AI for personalization: from predictive to generative modeling.

Gabriel Bénédict

AI for personalization: from predictive to generative modeling.

Academisch Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. G.T.M. ten Dam
ten overstaan van een door het College voor Promoties ingestelde
commissie, in het openbaar te verdedigen in
de Aula der Universiteit
op vrijdag 01 January 2024, te 14:00 uur

door

Gabriel Bénédict

geboren te Chavannes-Pres-Renens, Zwitserland

Promotiecommissie

Promotor:		
	Prof. dr. M. de Rijke	Universiteit van Amsterdam
Co-promotor:		
	dr. D. Odijk	RTL NL

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

This research was (partially) funded by Bertelsmann SE & Co. KGaA and by the Hybrid Intelligence Center, a 10-year program funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, `https://hybrid-intelligence-centre.nl`.

Copyright © 2024 Gabriel Bénédict, Amsterdam, The Netherlands
Cover by Off Page, Amsterdam
Printed by Off Page, Amsterdam

ISBN:

Certains pensent qu'ils font un voyage, en fait, c'est le voyage qui vous fait ou vous défait.

– Nicolas Bouvier

Acknowlegements

Gabriel Bénédict
Amsterdam
January 2024

Table of Contents

Chapter 1

Introduction

Define a streaming platform

personalization

pipeline

generative recommendation

multilabel classification

intent

RADio

1.1 Research Outline and Questions

We scope the thesis around four research questions, each corresponding to a chapter in the thesis.

A streaming platform ()

Personalization on streaming platforms is oftentimes perceived as a purely predictive phenomenon: we propose to view it as a comprehensive and responsible generative approach, throughout a pipeline. We introduce RecFusion to issue recommendations in a generative way with diffusion models, as part of the nascent Generative Information Retrieval field. For these recommendations, we propose a method to generate personalized stills from movies, with sigmoidF1. We show that the resulting interactions on platforms are also dependent on implicit data hidden from a web analytics platform, with our intent-satisfaction analysis. At the end of the pipeline, we propose to ensure normative diversity in the issued recommendations with our RADio metrics framework.

RQ 1. Can we use diffusion to do recommendation in the classical user-item matrix setting?

Is Bernoulli diffusion a suitable forward and backward process theoretically and empirically on binary data?

RQ 2. Is there a way we can generate personalized posters and stills for each item on a streaming platform?

Can we formulate loss function that accommodates for multilabel classification at training time and operates on the whole batch to balance confusion matrix entries?

RQ 3. Are users' intents together with their behavioral data useful signals to predict or explain satisfaction on a video streaming platform?

Can we use Bayesian posterior draws to meaningfully draw conclusions from data?

RQ 4. Can we formulate a divergence metric that measures the normative diversity of recommendations?

This work applies to news recommendation but trivially generalizes to any domain that has categories (e.g. video streaming with movie genres).

Can we formulate a divergence metric that is distributional and rank-aware?

1.2 Main Contributions

In this section, we summarize the main contributions of this thesis.

Theoretical Contributions

- Diffusion applied on unstructured data, where there is no spatial dependency
- Diffusion for binary and/or 1D data: A demonstration that KL divergence is also suited for binary data and that the Bernouilli Markov Process has the same properties as its Gaussian counterpart.
- A multilabel loss function that accounts for all examples in a batch
- An F1 score surrogate as a loss function.
- An account of the current limitations and underreporting of thresholding at inference time.
- a proposal of typical intents for a video streaming that we divide into explorative and decisive categories
- A diversity metric that is versatile to any normative concept and expressed as the divergence between two (discrete) distributions, rank-aware and mathematically grounded in distributional divergence statistics.

Artifact Contributions **[Gab: I found this expression here, but I am not sure it is commonly used]**

- Frequentist logistic regression model, we test Bayesian multilevel models for visualization and explanations, along with random forests for improved accuracy
- A reproducibility study from music to video streaming platforms of intent-satisfaction modeling (this time with code and synthetic data).
- In-app survey design for a medium size streaming platform (∼1 million users) and corresponding synthetic data.
- A metadata enrichment pipeline (e.g., sentiment analysis, named entity recognition) to extract normative concepts from news articles.

