Chapter 2 - Interpretability

1 Introduction

Interpretability (or explainability) is the degree to which a human can understand why a model made a particular decision. Alternatively, it is the degree to which a human can predict a model's result.

2 Importance of Interpretability

Interpretability helps us understand why a model made a particular prediction. This is useful in debugging the model, gleaning insights into how the model works (and thus how it can be improved), discovering model weaknesses, complying with regulations, removing prejudice from models, and providing explanations to customers.

Models that explain their predictions may be more socially acceptable and offer a better user experience than black box models.

Interpretability allows us to examine the following properties of the model: fairness (model should not be prejudiced), privacy (model is not exposing or using sensitive information), reliability/robustness (small changes in input should not cause big changes in output), and causality (model uses causal features for prediction).

We don't need interpretability for hobby projects, for well understood problems (e.g. optical character recognition), or when malicious actors might try to game the model (e.g. a security model that explains how it detects bad actors might give them hints on how to circumvent it).

3 Taxonomy of Interpretability Methods

Interpretability can be intrinsic (the model is simple, like a decision tree) or post-hoc (i.e. we use some algorithm to extract interpretations). Here are the possible outputs of an interpretation method.

A feature summary statistic is a number about a feature or feature pair. A feature summary visualization (e.g. partial dependence plot) shows the interpretation visually. Model internals like the splits of a decision tree or visualization of the filters learned in a convolutional neural net can also provide interpretability. The output could be a data point, as in class prototypes or the counterfactual method. Finally, the output could be an intrinsically interpretable model that approximates the original black box model.

Interpretation tools can be model specific or model agnostic. They can explain an individual prediction, a local group of predictions, or the global set of all predictions.

4 Scope of Interpretability

Algorithm transparency is about how a particular learning algorithm learns from data. For example, we understand how least squares works and roughly how a convolutional neural net works (i.e. successive

layers learn edges, shapes, parts, and objects). We don't understand how complicated neural nets or gradient boosted trees learn though.

Global, holistic model interpretability is about how a trained model makes decisions in terms of its features and weights. Which features are important? How do they interact? This is basically impossible to do if you have more than 2 or 3 features.

Global model interpretability on a modular level is about understanding a model by looking at its pieces (e.g. one decision tree in a random forest). Note that a linear model's weights cannot be interpreted by considering just one weight, you need to consider multiple features and their weights.

Local interpretability for a single prediction is about understanding why a particular prediction was made and what features were important. LIME works well here.

Local interpretability for a group of predictions is about understanding a group of predictions. You can either run a global method on the group or a local method on each prediction and aggregate.

5 Evaluation of Interpretability

One option is application level evaluation. Just deploy the method and see if users indicate that they like it. Alternatively, you can have some experts try to concoct explanations and see if your method produces similar explanations.

Another option is human level evaluation. Get some volunteers, show them a bunch of examples each with different possible explanations, and ask them to pick the best ones.

A third option is function level evaluation. If your explanations take the form of a decision tree, you could ensure that the decision trees are not too deep.

6 Properties of Explanations

Explanations can come in many forms. They may be text or a visualization that ties feature values to the prediction. They may be a set of data points (e.g. in k nearest neighbors). Or they may take the form of simple, intrinsically interpretable models like decision trees.

The important properties of explanation methods are expressive power (how rich is the "language" used to generate explanations?), translucency (how much does the method depend on looking inside the model?), portability (how many different kinds of models does the explanation method work with?), and algorithmic complexity (how much compute time is needed to create the explanation?).

For individual explanations, we have some important properties. Accuracy is how well the explanation predicts unseen data. If the explanation, is "The house was priced as expensive because it has many rooms", then this should be true of other homes too. Fidelity is how well the explanation reflects what is actually going on in the model. Fidelity can be local or global. Consistency is how similar explanations are between two different models trained for the same task. Note that if two models arrive at the same result using different features, it may actually be possible that both are correct and could have different explanations (Rashomon Effect). Stability is how similar explanations are for similar examples (i.e. slight changes in feature values should not change explanation a lot). Comprehensibility is how well humans understand the explanations (e.g. is the explanation short? can people predict what the explanation will be for a new data point?). Certainty is whether the explanation indicates the confidence level of the model (assuming the model outputs a score instead of a label). The degree of importance is whether the explanation shows how important each part of the explanation is. Novelty is whether the data point being considered is very different from typical data points. Representativeness is how many examples the explanation covers (e.g. a global explanation covers all of them).

7 Human-friendly Explanations

A good explanation has just one or two causes and contrasts the current situation with one where the prediction would have been different.

An explanation answers a "why" question.

Explanations are contrastive, so they should explain why a different prediction was not made (e.g. why was this loan application not accepted?). It should explain the difference between the current example and some reference example.

Explanations are selective, so pick out the one or two key causes.

Explanations are social, so make sure your audience will understand them.

Explanations focus on the abnormal, so point out what makes the current data point unique. If a feature value is atypical, point that out.

Explanations are truthful, so make sure that it reflects what is going on in the model.

Explanations are consistent with user's beliefs. For example, people know that larger houses are more expensive, so your explanation should not contradict that.

Explanations are general and probable, so they should explain other similar situations (unless the current data point is actually an abnormal case). You can measure this by feature support, which is the fraction of instances to which the explanation applies.