9 Jan 2013

eBay/google short course: Problem set 1 solutions

1. (a) From Bayes' Theorem in odds form, the odds value o_{PPV} associated with the PPV is

$$\begin{aligned} o_{\text{PPV}} = \frac{P(A|D)}{P(\text{not } A|D)} &= \left[\frac{P(A)}{P(\text{not } A)}\right]\left[\frac{P(D|A)}{P(D|\text{not } A)}\right] \\ &= \left(\frac{p}{1-p}\right)\left[\frac{\epsilon}{1-P(\text{not } D|\text{not } A)}\right] \\ &= \frac{p\epsilon}{(1-p)(1-\pi)}, \end{aligned} \tag{1}$$

and since probabilities pr are related to odds o through $pr = \frac{o}{1+o}$, the PPV in symbolic form is

$$P(A|D) = \frac{\frac{p\epsilon}{(1-p)(1-\pi)}}{1+\frac{p\epsilon}{(1-p)(1-\pi)}} = \frac{p\epsilon}{(1-p)(1-\pi)+p\epsilon}. \tag{2}$$

By a parallel calculation

$$\begin{aligned} o_{\text{NPV}} = \frac{P(\text{not } A|\text{not } D)}{P(A|\text{not } D)} &= \left[\frac{P(\text{not } A)}{P(A)}\right]\left[\frac{P(\text{not } D|\text{not } A)}{P(\text{not } D|A)}\right] \\ &= \left(\frac{1-p}{p}\right)\left[\frac{\pi}{1-P(D|A)}\right] \\ &= \frac{(1-p)\pi}{p(1-\epsilon)}, \end{aligned} \tag{3}$$

so that the NPV in symbolic form is

$$P(\text{not } A|\text{not } D) = \frac{\frac{(1-p)\pi}{p(1-\epsilon)}}{1+\frac{(1-p)\pi}{p(1-\epsilon)}} = \frac{(1-p)\pi}{p(1-\epsilon)+(1-p)\pi}. \tag{4}$$

(b)(i) Substituting $\epsilon = 0.95$ and $\pi = 0.98$ into (2) and (4) above gives

$$\begin{aligned} \text{PPV} &= \frac{0.95p}{0.02(1-p)+0.95p} = \frac{95p}{93p+2} \quad \text{and} \\ \text{NPV} &= \frac{0.98(1-p)}{0.05p+0.98(1-p)} = \frac{98(1-p)}{98-93p} \end{aligned} \tag{5}$$

Figure 1, which was produced by the `R` code below, plots the PPV and NPV in (5) for p from 0 to 0.1.

```
rosalind 262> R
```

