13/8/25

ASSIGNMENT 2

Unit - 2 Interpretability

Q.1 — Difference : Interpretability v/s Explainability

	Interpretability	Explainability
Definition	The extent to which a human can understand the internal mechanism of a system	The extent to which the internal mechanics of a system can be explained in human terms.
Focus	Simplicity & Transparency of the model	understanding the reasoning behind the model's predictions.
Model Type	Mostly applies to "white-box" models. (eg: linear regression, Decision Trees)	Often used for "black-box" models. (eg: NN, Random Forests).
User Involvement	Helps developers and data scientists	Aimed at end-users, regulators or non-technical stakeholders.
Goal	Understand how the model works	understand why the model make a specific decision.
Example	A linear regression where co-efficients are directly interpretable.	Explaining why a NN predicts a loan rejection using SHAP values.
Key Difference	Transparency	Comprehensibility.

Q.2 Discuss black box model and list interpretability methods to explain black box models.

→ A Black Box Model is a model whose internal workings are opaque, making it difficult to understand how it arrives at a specific prediction or decision. You can see the inputs and the final output, but the process in between is a "black box."

These models can achieve very high accuracy but lack transparency and it can be dangerous to blindly trust in critical applications like healthcare, finance or law.

Examples, Deep Neural Networks, Random Forests, Gradient Boosted Trees (XGBoost, lightBGM) and Support Vector Machines.

→ Interpretability Methods :-

1. Model-Specific Methods.

These are designed for specific algorithms. For example, feature importance in Decision Trees, visualization of Convolutional Filters in CNNs.

2. Model-Agnostic Methods. (work with any black box model)

(i) LIME [local Interpretable Model-agnostic Explanations]: Builds a simple model around a prediction to explain it locally.

(ii) SHAP (Shapely Additive exPlanations): uses game theary to assign feature importance for a prediction.

(iii) Partial Dependence Plot (PDP): shows how changing

one feature affects predictions on average.

(iv) Individual Conditional Expectation (ICE): Similar to PDP but at the individual instance level.

(v) Counterfactuals: Show what small input changes could flip the prediction.

(vi) Saliency Maps / Grad-CAM: Highlight which input regions (eg: in images) influenced the output.

(vii) Surrogate Models: Train a simple model to mimic the black box for ~~easier~~ easier explanation.

Q.3 Discuss linear Regression. Explain how linear regression is an interpretable model.

—4 Linear Regression is a supervised learning algorithm that models the linear relationship between a continuous dependent variable and one or more independent variables by fitting a line (or hyperplane) that best minimizes the distance between the actual data points and the predicted values. Its primary goal is to predict continuous outcomes based on input features.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$$

- Y = dependent variable (output)
- $X_1, X_2, \cdots, X_n$ = independent variables (features)
- β_0 = intercept
- β_1 = coefficients (weights)
- ϵ = Error term.

The algorithm finds the best-fit line by

minimizing the sum of squared errors (SSE) using Ordinary least Squares (OLS). Once the coefficients are estimated, it can predict outcomes for new data.

—4 Linear Regression is considered a highly interpretable model because:

(i) Coefficient Meaning: Each coefficient (B_i) shows how much the target variable changes with a one-unit increase in the corresponding feature, holding other features constant. For example, if $B_1 = 5$, then increasing x_1 by 1 increases 'y' by 5 units.

(ii) Transparency: The equation is easy to understand and visulize, making it straightforward to explain predictions to non-technical stakeholders.

(iii) Global Interpretability: The entire model behaviour can be analyzed at once, providing a global understanding of the relationship between inputs and outputs.

(iv) Feature Importance: The magnitude and sign of coefficients indicate the importance and direction of each feature's impact.

Q.5 Write a note on GAM.

