Ribeiro2016 Notes

October 6, 2020

1 Introduction

- Local trust: Trusting specific prediction enough to act on it
- Global trust: Trusting model to provide reasonable predictions over entire real-world input range.
- Validation metrics may not reflect what is actually valued in production
- They propose LIME as a add-on to any black-box prediction algorithm to provide local interpretability/trust
- Global trust/interpretability is built up from multiple, representative instances, selected via SP-LIME.

Can trust in the global performance of a model actually be build up from local trust? What about "unboundedly bad" worst-case predictions?

2 Case for explanations

- Intelligible explanations allows for human prior knowledge of the domain to be brought to bear on the question of trusting the model.
- Model explanations can help to catch problems of data leakage or dataset shift, which may go unnoticed if only validation metrics are used.
- Can be used for "model selection", in that models which underperform on easily-computed metrics may still be judged positively if explanations are provided.

2.1 Desired characteristics for Explainers

- Interpretable: provide qualitative connection between input and response, taking into account user limitations.
- A linear or otherwise transparent model may not be interpretable, if too coefficients are non-zero.

- If the features themselves are too complicated to interpret, the input variables in the explanations may need to be different.
- Local fidelity: explanation should capture well the local behavior of the black-box predictor.

3 Local Interpretable Model-Agnostic Explanations

- Goal of LIME is to produce an easily interpretable model over an interpretable representation of the input, which is locally faithful to the black-box predictor.

3.1 Interpretable data representations

- An interpretable representation of an instance is one where all the covariates are immediately comprehensible. In the paper, all interpretable representations x' are d' dimensional binary vectors, indicating presence or absence of word/super-pixel.
- Distinct form d dimensional real vectors x, which are used as the feature vectors input to the model to be explained.

What representations other than indicators are potentially interpretable?

3.2 Fidelity-interpretability trade-off

- G is class of possibly interpretable models (linear models, trees, rule lists). The input to $g \in G$ is the presence or absence of particular interpretable components, the vector x'.
- Complexity measure $\Omega(g)$ penalizes models in G as they become more difficult to interpret.
- Locality measure $\pi_x(z)$ gives closeness of instance z to instance x. Unfaithfulness loss $\mathcal{L}(f, g, \pi_x)$ measures how well g approximates the full classifier f, weighted by $\pi_x(z)$.
- LIME loss is to minimize over G the unfaithfulness loss $mathcalL$ plus the complexity penalty Ω around a specific sample instance x.

3.3 Sampling for local exploration

- Minimizing the LIME loss is made difficult as f must be treated as black box. Hence we need to approximate $mathcalL(f, g, \pi_x)$.

- This is done via sampling. For the specific instance x to be explained, move to it's interpretable rep x'. Find perturbed sample z' near x' by uniformly sampling non-zero components of x' to include. Map z' back to feature vector z, and obtain full model prediction $f(z)$. Use $f(z)$ as label for z'. Pick number of times to do this, and obtain dataset $\mathcal{Z}$. Minimize LIME loss on this dataset.

3.4 Sparse linear explanations

- Concretely, paper sets G be class of linear models over binary vectors, $\pi_x(z)$ is exponential kernel with distance function dependent on data type (cosine for text, L2 for images).
- Unfaithfulness loss is squared error loss weighted by π_x.
- Complexity penalty is set to be just a limit K on the number of non-zero weights in the explanation g. That is, only K of the indicators can have non-zero weight in the explanation for a specific point x, but which K can change from explanation to explanation.
- Minimizing a l_0 penalty is intractable, so approximate by running LASSO to select K indicators, then refitting for those K indicators with OLS.

Their sampling scheme requires $|\mathcal{Z}|$ forward passes through the full model f for each point to be explained. Also, possibly deeper networks also learn more non-linear decision boundaries and so are harder to faithfully approximate. Does this lead to unfavourable scaling?

How can we ensure that we can map z' sampled around x' back to a feature vector z? Will map $z \mapsto z'$ necessarily be invertible?

4 Submodular pick for explaining models

- The present a "jucidious selection" procedure to pick out a small number of individual instance explanations which exhibit a wide range of the full model's behavior. Goal is to build up global understanding from local explanations. Selection of instances is called pick step.
- For a set of n explained instances, an $n \times d'$ explanation matrix $\mathcal{W}$ is constructed, whose ijth entry is the size of the weight given to indicator j in the explanation of x_i.
- Global importances I_j suggest the overall importance of indicator j. For their text applications, they take the square root of the sum of the jth column in $\mathcal{W}$.
- Goal is to pick out B explanations into set V with best coverage, defined as sum of all the importances I_j where one one of the explanations in V has a non-zero weight for the jth indicator.

- Submodular pick algorithm is greedy forward maximization of coverage function. Each time, add instance to V who most increases the coverage (i.e. has weights on important indicators which haven't already been covered by other instances in V).
- Submodularity is a property of a set function, namely that the incremental effect of a a single element decreases as size of input set increases. Coverage is submodular, as if $V_1 \subset V_2$ then adding a particular instance i increases coverage by sum of the importances of non-zero W_{ij} for i minus the W_{ij} that were covered by the other instances already. That later term is increasing.