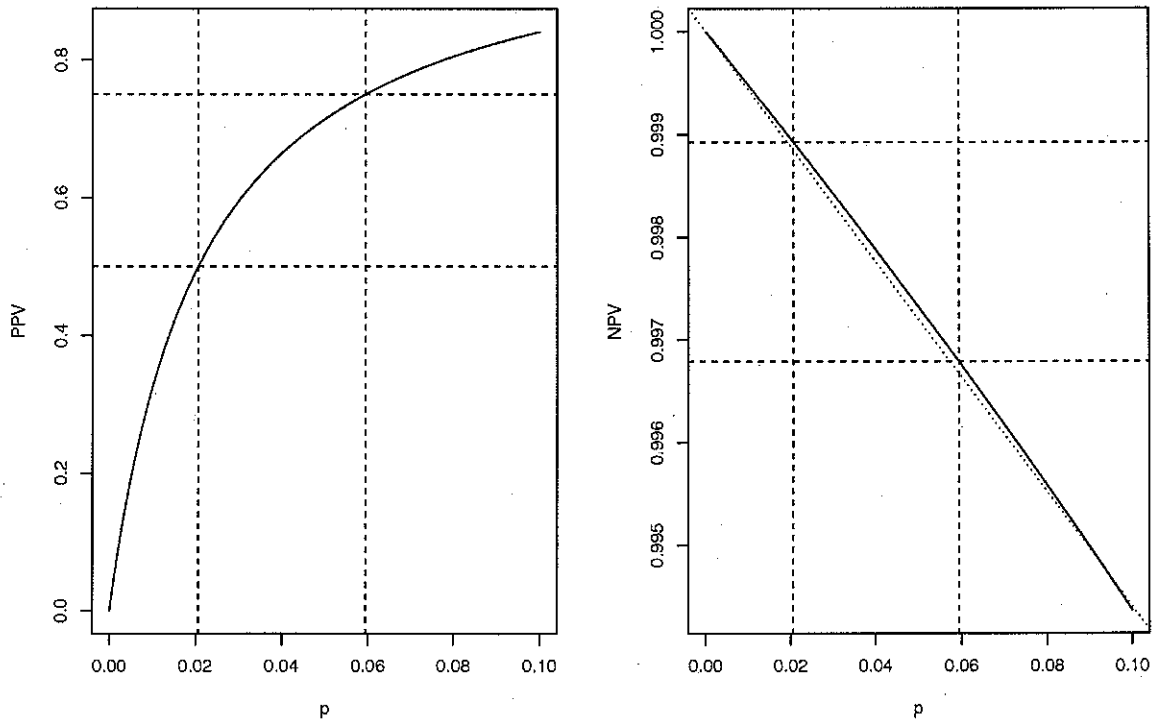


Figure 1: *PPV (left-hand plot) and NPV (solid line in right-hand plot) as a function of* p *in the interval* $[0,1]$ *when* $(\epsilon, \pi) = (0.95, 0.98)$*; the small dotted line on the right is* $(1-0.056p)$*. The horizontal and vertical dotted lines are relevant to parts (iii) and (iv) of the problem.*

```
R : Copyright 2001, The R Development Core Team
Version 1.2.1 (2001-01-15)

> postscript( "hwk1-1.ps" )
> par( mfrow = c( 1, 2 ) )
> p <- seq( 0.0, 0.1, length = 500 )
> plot( p, 95 * p / ( 93 * p + 2 ), type = 'l', xlab = 'p', ylab = 'PPV' )
> abline( h = 0.5, lty = 2 )
> abline( v = .02061855670, lty = 2 )
> abline( h = 0.75, lty = 2 )
> abline( v = .05940594059, lty = 2 )
> plot( p, 98 * ( 1 - p ) / ( 98 - 93 * p ), type = 'l', xlab = 'p',
    ylab = 'NPV' )
> abline( 1.0, -0.056, lty = 3 )
> abline( v = .02061855670, lty = 2 )
> abline( h = .9989270389, lty = 2 )
> abline( v = .05940594059, lty = 2 )
> abline( h = .9967880086, lty = 2 )
> dev.off( )
null device
          1
> q( )
```

```
Save workspace image? [y/n/c]: y
rosalind 285>
```

(ii) You can see that locally in this range the NPV is nearly linear (although that can't be true globally, because the NPV goes to 0 as $p \to 1$). Maple can help to figure out the equation of a line that closely approximates the NPV in this range, via a Taylor series expansion around 0.05, the midpoint of the relevant range:

```
rosalind 272> maple

    |\^/|     Maple V Release 5 (University of California, Santa Cruz)
._|\|   |/|_. Copyright (c) 1981-1997 by Waterloo Maple Inc. All rights
 \  MAPLE  /  reserved. Maple and Maple V are registered trademarks of
 <____ ____>  Waterloo Maple Inc.
      |       Type ? for help.

> taylor( 98 * ( 1 - p ) / ( 98 - 93 * p ), p = 0.05, 2 );
                                                              2
            .9973219067 - .05622991614 (p - .05) + O((p - .05) )

> simplify( .9973219067 - .05622991614 * ( p - .05 ) );

                          1.000133403 - .05622991614 p
```

The right-hand plot in Figure 1 contrasts the NPV (solid line) with the linear approximation $(1 - 0.056p)$ (dotted line), and it's evident that the latter tracks the former quite well in this range.

