Questions A

A1. Glassbox Models

- What are Glassbox models, and why are they considered interpretable?
- What are the pros and cons of using linear regression as a Glassbox model?

A2. Local Interpretable Model-agnostic Explanations (LIME)

- What is the general principle behind LIME?
- What are the main advantages and disadvantages of using LIME?

A3. Shapley Additive Explanations (SHAP)

- What is the origin of SHAP, and how does it relate to cooperative game theory?
- What are the key properties of Shapley values in SHAP?

Answers B

B1. Layer-wise Relevance Propagation (LRP)

- (slide 36) LRP explains model predictions by backpropagating relevance through layers to highlight important input features.
- (slide 39) LRP-0 redistributes relevance proportionally to each input's contribution to neuron activations, capturing function behavior but potentially producing complex explanations.

B2. LIME and SHAP Comparison

- LIME provides local explanations using surrogate models, while SHAP offers both local and global explanations by distributing prediction outcomes across feature subsets.
- (slide 33) SHAP is computationally expensive due to evaluating all feature subsets. Approximations like Monte Carlo sampling reduce complexity but increase variance.

B3. Decision Trees in Glassbox Models

- (Look at slide 20) Strengths: Interpretable, visualizable, captures non-linear relationships and feature interactions.
- Weaknesses: Sensitive to data changes, step-function outputs, unstable with small changes in input.

B4. Logistic Regression in Glassbox Models

- (slides 15-18) Logistic regression extends linear regression by using a logistic function to squeeze the output of a linear equation between 0 and 1, making it suitable for binary classification. The output can be interpreted as the probability of belonging to a particular class.
- It is the change in the log-odds of the outcome for a one-unit change in the corresponding feature. This can be interpreted as the impact of each feature on the probability of the outcome.

Questions B

B1. *Layer-wise Relevance Propagation (LRP)*

- ❑ What is the main goal of Layer-wise Relevance Propagation (LRP)?
- ❑ What is the principle behind the LRP-0 rule, and how does it work?

B2. *LIME and SHAP Comparison*

- ❑ How does LIME differ from SHAP in terms of explanation scope?
- ❑ What are the computational challenges associated with SHAP, and how are they addressed?

B3. *Decision Trees in Glassbox Models*

- ❑ What are the strengths and weaknesses of decision trees as Glassbox models?

B4. *Logistic Regression in Glassbox Models*

- ❑ How does logistic regression extend linear regression for classification problems?
- ❑ What is the interpretation of weights in logistic regression?

Answers A

A1. Glassbox Models

- (slide 11) Glassbox models (e.g., linear regression, logistic regression, decision trees) are inherently interpretable because their internal workings are transparent, allowing users to understand how predictions are made.
- (slide 14) **Pros:** Simple, interpretable, shows direct relationships between inputs and outputs
  Cons: Assumes linearity, poor for classification, fails with complex patterns.

A2. Local Interpretable Model-agnostic Explanations (LIME)

- (slide 22) LIME explains individual predictions by approximating the model locally with a simple, interpretable model like linear regression
- (slide 25-26) Advantages: Model-agnostic, works on different data types, offers a fidelity measure.
  Disadvantages: Instability, predefined complexity trade-off, possible manipulation of explanations.

A3. Shapley Additive Explanations (SHAP)

- (slide 28) SHAP is based on Shapley values from cooperative game theory, distributing the total model prediction fairly among input features.
- (slide 33) Efficiency: Sum of values equals the difference between prediction and mean.
  Symmetry: If two features contribute equally, their Shapley values are the same.
  Dummy: Features with no influence get a value of zero.
  Additivity: Shapley values can be summed for feature interactions.

Answers A (full)

A1. Glassbox Models

- (slide 11) Glassbox models are machine learning models that are inherently interpretable, such as linear regression, logistic regression, and decision trees. They are considered interpretable because their internal workings are transparent, allowing users to understand how predictions are made based on the input features.
- (slide 14) Pros: Linear regression is highly interpretable, as it provides a clear relationship between input features and the output through weighted sums. It is easy to understand and implement.
  Cons: It assumes a linear relationship between features and the output, which is often an oversimplification of real-world problems. It also performs poorly for classification tasks and cannot handle non-linear relationships.

A2. Local Interpretable Model-agnostic Explanations (LIME)

- (slide 22) LIME explains individual predictions by approximating the model locally around the instance of interest. It creates a simpler, interpretable model (like linear regression) that mimics the behavior of the complex model in the local region of the data point being explained.
- (slide 25-16) Advantages: LIME is model-agnostic, meaning it can be applied to any machine learning model. It is also versatile, working with different data types (tabular, text, images). The fidelity measure helps assess the reliability of the explanation.
  Disadvantages: LIME can be unstable, producing different explanations for very similar data points. It also requires predefining the complexity of the explanation, which can be a trade-off between interpretability and fidelity.

A3. Shapley Additive Explanations (SHAP)

- SHAP is based on Shapley values, which were originally developed in cooperative game theory to fairly distribute the payoff among players based on their contributions. In the context of machine learning, SHAP values distribute the prediction outcome among the features, reflecting their individual contributions.
- **Efficiency**: The sum of the Shapley values equals the difference between the prediction and the average prediction.
  Symmetry: If two features contribute equally, their Shapley values are the same.
  Dummy: A feature that has no influence on the prediction has a Shapley value of zero.
  Additivity: Shapley values can be added up for combined features.

Answers B (full)

B1. Layer-wise Relevance Propagation (LRP)

- to explain the relevance of input features for a model's prediction by propagating the output relevance back through the layers of the model. It is particularly used to highlight which parts of the input (e.g., pixels in an image) are most relevant for the prediction.
- It redistributes relevance proportionally to the contribution of each input to the neuron activations. It calculates the relevance of a neuron in a lower layer based on the relevance of neurons in the higher layer, weighted by their contributions. This rule is used in upper layers where the function is close to its gradient.

B2. LIME and SHAP Comparison

- LIME provides local explanations for individual predictions by approximating the model locally, while SHAP provides both local and global explanations by distributing the prediction outcome among all features based on their contributions across all possible subsets of features.
- SHAP can be computationally expensive because it requires evaluating all possible subsets of features, which grows exponentially with the number of features. To address this, SHAP often uses approximation methods like Monte-Carlo sampling to estimate the Shapley values, reducing the computational burden but potentially increasing variance.

B3. Decision Trees in Glassbox Models

- (Look at slide 20) Strengths: Decision trees are highly interpretable, as they can be visualized naturally, and their decision paths are easy to follow. They can capture non-linear relationships and feature interactions.
- Weaknesses: Decision trees can be unstable, as small changes in the data can lead to completely different trees. They also struggle with linear relationships and can create step functions, leading to a lack of smoothness in predictions.

B4. Logistic Regression in Glassbox Models

- (slides 15-18) Logistic regression extends linear regression by using a logistic function to squeeze the output of a linear equation between 0 and 1, making it suitable for binary classification. The output can be interpreted as the probability of belonging to a particular class.
- It is the change in the log-odds of the outcome for a one-unit change in the corresponding feature. This can be interpreted as the impact of each feature on the probability of the outcome.