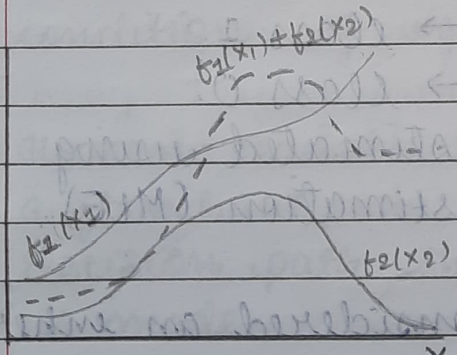
—4 Generalized Additive Models (GAMs) are a type of interpretable model that extend linear regression by allowing non-linear relationships between features and the target, while still

being additive and easy to understand.

$$g(E[Y]) = \beta_0 + f_1(X_1) + f_2(X_2) + \cdots + f_n(X_n) + \epsilon$$

- g = link function / y = predicted value
- β_0 = intercept
- $f_i(X_i)$ = smooth (non-linear) function learned from data

- Additive: Each feature contributes independently to the prediction.
- Interpretable: You can visualize each function $f_i(X_i)$ to understand how a feature affects the output.
- Flexible: Captures non-linear trends, unlike linear regression.
- Transparent: Still easy to explain to non-technical users compared to black box models.



For Example, in predicting bike rentals:
- f_1(Temperature) might be a curve showing rentals increase with temperature up to a point, then decreases if too hot.
- f_2(Humidity) could show a drop in rentals as humidity increases.

Q.4 Explain logistic Regression. Discuss what makes logistic regression an interpretable model.

→ Logistic Regression is a supervised learning algorithm used for binary classification, which means it predicts a categorical outcome with two possible values. Instead of predicting a continuous value, it estimates the probability that an instance belongs to a particular class.

$$P(Y=1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n)}}$$

- $P(Y=1|X)$ = probability of event occurring.
- β_0 = intercept
- β_i = coefficient of feature X_i

The algorithm uses the logistic (sigmoid) function to map predicted values between 0 and 1. A decision threshold (commonly 0.5) is applied:

① If probability ≥ 0.5 → class 1.
② If probability < 0.5 → class 0.

Coefficients are estimated using Maximum likelihood Estimation (MLE).

→ Logistic Regression is considered an inherently interpretable model because:

(i) Coefficient Meaning: Each coefficient (β_i) represents represents the change in log-odds of the outcome for a 1-unit increase in the feature, holding others constant.
Example: If $\beta_1 = 0.7$ → a 1-unit increase in X_1, multiples the odd by $e^{0.7} \approx 2.01$ times

(ii) Transparency: Provides a clear, mathematical link between features and outcome probabilities
(iii) Feature Importance: Magnitude of β → indicates importance
- Sign of β:
  Positive → increases probability
  Negative → decreases probability

(iv) Interpretability:
- Global: Same coefficients apply to all predictions
- Local: Individual predictions can be explained by feature contributions.

Q.6 Explain why decision tree is considered as an inherent interpretable model?

→ A Decision Tree is a supervised learning model used for both classification and regression. It splits data based on feature values into branches, forming a tree-like structure. At each node, a decision is made based on a condition (eg: Age > 30?)

(i) Human-readable structure: The model looks like a flowchart, which is easy to follow. Each decision path shows how the final prediction is made.
(ii) Clear feature usage: You can directly see which features are used and how they affect the outcome.
(iii) Local and global explanations:
- Globally: The entire tree can be visualized to understand the overall logic
- Locally: You can trace a specific condition by following the path from root to leaf.

(iv) No need for post-hoc methods : Do not require extra explanations techniques like LIME or SHAP.

(v) Visual Representation : Trees can be drawn or printed, which helps in explaining results to non-technical users.

eg: In a loan approval model :
- If the tree splits on "Salary > $30,000" and "Credit score > 700", it's clear that applicants with higher income and good credit are more likely to be approved.

Q.7 Explain PDP and discuss how it can be used to interpret machine learning model.

→ Partial Dependence Plot (PDP) is a model-agnostic interpretability ~~model~~ method used to show the average effect of one or two features on the predicted outcome of a machine learning model.