1.3 Thesis Overview

This thesis is organized into three parts, each part can be read independently.

The first part focuses on proposing new algorithms for explaining predictions from ML models. Specifically, we propose methods for generating counterfactual explanations for tree-based models (Chapter **??**), and for graph-based models (Chapter 2). These methods can be applied on any tree- or graph-based model, respectively.

The second part focuses on the interaction between ML explanations and the users who consume them. We propose a method for explaining errors in forecasting predictions (Chapter 4). To evaluate our method, we propose a user study with both objective and subjective components, where we contrast and compare the results between two types of users: researchers and practitioners.

In the third part of the thesis, we shift our focus from translating knowledge about individual predictions to transferring knowledge to the next generation of researchers. We propose a course setup for teaching about responsible AI topics to a graduate-level audience and reflect on our learnings from past implementations of the course at the University of Amsterdam (Chapter **??**).

1.4 Origins

We list the publications that are the origins of each chapter below.

Chapter 1 is based on the following paper:

- Gabriel Bénédict, Olivier Jeunen, Samuele Papa, Samarth Bhargav, Daan Odijk, and Maarten de Rijke. Recfusion: A binomial diffusion process for 1d data for recommendation, 2023.

GB wrote the first draft, code, experiments and mathematical formulations. GB was helped all along the way via discussion with the coauthors. Coauthors then edited the first draft together with GB. GB did most of the writing.

Chapter 2 is based on the following paper:

- Gabriel Bénédict, Hendrik Vincent Koops, Daan Odijk, and Maarten de Rijke. sigmoidf1: A smooth f1 score surrogate loss for multilabel classification. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL `https://openreview.net/forum?id=gvSHaaD2wQ`.

GB wrote the first draft, code, experiments and mathematical formulations. GB was helped all along the way via discussion with the coauthors. Coauthors then edited the first draft together with GB. GB did most of the writing.

Chapter 3 is based on the following paper:

- Gabriel Bénédict, Daan Odijk, and Maarten de Rijke. Intent-satisfaction modeling: From music to video streaming. *ACM Trans. Recomm. Syst.*, 1(3), aug 2023. doi: 10.1145/3606375. URL `https://doi.org/10.1145/3606375`.

GB wrote the first draft, code, experiments and mathematical formulations. GB was helped all along the way via discussion with the coauthors. Coauthors then edited the first draft together with GB. GB did most of the writing.

Chapter 4 is based on the following paper:

- Sanne Vrijenhoek, Gabriel Bénédict, Mateo Gutierrez Granada, Daan Odijk, and Maarten De Rijke. Radio – rank-aware divergence metrics to measure normative diversity in news recommendations. In *Proceedings of the 16th ACM Conference on Recommender Systems*, RecSys '22, page 208–219, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392785. doi: 10.1145/3523227.3546780. URL `https://doi.org/10.1145/3523227.3546780`.

GB, together with SV, wrote the first draft, code, experiments and mathematical formulations. GB was helped all along the way via discussion with the coauthors. Coauthors then edited the first draft together with GB. GB and SV did most of the writing.

The writing of this thesis also benefited from work on the following publications:

- Gabriel Bénédict, Ruqing Zhang, and Donald Metzler. Gen-ir @ sigir 2023: The first workshop on generative information retrieval, 2023.
- Shashank Gupta Maria Heuss Pooya Khandel Ming Li Fatemeh Sarvi Ali Vardasbi, Gabriel Bénédict. The university of amsterdam at the trec 2021 fair ranking track. *TREC Fair Ranking*, 2021.
- .

Chapter 2

Generative Recommendations with Diffusion

??: rq:cf-gnn!

Reproducibility

To facilitate the reproducibility of the work in this chapter, our code is available at `https://github.com/gabriben/recfusion`.

This chapter is under submission at International Conference on Learning Representations (ICLR) under the title “RecFusion: A Binomial Diffusion Process for 1D Data for Recommendation” [**?**].

