Module 1: Chapter 2

Molnar, Christoph. “Machine learning. A Guide for Making Black Box Models Explainable”, 2019.

Importance of Interpretability

- Interpretability is crucial for understanding the decisions made by machine learning models.
- It is essential for ensuring safety measures, detecting bias, and achieving social acceptance.
- Interpretability allows for models to be debugged and audited, enhancing their reliability and trustworthiness.

Explaining Decisions

- Fairness: Interpretability helps in ensuring that predictions are unbiased and do not discriminate against underrepresented groups.
- Privacy: It aids in protecting sensitive information in the data.
- Reliability: Interpretability ensures that small changes in the input do not lead to large changes in the prediction.
- Causality: It helps in checking that only causal relationships are picked up by the model.
- Trust: Models that explain their decisions are easier for humans to trust compared to black box models.

When Interpretability is Not Needed

- Interpretability is not required if the model has no significant impact, such as in low-risk or non-critical applications.
- In scenarios where the model's predictions do not have serious consequences, interpretability may not be necessary.
- For example, in a casual side project or in extensively studied and evaluated methods like optical character recognition, interpretability may not be crucial

Taxonomy of Interpretability Methods

- Interpretability methods can be classified based on various criteria.
- One important criterion is whether interpretability is achieved intrinsically or post hoc.
- Intrinsic interpretability refers to models that are considered interpretable due to their simple structure.
- Post hoc interpretability refers to the application of interpretation methods after model training.

Differentiation of Interpretation Methods

- Interpretation methods can be differentiated based on the result they provide.
- Feature summary statistics provide summary statistics for each feature, such as feature importance or pairwise feature interaction strengths.
- Feature summary visualization allows for the visualization of feature summaries, such as partial dependence plots.
- Model internals, such as learned weights or tree structures, can be interpreted for intrinsically interpretable models.
- Data points, such as counterfactual explanations or prototypes, can be used to make a model interpretable.
- Algorithm transparency provides knowledge of how the algorithm learns a model from the data and what kind of relationships it can learn.

Intrinsic vs. Post Hoc Interpretability

- Intrinsic interpretability refers to models that are considered interpretable due to their simple structure, such as decision trees or linear models.
- Post hoc interpretability refers to the application of interpretation methods after model training, such as permutation feature importance or LIME.
- Intrinsic interpretability is limited to simple models, while post hoc interpretability can be applied to any model.
- Post hoc interpretability methods can also be applied to intrinsically interpretable models to provide additional insights.

Model-Specific vs. Model-Agnostic Interpretation

- Model-specific interpretation tools are limited to specific model classes.
- Interpretation of regression weights in a linear model is a model-specific interpretation.
- Tools designed for the interpretation of specific models, such as neural networks, are also model-specific.
- Model-agnostic tools can be used on any machine learning model and are applied after the model has been trained.
- Agnostic methods analyze feature input and output pairs and do not have access to model internals such as weights or structural information.

Local vs. Global Interpretability

- The interpretation method's scope can be categorized as local, global, or somewhere in between.
- Local interpretation explains an individual prediction, providing insights into why a specific prediction was made.
- Global interpretation aims to explain the entire model behavior, providing an understanding of how the model makes decisions based on a holistic view of its features and learned components.
- The scope criterion determines whether the interpretation method focuses on individual predictions, the overall model behavior, or a combination of both.

Scope of Interpretability

- Algorithm Transparency
  - Focuses on how the algorithm learns a model from the data and the relationships it can learn.
  - Provides an understanding of the algorithm's workings, but not specific model predictions.
- Global, Holistic Model Interpretability
  - Aims to comprehend the entire model at once.
  - Requires knowledge of the trained model, algorithm, and data.
  - Focuses on understanding how the model makes decisions based on a holistic view of its features and learned components.

Scope of Interpretability...

- Local Interpretability for a Single Prediction
  - Addresses why the model made a certain prediction for an instance.
  - Provides insights into the factors influencing individual predictions.
- Local Interpretability for a Group of Predictions
  - Explores the reasons behind specific predictions for a group of instances.
  - Can be achieved through global model interpretation methods or explanations of individual instances.
- Evaluation of Interpretability
  - Involves assessing the transparency and interpretability of the model at different levels.
  - Considers the applicability of interpretability methods to specific tasks and user understanding.

