTM



LIME vs. Anchors – A Toy Visualization

Robustness of local interpretable methods

Robustness

- Similar inputs should produce similar explanations.
- There are many ways to formalize it.

Why robustness?

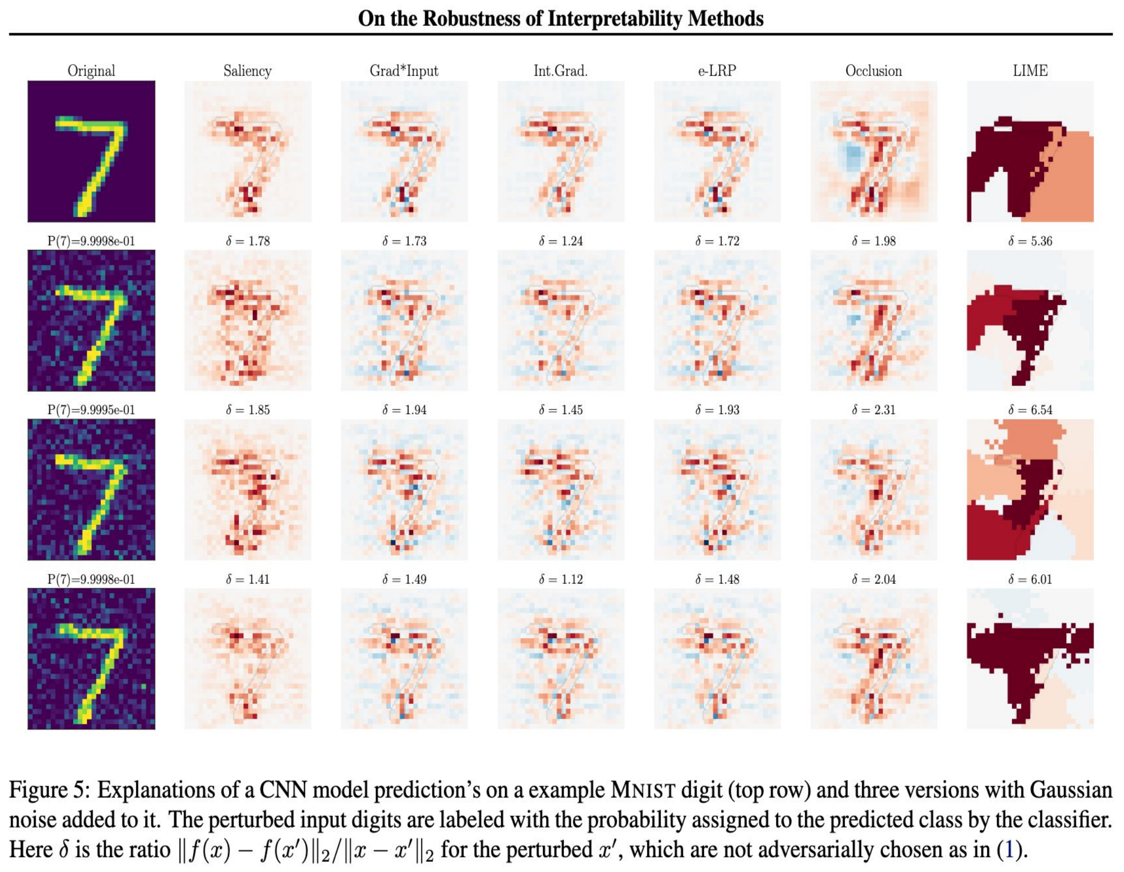
- In order for an explanation to be valid around a point, it should be constant in the close vicinity or should change marginally.

- if we seek an explanation that can be applied in a predictive sense around the point of interest as described above,then robustness of the simplified model implies that it can be approximately used in lieu of the true complex model,at least in a small neighborhood.

Previous work (by: David Alvarez-Melis, Tommi S. Jaakkola, 2018)

The authors notion of robustness concerns with the variations of a prediction's "explanation" with respect to changes in the input leading to that prediction.

Intuitively, if the input being explained is modified slightly—subtly enough so as to not change the prediction of the model too much—then we would hope that the explanation provided by the interpretability method for that new input does not change much either.



On the Robustness of Interpretability Methods

Figure 5: Explanations of a CNN model prediction's on a example MNIST digit (top row) and three versions with Gaussian noise added to it. The perturbed input digits are labeled with the probability assigned to the predicted class by the classifier. Here δ is the ratio $\|f(x) - f(x')\|_2/\|x - x'\|_2$ for the perturbed x', which are not adversarially chosen as in (1).

