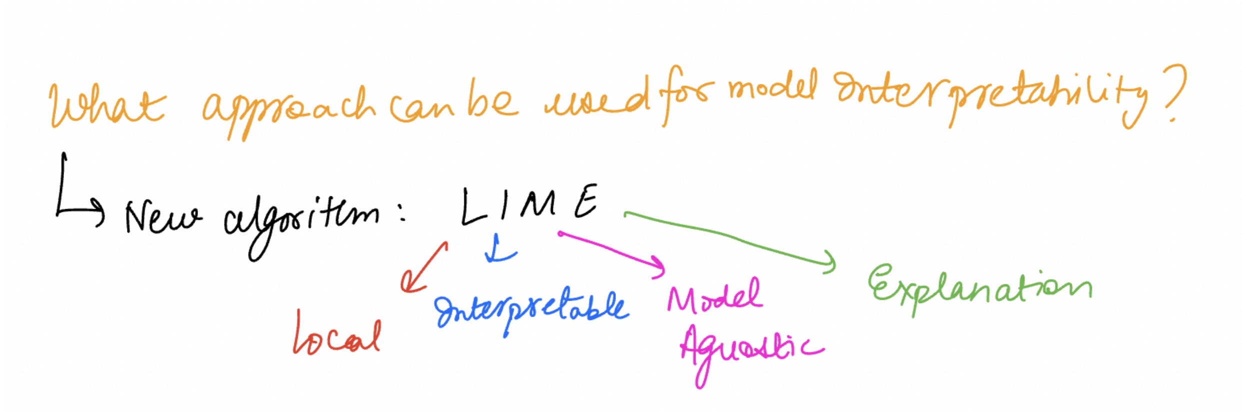
NN: Model interpretability: LIME

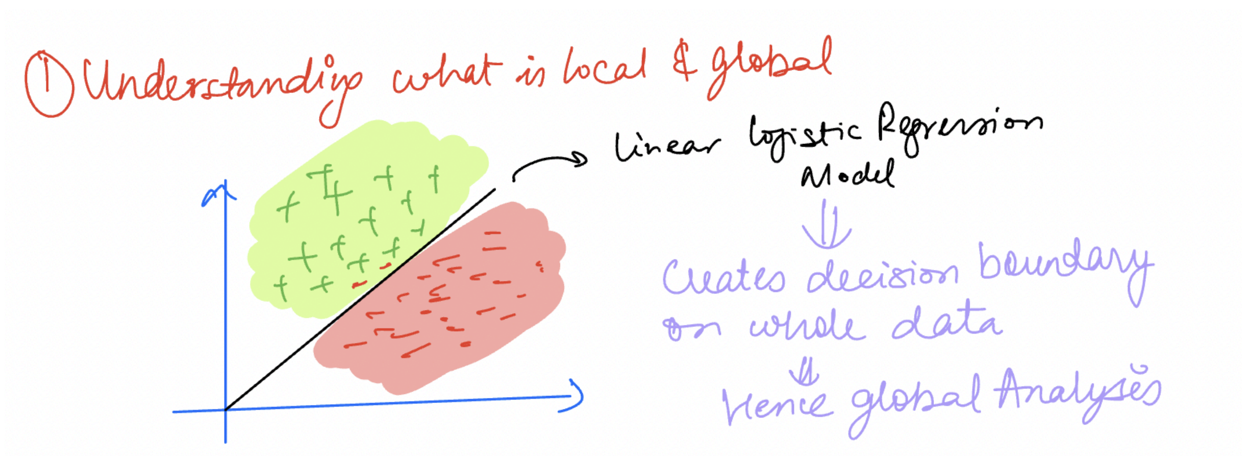
LIME

How to make interpretability for a Complex model like NN?



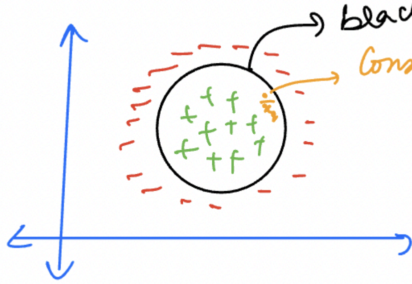
- Lime (Local Interpretable Model-agnostic Explanations) is used to explain predictions made by machine learning models.
- Lime is a popular technique for interpreting individual predictions of black-box models.
- It approximates the local behavior of a model by creating a local interpretable model, which helps us understand why a model makes a specific prediction for a given instance.