"How does a feature affect predictions, keeping other features constant ?", it answers.

$$PD(x_j) = \frac{1}{n} \sum_{i=1}^{n} f(x_j, x_{i,-j})$$

- f = prediction function of the model
- x_j = selected feature
- $x_{i,-j}$ = all the other features of i^{th} instance

It selects a feature (eg, "Age"), vary its value across a range and keep the other

features constant. Then observes the average change in model prediction.

The result is a 2D plot (or 3D for two features) showing how predictions change with the selected feature.

(i) Global understanding: Shows overall trend (eg. whether prediction increases or decreases as the feature increases).

(ii) Feature Influence: Helps identify if a feature has a positive, negative or non-linear effect on predictions.

(iii) Model debugging: Can reveal unexpected patterns or biases in the model's behaviour.

(iv) Transparency: Useful for explaining model behaviour to stakeholders.

eg: In a house price model:

- A PDP for "Size" may show that price increases with size up to 2000 sq. ft, but flattens beyond that — revealing the diminishing returns.

Q.8 Explain surrogate model. Discuss how it enhance interpretability in ML model.

-4 A surrogate Model is a simple, interpretable model (like a decision tree or linear regression) that is trained to mimic the behaviour of a more complex black-box model.

It acts as a proxy to explain how the original model works without modifying the actual complex model.

It trains a black box model (eg, Random Forest) and uses that model to generate predictions on input data. Then it trains a simple model (eg, decision tree) on the same inputs and the predicted outputs. Then use the surrogate model to interpret the black-box model.

(i) Simplifies understanding: Converts complex logic into simple rules or relationships that are easier to explain.

(ii) Global Interpretability: Gives a high-level overview of how features generally influence predictions.

(iii) Model Transparency: Useful when the original model is too complex to interpret directly.

(iv) Stakeholder Communication: Makes it easier to communicate model behaviour to non-technical audiences.

eg: Suppose a NN predicts loan approvals. A surrogate decision tree can show that most approvals happen when income > $50,000 and credit score > 700, making the decision logic clearer.

Q.9 Explain Saliency Maps:

Ans Saliency Maps are interpretability techniques mainly used in Deep Neural Networks, especially in Computer Vision, to highlight which parts of the input (eg: an image) had

the most influence on the model's prediction.

They help us understand what the model is "looking at" when making a decision.

For a given image and prediction, the model computes the gradient of the output with respect to each pixel. The result is a heatmap showing which pixels most affect the output - brighter areas = higher importance.

$$S(x) = \frac{\partial f(x)}{\partial x}$$

- $f(x)$ = model prediction score for class of interest.
- x = input image (pixels)
- $S(x)$ = saliency map showing pixel importance.

eg: In a image classification model that identifies cats vs. dogs, a saliency map may highlight the ears and eyes of the animal - indicating these features influenced the prediction.

Q.10 Discuss different evaluation criteria for interpretability.

-4 Interpretability refers to how easily humans can understand a model's decisions. Evaluating interpretability is important to ensure transparency, trust and usability of models.

(i) Simplicity : Simpler models (fewer features, shallow trees) are easier to interpret.

(ii) Transparency : White-box models (eg: linear

regression) are transparent and interpretable by design.

(iii) Fidelity: Measures how well an explanation method reflects the true behaviour of the original model.

(iv) Consistency: Similar inputs should ~~practice~~ produce similar explanations to maintain trust.

(v) Stability: Small changes in input should not cause big changes in the explanations.

(vi) Human understandability: Explanations should be clear and easily understood by non-experts.

eg: Suppose a loan approval model uses a surrogate decision tree to explain predictions.

If the tree shows:

- "If income > $50,000 and credit score > 700 → approve loan".

this rule is simple, transparent and understandable to both Data Scientists & loan officers.

- If similar applicants get similar rules, it shows consistency and stability.