Bayesians *Can* Learn from Old Data

William H. Jefferys
University of Texas at Austin, Department of Astronomy
University of Vermont, Department of Mathematics and Statistics
Email: bill@bayesrules.net

May 14, 2007

Abstract

In a widely-cited paper, Glymour [1] claims to show that Bayesians cannot learn from old data. His argument contains an elementary error. I explain exactly where Glymour went wrong, and how the problem should be handled correctly. When the problem is fixed, it is seen that Bayesians, just like logicians, *can indeed* learn from old data.

Outline of the Paper I first review some aspects of standard logic that are relevant to this paper. I then discuss the relationship between standard logic and standard probability theory, and in particular point out the fact that standard probability theory contains standard logic in the particular sense that for any argument that reaches a conclusion using standard logic, there exists a parallel argument (calculation) in standard probability theory that reaches the same conclusion, and furthermore, that any *valid* argument by any method (whether logical or Bayesian) must arrive at the same conclusion.

I then introduce a simple "toy example" that is nonetheless sophisticated enough to reveal the problem with Glymour's claim. The toy example is an extension of the example that Glymour used in his paper. I describe Glymour's argument, and use the toy example to show that his reasoning leads to a contradiction with ordinary logic, and therefore must be invalid. I then explain, again in terms of the toy example, exactly where Glymour's argument goes wrong, and how to correct it. I conclude with a summary of what we have learned.

1 Standard Logic

Standard logic tells us how to combine propositions $A, B, C, \ldots$ with logical operations such as $\wedge, \vee, \neg, \rightarrow, \ldots$ to obtain new and valid propositions. The propositional calculus allows us to calculate, using definite rules, the truth value of any proposition that has been constructed from other propositions using these

logical operations, given the truth values of the propositions from which they are constructed.

For example, given propositions A, B, we can calculate the truth value of the proposition $C = A \wedge B$ as follows: C is true if both A and B are true, otherwise it is false. Similarly, the truth value of the proposition $D = A \vee B$ is true unless both A and B are false.

Likewise, the truth value of the proposition $E = A \rightarrow B$ is true if A is false, otherwise it is equal to the truth value of B. That is, if A is true, then B must be true. If A is not true, then it doesn't matter what the truth value of B is, $A \rightarrow B$ is true.

An important feature of standard logic is that it is time-independent (Jaynes [3], p. 89). That is, it describes relationships between propositions that are independent of when we may learn the truth or falsity of the propositions themselves. For example, the truth-value of the expressions $\neg A$, $A \wedge B$, $A \vee B$, and $A \rightarrow B$ depend only on the truth-values of A and B, and not upon when we may have learned their truth-values.

2 Probability and Logic

Probability theory extends the basic notions of standard logic to a regime where the *degree of plausibility* of propositions is no longer just "true" or "false", but may be intermediate between the two. That is, to any proposition we can assign a number in the unit interval $[0, 1]$ that corresponds to our assessment of how likely it is that the proposition is true, where 1 means that we are certain the proposition is true and 0 means that we are certain that it is false. The larger the degree of plausibility, the more likely it is that we would regard the proposition as true.

A theorem of Cox [2] proves that, up to an isomorphism, standard probability theory is the unique extension of ordinary logic to this regime that satisfies certain obvious requirements necessary for the theory to yield consistent results. Jaynes ([3], p. 19) lists a set of three such requirements, which he calls *desiderata*:

1 If a conclusion can be reasoned out in more than one way, then every possible way must lead to the same result.

   An important aspect of this desideratum is that if a conclusion can be obtained using ordinary logic, then a *valid* calculation using probability theory must arrive at the same result. If a purported Bayesian calculation arrives at a result different from one that we can derive using standard logic, it must *ipso facto* be invalid. We will see below that Glymour's calculation fails this test.

2 The calculation takes into account all of the evidence relevant to the question. It does not arbitrarily ignore some of the information,

basing its conclusions only on what remains. It is, as Jaynes says, completely nonideological.

Glymour's calculation doesn't fail this test, but it does muddle the issue by failing use standard probability notation to indicate all the information that was taken into account. Indeed, this results in a basic confusion of models that turns out to be at the root of the problem with Glymour's calculation.

3 Equivalent states of knowledge are always represented by equivalent plausibility assignments. That is, if in two situations the state of knowledge is the same, then (except for possible relabeling of the propositions), the calculation must assign the same plausibilities to both.

Glymour's calculation fails this test as well.

It turns out that these three desiderata, along with the assumption that degrees of plausibility are represented by real numbers on the unit interval $[0, 1]$, are sufficient to derive standard probability theory as the unique embodiment of these sensible requirements of plausible reasoning.