Understanding Global Analysis



Understanding Local Analysis

Now, instead of entire dataset
we look at a particular datapoint



→ black box Model

Considering datapoint very close to boundary & making prediction only for this datapoint
⇓
Hence called Local Analysis

In Local Analysis:

↳ finding → $\boxed{\text{what is special about } \bar{x}_q \text{ point \& how varying it causes model predictions to change}}$

↳ called as Local Interpretability

Understanding locality of interest

How to vary $\bar{x}_q$ datapoint?

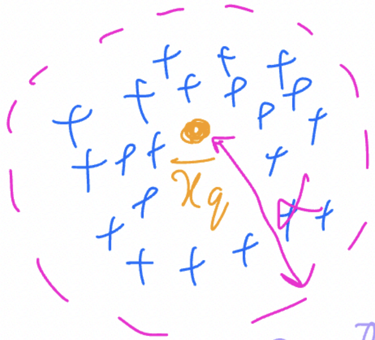
↳ for this we will create N datapoints which which will be neighbors to $\bar{x}_q$.

$$\therefore \bar{X} = \{\bar{x}_q + \varepsilon_i\}_{i=1}^{N=10,000}$$

→ Adds noise to our $\bar{x}_q$ datapoint

& with $||\varepsilon_i||^2 \le \alpha$

this means that all the datapoints will lie inside a circular region with radius $= \alpha$



creates locality of interest around $\bar{x}_q$

Note: The process is analogous to Nearest Neighbor

These Locality of interest points shows how flucuating features

- Affects the model's performance

Understanding model Agnostics

③ Understanding Model Agnostic

Remember for

* linear model ⇒ wts are checked for intrepretability

* Decision model ⇒ Depth of the tree is used for intrepretability

observe:

* for different models, different intrepretability techniques used

Is LIME used only for NN?

↳ No, LIME is a general intrepretability technique

* Hence can be used for any model

⇓

therefore LIME is model Agnostic

What has LIME had to offer on model interpretability?

1. A consistent model agnostic explainer [LIME].
2. A method to select a representative set with explanations [SP-LIME] to make sure the model behaves consistently while replicating human logic. This representative set would provide an intuitive global understanding of the model.

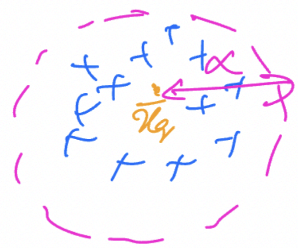
LIME explains a prediction so that even the non-experts could compare and improve on an untrustworthy model through feature engineering. An ideal model explainer should contain the following desirable properties:

Summarizing LIME process

Summarizing steps of LIME:

Step1: select a local point $\bar{x}_q$

step2: Create N datapoints within α radii of $\bar{x}_q$ by introducing noise to the local point



Step 3: Pass all the points in the new dataset to the model

step4: Make a new interpretable (simplier) model to learn on the newly created dataset

Note: This newly created interpretable model uses K-Lasso as loss function.

↳ finds top k features using L1 regularization

SHAP: Alternate way to find interpretability

SHAP: Another way to intrepret model

Suppose for ABC bank

↓ need?

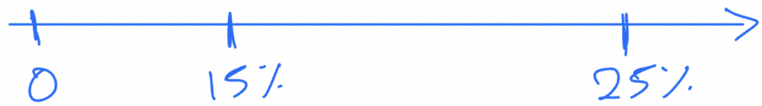
Model predicting if person will default or not for a loan payment

↓ observed

the baseline for people defaulting is 15%.

Suppose for a person B, the model predicts 25%, then why model prediction for person B differs from our baseline?

↳ this is what SHAP measures



interested region ⇒ what factors are influencing the Model

- SHAP values (SHapley Additive exPlanations) is a method rooted in cooperative game theory that enhances the transparency and interpretability of machine learning models.
- It provides a unified framework for attributing the contributions of each feature towards a prediction, addressing the limitations of traditional linear models and feature importance in tree-based models.
- Unlike linear models that rely on feature coefficients, which may be influenced by variable scales and fail to capture local importance, SHAP values offer a more comprehensive understanding of feature importance.
- They consider the impact of each feature when combined with different subsets of features, accounting for interactions and dependencies among them.
- SHAP values are based on the concept of Shapley values from cooperative game theory. By calculating the marginal contribution of a feature when added to or removed from coalitions of features, SHAP values ensure fairness and consistency in attributing importance.
- The insights provided by SHAP values are interpretable and nuanced, revealing the relative importance of each feature in a prediction. Visualizations, such as summary plots, individual feature importance plots, dependence plots, or force plots, allow for a detailed exploration of feature impact on predictions.