ML Explainability Summary

Mathematical Notes

October 27, 2025

Contents

1 Introduction to Explainability

1.1 Motivation for Explainability

- **Trust**: Users need to understand model decisions
- **Fairness**: Detect and mitigate bias
- **Debugging**: Identify model failures
- **Compliance**: Regulatory requirements (GDPR, AI Act)
- **Scientific understanding**: Gain insights from data
- **Model improvement**: Identify areas for enhancement

1.2 Types of Explainability

Definition 1.1. Global explainability provides understanding of the overall model behavior and decision-making process.

Definition 1.2. Local explainability explains individual predictions for specific instances.

Definition 1.3. Post-hoc explainability generates explanations after model training, without modifying the model.

Definition 1.4. Intrinsic explainability uses inherently interpretable models like linear regression or decision trees.

1.3 Properties of Good Explanations

- **Faithfulness**: How well does the explanation reflect the actual model behavior?
- **Stability**: Are explanations consistent across similar inputs?
- **Completeness**: Does the explanation capture all relevant factors?
- **Simplicity**: Is the explanation easy to understand?
- **Contrastiveness**: Does it explain why this prediction rather than alternatives?

2 Feature Importance Methods

2.1 Permutation Importance

Definition 2.1. Permutation importance measures the increase in prediction error when a feature is randomly shuffled, breaking its relationship with the target.

For feature j:

$$\mathrm{PI}_j = s - \frac{1}{K}\sum_{k=1}^{K} s_{k,j}$$

where s is the baseline score and $s_{k,j}$ is the score when feature j is permuted in permutation k.

2.2 SHAP Values

Definition 2.2. SHAP (SHapley Additive exPlanations) values satisfy the efficiency, symmetry, dummy, and additivity axioms:

$$\phi_i(f, \mathbf{x}) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(\mathbf{x}_{S \cup \{i\}}) - f_S(\mathbf{x}_S)]$$

where F is the set of all features.

2.3 Integrated Gradients

Definition 2.3. Integrated Gradients computes the integral of gradients along the path from baseline to input:

$$\mathrm{IG}_i(\mathbf{x}) = (x_i - x_i') \int_{\alpha=0}^{1} \frac{\partial f(\mathbf{x}' + \alpha(\mathbf{x} - \mathbf{x}'))}{\partial x_i} d\alpha$$

where $\mathbf{x}'$ is the baseline input.

2.4 Layer-wise Relevance Propagation (LRP)

Definition 2.4. LRP propagates relevance scores backward through neural network layers to identify important input features.

3 Local Explainability Methods

3.1 LIME (Local Interpretable Model-agnostic Explanations)

Definition 3.1. LIME approximates the model locally around a prediction using a simple interpretable model:

$$\xi(x) = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

where G is a class of interpretable models, π_x is a proximity measure, and $\Omega(g)$ penalizes complexity.

3.2 SHAP Local Explanations

Definition 3.2. SHAP local explanations provide feature attributions for individual predictions using the Shapley value framework.

3.3 Anchors

Definition 3.3. Anchors find the minimal set of features that, when present, guarantee a specific prediction with high confidence.

4 Gradient-Based Methods

4.1 Grad-CAM

Definition 4.1. Grad-CAM generates visual explanations by computing gradients of the target class with respect to feature maps:

$$L^c_{Grad-CAM} = \text{ReLU}\left(\sum_k \alpha^c_k A^k\right)$$

where $\alpha^c_k = \frac{1}{Z}\sum_i \frac{\partial y^c}{\partial A^k_{ij}}$ and A^k are the feature maps.

4.2 Guided Grad-CAM

Definition 4.2. Guided Grad-CAM combines Grad-CAM with guided backpropagation for finer-grained visualizations.

4.3 SmoothGrad

Definition 4.3. SmoothGrad reduces noise in gradient-based explanations by averaging gradients over multiple noisy versions of the input.

4.4 Integrated Gradients

Definition 4.4. Integrated Gradients satisfies the sensitivity and implementation invariance axioms for attribution methods.

5 Attention-Based Explanations

5.1 Attention Visualization

Definition 5.1. Attention weights in transformer models can be visualized to show which input tokens the model focuses on for each prediction.

5.2 Attention Rollout

Definition 5.2. Attention rollout aggregates attention weights across all layers to understand the flow of information.

5.3 Attention Flow

Definition 5.3. Attention flow traces how information flows through attention mechanisms to identify important input regions.

6 Surrogate Models

6.1 Global Surrogate Models

Definition 6.1. A **global surrogate model** is a simpler, interpretable model trained to mimic the behavior of a complex model across the entire input space.

6.2 Local Surrogate Models

Definition 6.2. A **local surrogate model** approximates the complex model's behavior in a specific region around a given input.

6.3 Decision Trees as Surrogates

Definition 6.3. Decision trees can serve as surrogate models by learning rules that approximate the complex model's decision boundaries.

7 Interpretable Models

7.1 Linear Models

Definition 7.1. Linear models are inherently interpretable as coefficients directly indicate feature importance and direction of influence.

