explainable ML section 3

Kristin Chen

10/15/2020

- machine learning: An algorithm trains a model that produces the predictions.
- what is interpretability?
- why interpretability is important?
- how to interpret: to explain a ML model, you need the trained model, knowledge of the algorithm and the data.
- Taxonomy of interpretability
  - intrinsic vs post-hoc: interpretability is achieved
    - intrinsic: by restricting the complexity of the machine learning model that are considered interpretable due to their simple structure, such as short decision trees or sparse linear models.
    - post-hoc: by applying methods that analyze the model after training
  - results of interpretation methods:
    - feature summary statistic: provide summary statistics for each feature. Some methods return a single number per feature, such as feature importance, or a more complex result, such as the pairwise feature interaction strengths, which consist of a number for each feature pair.
    - feature summary visualization: some feature summaries are actually only meaningful if they are visualized. Partial dependence plots are curves that show a feature and the average predicted outcome.