Our Formulation

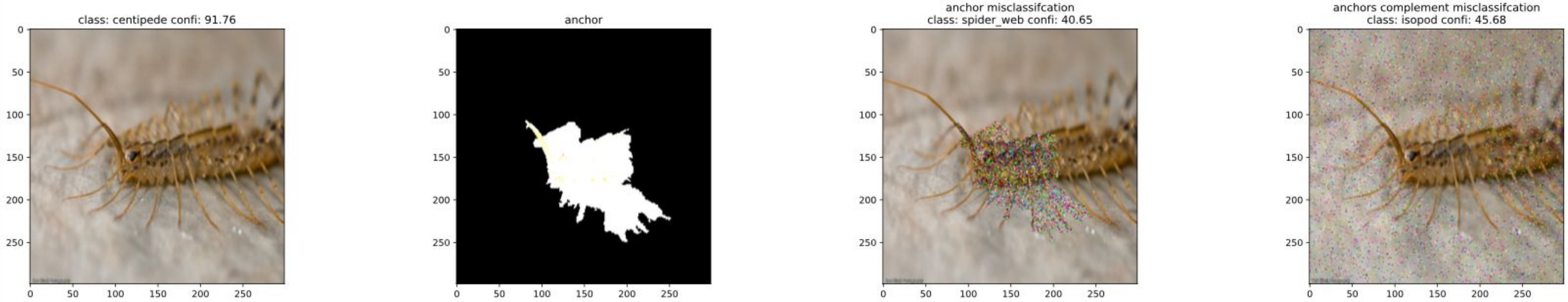
Using the concept of anchor (set of important features), we make the following observations.

1. Intuitively, if the anchor indeed depicts important features in the local neighbourhood (meaning the explanation is robust) then the prediction of the original model should be more volatile to change in the anchor as compared to changes in anchor complement.

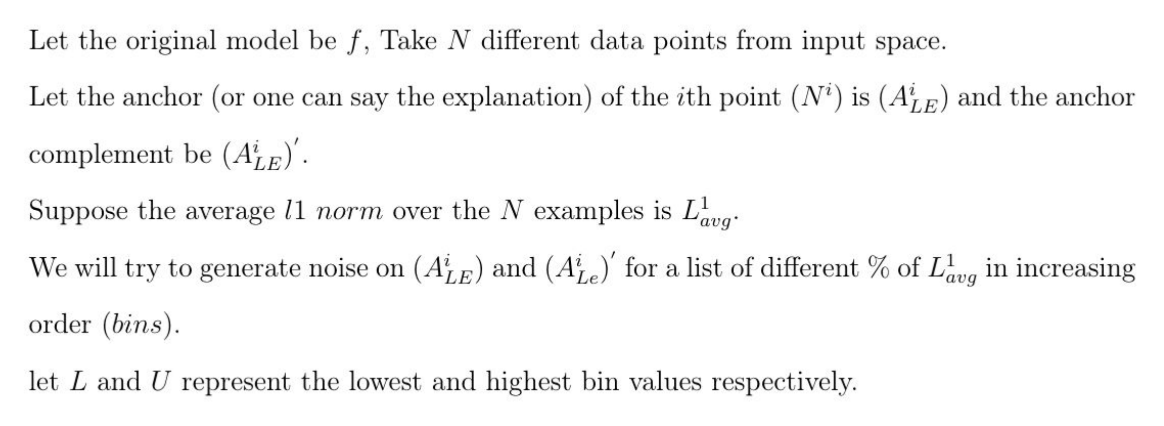
2. If we demand the explanation to be robust we expect that a small change in anchor should bring a small change in original models prediction. Likewise, a small change in anchor complement should bring a relatively bigger change in original models prediction.

Recipe to produce robustness score:

1. Generate a fixed amount of noise (% of L1 avg) to anchor and anchor complement.
2. See if at some point (% of L1 avg) the model’s prediction is different from original.
3. Use these amounts to define robustness_1
4. Use the divergence of original prediction from misclassified anchor and misclassified anchor complement to define robustness_2.
5. Combine these two formalisation (robustness_1 and roubustness_2) to get final robustness score.


class: centipede confi: 91.76
anchor
anchor misclassifcation
class: spider_web confi: 40.65
anchors complement misclassifcation
class: isopod confi: 45.68
0
50
100
150
200
250

An example of getting mis-classification for the method Anchor



Let the original model be f, Take N different data points from input space.

Let the anchor (or one can say the explanation) of the ith point (N^i) is (A^i_{LE}) and the anchor complement be $(A^i_{LE})'$.

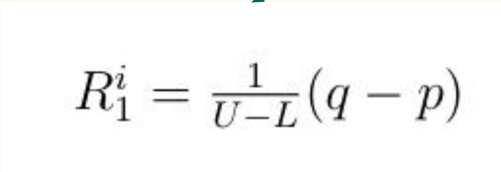
Suppose the average $l1\ norm$ over the N examples is L^1_{avg}.

We will try to generate noise on (A^i_{LE}) and $(A^i_{Le})'$ for a list of different % of L^1_{avg} in increasing order $(bins)$.

