Questions and answers about philosophy of science, causation, and human/machine learning

Yuan Qi
MIT Media Lab
Cambridge, MA, 02139

1 Question

(A) Explain how the tools of statistical learning (Bayesian or otherwise) might be useful in thinking about paradigm shifts in science. Is a paradigm shift just the replacement of one model or hypothesis by another, or does it involve something more? (B) What would be the analogy of a Kuhnian paradigm shift in human learning or machine learning? Give at least one specific example of a phenomenon of human learning, and also a phenomenon from machine learning, that can be thought of productively in terms of paradigm shifts.

Answer:

Part (A): From the philosophy of science perspective, when some observations do not match up with a current paradigm, people begin to propose new theories to explain them. We can use the following Bayesian formula to model the confirmation of a new theory $T_i, i = 1, \ldots, J$, given the evidence E and the background knowledge B that may conclude the current paradigm.

$$P(T_i|E, B) = \frac{P(E|T_i, B)P(T_i|B)}{P(E|B)} \tag{1}$$

Then the new theory T_i can be combined with our background knowledge B to explain the abnormal evidence E. However, as we obtain more and more abnormal observations, we might propose a new theory T_j which is quite different from, or even contradict to B, in order to well explain the evidence. However, this may result in small $P(T_i|B)$ and $P(E|T_i, B)$, as well as a small posterior $P(T_j|E, B)$. Thus, we might try to throw away our current paradigm in B to obtain B' such that we can obtain a new large posterior for the theory T_j as follows:

$$P(T_i|E, B') = \frac{P(E|T_i, B')P(T_i|B')}{P(E|B')} \tag{2}$$

Then, we realize a paradigm shift happened.

From a more statistical point of view, Bayesian inference on dynamic models where we may interpret hidden states as paradigms at different times can be naturally applied to model paradigm shifts. Bayesian estimates of hidden states in a dynamic model may go through big changes along time, including both changes of the structural relation between hidden variables and changes of estimated parameters of hidden variables. These changes may be thought as paradigm shifts.

Clearly, from a Bayesian point of view, a paradigm shift is not just the replacement of one model or hypothesis by another. It requires the interaction between observational data, new paradigms, and old paradigms in a probabilistic way. Furthermore, instead of just replacing old paradigms, a new paradigm often unifies and generalizes several old ones.

Part (B): In human learning, people can change their opinion dramatically based on new observations, in a way similar to a Kuhnian paradigm shift. Let us consider the explaining away effect. For example, a person may think it just rained when he saw the lawn outside his window was wet in a summer. But after he noticed the sprinkler was on, he might totally change his opinion about the cause of the wet lawn from the rain to the sprinkler, as a sort of paradigm shift.

In machine learning, the output of a learning algorithm might change significantly when training data are not stationary. For example, a trained Gaussian mixture model might change significantly by varing the number of Gaussians, instead of just changing the weights, and the mean and variance parameters, when there are new training data points available.

From a higher level, human and machine learning theories themselves are undergoing paradigm shifts. For example, the Bayesian framework, proposed in [9], gives a new universal framework that successfully accomplishes different tasks in human learning such as concept learning and inductive generalization, outperforms previous problem-centric models , and subsumes Tversky's set-theoretic models and Shepard's exponential decay functions. In this sense, the Bayesian framework can be thought of a paradigm shift from previous paradigms where many models are proposed in a task-centric way and cannot be used for solving other problems, though, on the other hand, there are still some unsolved problems in the Bayesian framework, e.g., how to construct priors and likelihoods in general.

In machine learning, the change from the connetionist, i.e., neural networks, to Bayesian algorithms such as Gaussian processes and Bayes Point Machines can be viewed a paradigm shift, because Bayesian algorithms can handle problems which cannot be solved by neural networks, e.g., tuning hyperparameters without doing expensive cross-validations, can give more insights into the problem than neural networks, and have outperformed neural networks in many applications.

