Hackathon Subject Description (Saint-Gobain): Causal Discovery from Sequential Data

June 12, 2023

Are you ready to dive into causal discovery in sequential data? Join us for an exciting Hackathon, where we will explore the fascinating field of causal reasoning using a rich dataset provided by Saint-Gobain, a leading materials and solutions company.

1 Context

Saint-Gobain Distribution Bâtiment France (SGDBF) is dedicated to developing innovative solutions that enhance well-being and contribute to a sustainable future. They have gathered extensive data on client interactions and actions as part of their efforts. This sequential data provides valuable insights into clients' identities and behaviors, and understanding the causal relationships within this data can unlock numerous possibilities for improving customer understanding and anticipating future actions.

Practical Application: One specific area of interest for Saint-Gobain is understanding building site dynamics. Many customers request Saint-Gobain for the construction products they need to meet their customers' site requirements. However, craftsmen only purchase some of their products from Saint-Gobain, not all of them. This can be interpreted as an observed purchasing trajectory that is not complete but that we would like to complete causally. Causal discovery techniques can help uncover the latent causal factors behind the incomplete, partially expressed needs, enabling Saint-Gobain to assess its capacity to respond to the global need and prepare advantageous offers.

2 Problem

In this Hackathon, we aim to unravel the causal structures hidden within the sequential data of Saint-Gobain's clients. By identifying causal relationships, we can better understand clients' needs and behaviors, developing more relevant and efficient recommendation systems. Furthermore, capturing the structure that governs the sequentiality of client actions allows us to anticipate clients' future needs and provide tailored solutions.

Dataset Details: The dataset provided for this Hackathon consists of a history of transactions, where a unique ID represents each client. Clients make purchases on specific dates, buying various products. Products can be grouped into multiple bundles based on similar needs. Transactions involve purchasing multiple products, and there is a causal relation between these actions. For example, the order of purchasing plasterboards might be motivated by a hidden reason, such as a kitchen renovation

project, leading to subsequent purchases of related products like rails, screws, and glue.

In this Hackathon, our primary goal is to perform causal discovery on the sequential data of purchases provided by Saint-Gobain. We aim to identify the temporal causal relationships between products or product families and whether such relations are stationary or non-stationary. Additionally, we seek to explore the appropriate level of granularity for learning causal graphs, whether at the article level, sub-family level, or family level.

Exploring Granularity Levels: One of the key considerations in learning a causal graph from transaction data is determining the appropriate level of granularity. Granularity refers to the level of detail at which products or product families are considered individual entities. Each level of granularity provides a different perspective on the causal relationships and poses its challenges:

- **Article Level:** Individual articles are treated as distinct entities at the finest granularity level. This level of detail allows for a more precise understanding of the causal relationships between specific products. However, it may pose challenges regarding data sparsity and the need for large data to capture meaningful causal patterns.
- **Sub-Family Level:** Sub-families group similar products based on specific attributes or characteristics. Analyzing causal relationships at the sub-family level provides a balance between granularity and data availability. It allows for a more manageable number of variables while still capturing relevant causal connections.
- **Family Level:** At the highest level of granularity, products are grouped into broader families based on their general characteristics. Analyzing causal relationships at the family level provides a more holistic understanding of the causal structure. It reduces the dimensionality of the data and simplifies the analysis. However, it may overlook subtle differences and variations within product families that could be relevant for causal discovery.

Challenges of Causal Discovery on a Large Number of Variables: Performing causal discovery on a large number of variables that evolve presents several challenges:

- **Sparsity:** With many variables, sparse data becomes more prevalent. Sparse data hinders the accurate identification of causal relationships, as there may be insufficient instances of specific combinations of variables to establish robust causal connections.
- **Computational Efficiency:** Dealing with many variables requires efficient algorithms and computational resources to perform causal discovery in a reasonable timeframe. Scalable methods that can handle the computational burden are essential.
- **Interpretability:** Interpreting causal relationships in a large-scale context can be challenging. As the number of variables grows, understanding the underlying mechanisms and causal pathways becomes more complex. Ensuring the interpretability of the discovered causal graph is crucial for actionable insights.

3 Challenges of Learning a Causal Graph:

1. **Temporal Dependencies:** The sequential nature of the transaction data introduces temporal dependencies that need to be properly accounted for. Discovering causal relationships becomes challenging as the order of actions can significantly impact subsequent purchases. Suitable methods for handling temporal dependencies and capturing causal directionality are essential.
2. **Hidden Factors:** Uncovering causal relationships requires identifying and accounting for hidden factors influencing purchasing decisions. These hidden factors could be external events, customer preferences, or contextual factors contributing to the causal structure.

3. **Data Sparsity:** Transaction histories can be sparse in real-world datasets, especially for certain products or product families. Sparse data can hinder the discovery of causal relationships and pose challenges in accurately inferring causality. Advanced techniques are required to handle sparse data and effectively leverage available information.

Practical Implications of Learning a Causal Graph:

1. **Enhanced Recommendation Systems:** By uncovering causal relationships within the transaction data, companies like Saint-Gobain can develop advanced recommendation systems that go beyond simple associations. Understanding the causal drivers behind purchasing decisions allows for more accurate and personalized recommendations, improving customer satisfaction and driving business growth.

2. **Anticipating Customer Needs:** Learning the causal structure of transaction data enables businesses to anticipate customer needs and proactively offer relevant products or services. By identifying the trigger products or events that lead to subsequent purchases, companies can predict future demands and tailor their offerings accordingly, increasing customer loyalty and profitability.

3. **Strategic Decision-Making:** Causal graphs derived from transaction data provide valuable strategic decision-making insights. Understanding the causal drivers behind customer behavior allows companies to make informed decisions about product development, pricing strategies, marketing campaigns, and partnerships, ensuring effective resource allocation and maximizing business outcomes.

By addressing the challenges of learning a causal graph and leveraging the practical implications, businesses can unlock the potential of their transaction data to gain a deeper understanding of customer behavior, optimize operations, and drive innovation in their respective industries. Join us at the Hackathon, and let's unravel the causal mysteries hidden within the sequential data provided by Saint-Gobain!

Contact:
Mouad El Bouchattaoui, mouad.elbouchattaoui@saint-gobain.com.
Laura Maag, laura.maag@saint-gobain.com