In particular it turns out, as a consequence of Jaynes' desideratum #1 and Cox's theorem, that standard probability theory contains standard logic as a subset. This means that for every calculation that can be made using standard logic, (that is to say, where all of the propositions are either definitely true or definitely false), there is a corresponding calculation in standard probability theory that will arrive at the *same* result, and no *valid* calculation in standard probability theory can yield a different result.

Let $\mathcal{P}$ be the disjunction of one or more propositions, and Z another proposition. Then $\mathcal{P} \models Z$ states that the truth of Z validly follows from the truth of $\mathcal{P}$. If in addition $\mathcal{P}$ is in fact true, then the argument $\mathcal{P} \models Z$ is also *sound*.

In particular, I note the following correspondences: The argument $\{A, A \rightarrow B \models B\}$ is sound if, and only if the argument $\{A, P(B|A) = 1 \models B\}$ is sound.

Clearly, if $P(B|A) = 1$ then $P(\neg B|A) = 0$ by standard probability theory. This last result comes from the identity $P(B|A) + P(\neg B|A) = 1$, which is equivalent to the tautology $A \rightarrow (B \vee \neg B)$.

3 A Toy Example

We consider a situation where there are precisely two theories under consideration, say T and $T' = \neg T$, and only two observations of evidence are possible, that is E and $E' = \neg E$. We furthermore presume that $T \rightarrow E$ and $T' \rightarrow E'$. This means that if theory T is true, we must observe evidence E, and if theory T' is true, then we must observe evidence E'.

For example, let T be the theory of general relativity, and T' be pure Newtonian mechanics. Let E be the (in this case old) evidence that the motion of Mercury's perihelion is anomalous (cannot be explained under Newtonian mechanics). If we ignore the infinitesimally low-probability situation that observational errors have somehow transformed a truly non-anomalous perihelion motion into an apparently anomalous one (see Pennock [4] for a discussion) then we see immediately that in our toy example $T \rightarrow E$ and $T' \rightarrow E'$.

It is important to recognize that these relationships are *defined by the theory*, independently of any data that may have been observed and independently of when those data may have been observed. The relationships are therefore *time-independent*. It is *always* the case that Newtonian theory entails that no anomalous perihelion motion will be observed, and *always* the case that general relativity entails that anomalous perihelion motion will be observed.

If we observe evidence E, then standard logic says $T' \rightarrow E'$, so $\neg T \rightarrow \neg E$. It follows that $E \rightarrow T$ and $E \rightarrow \neg T'$. Hence observing E rules out T' and confirms T.

Note that this result was obtained by an appeal to nothing but standard logic. Since standard logic is just a calculus on the truth-values of the propositions, and does not depend on when we observe evidence E, it follows that we can certainly learn from old evidence if we use only logic. But, as pointed out above, Jaynes' desideratum #1, together with Cox's theorem, implies that the same result *must* be obtainable by a valid application of probability theory. If a calculation using probability obtains a different result, it is certainly not a valid calculation.

Translated into the language of probability theory, the result $E \rightarrow \neg T'$ is equivalent to $P(\neg T'|E) = P(T|E) = 1$ and $P(T'|E) = 0$. Any purported Bayesian calculation that does not arrive at this result must be invalid. Note also that when we translate the initial assumptions of this toy example into standard probability notation we can calculate the *likelihood* as $P(E|T) = 1$ and $P(E|T') = 0$ for use when we observe E, and $P(E'|T) = 0$ and $P(E'|T') = 1$ for use when we observe E'. Since all of these probability assignments are simply translations of statements of ordinary logic into the language of probability theory, they are time-independent, that is, their values are independent of when we happen to observe the evidence.

4 Glymour's Argument

Glymour argues that the Bayesian cannot learn from old evidence E. The argument goes as follows: Since we know the old evidence E to be true (we have observed it, after all), Glymour claims that

$$P(E) = 1 \quad ??? \tag{1}$$

I put question marks here because I believe this equation to be wrong.

Nonetheless, if we grant Eq. (1), the rest of Glymour's alleged proof goes through easily. Since $P(E) = 1$, it follows from standard probability theory that $P(E|X) = 1$ for all propositions X. In particular, $P(E|T) = 1$. Therefore, by Bayes' theorem,

$$P(T|E) = \frac{P(E|T)}{P(E)} P(T) = P(T)$$

and since the posterior probability is equal to the prior probability, we haven't learned anything.

