CS6462

Probabilistic and Explainable AI

Lesson 23

Meta-Learning Causal Structures

by Emil Vassev

March 24, 2025

Causal Direction

Causality & causal inference:

- *basic goal*: learn causality from data:
  - what is the cause and what is the effect
  - correlations between variables should be enough for most of the cases
- *causal inference:* go one step further and figure out what would happen if we decide to change some of the underlying assumptions in our model

Inferring causal direction:

- the most popular tool is proper trial designs of experiments - *randomized control trial (RCT)*
- RCT are not universal, because cannot be conducted in many scenarios: can be too costly or infeasible due to the complexity of real-world systems
- the development of more causality in Machine Learning is a necessary step in building more human-like machine intelligence

Meta-Learning to Infer Causal Direction

Meta-learning in ML:

- learning algorithms learn from other learning algorithms
- use of machine learning algorithms that learn how to best combine the predictions from other machine learning algorithms
- *meta*: raising the level of abstraction one step higher + additional model information

Meta-learning for causal direction:*

- apply learned models assuming different causal directions to data with a changed transfer distribution
- meta-learning objective concerned with assumptions on data distribution and data changes due to moving from a <u>training distribution</u> to a <u>transfer distribution</u>, possibly resulting from some neuron's actions (part of the neural network we train)
- the <u>correct causal model</u> needs to adjust its transfer distribution only, and thus it adapts faster, which allows us to determine the underlying causal directions

*Y. Bengio et al. 2019. A Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms

Learning a Causal Graph with Two Variables

Learning meta-model:

- Bayesian network of two discrete random variables ***A***, ***B*** (each taking ***n*** possible values)
- training samples ***(a; b)*** from a pair of related distributions: training distribution and transfer distribution
- based only on samples from a single training distribution, both models ***A* → *B*** (A causes B) and ***B* → *A*** tend to perform
- determining if variable ***A*** causes variable ***B*** or vice-versa
- parametrizations of models (***A* → *B*** and ***B* → *A***) to estimate their joint distribution:

  $P_{A\to B}(A,B) = P_{A\to B}(A) * P_{A\to B}(B|A)$ - for training and transfer distributions

  $P_{B\to A}(A,B) = P_{B\to A}(B) * P_{B\to A}(A|B)$ - for training and transfer distributions

  four distribution models (training and transfer distribution models per graph)

- ***A*** and ***B*** are completely observed: maximum likelihood estimator is used to independently train all four models

Learning a Causal Graph with Two Variables (cont.)

Distributions:

- model ***A → B***
  - training distributions: $P^0_{A\to B}(A,B) = P^0_{A\to B}(A) * P^0_{A\to B}(B|A)$
  - transfer distributions $P^1_{A\to B}(A,B) = P^1_{A\to B}(A) * P^1_{A\to B}(B|A)$
- model ***B → A***
  - training distributions: $P^0_{B\to A}(A,B) = P^0_{B\to A}(B) * P^0_{B\to A}(A|B)$
  - transfer distributions $P^1_{B\to A}(A,B) = P^1_{B\to A}(B) * P^1_{B\to A}(A|B)$
- models trained with $P^0_{A/B\to B/A}$ first and then moved to $P^1_{A/B\to B/A}$

Learning a Causal Graph with Two Variables (cont.)

Result likelihood distributions:

- both network models are meta-trained on both training and transfer distributions for T steps with resulting likelihoods:

$$L_{A\to B} = \prod_{i=1}^{T} P_{A\to B,i}(A_i, B_i)$$

$$L_{B\to A} = \prod_{i=1}^{T} P_{B\to A,i}(A_i, B_i)$$

- models are trained following the steps:

1) models of ***A* → *B*** and ***B* → *A*** are trained using their training distribution
2) relationship between ***A*** and ***B*** is learned using the training results
3) distribution of both models is moved from a training to a transfer distribution
4) both models are retrained on the new data and the resulting likelihoods are recorded
5) based on the results, we evaluate the direction of the causal relationship

Learning a Causal Graph with Two Variables (cont.)

Loss function:

- the loss function of both training steps (over training and transfer distributions):

  $R(\alpha) = -\ln[\sigma(\alpha) * L_{A\to B} + (1 - \sigma(\alpha)) * L_{B\to A}]$

- $R(\alpha)$ computed with α denoting a structural parameter defining the causal direction and $\sigma(\alpha)$ the sigmoid transformation
- α is optimized to minimize $R(\alpha)$

Loss gradient:

- $\frac{dR}{d\alpha} = \sigma(\alpha) - \sigma(\alpha + \ln(L_{A\to B}) - \ln(L_{B\to A}))$
- $\frac{dR}{d\alpha} > 0$ *if* $L_{A\to B} < L_{B\to A}$ - i.e., if ***B → A*** is better at explaining the transfer distribution than ***A → B***

Summary

Meta-Learning Causal Structures

Causal Directions

Meta-Learning to Infer Causal Direction

Learning a Causal Graph with Two Variables

Next Lesson:

- Structural Equation Modeling

Thank You!

Questions?