Probabilistic Methods for Plausibility, Robustness, and Diversity in Counterfactual Explanations

Artificial intelligence (AI) increasingly influences our lives, from recommending products to aiding medical diagnoses. However, many AI systems operate as "black boxes," making their decisions hard to understand. For instance, if an AI denies a loan, users often don't know why or how to improve their chances. This lack of transparency hinders trust in AI.

Counterfactual explanations tackle this challenge by offering "what if" scenarios, showing how small changes to input data could lead to different outcomes. For example, increasing income or reducing debt might change a loan rejection to an approval. However, good counterfactuals must be plausible (realistic), robust (stable under slight changes), and diverse (offering multiple viable options).

Figure 1 illustrates counterfactuals navigating an AI model's decision boundary, moving from the original class (e.g., loan rejected) to the desired class (e.g., loan approved). The figure highlights examples of plausible and implausible suggestions and high-density regions where feasible counterfactuals are more likely.

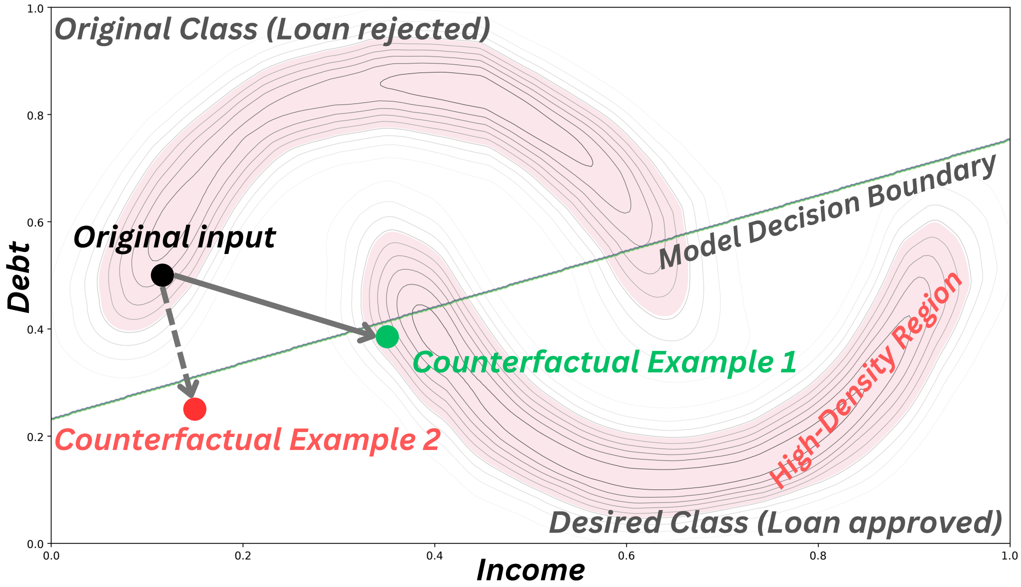

Figure 1: Counterfactual explanations in AI decision-making. The plot shows the decision boundary between two classes: original (loan rejected) and desired (loan approved). The black dot represents the original input, while green and red dots illustrate plausible and implausible counterfactuals, respectively. High-density regions indicate areas where realistic counterfactuals are more feasible.

This research project aims to enhance the quality of counterfactual explanations by leveraging probabilistic methods to address three essential aspects: plausibility, robustness, and diversity. By capturing uncertainty and variability in data, probabilistic models enable the generation of realistic and practical suggestions. For instance, instead of proposing an impractical income increase, these methods can identify plausible adjustments based on real-world data patterns.

The project focuses on developing advanced techniques that generate counterfactuals capable of balancing these three aspects effectively. The resulting methods will be versatile, applicable to diverse fields such as healthcare and finance, and will provide practical tools and guidelines for building more transparent AI systems. Ultimately, this work aims to foster trust in AI by making its decisions clearer, actionable, and more aligned with human understanding, bridging the gap between complex algorithms and their users.