neurips

explaining machine learning predictions: state-of-the-art, challenges, opportunities

---------------------------------------------------------

model understanding can facilitate bias detection

defendant details -->model-->risky to release

but what if it was just biased by gender and race features (training data for example)

loan applicant details -->model-->denied.

Should provide applicant with more information as to why so it helps!

why model understanding

utility: debugging, bias detection, recourse, if and when to trust predictions, vet models to assess suitability for deployment

stakeholders: end users, decision makers, agencies, researchers

achieving model understanding

if you can build a simple model do it. else : post-hoc explanations

explanation: interpretable description of the model behavior

- should be faithful to the classifier/model

- should be understandable by the user (depends on the user / domain)

-> send parameters? send example predictions? summarize with rule tree? select most important feature / points? describe how to flip model prediction?

local vs global

- global explanation might be too complicated!

local: help unearth biases in the local nieghborhood of a given instance;

help vet if individual predictions are made for the right reasons

global: help shed light on big picture biases affecting larger subgroups;

help vet if the model at a high level is suitable for deployment

---------------------------------------------------------

local explanations

-feature importances

-rule based

-saliency maps

-prototypes/example based

-counterfactuals

global explanations

- collection of local explanations

-representation based

=model distillation

- summaries of counterfactuals

why model agnostic:

- not restricted to specific models

- easy: no library /code base ties

- study models you dont have access to

LIME: sparse linear explanations

sparse: zero out as many features as possible.

allows to identify important dimensions (sparse) and present their relative importance (linear)

on images: create perturbations and see what is the outcome

this creates a new dataset. Run locally weighted regression,

then you get an explanation: important coefficients

parameters

- how to perturb? distance / similarity? how local do you want to be? how to express explanation?

SHAP: shapley avlues as importance

marginal contribution of each feature towards the prediction, averaged over all possible permutations

fairly attributes the prediction to all the features

e..g: 3 features all binary, the contribution of feature 3 for 1 permutation can be tracked by computing the outcome with feature 3 on and with feature 3 off

anchors: sufficient conditions

- identify conditions under which the classifier has the same prediction

if feature 1> feature 2 then outcome = X for example

saliency maps:

which parts are most relevant for the model's prediction

-->heatmap

=feature attribution = heatmap

linear model detour: sensitivity:

how much does a unit change in an input dimension induce in the output?

(derivative of output wrt input)=^weight vector

how can we apportion the output across all the input dimension?

(weight\*input)

input gradient: noisy and difficult to interpret;gradient saturation

smoothgrad: copy your input n times with gaussian noise each time --> average input gradients of noisy inputs

integrated gradients: path integral from baseline to outcome of interest; path integral: sum of interpolated gradients

gradient-input: element wise product of input-logit-gradient and input

modified backprop approaches: compute feature relevance by modifying backpropagation

guided backpropagation: zeroes out negative gradients during backprop iteration (->more sparse)

layer relevance propagation: LRP: compute feature relevance iteratively and propagate. Different propagation rules can be specified

REFERENCE PAPER: Samek and Montavon et al., 2020

Prototype approaches: explain a model with synthetic or natural input "examples"

insights: what kind of input is the model most likely to misclassify? which training samples are mislabelled? which input maximally activates an intermediate neuron?

which training data points have the most influence on the test loss?

influence function: classic tool used in robust statistics for assessing the effect of a sample on regression parameters

-->Cooks distance: instead of refitting model for every data point, cooks distance provides analytical alternative (Cook and Weisberg, 1980)

koh and liang 2017: cook distance extension insight to modern machine learning setting

--> compute self influence to identify mislabelled examples

--> diagnose possible domain mismatch

scalability: meh.

non convexity: meh

--> representer points, yeh et al 2018

-->tracIn, pruthi et al, 2020

activation maximization: which examples strongly activate a neuron / function

either by using natural or synthetic examples

counterfactuals

high stakes decisons-> provide recourses!

=what features need to be changed and by how much to flip a model's prediction = reverse an outcome?

solutions differ on : how to choose among candidate counterfactuals? how much access is needed to the underlying predictive model?

minimum distance between input and counterfactual: can use normalized manhattan distance

feasible and least cost counterfactuals; cost is modeled as total log percentile shift: changes become harder when starting off from a higher percentile value

can also include a set of feasible counterfactuals (end user input)

(REFERENCE PAPER Ustun et al 2019)

but only on linear classifiers... so use LIME first?

changing 1 feature might affect the rest / some other features!

-> can use causally feasible counterfactuals: find features and graph, if feature has no parent: compute something

(requires knowledge of full causal graph. partial knowledge should work fine, but meh )

- data manifold closeness: generated counterfactual should be close to the original data distribution

- sparsity: ideal to change small number of features in the counterfactual

---------------------------------------------------------

global explanations

collection of local explanations

- create local explanation, then pick a subset of them

SP-LIME:

cant pick all explanations->should be representative and diverse

(summarize overal model behavior but shouldnt be redundant)

sp lime uses submodular optimization and greedily picks k explanations

SP-ANCHOR

very similar

representation based :

derive model udnerstanding by analyzing intermediate representations of a DNN

- determine model's reliance on concepts that are semantically meaningful to humans