Chapter 3

Metrics as Losses

In this chapter, we address the following research question:

??: rq:focus!

Reproducibility

To facilitate the reproducibility of this work, our code is available at `https://github.com/gabriben/metrics-as-losses`.

This chapter was published at the Transactions of Machine Learning Research (TMLR) under the title “sigmoidF1: A Smooth F1 Score Surrogate Loss for Multilabel Classification” [2].

Chapter 4

In this chapter, we address the following research question:

??: rq:mcbrp!

Reproducibility

To facilitate the reproducibility of the work in this chapter, our code is available at `https://github.com/a-lucic/mcbrp`.

This chapter was published at the ACM Transactions on Recommender Systems (TORS) under the title “Intent-Satisfaction Modeling: From Music to Video Streaming” [3]

Chapter 5

Normative Diversity

In this chapter, we address the following research question:

??: rq:pedagogy!

This chapter was published at the ACM Conference on Recommender Systems (RecSys 2022) under the title “RADio – Rank-Aware Divergence Metrics to Measure Normative Diversity in News Recommendations” [6], , where it won a best paper runner up award.

Chapter 6

Conclusions

In this thesis, we have investigated explainability in ML from three viewpoints: (i) algorithms, (ii) users, and (iii) pedagogy. In this final chapter of the thesis, we revisit the research questions from Chapter 1, state our main findings in Section 6.1, and identify directions for future work in Section 6.2.

6.1 Main Findings

In this section, we describe our main findings across the three parts of the thesis.

6.1.1 Algorithms

The first part of this thesis focused on investigating behavior-based explanations in order to explain individual predictions from specific types of ML models. In Chapter **??**, we asked our first research question:

rq:focus! rq:focus!

The answer to **RQ1** is yes: we are able to explain predictions for tree ensembles in a counterfactual manner by including differentiable approximations of tree-based models within a standard gradient-based optimization framework. In the majority of experimental settings, our method outperforms existing baselines in terms of (i) the number of counterfactual examples produced, (ii) the average distance between the counterfactual examples and the original examples, and (iii) the proportion of counterfactual examples that are closer to the original examples. Our method is flexible since it can produce different types of counterfactual explanations depending on which distance function we choose to include in the loss function. In practice, this allows the user to customize the explanations depending on the use case.

We then turned to our next research question:

rq:cf-gnn! rq:cf-gnn!

The answer to **RQ2** is yes: we can adapt our method from **RQ1** to the graph setting by introducing a binary perturbation matrix that is multiplied element-wise with the adjacency matrix in order to remove edges from the graph. Since this was one of the first methods for generating counterfactual explanations for GNNs, we also had to design a corresponding experimental setup to evaluate it.

In the majority of experimental settings, we found that our method outperformed the baselines in terms of (i) accuracy, (ii) the number of counterfactual examples produced, (iii) the number of edges removed, and (iv) the proportion of the subgraph neighborhood that was perturbed.

6.1.2 Users

In the second part of this thesis, we continued our work on behavior-based explanations, but shifted our predominantly algorithmic focus to also account for the users who consume the explanations. We investigated the following research question:

rq:mcbrp! rq:mcbrp!

The answer to **RQ3** is yes: we developed an explanation method based on the needs of real-world analysts in order to help them understand large errors in sales forecasting predictions. We designed a user study to evaluate our method and found that for the vast majority of users, our explanations were both interpretable and actionable. We also found that most users believed the explanations helped them understand large errors in predictions, but they did not have an impact on other aspects such as trust or confidence in the model.

6.1.3 Pedagogy

In the third part of this thesis, we transitioned from creating knowledge about ML model predictions to communicating knowledge about responsible ML practices. Process-based explanations play an important role here, specifically in the context of documentation practices which help ensure research is conducted in a responsible and reproducible manner. We asked our final research question:

rq:pedagogy! rq:pedagogy!