2 Question

Clark Glymour [3] has a number of reasons for not being a Bayesian. How might these arguments be relevant to human or machine learning? That is, to what extent or in what ways should a theorist of human or machine learning be dissuaded from adopting a Bayesian approach, based on Glymour's arguments?

Answers:

First, Glymour argues it is not easy to show why beliefs come in grades and why beliefs conform to the probability calculus. Though I do not totally agree with him , his argument does touch one subtle and important issue for a Bayesian human or machine learning practitioner: how to assign priors? In some applications, we can easily apply these text book priors, such as conjugate priors, Jeffery's priors, scale-invariant priors, etc. to reflect our prior knowledge or ignorance. On the other hand, we may find difficult to construct a meaningful and consistent informative or non-informative prior in many other applications. Associated with the prior design is the problem of how to construct an appropriate hypotheses space. Though some work has been done in this direction, it is still far from being complete and satisfying in general. Thus, in practice, people might turn to other techniques when facing these difficulties.

Second, Glymour argues the evaluation of all likelihoods, especially the catch-all term, is difficult in many situations. I agree with him on this point. The evaluation of a normalization factor, including the evaluation of the the catch-call term, can be difficult and time-consuming. On one hand, this is part of the reason why many researchers work on efficient Bayesian approximation techniques such as MCMC and variational methods. On the other hand, this also dissuades the use of Bayesian techniques in large-scale complicated problems.

Third, Glymour thinks we need a general principle restricting conditional probabilities, i.e., likelihoods. He said, " Presumably any such principles will have to make use of relations of content or structure between evidence and hypothesis". In other words, it remains as a problem how to choose likelihoods in a principled way. In practice, cook-book Bayesian practitioners will pick a convenient likelihood function, e.g. a Gaussian, no matter what problems they are facing and use the chosen likelihood to obtain a so-called “optimal” solution. This is problematic in the sense (1) by this way we might end up with absurd inference results, and (2) even when the inference results seem reasonable, we cannot simply interpret the obtained posterior as the degree of the confirmation. The posterior needs calibrated before this kind of interpretation. Though some principle exists, e.g. the size principle [10], and non-parametric Bayesian methods have been developed, building a good likelihood function remains a problem in many applications and may stop people to adopt a Bayesian approach.

Fourth, Glymour thinks “there are elementary but perfectly common features of the relation of theory and evidence that the Bayesian scheme cannot capture without serious-and perhaps not very plausible-revision”. This is true to some extent. For example, pure probabilistic inference is not sufficient for inferring causal relations as argued by Pearl [7]. Also, probabilistic relation is non-monotonic, which in general cannot easily capture a relation that says “as long as A is true, B is true”, while relational logic can easily represent such a relation. This limitation may make it hard to use purely probabilistic methods on relational data.

Glymour’s above arguments do make sense in either theoretical or practical point of view. But I think his following arguments do not hinder the use of Bayesian methods in practice:

Glymour argues Bayesian methods cannot rationalize preference towards simple models. First, simplicity is actually preferred in many Bayesian methods, such as automatic relevance determination [5] and Bayesian models of inductive generalization [9]. Second, I agree with Domingos [1] that simplicity is desirable in itself for real-world applications, but does not necessarily indicate a model with good generalization performance. So we do not need to rationalize the preference towards simple models anyway.

Also, Glymour thinks since the conjunction of observation data can have higher probability than any possible hypothesis, why we need a hypothesis at all according to Bayesian confirmation theory? I think instead of criticizing Bayesian methods, he actually points out a problem of people using Bayesian theorem in philosophy of science: they only consider the posterior of a hypothesis. Instead, I argue we should also consider the predictive posterior of a new hypothesis for generalization purposes. In this sense, I think, only using posteriors for choosing a new hypothesis is more like a maximum likelihood approach rather than a Bayesian approach. For example, we can fit each data point with a Gaussian having near 0 variance to achieve a larger likelihood than any other Gaussian mixture. However, the first Gaussian mixture will generalize poorly on predicting a new data point’s probability.

