Explainable AI (XAI) Assignment

Ali Muhammad Asad - aa07190

1 On Shapley Values and Factorial Growth

Shapley values attribute feature contributions by averaging marginal impacts across all possible orderings. So for n examples, the exact computation would involve $O(n!)$ evaluations of the model. So if a model has 7 features, there would be $7! = 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 5040$. This is explained in section 6.5.2 of the reference paper [1].

2 Causal Inference and Robustness

I selected the Invariant Risk Minimization (IRM) explained in section 6.8.2 of the reference paper [1]. IRM improves generalization by learning invariant causal features across environments. It minimizes risk $R(\omega \cdot \Phi)$ subject to $\omega \in \text{argmin } R(\omega \cdot \Phi), \forall e \in E$, ensuring the predicator ω is optimal across environments, focusing on causal representations Φ. Now spurious correlations occur when features are predictive in some environments but not others. IRM mitigates this by enforcing invariance, ensuring the model relies on causal, not environment-specific, features, enhancing robustness to distributional shifts. For example, in credit scoring, consider regions where "job type = sales" correlates with defaults in Region A but not Region B, while true causal factors are "income stability" and "credit history." IRM ensures the model focuses on these invariant features, avoiding spurious correlations, thus improving robustness across regions.

3 Local vs Global Explanation Dilemma

Local explanations (e.g., LIME, SHAP) focus on individual predictions, missing global patterns. For instance, in hiring, local explanations might show why one candidate was rejected, but not reveal systematic bias against a demographic group, as they lack aggregation across instances. Global explanations (e.g., feature importance) provide an overview but may obscure individual nuances. For example, a model might favor candidates with certain degrees, but local explanations could show exceptions where candidates without those degrees were favored due to unique qualifications. This dilemma highlights the need for a balance between local and global explanations to ensure fairness and transparency in AI systems.

A strategy could be to combine SHAP for local feature attributions with global analysis by grouping instances by protected attributes (e.g., race, gender) and computing average SHAP values per subgroup. Use Partial Dependence Plots (PDPs) for each subgroup to detect differential treatment, ensuring both detailed insights and broad fairness checks.

4 LLMs and Prompt-driven Interpretability

A two-step strategy could be as so:

1. Prompt: "Solve this problem step by step and box your final answer: [problem]." This encourages chain-of-thought reasoning, where the model explains its thought process, enhancing interpretability.
2. Prompt: "Now, explain why each step is necessary and correct, and consider alternatives." This tests conssitency and depth, revealing if reasoning is grounded and robust.

Direct prompting may yield confabulated explanations, not reflecting internal processes, as LLMs mimic plausible text. Section 5.7 of the reference paper [1] notes that LLMs can still generate errors or skip details, thus limiting interpretability and insight.

5 Quantifying Explanation "Quality"

Metrics:

- **Fidelity:** Measures explanation accuracy, defined as $FS = E[1(f(x) = g(x))]$ for classification or R^2 for regression, reflecting model behaviour.
- **Comprehensibility:** This is subjective, defined as $ES = 1/(1 + Complexity(E))$ where $Complexity(E)$ is the based on features or depth, assessing human understanding.

Experiment:

- Dataset: Iris (4 features, 3 classes).
- Black-box method: Random Forest (100 trees)
- Method: LIME for explanations
- Compute Fidelity as average R^2 on 500 perturbed instances per test case; Comprehensibility as $ES = 1/(1 + \text{features with } —\text{coefficient}— > 0.01)$, averaged.
- Expected: Higher fidelity with more trees, but lower comprehensibility due to complexity.

If the black-box model is replaced with an intrinsically interpretable model, we can expect the fidelity to be perfect since the model is interpretable inherintly. We can also expect the comprehensibility to be higher, reflecting simpler rules. Black-box trades comprehensibility for accuracy; interpretable model offers perfect fidelity, higher comprehensibility, but potentially lower accuracy.

References

[1] Weiche Hsieh, Ziqian Bi, Chuanqi Jiang, Junyu Liu, Benji Peng, Sen Zhang, Xuanhe Pan, Jiawei Xu, Jinlang Wang, Keyu Chen, Caitlyn Yin, Pohsun Feng, Yizhu Wen, Xinyuan Song, Tianyang Wang, Junjie Yang, Ming Li, Bowen Jing, Jintao Ren, and Ming Liu. A comprehensive guide to explainable ai: From classical models to llms, 12 2024.