We answered **RQ4** by developing a course that was centered on a reproducibility project, where students worked in groups to reimplement algorithms from top-AI conferences on responsible AI topics. We shared our experiences with teaching the course over two academic years and suggested best practices for implementing similar reproducibility courses in the future.

6.2 Future Directions

In this section, we describe some limitations of the methods proposed in this thesis and identify potential avenues for future work.

Limitations of counterfactual explanations

Researchers have raised concerns about the hidden assumptions behind the use of counterfactual examples [**?**], as well as potentials for misuse [**?**]. When explaining ML models using counterfactual examples, it is important to account for the context in which the systems are deployed. Counterfactual explanations are not a guarantee to achieving recourse [**?**] – changes suggested should be seen as candidate changes, not absolute solutions, since what is pragmatically actionable differs depending on the end user and context. While existing research from the cognitive sciences has shown that humans are able to interpret counterfactual explanations, the notion of what constitutes a *minimal* perturbation is not clear [**?**]. Further research into the interpretability and cognitive efficacy of counterfactual explanations could help the field better understand the appropriate criteria to optimize for.

Accommodating different types of perturbations

In Chapter 2, we proposed a method for generating counterfactual explanations for GNNs. In its current form, our method is limited to performing edge deletions for node classification tasks. Given that many graph datasets also include node features, future work should involve incorporating node feature perturbations in our framework. We could also extend our method to accommodate graph classification tasks.

Including additional criteria in loss functions

In Chapter **??**, we proposed a method for generating flexible counterfactual explanations for tree-based models using gradient optimization techniques. The flexibility comes from varying the distance function used in the loss function, which results in different types of counterfactual explanations depending on which distance function is chosen. Future work could involve trying alternative distance functions or including additional criteria in the loss function, such as proximity to other points in the dataset or stability of the counterfactual example. This could also be applied to the method proposed in Chapter 2.

Evaluating with users

Although the counterfactual explanations proposed in Chapters **??** and 2 perform well on various distance metrics, we should conduct a user study to evaluate how useful they are in practice. We could build on our existing user study design from Chapter 4 to test how (i) varying the distance functions, and (ii) introducing new components into the loss functions impacts user preferences for counterfactual explanations.

Improving trust in explanations

In Chapter 4, we proposed a method for explaining errors in forecasting predictions based on identifying unusual feature values. We find that although explanations from our method help users understand large errors in predictions, they do not have an impact on users' trust, deployment support, or perception of the model's performance. Future work could place more emphasis on trying to improve these aspects, for example by allowing a predictive model to abstain from prediction when a particular instance has unusual feature values beyond a certain threshold.

Developing robust protocols for XAI evaluation

In general, we believe it is crucial for the ML community to invest in developing more rigorous evaluation protocols for XAI methods, both in terms of user studies as well as formal metrics. The XAI community could pursue collaborations with researchers from human-computer interaction to design human-centered user studies about evaluating the utility of XAI methods in practice. To design metrics, the XAI community could try borrowing ideas from information theory or collaborating with ML evaluation researchers in order to ensure that the explanations we generate are truly representative of the model's behavior.

Teaching reproducibility as a fundamental component of ML research

Reproducibility mechanisms such as checklists and challenges can help promote reproducible research practices, but we do not believe they alone are enough to cause a shift in the ML community. We believe the key to fostering reproducible research starts in the classroom. It is important to teach the next generation of ML researchers that reproducibility is not an afterthought, but rather a fundamental component of conducting ML research responsibly. In addition to conducting reproducibility projects, we could also introduce reproducibility components in programming assignments across all courses within an ML study program.

Identifying consistent terms for explainability

Due to growing collection of XAI literature, there are many definitions for various distinct but related concepts such as explainability, interpretability, transparency, and intelligibility. As a community, we should make an effort to standardize the terms we use in order to facilitate easier communication, especially with researchers from non-ML disciplines. Developing XAI that is useful in practice requires interdisciplinary collaboration, which is more straightforward if we can all speak the same language. This could be achieved through a workshop-style event with researchers working on XAI to consolidate a standardized terminology.