5 Counterexample to Glymour's Argument

We see immediately that Glymour's calculation fails to satisfy Jaynes' desideratum #1, for we have proved that for our toy problem, knowledge of E together with standard logic leads to the conclusion that T is true and T' is false, regardless of what we may have thought before we did the calculation. But Glymour's calculation allows for no such conclusion: If for example we had adopted $P(T) = 1/2$, Glymour's calculation tells us that $P(T|E) = 1/2$, in blatant contradiction to the calculation from ordinary logic. The equation $P(T|E) = 1/2$ says that E does *not* entail T, whereas logic says that E *does* entail T. Since Cox's theorem guarantees that any *valid* calculation using probability theory must arrive at the same conclusion we obtained using standard logic, this fact in itself proves that Glymour's argument cannot be valid.

It is not hard to pinpoint the source of the problem, again using the toy example as a guide. If $P(E) = 1$, then it follows that $P(E|X) = 1$ for any proposition X; in particular, $P(E|T') = 1$, or translated into the language of logic, $T' \rightarrow E$. That is, according to Glymour's reasoning, if we know that E is true, we must conclude that *Newtonian physics entails that we will observe anomalous motion of the perihelion of Mercury.* But this is absurd. Purely as a matter of logic, and as a consequence of physical theory, Newtonian physics entails that we will *never* observe anomalous perihelion motion for Mercury, that is, $T' \rightarrow E'$, which is equivalent to $P(E'|T') = 1$ or equivalently $P(E|T') = 0$, as demonstrated above.

Thus we have from Glymour's argument $P(E|T') = 1$, and at the same time we have from standard logic $P(E|T') = 0$. Equivalently, we have from Glymour's argument that $T' \rightarrow E$, whereas logic says that $T' \rightarrow E' \neq E$. Both cannot be true, and since the calculation using standard logic is clearly correct, it follows that Glymour's argument has yielded a contradiction, and cannot be valid.

The problem arises from Glymour's assumption that $P(E) = 1$. Without that assumption, the rest of his alleged proof fails.

6 Glymour's Friend

Physicists are familiar with "Wigner's Friend," a thought experiment named for the late physicist Eugene Wigner, that is designed to help us think about when and under what circumstances the "collapse" of states in quantum mechanics takes place. In this thought experiment, Wigner and his "friend" have different states of knowledge, until Wigner's friend informs Wigner of certain facts, so that they end up with the same state of knowledge, and thus should have the same conclusions. The details of the physics aren't important here, but the idea that people arrive at the same state of knowledge having started with different states of knowledge is the key idea that I want to carry over to the present problem.

Let me introduce Glymour's friend Tom. Tom is ignorant of E. Therefore, when Glymour explains the toy problem to Tom, Tom can set priors and set up the problem without knowing that E is true. After he has done this, Tom can tell Glymour what his priors are. Suppose the priors are the same as the ones that Glymour has already adopted, and that $P(T) \neq 1$. Then both are starting with the same priors.

Now Glymour informs Tom that E is true, and Tom performs the standard Bayesian analysis (as he may, since he was igonorant of E up to this point, so the data are for him "new," not "old"), and arrives at the result that $P(T|E) = 1$. Glymour performs the calculation that he advocates (since for him the data are "old") and arrives at $P(T|E) = P(T) \neq 1$.

This violates Jaynes' desideratum #3, since at this point both parties have the same state of knowledge, yet they have assigned different plausibilities to $T|E$. Since the axioms of probability theory, in virtue of Cox's theorem, cannot violate Jaynes' three desiderata when used validly, we have again arrived at a contradiction. Since it is clear that Tom does not view E as "old" data, and therefore is entitled to carry out the standard Bayesian calculation, his conclusions must be correct and Glymour's wrong.

7 Where Glymour Went Wrong

Jaynes ([3], pp. 473, 484) points out an important fact: *A fruitful source of error and even apparent paradoxes in probabiliy theory is to fail to condition properly and explicitly on all background information used.* By failing to include such background information, one can imagine that one is discussing one model while actually discussing another model. Since the intended model may be different from the one you think you are describing, it is easy to arrive at apparent contradictions.

In the present case, the source of the error is embarrassingly obvious. Recall that Eq. (1) was derived in the light of knowledge of the old evidence E and *actually used* that information as background information, even though

this dependence was not explicitly noted in the equations. Following Jaynes' advice above, standard notational convention demands that we call out this fact explicitly. If we do this, we obtain the correct Eq. (2):

$$P(E|E) = 1 \quad !!! \tag{2}$$

The rest of the proof translates as follows:

$$P(E|E,T) = 1 \tag{3}$$

$$P(T|E,E) = \frac{P(E|E,T)}{P(E|E)} P(T|E) \tag{4}$$

But of course, $P(T|E,E) = P(T|E \wedge E) = P(T|E)$ by standard logic. Thus we see that when the conditioning that is implicit but unstated in Eq. (1) is explicitly recognized in Eq. (2), what Glymour has actually proved is the (well-known) fact that the Bayesian machinery, quite sensibly, prevents us from using the same evidence twice. He has *not* proved that a Bayesian cannot learn from old evidence, only that he cannot validly manipulate the Bayesian machinery to get additional information out of information that has already been used.