3 Question

How do Pearl’s recent analyses [6] of causality bear on the philosophy of science? Give at least two examples of classic problems in philosophy of science that may be illuminated by his approach, and contrast the analyses his approach suggests with the analyses traditionally given by philosophers.

Answers:

Pearl’s recent analyses bear on the the philosophy of science through the following key factors:

1. treating causation as behaviors under interventions.
2. interpreting intervention as local surgeries on mechanisms.
3. treating mechanisms as causal diagrams and structural equations.

To this end, he defines the *do* operator such that $do(X = x)$ represent the minimal change to a casual model necessary for establishing the antecedent $X = x$ for all possible backgrounds. Furthermore, he developed the calculus of the *do* operator, which allows one to compute $P(\text{Lung Caner}|do(\text{Smoke}))$ without controlled experiments.

Pearl's approach can illuminate several classic problems in phyilosophy of science:

1. Evaluation of conterfactuals. Conterfactuals are important in causal-effect analysis. According to [7], philosophers R. Stalnaker and D. Lewis construct a closet-world semantics to evaluate conterfactuals. The closet-world semantics does not answer what choice of distance measure and what mental representation of inter-world distance we shall choose. In order to match similarity measures to our conception of causal laws, Lewis set up a system of weights and priorities among various aspects of similarity. Pearl comments that these weights and priorities are "post hoc and still yield counterintuititive inferences" [7]. In contrast, Pearl's approach allows us to compute counterfactuals using the *do* operator and causal models, e.g. via the twin network method. Similarities and priorities may be read into the *do* operator if needed as an afterthought. Moreover, causal diagrams offer a mental representation that parsimonious encodes causal knowledge and allows efficient inference algorithms about probabilities, counterfactuals, and causes.

2. Single-event v.s. type causes The relation between an actual cause at a single-event level and a general cause at a type-level has been controversial in philosophy of science. The debate has led to theories that view type and single-event claims as two distinct species of causal relations. In contrast, Pearl's structural account treat type and single-event claims in a unifying framework, in which these two claims only differ in the details of scenario-specific information provided to a structural causal model. In a type level, an intermediate level between these two extreme levels, and a single-event level, we specify the background variables U to different extents and, correspondingly, we have causal models based on $P(u)$, $P(U|e)$ where e stands for evidence, and $U = u$. Then probable sufficiency and necessity defined in structural causal models are close to type and single-event level claims respectively.

3. Overdetermination By using conterfactuals, Pearl's structural account can present causal necessity, which is very useful for actual cause, while it is hard for pure probabilistic approaches by Cartwright etc., which denies conterfactuals, to deal with actual causes. Furthermore, by constructing the notion of sustenance, Pearl's structural account can handle the simultaneous presentation of multiple causes, i.e., overdetermination, while conterfactual, proposed by Hume, or conterfactual dependence chain, proposed by Lewis, fails in case of overdetermination.

4 Question

To what extent can human inference about properties of categories or cause-effect relations be described in terms of rational statistical inference? Do the fields of machine learning or philosophy of science have anything to gain from the study of how human inductive inference works?

Answers:

Bayesian inference can well describe human inference about properties of categories or cause-effect relations as demonstrated in [9, 11, 10]. In [9], the authors use unsupervised learning methods to construct hypothesis spaces and propose priors and likelihoods based on learned hypothesis spaces to carry on Bayesian inference. Furthermore, the Bayesian Occam's razor is reflected in both priors and likelihoods. This approach outperforms previous similarity-based approaches in experiments. In [11], the authors formulate classical ΔP and causal power models as parameter inferences in causal graphical models, and propose a new causal support model as structural inference. More specifically, they show how human causal induction about $P(C \rightarrow E)$, where C and E are two events, can be well modeled by causal support, the log posterior odds of two causal graphical models with and without the edge link $C \rightarrow E$ respectively.