7.2 Generalized Additive Models (GAMs)

Definition 7.2. GAMs model the target as a sum of smooth functions of individual features:

$$g(\mathbb{E}[Y]) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p)$$

7.3 Decision Trees

Definition 7.3. Decision trees provide interpretable rules through recursive binary splits based on feature thresholds.

7.4 Rule-Based Models

Definition 7.4. Rule-based models express predictions as logical rules that are easy to understand and validate.

8 Counterfactual Explanations

8.1 Counterfactual Generation

Definition 8.1. A **counterfactual explanation** answers: "What would need to change in the input to get a different prediction?"

8.2 Minimal Changes

Definition 8.2. Minimal counterfactuals find the smallest changes to the input that would result in a different prediction.

8.3 Actionable Counterfactuals

Definition 8.3. Actionable counterfactuals consider only changes that are feasible in the real world.

9 Causal Explainability

9.1 Causal Inference

Definition 9.1. Causal inference methods aim to understand cause-and-effect relationships rather than just correlations.

9.2 Interventional Explanations

Definition 9.2. Interventional explanations show how changing specific features would affect the prediction, accounting for causal relationships.

9.3 Causal Discovery

Definition 9.3. Causal discovery algorithms learn causal graphs from observational data to understand feature relationships.

10 Concept-Based Explanations

10.1 Concept Activation Vectors (CAVs)

Definition 10.1. CAVs represent human-interpretable concepts as directions in the neural network's activation space.

10.2 Testing with Concept Activation Vectors (TCAV)

Definition 10.2. TCAV quantifies the influence of concepts on model predictions using CAVs.

10.3 Concept Bottleneck Models

Definition 10.3. Concept bottleneck models explicitly model the relationship between input features and human-interpretable concepts.

11 Evaluation of Explanations

11.1 Faithfulness Metrics

- **Deletion**: Remove important features and measure performance drop
- **Insertion**: Add important features and measure performance gain
- **ROAR**: Remove and retrain to evaluate explanation quality

11.2 Stability Metrics

- **Consistency**: Similar inputs should have similar explanations
- **Continuity**: Small input changes should not cause large explanation changes

11.3 Human Evaluation

- **Comprehensibility**: How well do humans understand the explanation?
- **Trustworthiness**: Do explanations increase user trust?
- **Actionability**: Can users act on the explanations?

12 Challenges and Limitations

12.1 Adversarial Explanations

Definition 12.1. Adversarial explanations can be manipulated to hide model biases or create misleading interpretations.

12.2 Computational Complexity

Many explainability methods are computationally expensive, especially for large models and datasets.

12.3 Evaluation Challenges

- Lack of ground truth for explanations
- Difficulty in measuring explanation quality
- Subjectivity in human evaluation

12.4 Model-Specific Limitations

- Some methods only work with specific model types
- Gradient-based methods require differentiable models
- Attention-based methods only apply to attention-based architectures

13 Applications

13.1 Healthcare

- Medical diagnosis explanations
- Treatment recommendation reasoning
- Drug discovery insights
- Clinical decision support

13.2 Finance

- Credit scoring explanations
- Fraud detection reasoning
- Risk assessment transparency
- Regulatory compliance

13.3 Autonomous Systems

- Self-driving car decision explanations
- Robot behavior understanding
- Safety-critical system validation

13.4 Legal and Compliance

- Algorithmic decision explanations
- Bias detection and mitigation
- Audit trail generation
- Regulatory reporting

14 Ethical Considerations

14.1 Fairness and Bias

- Detecting algorithmic bias
- Ensuring fair explanations across groups
- Mitigating discriminatory practices

14.2 Privacy

- Protecting sensitive information in explanations
- Differential privacy in explanation generation
- Avoiding data leakage through explanations

14.3 Transparency vs. Security

- Balancing transparency with model security
- Preventing adversarial attacks through explanations
- Protecting intellectual property

15 Future Directions

15.1 Interactive Explanations

- Dialogue-based explanation systems
- User-guided explanation generation
- Iterative refinement of explanations

15.2 Multi-Modal Explanations

- Combining text, visual, and numerical explanations
- Cross-modal explanation consistency
- Personalized explanation formats

15.3 Automated Explanation Generation

- Natural language explanation generation
- Automated explanation quality assessment
- Self-explaining models

15.4 Regulatory Compliance

- Standardized explanation formats
- Automated compliance checking
- Industry best practices

16 Important Algorithms

16.1 Explanation Generation

- SHAP
- LIME
- Grad-CAM
- Integrated Gradients
- Permutation importance
- Counterfactual generation

16.2 Evaluation Methods

- Faithfulness testing
- Stability analysis
- Human evaluation protocols
- Automated quality metrics

17 Key Theorems

17.1 Shapley Value Properties

Theorem 17.1. The Shapley value is the unique solution that satisfies efficiency, symmetry, dummy, and additivity axioms.

17.2 Explanation Completeness

Theorem 17.2. For any explanation method that satisfies the efficiency axiom, the sum of feature attributions equals the difference between the prediction and the baseline.

17.3 Uniqueness of Integrated Gradients

Theorem 17.3. Integrated Gradients is the unique attribution method that satisfies sensitivity, implementation invariance, and completeness axioms.