We now see that $P(E)$ and $P(E|E)$ are entirely different. $P(E|E)$ has already used evidence E, whereas according to the standard notational convention, $P(E)$ has *never* used evidence E, not even once. $P(E)$ is in fact entirely ignorant of our knowledge of E. Thus, there is no reason to suppose that $P(E) = 1$, and indeed, it is usually not.

Note that the right-hand side of Eq. (4) has its as prior $P(T|E)$, *not* $P(T)$. In other words, the prior in Eq. (4) must be constructed from full knowledge of E; it is *not* the same as $P(T)$, which is (of course) ignorant of E. One cannot substitute $P(T)$ for $P(T|E)$ in Eq. (4); the resulting equation is not a valid equation in probability theory.

In order to calculate the value of $P(T|E)$ for substitution into Eq. (4), we have to start from $P(T)$ and then apply Bayes' theorem in the usual way, where in this case the right hand side is calculated *unconditioned* on E (which is to say, the right-hand side is ignorant of any knowledge we may have about E). In this case, $P(E)$ does not know that E has been observed, and is correctly calculated from the priors and the *time-independent* likelihood as follows:

$$P(E) = P(E|T)P(T) + P(E|T')P(T') \tag{5}$$

Thus, in the toy example, where $P(E|T) = 1$ and $P(E|T') = 0$,

$$P(E) = P(T), \tag{6}$$

which is in general *not* equal to 1 unless the prior $P(T) = 1$, which is usually not the case.

This tells us the correct way to do the Bayesian calculation, in the case where E has been observed as old data. We still have to assign priors $P(T)$ and

$P(T')$, and this must be done without taking E into account. Although this step might pose some problems of its own (assignation of priors in general requires careful thought), any such problems are unrelated to Glymour's argument, so I will pass over this issue. Suppose, for example, we have assigned $P(T) = \alpha$, $P(T') = 1 - \alpha$, where $\alpha \in (0, 1)$. Then the Bayesian calculation goes through in the usual way as follows:

$$P(T|E) = \frac{P(E|T)}{P(E)} P(T) = 1 \tag{7}$$

since in the toy example example $P(E) = P(T)$ and $P(E|T) = 1$. Note that independent of α, we obtain the same result as we did from the calculation using ordinary logic. Thus, Jaynes' desideratum #1 is satisfied: No matter how we do the calculation, whether by ordinary logic or by a *valid* application of probability theory, Cox's theorem guarantees that we *must* arrive at the same result.

8 Summary and Conclusions

As Jaynes ([3], p. 89) points out, probability theory, like logic, is time-independent. All of the relationships in probability theory are *logical* relationships and have nothing to do with the order in which we happen to learn about the evidence or write down the Bayesian equations. When we calculate $P(T|E)$ from $P(T)$, it does not matter when we have actually observed E; the relationship between the two is purely a logical relationship, and the quantities that go into the calculation (likelihoods, priors) will be the same, regardless of when E is observed. As my colleague Tom Loredo observed when I showed him Glymour's argument, "Time plays the same role in probability theory as it does in logic: That is to say, no role whatsoever."[5]

In a *valid* Bayesian calculation, there is one and only one way to take into account ones knowledge of a particular piece of data, and that is to condition on that piece of data. Furthermore, it is essential that this conditioning be called out *explicitly* in the notation, as Jaynes advises. Conditioning on a piece of data without explicitly calling it out in the notation, as Glymour did, is a reliable route to disaster.

Glymour's error resulted from a failure to follow these basic principles. Using the principles of his "proof" I was able to derive a contradiction that seems not to have been noticed up to this point, but which is sufficient to demonstrate that Glymour's alleged proof is invalid. The bottom line is that Bayesians can and do learn from old data, when they do the calculation carefully and correctly.

9 Acknowledgements

References

[1] Glymour, Clark N. (1980), "Why I Am Not a Bayesian," in *Theory and Evidence* pp. 86. Princeton, N. J.: Princeton University Press.

[2] Cox, R. T. (1946), "The Algebra of Probable Inference," *American Journal of Physics* **14**, 1–13.

[3] Jaynes, E. T. 2003, *Probability Theory: The Logic of Science.* Cambridge: Cambridge University Press.

[4] Pennock, R. 19xx: [Rob: I haven't gotten the paper you promised yet; I think this is the right place to cite it but won't know until you send it to me.]

[5] Loredo, T. 2006. *Private communication.*