In other words, causal support is basically a log Bayes factor. They also propose to use χ^2 statistic to approximate causal support for simple computation. Their approaches outperform ΔP and causal power models in human subject experiments.

However, these two approaches have their limitations in statistical inference. In the following paragraphs, I attempt to point out some of these limitations. For Bayesian models of human inductive generalization [11, 10], the possible limitations may include:

1. Though showing the nice size principle, the proposed likelihood functions all have a rectangle shape and do not handle observation noises in the data. A likelihood function with a soft-boundary shape, e.g. a bell-shape pdf, will be able to model observation noises though it may make numerical computations more complicated.

2. The proposed likelihoods do not model labeling noises either. In other words, they do not handle the case some of given examples are actually negative ones instead of positive ones. A sensible attempt might be to model the likelihoods as a linear combination of two likelihoods based on correct and wrong labeling respectively with a prior on a latent labeling variable.

3. The hyperparameters in the prior are manually tuned to maximize the correlation between model and data in [11]. What if the hyperparameters in the prior be tuned automatically by the empirical Bayes approach? Will this also result in a good match between statistical predictions and human subject's confirmation judgments?

4. The generated hypothesis space is sensitive to the distance metric or similarity measure used in clustering. A different distance metric may result in quite different results as reported in [11]. How to choose an appropriate distance metric in practice remains a problem. Is it possible to learn the hypothesis spaces in a principled way? The results on kernel learning in machine learning might be useful for constructing hypothesis spaces in human learning. On the other hand, the optimal solution to this problem might be beyond the capability of pure statistical inference. Visualization and domain knowledge seem very important here. To some extent, as Glymour said a practitioner's insights to the problem seem more important than the use of the probability calculus.

For structural models of human causal induction [11], the possible limitations may include:

1. There is no conterfactual consideration in the causal support model, which makes it hard for causal support to find actual causes in preemption events.

2. Also causal support may not be sufficient to obtain actual causes in overdetermined events.

3. In a deeper level, Pearl's argument that causal-effect relation is quasi-deterministic, not purely probabilistic, might be true. In this sense, statistical structure models may be illuminating and very useful, but not sufficient for modeling cause-effect relations by themselves.

From the study of how human inductive inference works, we can obtain inspiration to both machine learning and philosophy of science.

1. First, using clustering techniques to generate hypothesis spaces gives a new direction of how to combine supervised and unsupervised learning techniques. Semi-supervised learning has become a hot research topics in the machine learning community. Machine learning researchers try to efficiently combine information from the location distribution of the data and the labeling distribution of the data to improve the generalization performance of trained algorithms. Different from transduction and other semi-supervised learning techniques, the work in modeling human inductive generalization offers an interesting new direction.

2. The taxonomic hypothesis spaces also provide a new principled way for designing good priors.

3. These priors also implement Occam's razor. This will help to generate sparse learning algorithms, which leads to fast computation for large scale problems.

4. The work in modeling human inductive generalization can inspire algorithms that can learn from only positive examples, while popular discriminative approaches in machine learning cannot work in this case.

5. The causal support model measures causal-effect relations by using two graphical models with different structures. This is similar to Pearl's twin network method in the sense both methods incorporate two different model structures, though they are different in many other aspects. It would be interesting to see if causal support can be used in Pearl's framework and applied to defining actual causes as well as probability of necessity and sufficiency for inferring causes.

6. It would be also interesting to see how causal model can be applied to define explanations. Philosophers and statisticians have proposed the log likelihood ratio $\frac{P(e|c)}{P(e|c')}$ as the proper measure of the degree to which c is a better explanation than c'. Pearl has proposed maximizing the posterior $P(c|e)$ as the model of explanations. Both measures have their faults and have been criticized [7]. Due to the strong relation between causes and explanations, it would be very interesting to see how well causal support can model explanations.