Evaluation of Interpretability

- Lack of Consensus and Measurement
  - Interpretability in machine learning lacks a clear definition and measurement criteria.
  - Initial research and attempts to formulate evaluation approaches are underway.
- Levels of Evaluation
  - **Application Level Evaluation (Real Task)**
  - Involves integrating explanations into the product and testing by end users.
  - Requires a robust experimental setup and comparison to human performance.
  - **Human Level Evaluation (Simple Task)**
  - Simplified application level evaluation conducted with laypersons.
  - Involves presenting different explanations to users for assessment.
  - **Function Level Evaluation (Proxy Task)**
  - Does not require human involvement and can be based on proxy tasks.
  - Utilizes known human-level evaluation results for specific model classes.

Properties of Explanation Methods

- Expressive Power
  - Describes the "language" or structure of the explanations generated by the method.
  - Examples include IF-THEN rules, decision trees, weighted sums, or natural language.
- Translucency
  - Measures how much the explanation method relies on inspecting the machine learning model.
  - Ranges from high translucency (relying on model parameters) to zero translucency (only manipulating inputs).
- Portability
  - Defines the range of machine learning models compatible with the explanation method.
  - Methods with low translucency have higher portability due to treating the model as a black box.

Properties of Individual Explanations

- Accuracy
  - Evaluates how well an explanation predicts unseen data.
  - High accuracy is crucial, especially if the explanation is used for predictions in place of the machine learning model.
- Fidelity
  - Assesses how well the explanation approximates the prediction of the black box model.
  - High fidelity is essential for the usefulness of an explanation.
- Consistency
  - Examines the variation of explanations between models trained on the same task and producing similar predictions.
  - Consistency is desirable when models rely on similar relationships.
- Stability
  - Focuses on the similarity of explanations for similar instances within a fixed model.
  - High stability indicates that slight variations in features do not substantially change the explanation.

Properties of Individual Explanations (Contd.)

- Comprehensibility
  - Evaluates the human understanding of explanations.
  - Considers the audience, size of the explanation, and the comprehensibility of the features used.
- Certainty
  - Addresses the reflection of the model's certainty in the explanation.
  - Includes the model's confidence in predictions and the certainty of different instances.
- Degree of Importance
  - Reflects the importance of features or parts of the explanation.
  - Aims to clarify which conditions or features are most influential in the explanation.

Properties of Individual Explanations (Contd.)

- Novelty
  - Reflects whether a data instance comes from a "new" region.
  - Considers the accuracy and usefulness of explanations for instances outside the training data distribution.
- Representativeness
  - Considers the coverage of explanations.
  - Can cover the entire model or represent individual predictions, impacting the breadth of interpretability.
- Diversity
  - Evaluates the diversity of explanations generated by the method.
  - Aims to provide multiple explanations for the same prediction to increase user trust and understanding.
- Actionability
  - Measures the usefulness of the explanation for decision-making.
  - Considers whether the explanation provides actionable insights for the user.

What Constitutes a Good Explanation?

- Explanations are **Contrastive**
  - Discusses the importance of contrasting the current situation with a situation in which the event would not have occurred.
  - Interpretation for machine learning: Explanations should highlight the difference between the predicted outcome and the alternative outcome to provide a clear understanding of the model's behavior.
- Explanations are **Selected**
  - Explores the idea that explanations are selective and focus on one or two relevant causes.
  - Interpretation for machine learning: Explanations should prioritize the most relevant features or factors that contribute to the model's predictions to increase interpretability.
- Explanations are **Social**
  - Discusses the social nature of explanations and the importance of considering the audience and context.
  - Interpretation for machine learning: Explanations should be tailored to the user's background, knowledge, and experience to ensure maximum understanding and engagement.

What Constitutes a Good Explanation? 2

- Explanations Focus on the **Abnormal**
  - Explores the idea that good explanations focus on abnormal or unexpected events.
  - Interpretation for machine learning: Explanations should highlight the unusual or unexpected features or factors that contribute to the model's predictions to provide a clear understanding of the model's behavior.
- Explanations are **Truthful**
  - Discusses the importance of providing truthful and accurate explanations.
  - Interpretation for machine learning: Explanations should reflect the actual behavior of the model and provide evidence and justification for the model's predictions.
- Explanations are **Consistent** with Prior Beliefs
  - Explores the idea that good explanations are consistent with prior beliefs and expectations.
  - Interpretation for machine learning: Explanations should align with the user's prior knowledge and expectations to increase understanding and engagement.

What Constitutes a Good Explanation? 3

- Explanations are **General and Probable**
  - Discusses the importance of providing general and probable explanations.
  - Interpretation for machine learning: Explanations should provide a general understanding of the model's behavior and highlight the most probable factors or features that contribute to the model's predictions.
- Summary of Key Points
  - Summarizes the key ideas and interpretations for machine learning.
  - Emphasizes the importance of providing clear, accurate, and relevant explanations that are tailored to the user's needs and preferences.