(iii, iv) The left part of Figure 1 shows that the PPV is 0.5 or more when the disease prevalence p is greater than about 0.02, and it's 0.75 or more for p bigger than about 0.06. Maple can be used to pin this down a bit more precisely, and to work out what the NPV would be for the relevant values of p:

```
> PPV := p -> 95 * p / ( 93 * p + 2 );

                                        p
                        PPV := p -> 95 --------
                                      93 p + 2

> solve( PPV( p ) = c, p );

                                    c
                                -2 ----------
                                   -95 + 93 c

> invPPV := c -> solve( PPV( p ) = c, p );
```

```
                    invPPV := c -> solve(PPV(p) = c, p)

> invPPV( 0.5 );

                                .02061855670

> invPPV( 0.75 );

                                .05940594059

> NPV := p -> 98 * ( 1 - p ) / ( 98 - 93 * p );

                                          1 - p
                        NPV := p -> 98 ---------
                                       98 - 93 p

> NPV( .02061855670 );

                                .9989270389

> NPV( .05940594059 );

                                .9967880086

> PPV( 0.1 );

                                .8407079646

> NPV( 0.1 );

                                .9943630211
```

(v) As Maple also notes (above), at $p = 0.1$ the PPV is about 84% and the NPV is about 99.4%. This NPV looks terrific, but the population prevalence of HIV positivity is now 10%; if the population had (say) 10,000,000 people in it and all of them were tested with ELISA, 6,000 people (0.6% of 10%) would be told they don't have HIV when in fact they do. A PPV of 84% is much better than the value we saw in class for $p = 0.01$ (about 32%), but at a population prevalence of 10% ELISA is still yielding about 1 in 6 false positives, which is not wonderful (even today people who are told they're HIV+ may react drastically).

(c) Holding p constant at 0.01 and letting ϵ and π vary produces the expressions

$$\text{PPV} = \frac{\epsilon}{99(1-\pi)+\epsilon} \quad \text{and} \quad \text{NPV} = \frac{99\pi}{(1-\epsilon)+99\pi}. \tag{6}$$

(i) You can see from the left part of (6) that the PPV is an increasing function of π, so that the inequality $\pi \le 0.98$ corresponds to the inequality $\text{PPV} \le \frac{\epsilon}{99(1-0.02)+\epsilon} = \frac{\epsilon}{1.98+\epsilon}$. As

$\epsilon \to 1$ this expression goes to $\frac{1}{2.98} \doteq 0.336$, so that if π is not allowed to go any higher than 0.98 the best we can do is to achieve a sensitivity of only about $\frac{1}{3}$.

(ii, iii) Plugging $\epsilon = 0.95$ into the left part of (6) and simplifying yields

$$\text{PPV} = \frac{95}{9995 - 9900\pi}, \tag{7}$$

which when substituted into the inequality PPV $\geq t$ for some desired PPV target t can be solved to yield

$$\pi \geq \frac{1999t - 19}{1980t}. \tag{8}$$

For $t = 0.75$ this requires $\pi \geq 0.9968$. From the right part of (6), at $\pi = 0.9968$ and $\epsilon = 0.95$ the NPV is 0.9995. The main conclusion from all of this is that if you want to change ELISA to perform better in the cohort of people whose prior probability of being HIV+ is only 1%, you need to try to get the specificity up to a fantastically high level and hope that this is not accompanied by too much of a decrease in the sensitivity.

2. (a) Mr. Soltau's quoting of 3 to 1 odds *against* Tiger Woods means that in his opinion Tiger has less than a 50% chance of winning; in fact he's trying to say that if the tournament were repeated Tiger would lose 3 times for every 1 time he wins, so "3 to 1 against" means a probability in *favor* of Woods winning of $0.25 = \frac{1}{1+3}$. Thus when somebody quotes odds o *against* something happening the probability p it *will* happen is $p = \frac{1}{1+o}$ (this formula is different from the odds expression we used in class because in sports people always quote the odds *against* something happening and in class we were talking about the odds in *favor* of it happening). When you use this new formula to convert Mr. Soltau's odds against each of the 24 players to probabilities and sum them, you get a value bigger than 1 (about 1.08), and that doesn't even include the other 65 or so players in the field. Mr. Soltau has violated the most basic of the rules that coherent probabilities obey: that they sum to 1.