Final thoughts

Overall, our main advice for future work is to continue prioritizing explainability in the ML community, whether it is *behavior-based* or *process-based.* For both types of explanations, we should explore developing explainability techniques that cater to different types of users with varying levels of granularity, as well as robust mechanisms for evaluation. As a community, we need to prioritize both correctness and interpretability of explanations – incorrect explanations that are interpretable do not provide the user with any concrete information, and correct explanations that are uninterpretable are not useful to the user. To promote correctness, we need to first identify what it means for an explanation to be “correct” and create datasets that allow us to explore this task explicitly. To promote interpretability, we need to approach the explainability problem from an interdisciplinary perspective, and suggest that XAI researchers spend more time connecting to the communities they are designing the explanations for.

Bibliography

[1] Shashank Gupta Maria Heuss Pooya Khandel Ming Li Fatemeh Sarvi Ali Vardasbi, Gabriel Bénédict. The university of amsterdam at the trec 2021 fair ranking track. *TREC Fair Ranking*, 2021.

[2] Gabriel Bénédict, Hendrik Vincent Koops, Daan Odijk, and Maarten de Rijke. sigmoidf1: A smooth f1 score surrogate loss for multilabel classification. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL `https://openreview.net/forum?id=gvSHaaD2wQ`. (Cited on page 7.)

[3] Gabriel Bénédict, Daan Odijk, and Maarten de Rijke. Intent-satisfaction modeling: From music to video streaming. *ACM Trans. Recomm. Syst.*, 1(3), aug 2023. doi: 10.1145/3606375. URL `https://doi.org/10.1145/3606375`. (Cited on page 9.)

[4] Gabriel Bénédict, Olivier Jeunen, Samuele Papa, Samarth Bhargav, Daan Odijk, and Maarten de Rijke. Recfusion: A binomial diffusion process for 1d data for recommendation, 2023.

[5] Gabriel Bénédict, Ruqing Zhang, and Donald Metzler. Gen-ir @ sigir 2023: The first workshop on generative information retrieval, 2023.

[6] Sanne Vrijenhoek, Gabriel Bénédict, Mateo Gutierrez Granada, Daan Odijk, and Maarten De Rijke. Radio – rank-aware divergence metrics to measure normative diversity in news recommendations. In *Proceedings of the 16th ACM Conference on Recommender Systems*, RecSys '22, page 208–219, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392785. doi: 10.1145/3523227.3546780. URL `https://doi.org/10.1145/3523227.3546780`. (Cited on page 11.)

Summary

Model explainability has become an important problem in artificial intelligence (AI) due to the increased effect that algorithmic predictions have on humans. Explanations can help users understand not only why AI models make certain predictions, but also how these predictions can be changed. In the first part of this thesis, we investigate counterfactual explanations: given a data point and a trained model, we want to find the minimal perturbation to the input such that the prediction changes. We frame the problem of finding counterfactual explanations as a gradient-based optimization task and first focus on tree ensembles. We extend previous work that could only be applied to differentiable models by incorporating probabilistic model approximations in the optimization framework, and find that our counterfactual examples are significantly closer to the original instances than those produced by other methods specifically designed for tree-based models.

We then extend our method for generating counterfactual explanations for tree ensembles to accommodate graph neural networks (GNNs), given the increasing promise of GNNs in real-world applications such as fake news detection and molecular simulation. We do so by introducing a perturbation matrix that acts on the adjacency matrix in order to iteratively remove edges from the graph, and find that our method primarily removes edges that are crucial for the original predictions, resulting in minimal counterfactual explanations.

In the second part of this thesis, we investigate explanations in the context of a real-world use case: sales forecasting. We propose an algorithm that generates explanations for large errors in forecasting predictions based on Monte Carlo simulations. To evaluate, we conduct a user study with 75 users and find that the majority of users are able to accurately answer objective questions about the model's predictions when provided with our explanations, and that users who saw our explanations understand why the model makes large errors in predictions significantly more than users in the control group. We also conduct an in-depth analysis of the difference in attitudes between practitioners and researchers, and confirm that our results hold when conditioning on the users' background.

