Initial Plan: Algorithms for Generating Robust and Diverse Counterfactuals

Abbie Griffiths (22026335)

2nd Februray 2025

Module: CM3203
Supervisor: Nico Potyka

School of Computer Science and Informatics at Cardiff University

Contents

1 Project Description

The increased occurrence of black-box machine learning models being implemented into decision making systems has created a new need for explanations behind these decisions. This project focuses on the field of counterfactual explanations. These are used to explain the decision process machine learning models make, allowing the user to see what changes would alter the outcome. A great example of an industry using counterfactual explanations is finance, more specifically, loan application scenarios. When an application is rejected, counterfactual explanations will identify and explain the minimum amount of changes needed to approve the application, such as increasing income or reducing debt levels.

While counterfactual explanations are intuitive, they struggle with robustness, as a similar input can receive very different explanations, leading to user confusion and potential vulnerability to adversarial attacks [1]. Current approaches address this issue by generating multiple diverse counterfactual explanations rather than a single one. However, the computational problem is quite challenging, as proposed in the heuristic to solve this problem in Promoting Counterfactual Robustness through Diversity [2] which works quite well compared to DICE, the state-of-the-art model [3], but there is still the need for improvement.

This project is fundamentally an optimisation problem as there is an input x, for instance a loan application with information such as income and debt, and a new input x' needs to be found so that the model changes its decision from a rejection to an acceptance. This needs to be achieved with minimal changes to x. Therefore, the mathematical formulation would be something like minimise distance(x,x'), using the Euclidean distance, subject to x' obtaining the desired outcome. The distance is taken to quantify the robustness of the counterfactual explanation, and the aim is to keep the distance minimal between x and x' so that it is easier for the user to make the necessary changes to acquire the output they want.

Consequently, for the algorithmic approach, I will start with the original input of x and define the search space x can have. I will then use a distance metric to measure the similarity x and x' have. I will be looking at different search strategies to find the counterfactuals, such as Genetic Algorithms.

This project is a sufficiently challenging topic as its an optimisation problem that needs to balance multiple objectives of similarity, diversity and computational efficiency. The scope of this project is constrained to be the development of new search strategies, with the evaluation being how effective my proposed search strategy is compared to current ones, therefore making this a feasible field of exploration that can be completed within the time frame.

2 Aims and Objectives

The overall aim of this project it to develop an improved algorithm for generating robust and diverse counterfactual explanations for black-box machine learning models, specifically focusing on improving computational efficient and explanatorily quality using search strategies.

Specific objectives:

1. Design a search strategy that balances the trade-off between similarity and diversity in counterfactual explanations.

- **Risk:** The new search strategy may not show signs of significant improvement compared to existing approaches.
- **Mitigation:** Develop another search strategy and identify the one that shows the most promise or systematically improve the first strategy by analysing why it does not show much improvement.

2. Establish run times for the proposed algorithm.

- **Risk:** Formal proofs may be challenging.
- **Mitigation:** Preform systematic runtime experiments.

3. Implement the search strategy.

- **Risk:** Could be challenging to implement due to the structure of the algorithm.
- **Mitigation:** Start with a simple prototype and add features incrementally to ensure it works.

3 Feasibility

I foresee no significant issues that may affect the feasibility of the project at this time. Since no humans or personal data is involved, no ethical approval will be required for this project. For resource requirements, all required software tools (Python and optimisation libraries) are freely available. For legal considerations regarding this project, open-source libraries and frameworks will be used. Regarding the technical feasibility for this project, the computing resources required are easily accessible, and if the initial approach does not work, there are other alternatives to trial. Additionally, the existence of current approaches that generate counterfactual examples demonstrates that the basic concept is implementable.

4 Work Plan

Week 1 - 3 :

- Research and background reading, reviewing existing counterfactual algorithms and implementations.
- Analysis current robustness issues.
- Begin literature review write up.

- Milestones: Abstract and Introduction sections completed.

Week 4 - 5 :

- Algorithm Design Phase
- Begin to design a search strategy, such as Genetic Algorithms.
- Begin write up for the Methodology and Algorithm Design section.

- Milestones: Literature review completed.

Week 6 (and reading week) :

- Initial Implementation phase.
- A prototype/search strategy created with basic functionality.

- Milestones: Methodology and Algorithm Design first section completed and basic prototype created.

Week 8 - 9 :

- Create framework for runtime analysis.
- Implementation improvement phase conducted by analysing the areas of weakness in the current search strategy proposed.
- Write up of Implementations section with the improvements made.

- Milestones: Implementation section write up.

Week 10 - 11 :

- Testing and Evaluation of search strategy. For example, with Genetic Algorithms the fitness of the algorithm can be used as quantitative measure to evaluate.
- Analysis of runtime.
- Compare algorithm and results with existing approaches.
- Testing and Evaluating report section write up started.

- Milestones: Testing completed and Testing and Evaluation section written.

Week 12 :

- Complete report writing.

- Milestones: Conclusion written and final send off.

References

[1] Dylan Slack et al. “Counterfactual explanations can be manipulated”. In: *Advances in neural information processing systems* 34 (2021), pp. 62–75.

[2] Francesco Leofante and Nico Potyka. “Promoting counterfactual robustness through diversity”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 19. 2024, pp. 21322–21330.

[3] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. “Explaining machine learning classifiers through diverse counterfactual explanations”. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* ’20. Barcelona, Spain: Association for Computing Machinery, 2020, pp. 607–617. ISBN: 9781450369367. DOI: 10.1145/3351095.3372850. URL: https://doi.org/10.1145/3351095.3372850.