5 Simulated user experiments

5.1 Experiment setup

- Datasets used are sentiment analysis of reviews (classification as positive or negative). 2 datasets, books and dvds, 2000 instances each.
- Classification models fit are: decision trees, logistic regression with L2 regularization, nearest neighbors, SVMs with RBF kernels. Features used for that were bag-of-words (each review was processed as a vector of word presence indicators).
- Also fit was a random forest classifier, using average word2vec vectors as features. (Each word in the review was mapped into a high-dimensional dense vector via a pre-learned mapping. The word embeddings were then averaged to get a paragraph embedding.)
- Individual predictions are explained via LIME, via parzen, via a greedy procedure, and randomly. Parzen windows are another term for kernel density estimation. The greedy procedure used takes an instance x, selects and removes features from x which contribute most to the predicted class until the class changes or K features are selected and removed.
- Random pick (where the explained and presented instances are chosen uniformly at random) is compared to submodular pick for the pick step.

5.2 Are explanations faithful to the model

- Faithfulness of explanatory techniques is tested on the classifiers which are themselves interpretable. Logistic regression and decision trees are trained so that max # of features used is 10, called the gold set of features. For each instance in test set, LIME, greedy, parzen, and random are used to identify which features should be used to explain that instance's predicted. The fraction of gold features recovered is then measured and compared in Figure 6.

- LIME performs the best, parzen and greedy perform comparably on explaining logistic regression but greedy performs poorly at explaining decision trees.

5.3 Should I trust this prediction?

- Random set of features (25% of total) is labeled as untrustworthy, and if a prediction on the test set changes when those untrustworthy features is removed, prediction is labelled as untrustworthy.
- For explanation g, if prediction from g changes when interpretable representations of untrustworthy features are removed, then explanation is untrustworthy.
- Comparison of trustworthiness labels from full model and from explanations are compared via $f1$ score. LIME performs the best.

5.4 Can I trust this model

- Noise features are added, which correlate with the label on the training/validation sets but not on the test set.
- Competing pairs of random forest classifiers are trained until they have the same validation accuracy but their test set accuracies differ substantially, indicating that one classifier is relying on the noise features to achieve validation accuracy.
- Explanations of B instances are given, selected either by random pick or submodular pick for either LIME or greedy explainers. Any interpretable representations of noise features which appear in one of the B instances are flagged as untrustworthy. Trustworthiness of instances is then assessed as before, and model which has fewer untrustworthy validation predictions is selected. That choice is compared to which of the two models has the higher test set accuracy.
- They find that LIME generally outperforms greedy on this test, and submodular pick improves upon random pick.

6 Evaluation with human subjects

6.1 Experiment setup

- Classifiers are trained on documents from 20 newsgroups dataset, to identify if document came from "Christianity" or "Atheism" newsgroup. Dataset contains features that don't generalize but are very predictive of document newsgroup, and so hopefully model explanations can help to identify this problem.

- Also produce a real-world "Religion" dataset which also contains documents to classify as about "Christianity" or "Atheism" but don't contain those non-generalizable features. If classifier trained on 20 newsgroup performs well on this test dataset, then it must be using semantic content to classify instead of data-specific issues.
- Full model used is SVM with radial basis function kernel trained on the 20 newsgroup data.

6.2 Can users select the best classifier?

- Human subjects are recruited via Mechanical Turk, and they are tasked to choose between two classifiers: a SVM trained on the original newsgroup dataset, and one trained on a cleaned version of 20 newsgroups where the data-specific issues had been removed.
- To decide between the two, explanations of 6 predictions with 6 non-zero indicator coefficients are presented to the Turkers from each of the two classifiers. Explanations are produced by either greedy or LIME, and the instances explained are selected via either RP or SP. The Turkers are asked to choose which of the two classifiers is most likely to perform well in the real world.
- They find that submodular pick selection of instances substantially improves Turker accuracy in picking the better classifier for both greedy and LIME, and that LIME gives better explanations than greedy.

6.3 Can non-experts improve a classifer?

- They go through several iterations in which explanations from a model are presented to a group of Turkers, who observe explanations of 10 instances. The Turkers mark words which should be removed from being input to the classifier. The classifier is then retrained on a dataset with these words removed.
- The models iteratively improve in performance on the real world "Religion" dataset, indicating that model explanation techniques allow for non-ML experts to aid in feature engineering.

6.4 Do explanations lead to insights?

- They conduct an experiment on images, where 20 images of huskies and wolves are selected, where all the wolf images have snow in the background while pictures of huskies did not. Features for the full model were the activations at the first max-pooling layer from Inception. This is used to train a bad classifier, which only looks at presence of snow.

- Experiment is conducted where experiment subjects were unwitting grad students. Grad students are shown test images along with husky-vs-wolf classifications and asked if they would trust the model to perform well in the real world, why, and how the classifier is reaching it's classifications.
- Then, they are shown the same images with explanations. Far fewer trust the classifier afterwards, and far more identify the possibility of snow as a spurious feature. Therefore, the explanations allowed the students to gain insight into the workings of the model.