5 Question

How might representations that combine first-order logic with probability theory (e.g., PRMs [2] or stochastic logic programs [8]) be relevant for addressing problems in philosophy of science or human cognition? Give at least one example of a problem from each of these areas that could be illuminated by first-order probabilistic approaches.

Answers:

The representations that combine first-order logic with probability theory can capture more relational information easily and have more representation power than probabilistic graphical models and logics. These representations may illuminate philosophy of science in the following areas:

1. First-order probabilistic approaches naturally match Pearl's argument on the quasideterministic property of causal-effect relations. So these approaches could be used as causal structure models, which can more efficiently represent deterministic relations, seamlessly blend deterministic and probabilistic causal knowledge.

2. Stochastic logic programs (SLP) allows one to do causal inference both on ungrounded predicates and on those grounded predicates. Thus, SLPs could deal with for causes from a type level, to an intermediate level, to a single-event levels.

3. First-order probabilistic representations can represent context-specific stochastic relations, which is beyond the representation power of causal Bayesian networks. This gives first-order probabilistic representations more power to cope with actual causes which is generally defined in specific contexts.

4. The capability of representing context-specific relations also lends first-order probabilistic approaches more power to model explanation since our background and interest may help us define what will be considered as an explanation for given events. Different backgrounds may lead to different explanations.

First-order probabilistic approaches may illuminate human learning in the following areas:

1. First-order probabilistic approaches, such as SLPs and Bayesian logic programs [4], can be used to construct query-sensitive hypothesis spaces for human learning. This allows us to generate context-specific hypothesis spaces, which may be very useful for human inductive generalization.
2. First-order probabilistic approaches can also be used to incorporate qualitatively relational knowledge, e.g. linguistic knowledge, into our hypothesis spaces for inductive generalization.

References

[1] Pedro Domingos. The role of occam's razor in knowledge discovery. *Data Mining and Knowledge Discovery*, 3(4):409–425, 1999. `http://www.cs.washington.edu/homes/pedrod/dmkd99.ps.gz`.

[2] Nir Friedman, Lise Getoor, Daphne Koller, and Avi Pfeffer. Learning probabilistic relational models. In *IJCAI*, pages 1300–1307, 1999.

[3] C. Glymour. Why I am not a Bayesian. In Martin Curd and Jan A. Cover, editors, *Philosophy of Science: The Central Issues*. W.W. Norton and Company, January 1998.

[4] Kristian Kersting and Luc De Raedt. Bayesian logic programs. In J. Cussens and A. Frisch, editors, *Proceedings of the Work-in-Progress Track at the 10th International Conference on Inductive Logic Programming*, pages 138–155, 2000.

[5] D. J. MacKay. Probable networks and plausible predictions - a review of practical Bayesian methods for supervised neural networks. In *Network (IOPP)*, 2000.

[6] Judea Pearl. Reasoning with cause and effect. In *International Joint Conference in Artifical Intelligence (IJCAI) Award Lecture*, 1999. `http://singapore.cs.ucla.edu/IJCAI99/`.

[7] Judea Pearl. *Causality: models, reasoning and inference*. Cambridge University Press, 2000.

[8] Muggleton S. Stochastic logic programs. In L. De Raedt, editor, *Advances in Inductive Logic Programming*, pages 254–264. IOS Press, 1996.

[9] N.E. Sanjana and J.B. Tenenbaum. Bayesian models of inductive generalization. Submitted to NIPS, 2002.

[10] J. B. Tenenbaum. Bayesian modeling of human concept learning. In *NIPS 11*, Cambridge, MA, 1998. MIT Press. `http://citeseer.nj.nec.com/tenenbaum99bayesian.html`.

[11] Joshua B. Tenenbaum and Thomas L. Griffiths. Structure learning in human causal induction. In *NIPS 13*, pages 59–65, 2000.