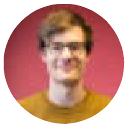
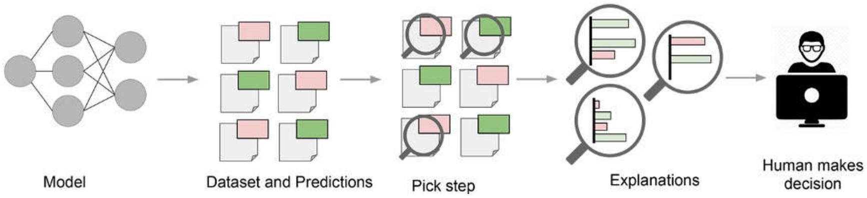


Lars Hulstaert Follow
Data Scientist at Microsoft. Previously Masters student at Cambridge, Engineering student in Ghent. I like connecting the dots.
Jul 11 · 5 min read

Understanding model predictions with LIME

In my previous post on model interpretability, I provided an overview of common techniques used to investigate machine learning models. In this blog post, I will provide a more thorough explanation of LIME.

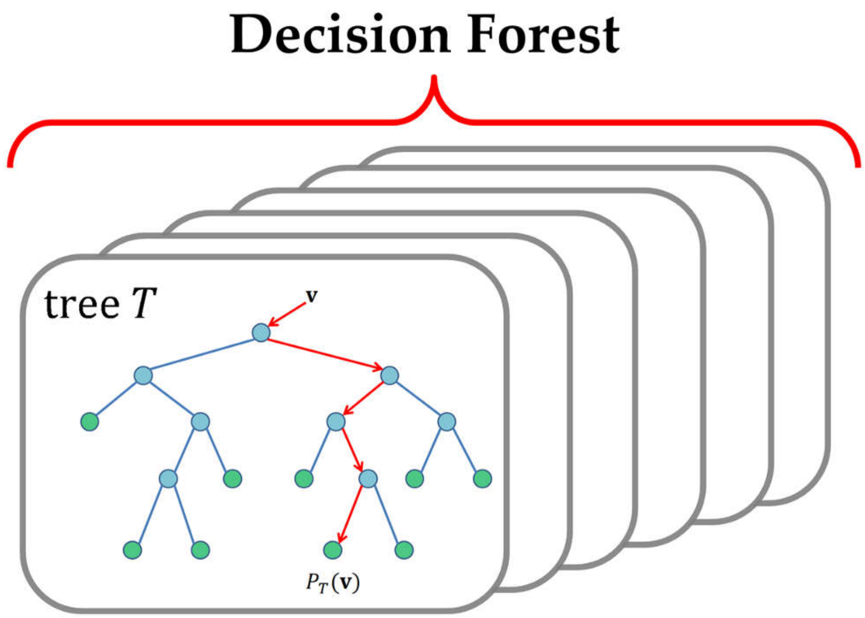


LIME explaines of model predictions at the data sample level. It allows end-users to interpret these predictions and take actions based on them. Source

Why is it necessary to understand interpretability methods?

If you trust a technique with explaining the predictions of your model, it is important to understand the underlying mechanics of that technique, and any potential pitfalls associated with it. Interpretability techniques are not fault proof, and without a good understanding of the method, you are very likely to base your assumptions on falsehoods.

A similar but significantly more thorough investigation was done in the following blog post on random forest importance's. Feature importance is often used to determine which features play an important role in the model predictions. Random forests provide an out-of-the-box method to determine the most important features in the dataset and a lot of people rely on these feature importance's, interpreting them as a 'ground truth explanation' of the dataset.

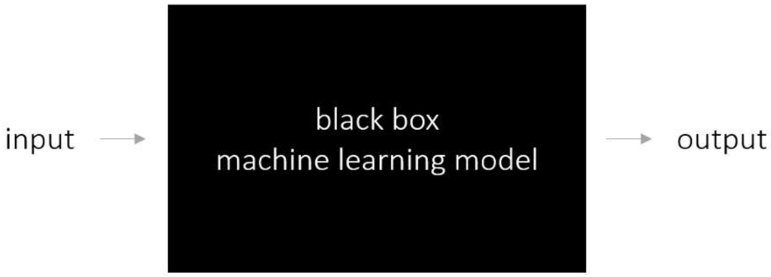


A decision or random forest consists of multiple decision trees. By investigating which features are used to construct the 'best' trees, it is possible to get an estimate of the feature importance. Source

The authors investigated two random Forest (RF) implementations and the standard measures of feature importance they provide. The authors show that permutation importance provides more robust estimates when variables are strongly correlated, compared to random forest importance's. I highly recommend to read their blog post for a thorough understanding of the findings.

LIME

LIME is model-agnostic, meaning that it can be applied to any machine learning model. The technique attempts to understand the model by perturbing the input of data samples and understanding how the predictions change.

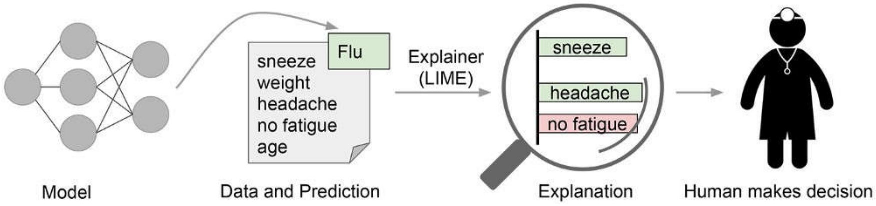


LIME assumes a black box machine learning model and investigates the relationship between input and output, represented by the model.