(b) Suppose that I'm allowed to simultaneously place bets on any or all of the following propositions (there are other more complicated possibilities, but they aren't needed): {W wins}, {W loses}, {M wins}, {M loses}, where W stands for Tiger Woods and M for Phil Mickelson. By trial and error (and paying attention to what doesn't work and why) you can verify that there are lots of combinations that will not embarrass Mr. Soltau, but there is one combination that will: suppose I bet $\$A > 0$ on {W loses} and simultaneously bet $\$B > 0$ on {M loses}. Then from Mr. Soltau's posted odds the following possibilities emerge:

Outcome	My Net Gain On Part 1 of My Bet	My Net Gain On Part 2 of My Bet	My Overall Net Gain
W wins	$-\$A$	$+\$B$	$\$(B - A)$
M wins	$+\$3A$	$-\$B$	$\$(3A - B)$
Somebody else wins	$+\$3A$	$+\$B$	$\$(3A + B)$

To make Dutch book against Mr. Soltau I want all three of the overall net gain values in this table to be positive. The third one is already positive from the assumption that

$\$A > 0$ and $\$B > 0$, and the other two inequalities are satisfied by any $\$A$ and $\$B$ with $0 < \$A < \$B < \$3A$. If I'm allowed to bet against one or more golfers, Mr. Soltau is in trouble.

(c) If I can't bet against anybody, obviously the solution to (b) is no longer available to me, and single bets by themselves that either Woods or Mickelson will win don't produce a Dutch book either, so all that's left to try is the simultaneous bet ($\$C > 0$ on {W wins} and $\$D > 0$ on {M wins}). The table for this bet is as follows:

Outcome	My Net Gain On Part 1 of My Bet	My Net Gain On Part 2 of My Bet	My Overall Net Gain
W wins	$+\$C$	$-\$D$	$\$(C-D)$
M wins	$-\$3C$	$+\$D$	$\$(D-3C)$
Somebody else wins	$-\$3C$	$-\$D$	$-\$(3C+D)$

But this is hopeless, too: I can certainly make my overall net gain under the first outcome positive by taking $\$C > \$D$, but then the second inequality is not possible at the same time as the first (you can't find positive numbers $\$C$ and $\$D$ so that $\$3C < \$D < \$C$), and of course if $\$C$ and $\$D$ are both positive there is no way to make $-\$(3C + D) > 0$. This is basically why people like Mr. Soltau (and actual bookies at racetracks or Las Vegas casinos) can get away with posting what amounts to sets of *un-normalized* odds like the ones in the table from the web page: as long as nobody is allowed to place bets *against* the individual competitors in the contest being handicapped (golfers, horses, football teams, ...), the fact that the probabilities derived from the posted odds add up to more than 1 can't be used to embarrass the person announcing the odds.

3. (a) Let Yi stand for the proposition that {you initially choose door i}; let Mj = {Monte Hall then opens door j}; let Ck = {the car really is behind door k}; and (without loss of generality) suppose you initially choose door 1 (say) and Monte Hall shows you that a goat is behind door 2 (say). Then we want to compare $P(C1|M2, Y1)$ with $P(C3|M2, Y1)$ (clearly, by the rules Monte Hall is supposed to obey, $P(C2|M2, Y1) = 0$). This problem is a lot like the HIV case study: the true location of the car plays the role of the unknown, and Monte Hall's revealing what's behind a door plays the role of data. We want to compute probabilities of the form P(unknown|data), and the rules governing Monte Hall's behavior define probabilities (analogous to sensitivity and specificity values) that involve conditioning in the other order: P(data|unknown). Reversing the order of conditioning is a job for Bayes' Theorem, and the odds form of this theorem is often the easiest to use, as in problem 1, because you don't have to compute the normalizing constant. So here goes:

$$\frac{P(C1|M2,Y1)}{P(C3|M2,Y1)} = \left[\frac{P(C1)}{P(C3)}\right]\left[\frac{P(M2,Y1|C1)}{P(M2,Y1|C3)}\right]. \tag{9}$$