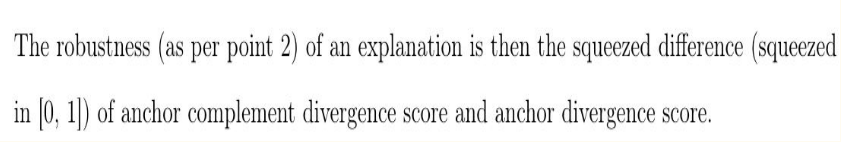
let L and U represent the lowest and highest bin values respectively.

Observation 1

Observation 2

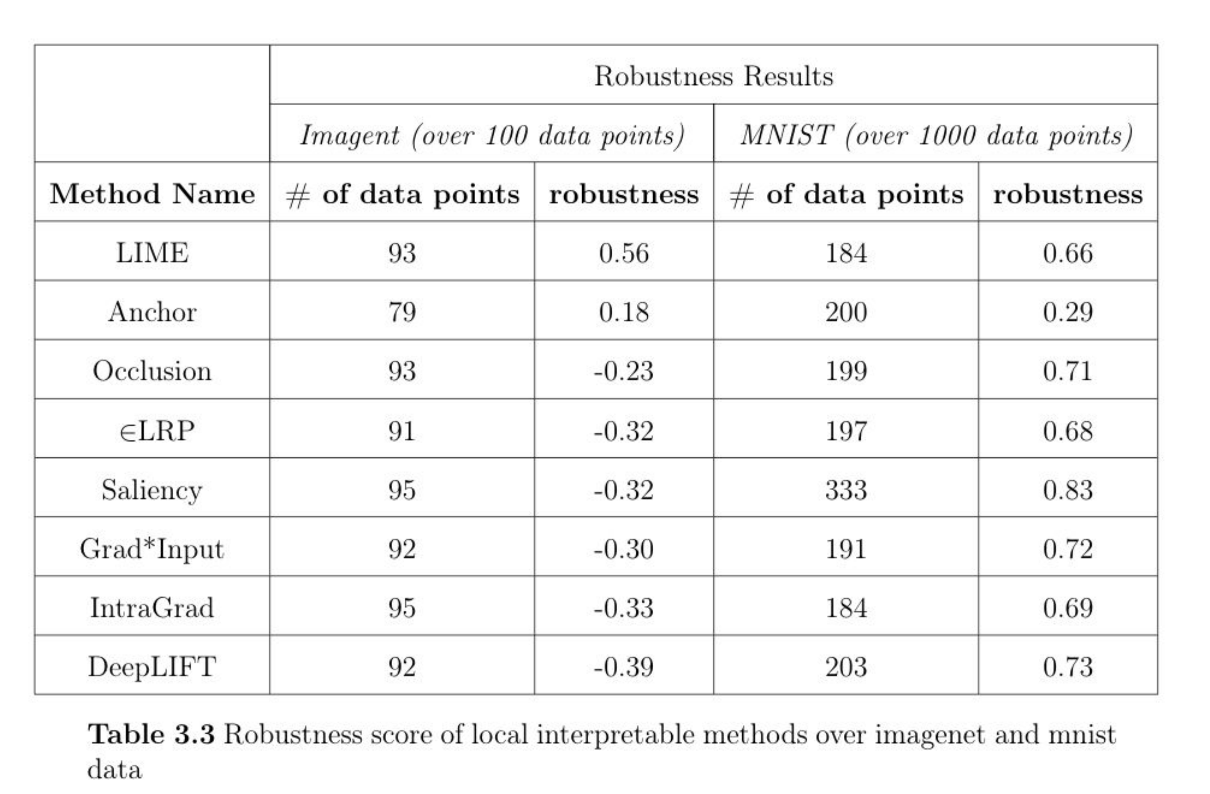


$$R_1^i = \frac{1}{U-L}(q-p)$$



The robustness (as per point 2) of an explanation is then the squeezed difference (squeezed in [0, 1]) of anchor complement divergence score and anchor divergence score.

Experiment results



	Robustness Results			
	Imagent (over 100 data points)		*MNIST (over 1000 data points)*	
Method Name	**# of data points**	**robustness**	**# of data points**	**robustness**
LIME	93	0.56	184	0.66
Anchor	79	0.18	200	0.29
Occlusion	93	-0.23	199	0.71
∈LRP	91	-0.32	197	0.68
Saliency	95	-0.32	333	0.83
Grad*Input	92	-0.30	191	0.72
IntraGrad	95	-0.33	184	0.69
DeepLIFT	92	-0.39	203	0.73

Table 3.3 Robustness score of local interpretable methods over imagenet and mnist data

Conclusion

- The perturbation based local interpretable methods have the potential to be a good explanations.
- Need to tweak the parameters for each dataset.
- Gradient based methods though work faster, but performs badly for high dimensional data.

Open problems

- The perturbation based methods are slow, especially for Image data.
- Additional notions of robustness found in literature would make for interesting complementary evaluation metrics to the one proposed here.

Thank You

Q/A