In the final part of the thesis, we explain the setup for a technical, graduate-level course on responsible AI topics at the University of Amsterdam, which teaches responsible AI concepts through the lens of reproducibility. The focal point of the course is a group project based on reproducing existing responsible AI algorithms from top AI conferences and writing a corresponding report. We reflect on our experiences teaching the course over two years and propose guidelines for incorporating reproducibility in graduate-level AI study programs.

Samenvatting

De uitlegbaarheid van voorspellende modellen is een belangrijk probleem geworden in kunstmatige intelligentie (KI) vanwege het toegenomen effect dat algoritmische voorspellingen hebben op mensen. Een uitleg kan gebruikers niet alleen helpen om te begrijpen waarom KI-modellen bepaalde voorspellingen doen, maar ook hoe deze voorspellingen beïnvloed kunnen worden. In het eerste deel van dit proefschrift onderzoeken we contrafeitelijke verklaringen: we willen, gegeven een datapunt en een getraind model, de minimale verandering van de input vinden die de voorspelling verandert. We formuleren het probleem van het vinden van contrafeitelijke verklaring als een op gradiënten gebaseerde optimalisatietaak en richten ons eerst op *tree ensembles*. We bouwen voort op eerder werk dat alleen kon worden toegepast op differentieerbare modellen, door probabilistische modelbenaderingen op te nemen in het optimalisatiekader, en komen tot de bevinding dat onze contrafeitelijke voorbeelden significant dichter bij het oorspronkelijke datapunt liggen dan de voorbeelden die geproduceerd worden door andere methoden, die specifiek zijn ontworpen voor modellen die op bomen zijn gebaseerd.

Vervolgens breiden we onze methode voor het genereren van contrafeitelijke verklaringen uit voor *tree ensembles* zodat die ook werkt voor *graph neural networks* (GNNs), gezien de toenemende belofte van GNNs voor toepassingen in de echte wereld, zoals de detectie van nepnieuws en moleculaire simulatie. We bereiken dit door een perturbatiematrix te introduceren die de elementen uit de bogenmatrix vermenigvuldigt om iteratief zijden van de graaf te verwijderen, en komen tot de bevinding dat onze methode voornamelijk zijden verwijdert die cruciaal zijn voor de oorspronkelijke voorspelling, wat resulteert in een minimale contrafeitelijke verklaring.

In het tweede deel van dit proefschrift onderzoeken we verklaringen in de context van een praktijkvoorbeeld: verkoopprognoses. We introduceren een algoritme dat verklaringen genereert voor grote fouten bij het doen van regressievoorspellingen op basis van Monte Carlo-simulaties. We evalueren door middel van een gebruikersonderzoek met 75 deelnemers. We komen tot de bevinding dat de meerderheid van de gebruikers in staat is accuraat antwoord te geven op meerkeuzevragen over de voorspellingen van het model wanneer zij voorzien zijn van onze uitleg. Daarnaast komen we tot de bevinding dat gebruikers die onze uitleg zagen, significant vaker dan gebruikers uit de controlegroep begrijpen waarom het model grote fouten maakt in voorspellingen. We voeren ook een analyse uit van het verschil in attitudes tussen praktijkbeoefenaars en onderzoekers, en bevestigen dat onze resultaten stand houden gegeven de

achtergrond van de gebruiker.

In het laatste deel van dit proefschrift behandelen we de opzet van een technisch vak op masterniveau over verantwoord gebruik van KI, gegeven aan de Universiteit van Amsterdam. In dit vak worden verantwoorde KI-concepten vanuit het oogpunt van reproduceerbaarheid gedoceerd. Het speerpunt van de cursus is een groepsproject dat gebaseerd is op het reproduceren van bestaande verantwoorde KI-algoritmen van vooraanstaande KI-conferenties en het schrijven van een bijbehorend rapport. We reflecteren op onze ervaringen met het geven van de cursus gedurende twee jaar en stellen richtlijnen voor voor het opnemen van reproduceerbaarheid in KI-studieprogramma's op masterniveau.