NUNI Online: Interpretable ML, 6 Oktober 2023

==========

==========

SLIDE #8

==========

1. Intrinsic or post hoc?

This criteria distinguishes whether interpretability is achieved by restricting the complexity of the machine learning model (intrinsic) or by applying methods that analyze the model after training (post hoc). Intrinsic interpretability refers to machine learning models that are considered interpretable due to their simple structure, such as short decision trees or sparse linear models. Post hoc interpretability refers to the application of interpretation methods after model training. Permutation feature importance is, for example, a post hoc interpretation method.

2. Result of the interpretation method

a. Feature summary statistic → feature importance, or the pairwise feature interaction strengths,

b. Feature summary visualization → partial dependence plots

c. Model internals (e.g. learned weights) →The interpretation of intrinsically interpretable models falls into this category. Examples are the weights in linear models or the learned tree structure (the features and thresholds used for the splits) of decision trees.

d. Model-specific or model-agnostic? →Model-specific interpretation tools are limited to specific model classes. The interpretation of regression weights in a linear model is a model-specific interpretation. Tools that only work for the interpretation of e.g. neural networks are model-specific. Model-agnostic tools can be used on any machine learning model and are applied after the model has been trained (post hoc). These agnostic methods usually work by analyzing feature input and output pairs.

e. Local or global? →Does the interpretation method explain an individual prediction or the entire model behavior?

==========

SLIDE #9

==========

- Interpretable models mempunyai big disadvantage that predictive performance is lost compared to other machine learning models and you limit yourself to one type of model.
- The disadvantage of model-specific interpretation methods is that it also binds you to one model type and it will be difficult to switch to something else.

==========

SLIDE #12

==========

- The interpretation method can work with any machine learning model, such as random forests and deep neural networks.
- You are not limited to a certain form of explanation. In some cases it might be useful to have a linear formula, in other cases a graphic with feature importances.
- The explanation system should be able to use a different feature representation as the model being explained. For a text classifier that uses abstract word embedding vectors, it might be preferable to use the presence of individual words for the explanation.

==========

SLIDE #13

==========

- This multi-layered abstraction also helps to understand the differences in approaches between statisticians and machine learning practitioners.
- Statisticians deal with the Data layer, such as planning clinical trials or designing surveys.
- They skip the Black Box Model layer and go right to the Interpretability Methods layer.
- Machine learning specialists also deal with the Data layer, such as collecting labeled samples of skin cancer images or crawling Wikipedia. Then they train a black box machine learning model. The Interpretability Methods layer is skipped and humans directly deal with the black box model predictions.

- It’s great that interpretable machine learning fuses the work of statisticians and machine learning specialists.

==========

SLIDE #17

==========

- For example, when applied to a linear regression model, partial dependence plots always show a linear relationship.
- Perlihatkan Xournal++

==========

SLIDE #18

==========

- The feature(s) in S are those for which we want to know the effect on the prediction.
- The feature vectors xS and xC combined make up the total feature space x.
- Partial dependence works by marginalizing the machine learning model output over the distribution of the features in set C, so that the function shows the relationship between the features in set S we are interested in and the predicted outcome.
- By marginalizing over the other features, we get a function that depends only on features in S, interactions with other features included.

==========

SLIDE #19

==========

- Perlihatkan contoh pakai Libre Calc

==========

SLIDE #20

==========

- Perlihatkan contoh pakai Libre Calc

==========

SLIDE #22

==========

- The largest differences can be seen in the temperature. The hotter, the more bikes are rented. This trend goes up to 20 degrees Celsius, then flattens and drops slightly at 30.
- Marks on the x-axis indicate the data distribution.

==========

SLIDE #23

==========

- PDPs for the bike count prediction model and the season.
- Unexpectedly all seasons show similar effect on the model predictions, only for winter the model predicts fewer bicycle rentals.

==========

SLIDE #24

==========

- Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos.
- When these people fill up a form providing their email address or phone number, they are classified to be a lead.
- Moreover, the company also gets leads through past referrals.
- Once these leads are acquired, employees from the sales team start making calls, writing emails, etc.
- Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

==========

SLIDE #29

==========

- For age, the PDP shows that the probability is low until 40 and increases after.
- The more years on hormonal contraceptives the higher the predicted cancer risk, especially after 10 years.
- For both features not many data points with large values were available, so the PD estimates are less reliable in those regions.

==========

SLIDE #30

==========

- PDP of cancer probability and the interaction of age and number of pregnancies.
- The plot shows the increase in cancer probability at 45.
- For ages below 25, women who had 1 or 2 pregnancies have a lower predicted cancer risk, compared with women who had 0 or more than 2 pregnancies.
- But be careful when drawing conclusions: This might just be a correlation and not causal!

==========

SLIDE #32

==========

- The partial dependence function at a particular feature value represents the average prediction if we force all data points to assume that feature value.
- In my experience, lay people usually understand the idea of PDPs quickly.
- If the feature for which you computed the PDP is not correlated with the other features, then the PDPs perfectly represent how the feature influences the prediction on average.
- We intervene on a feature and measure the changes in the predictions.
- In doing so, we analyze the causal relationship between the feature and the prediction.
- The relationship is causal for the model – because we explicitly model the outcome as a function of the features – but not necessarily for the real world!

==========

SLIDE #33

==========

- This is not the fault of PDPs, but of the 2-dimensional representation (paper or screen) and also of our inability to imagine more than 3 dimensions.
- Omitting the distribution can be misleading, because you might overinterpret regions with almost no data.
- This problem is easily solved by showing a rug (indicators for data points on the x-axis) or a histogram.
- It is assumed that the feature(s) for which the partial dependence is computed are not correlated with other features.
- For example, suppose you want to predict how fast a person walks, given the person's weight and height.
- For the partial dependence of one of the features, e.g. height, we assume that the other features (weight) are not correlated with height, which is obviously a false assumption.
- For the computation of the PDP at a certain height (e.g. 200 cm), we average over the marginal distribution of weight, which might include a weight below 50 kg, which is unrealistic for a 2 meter person.
- In other words: When the features are correlated, we create new data points in areas of the feature distribution where the actual probability is very low (for example it is unlikely that someone is 2 meters tall but weighs less than 50 kg).
- One solution to this problem is Accumulated Local Effect plots or short ALE plots that work with the conditional instead of the marginal distribution.
- PD plots only show the average marginal effects. Suppose that for a feature half your data points have a positive association with the prediction – the larger the feature value the larger the prediction – and the other half has a negative association – the smaller the feature value the larger the prediction.
- The PD curve could be a horizontal line, since the effects of both halves of the dataset could cancel each other out. You then conclude that the feature has no effect on the prediction. By plotting the

individual conditional expectation curves instead of the aggregated line, we can uncover heterogeneous effects.