Monday, May 21, 03:09

Your Submission STUD-D-18-00059

by Studia Logica (STUD) I em@editorialmanager.com

Dear Dr. Schulte,

We have received the reports from our advisors on your manuscript, "Causal Learning With Occam's Razor", submitted to Studia Logica

Based on the advice received, I have decided that your manuscript can be accepted for publication after you have carried out the corrections as suggested by the reviewer(s).

Below, please find the reviewers' comments for your perusal.
You are kindly requested to also check the website for possible reviewer attachment(s).

Please make sure to submit your editable source files (i. e. Word, TeX).

Please submit your revised manuscript online by using the Editorial Manager system.

Your username is: oschulte
If you forgot your password, you can click the 'Send Login Details' link on the EM Login page at stud.editorialmanager.com/

I am looking forward to receiving your revised manuscript before 23 Jul 2018.

With kind regards,
Jacek Malinowski
Editor in Chief
Studia Logica

Comments for the Author:

Reviewer #1: As editor of the Special Issue on Ockham's Razor and Simplicity, I collected the following two reports for this author.

****
Referee 1.

The submission demonstrates that an Occam learner is the only learner that minimizes the number of mind changes, and the time to convergence. This is a deep and important result. The theorem is shown under the following conditions:

-- the number of hypotheses is finite (this restriction is lifted in Luo and Schulte (2006))
-- data streams are such that every stream eventually displays all evidence items consistent with the hypothesis.

The first condition ensures that every hypothesis is eventually uniquely simplest, if true. Perhaps it is worth mentioning what happens in the infinite case if there are always several simplest hypotheses compatible with the data.

The second condition rules out "non-homogeneous", or "ragged" hypotheses, e.g. we cannot have a hypothesis H such that if H is true, then the data stream will either display only 1 or only 2. If the learning problem included another hypothesis H', on which a data stream will eventually display a 1 and a 3, then it seems that H ought to be simpler than H', and that a similar Occam necessity result will hold. In general, not every non-homogenous problem can be refined to a homogenous one, though I am not sure whether such a thing holds if the problem is finite.

The results are profitably applied to learning causal graphs from a stream of true conditional dependence facts, and to learning probability estimation diagrams. Clearly, there is some idealization here: conditional independence facts are not learned infallibly, but are subject to statistical error. A few sentences defending the idealization may be helpful.

Additional comments, typographical and otherwise, are included in the attached.

****
Referee 2.

Referee report on Causal Learning with Occam's Razor

This is a nice paper, and I recommend publication. There are some detailed stylistic suggestions in the accompanying markup that I think would improve the paper's impact, but I leave them to the authors' judgment.

First, I should say that the paper's value lies mainly in providing, as it says, a "gentle" introduction to learning theoretic work on occam's razor, to probabilistic work on causal discovery, and to demonstrating a strong connection between the two. It takes some doing to set all of that up for a general audience of non-specialists.

The paper also breaks some new ground. Standard causal diagrams do not determine which values of one variable cause which values of another. That extra structure is represented by "probability estimation diagrams (PEDs), which delve into the deeper structure of which values of a variable cause which values of another. The paper easily extends Schulte's approach to to PEDs.

You specifically asked me to comment on the relevance of the paper to the topic of Ockham's razor. From the preceding description, it is clear that the paper is on topic.

Here are some of the main issues that emerge in the markup:

The assumed approach to simplicity covers only very tidy problems in which every world in the simpler hypothesis is a boundary point of the more complex hypothesis. But the examples considered call for nothing more, so no harm is done, given the intended role of the paper as a gentle introduction for a general audience.

The paper does not fully solve the problem of causal inference, because it resorts to the artifice of an infallible oracle for dependence and non-identity of probabilities, rather than using statistical tests. There is a close analogy, but close is no cigar—there is a great deal that has to be done just right to connect topology in the right way with statistics, and that work is currently being done (e.g., Genin at CMU). It would be best to be candid about the gap.

Extending the preceding point, I found it puzzling that the authors consider both oracles for probabilistic independence and oracles for probabilistic dependence. Given the way the authors set up their logical learning model, the only plausible analogy to statistics is that the oracle is for dependence, since failure to detect independence at a given sample size does not imply that it will not be detected at a much larger sample size. These things are not a mere matter of convention, or of trying harder, etc. It's the way the problem of probabilistic inference is. To make the matter sound conventional turns a victory for the authors into a surrender.

Again, the authors' discussion of competing approaches like AIC and BIC is vitiated to some extent by the fact that those methods have to strike a balance between simplicity and fit, whereas the authors essentially assume that no such balance is even necessary. Of course, any tests employed to implement the assumed oracle will strike some such sort of balance, which will probably be different from both. The relevant point (which is also missing from that discussion), is that the authors' underlying argument for Ockham's razor rests on an entirely different foundation from either AIC or BIC. That is something that could very usefully have been explained to a general audience. The silence on that crucial point at the crucial place is almost deafening.

Assuming that the paper is a gentle introduction for a general audience, it was a tactical mistake, in my opinion, to include the fundamental motivating examples for the causal formalism only in a technical appendix. That puts the general reader in the position of accepting a lot of formalism entirely on faith that it matters for something. Also, there is no discussion of how to use a causal diagram to infer the effects of novel policies, which is the whole point of discovering them.

Similarly, the main motivation for the PED formalism occurs only after the definitions and results for them have been presented. I would again put the motivational stuff first.

The main technical result is that the length of the longest refinement path to a partition in a lattice is the cardinality of the partition. The authors prove it directly by induction. But it seems to be an easy consequence of the familiar fact that finite partition lattices are graded posets https://en.wikipedia.org/wiki/Graded_poset. A graded poset has a rank function defined directly on the poset that respects the partial order and that increments once for immediate successors. The incrementing property gives a lower bound on maximum path length, and the fact that the ranking function is defined directly on partitions, rather than on paths, provides the upper bound. In any event, including a proof of such an elementary result in a technical appendix is an affectation that distracts from the genuine merits of the paper, which are principally expository. That the underlying mathematical ideas are so easy and natural is a feature, rather than a bug, in my opinion.

There are some little typos here and there as noted in the attached PDF markup.