HUL315 Assignment-1

Amaiya Singhal 2021CS50598

1. Importance of causality in data analytics

Let us have a look at the Wikipedia definition of Data Analysis.

> " Data analysis is the process of inspecting, cleansing, transforming, and modeling data with the goal of discovering **useful information**, **informing conclusions**, and **supporting decision-making**."

Essentially, the goal of data analytics is to discover trends and information hidden in the data which can be used to draw meaningful conclusions and influence decision making. The goal is not to just make predictions, but to find out how changing one variable effects another. This is where causality comes in.

$$\hat{y} = m * x + c$$

- In machine learning we are more concerned with how close our prediction $\hat{y}$ is to the actual y.
- Causality is more about finding out that **m** that relates x and y. We want to answer the question: How much does changing the value of x, change the value of y?

But why not correlation?

While correlation might indicate a statistical association between two variables, causality helps us identify the factors that actually influence an outcome. Causality refers to the relationship between cause and effect, when other factors are unchanged. It enables us to move beyond mere correlations and helps us understand the fundamental relationships between variables.

Example: Analyzing the Effect of Urbanization on Air Quality

On looking at the data we would see a strong correlation between the two variables but there is a confounding variable **Industralisation** which can influence both urbanization and air quality and hence the correlation is not a true measure of the relationship between the variables.

Real world importance of causality

Causality contributes to effective intervention strategies. By identifying the causal factors behind a particular outcome, organizations can develop targeted interventions to influence the desired results.

This is particularly evident in fields like public health, where understanding the causal links between lifestyle factors and diseases can inform preventive measures. It empowers businesses, researchers, and policymakers to make more accurate predictions, design effective interventions, and ultimately harness the full potential of data for informed decision-making.

Conclusion

To conclude, the importance of causality in data analytics lies in its ability to move beyond surface-level patterns and correlations, providing a deeper understanding of the mechanisms driving outcomes.

2. Ways in which causal framework can be incorporated into Machine Learning

Machine learning has traditionally relied on detecting statistical relationships in observational data rather than uncovering the causal mechanisms that govern the data-generating process. While observational data shows that two events are related, only causal relationships can explain why.

To quote Judia Pearl, a pioneer in this field who has revolutionized the understanding of causality in statistics

> " Correlation does not imply causation. But causation does imply correlation."

Incorporating causality into machine learning can improve **generalization**. Machine learning tries to give you a picture of what the future will look like, but that future has been subjected to information from the past. This does not robustly consider uncertainties outside the training dataset that may confound the impact of the various input variables on the output to be predicted.

> **Generalization**: Model's ability to adapt properly to new, previously unseen data, drawn from the same distribution as the one used to create the model.

Some ways in which causality has been included in machine learning include but are not limited to:

Randomized Controlled Experiments (RCEs)

In RCEs, a population of individuals is split into two groups: **treatment** and **control**, by administering treatment to one group and nothing to the other, we can measure the outcome of both groups. Assuming that the treatment and control groups aren't too dissimilar (which is ensured by **random sampling**), we can infer whether the treatment was effective based on the difference in outcome between the two groups.

However, we can't always run such experiments. Often, it is not possible to control all the variables. Instead, we have to rely on observational data.

Causal Markov Condition

In the context of machine learning, particularly in causal modeling using graphical models like **Bayesian Networks**, the Causal Markov Condition imposes constraints on the relationships among variables.

> Causal Markov (CM) condition states that, conditional on the set of all its direct causes, a node is independent of all variables which are not effects or direct causes of that node

Enforcing this condition aids in the identification of causal relationships from observational data and allows for more accurate predictions when interventions are made.

Conclusion

Ultimately, embracing causality in machine learning through the frameworks such as the ones discussed above advances the field by providing a deeper understanding of the true drivers of relationships, enhancing the model's interpretability, and enabling more reliable predictions in diverse, real-world scenarios.

As the integration of causality continues to evolve, it holds the promise of creating more robust and adaptable machine learning models.