Deep learning methods [...] and non model agnostic :/

model distillation

process of approximating a given predictive model using simpler interpretable models which mimic the outcomes

generalized additive models as global explanations: output is a set of shape functions for predicting outcome

shape functions +/- PDP but created using GAMs

GAM: shap functions of individual features+ higher order feature interaction terms

fit GAM to predictions of black box to obtain global shap functions

decision sets/trees as global explanations

two level decision set

can customize subgroupds: e.g. if feature A = this and feature B = that then...

and next one is if feature B = other choice then...

desiderata and optimization problem

fidelity: describe model behavior accurately

-->maximize same outcome decision

unambiguity: minimize number of duplicate rules applicable to each instance

simplicity: minimize the number of conditions in rules, constraints on number of rules and subgroups

customizability: outer rules should only comprise of features of user interest

summaries of counterfactual explanations

use rule sets as above

-recourse correctness: minimize number of applicants for whom prescribed recourse does not lead to desired outcome

recourse coverage: minimize number of applicants for who recourse does not exist

minimal recourse cost: minimize total feature costs as well as magnitude of changes in feature values

interpretability of summaries: constraints on number of rules, conditions in rules, and number of subgrops

customizability: outer rules should only comprise of features of stakeholder interest

-->smooth local search algorithm

DIFFERENT DATA MODALITIES

- structured data (tabular)

challenges:

different types of variables can necessitate different similarity or perturbation functions

gradients may not always be meaningful (e.g. if categorical)

depending on domain/task: data could be high or low dimensional

evaluating post hoc explanations:

understand the behavior

- how important are selected features: remove and see outcome probability. or add feature and see outcome probability

- do the same on training data: add / remove influential training data

useful for debugging: all done using end - user feedback so meh

-detect problems in classifier:

- comparing classifiers

- fixing features of classifiers

- finding errors in training data

limitations of evaluating explanations

-evaluation setup is often simple / unrealistic

instances / perturbations create out of domain ponits, bugs are obvious artefacts

- sometimes flawed: model vs human explanations...

- automated metrics can be optimized

- user studies are not consistent

- conclusions are difficult to generalize

-----------------------------------------------------

-faithfulness: do explanations capture model-based discriminative signals?

measure sensitivity to model parameters: if parameter settings change, explanations should change

-fragiility: post hoc explanations can be easlly manipulated

scaffolding attack on LIME and SHAP: hide classifier dependence on feature X

adversarial attack: minimally modify the input with a small perturbation without changing the model prediction

but almost any method can be attacked....

in our case Clinicolab: why would doctors manipulate their data... its their problem

- stability; post-hoc explanations can be unstable to small, non-adversarial, perturbations to the input

alvarez et al., 2018: perturbation methods like LIME can be unstable; Yeh et al., 2019 provide boundaries at wich they become unstable

explanations can be highly sensitive to hyperparameters such as random seed, number of perturbations, etc; bansal, agarwal, nguyen 2020

- model debugging: tabular LIME improves forward and counterfactual simultability (Hase et al., 2020)

ribeiro et al: good explanations;lapuschkin et al 2020: good explanations

useful in practice?

high fidelity explanations can be misleading; Lakkaraju et Bastani 2019; rule based method

in a bail adjudication task, misleading high-fidelity explanations improve end-user (domain expert) trust... even for high fidelity explanations

in a housing price prediction, amazon mechanical turkers are unable to use linear model coefficients to diagnose model mistakes; poursabzi sangdeh et al., 2019

-->end users not able to debug... Sniffles

in a dog breeds classification task, users familiar with ML rely on labels instead of saliency maps for diagnosing model errors; adebayo et al.,

users found natural images more helpful than feature visualization in deciding whether an image strongly activated a neuron; borowski and zimmerman et al., 2020

rigorous user studies and pilots with end users can continue to help provide feedback to researchers on what to address (alqaraawi et al., 2020, bhatt et alk., 2020, kaur et al., 2020)

modeling uncertainty in post hoc explanations

bayesian version of LIME / SHAP with closed form solutions

lime analysis

garreau et al., 2020:

obtained closed form solutions of the average coefficients of the surrogate model (explanation output by LIME)

coefficients obtained are proportional to the gradient of the function to be explained

local error of surrogate model is bounded away from zero with high probability (UH OH)

rigorous evaluation of the utility of post hoc explanations

kaur et al., 2020; lakkaraju et al.; 2020

domain experts and end users seem to be over trusting explanations and the underlying models based on explanations

- law schools students trusted underlying model 9.8 times more when shown a misleading explanation which "white-washes" the model

- data scientists over trusted explanations without even comprehending them - "participants trusted the tools because of their visualizations and their public availability"

ustun et al., 2019; gupta et al., 2019

-it is commonly hypothesized that post hoc explanations can help with detecting model biases;

need for more rigorous theoretical and empirical studies to quantitavely evaluate this hypothesis

-can post hoc explanations help detect unfairness?

intersection with fairness:

fairness shapley values: attribute unfairness in model predictions to individual features using shapley values framework

for different definitions of fairness, it is possible to choose shapley valu functions which explain the unfairness

begley et al., 2020

parting thoughts:

leavitt and morcos, 2020; roth and kearns, 2019

when introducing a new explanation method:

who are the target end users that the method will help

a clear statement about what capability and / or insight the method aims to provide to its end users

careful analysis and exposition of the limitations and vulnerabilities of the proposed method

rigorous user studies (preferably with actual end users) to evaluate if the method is achieving the desired effect

use quantitative metrics (and not anecdotal evidence) to make claims about explainability