Model-specific approaches aim to understand the black model machine learning model by analysing the internal components and how they interact. In deep learning models, it is e.g. possible to investigate activation units and to link internal activations back to the input. This requires a thorough understanding of the network and doesn’t scale to other models.

LIME provides local model interpretability. LIME modifies a single data sample by tweaking the feature values and observes the resulting impact on the output. Often, this is also related to what humans are interested in when observing the output of a model. The most common question is probably: why was this prediction made or which variables caused the prediction?

Other model interpretability techniques only answer the question above from the perspective of the entire dataset. Feature importance’s explain on a dataset level which features are important. It allows you to verify hypotheses and whether the model is overfitting to noise, but it is hard to diagnose specific model predictions.

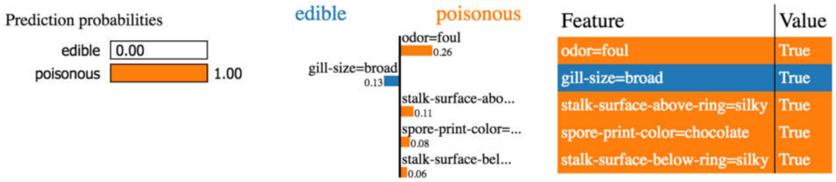


LIME attempts to play the role of the ‘explainer’, explaining predictions for each data sample. Source

Intuition behind LIME

A key requirement for LIME is to work with an interpretable representation of the input, that is understandable to humans. Examples of interpretable representations are e.g. a BoW vector for NLP, or an image for computer vision. Dense embeddings on the other hand or not interpretable, and applying LIME probably won't improve interpretability.

The output of LIME is a list of explanations, reflecting the contribution of each feature to the prediction of a data sample. This provides local interpretability, and it also allows to determine which feature changes will have most impact on the prediction.



Feature	Value
odor=foul	True
gill-size=broad	True
stalk-surface-above-ring=silky	True
spore-print-color=chocolate	True
stalk-surface-below-ring=silky	True

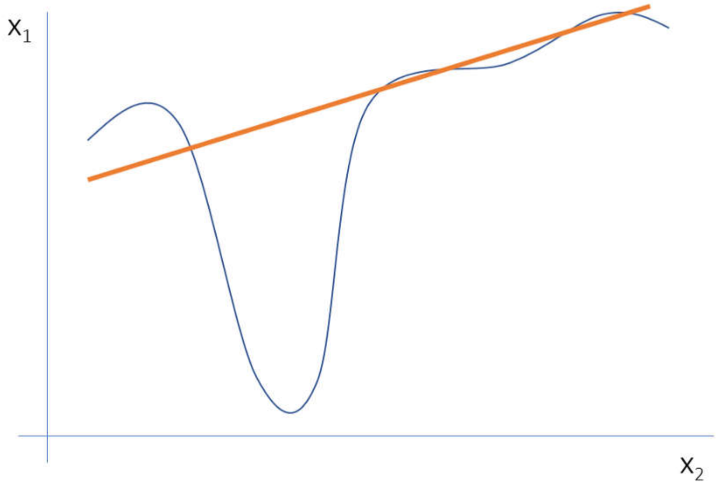
An example of LIME applied to a classic classification problem. Source

An explanation is created by approximating the underlying model locally by an interpretable one. Interpretable models are e.g. linear models with strong regularisation, decision tree's, etc. The interpretable models are trained on small perturbations of the original instance and should only provide a good local approximation. The 'dataset' is created by e.g. adding noise to continuous features, removing words or hiding parts of the image. By only approximating the black-box *locally* (in the neighborhood of the data sample) the task is significantly simplified.

Potential pitfalls

Although the general idea of LIME sounds easy, there are a couple of potential drawbacks.

In the current implementation, only linear models are used to approximate local behaviour. To a certain extent, this assumption is correct when looking at a very small region around the data sample. By expanding this region however, it is possible that a linear model might not be powerful enough to explain the behavior of the original model. Non-linearity at local regions happens for those datasets that require complex, non-interpretable models. Not being able to apply LIME in these scenario's is a significant pitfall.



A linear approximation of the local behaviour for two features is not a good representation and won't capture the highly non-linear behaviour of the model.

Secondly, the type of modifications that need to be performed on the data to get proper explanations are typically use case specific. The authors gave the following example in their paper:
For example, a model that predicts sepia-toned images to be retro cannot be explained by presence or absence of super pixels.

Often, simple perturbations are not enough. Ideally, the perturbations would be driven by the variation that is observed in the dataset. Manually steering the perturbations on the other is probably not a great idea, as it most likely would introduce bias into the model explanations.

Conclusion

LIME is a great tool to explain what machine learning classifiers (or models) are doing. It is model-agnostic, leverages simple and understandable idea's and does not require a lot of effort to run. As always, even when using LIME, it is still important to correctly interpret the output.

If you have any questions on interpretability in machine learning, I'll be happy to read them in the comments. Follow me on Medium or Twitter if you want to receive updates on my blog posts!