Now the first thing to observe is that $P(C1) = P(C3) = \frac{1}{3}$, so that the prior odds are 1:1 and they disappear from the equation, leaving only the Bayes factor (the rightmost

term in brackets in (9)). Then use the product rule for working with *and*, for example $P(M2, Y1|C1) = P(Y1|C1)P(M2|Y1, C1)$, to produce

$$\frac{P(C1|M2, Y1)}{P(C3|M2, Y1)} = \frac{P(M2, Y1|C1)}{P(M2, Y1|C3)} = \frac{P(Y1|C1)P(M2|Y1, C1)}{P(Y1|C3)P(M2|Y1, C3)}. \tag{10}$$

Next make the observation that your initial (random) choice of a door and the actual (random) location of the car are independent, so that $P(Y1|C1) = P(Y1) = P(Y1|C3)$, meaning that

$$\frac{P(C1|M2, Y1)}{P(C3|M2, Y1)} = \frac{P(M2|Y1, C1)}{P(M2|Y1, C3)}. \tag{11}$$

Now observe the following: (i) if you initially choose door 1 and the car really is behind that door, Monte Hall will choose to reveal door 2 with probability $\frac{1}{2}$ (because in this circumstance he is required to randomize between doors 2 and 3), whereas (ii) if you initially choose door 1 and the car really is behind door 3, Monte Hall will choose to reveal door 2 with probability 1 (since there is nothing else he can do in that case). So finally

$$\frac{P(C1|M2, Y1)}{P(C3|M2, Y1)} = \frac{\frac{1}{2}}{1} = \frac{1}{2}, \tag{12}$$

meaning that you're twice as likely to get the car if you switch as if you stay with your initial choice, i.e., $P(C3|M2, Y1) = \frac{2}{3}$.

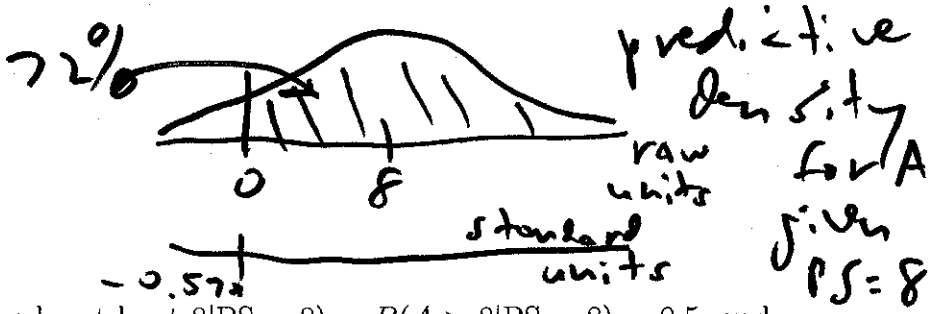
(b) There are a variety of ways to explain this intuitively; the one I currently like best is as follows: on the occasions—which occur a third of the time—on which you guess the location of the car correctly by your initial choice of door, Monte Hall gives you no information as to where the car really is by opening a door for you, but on the other occasions—which occur $\frac{2}{3}$ of the time—on which you initially guessed wrong, he implicitly tells you with certainty where the car is (namely, behind the door you did *not* initially choose), so you should switch. The form of the argument leading to equation (12) above basically just quantifies this intuition. (For those of you who play bridge, the reasoning in this problem is similar to that in something called the *Law*, or *Principle*, of *Restricted Choice*, which can help you to locate honor cards in the opponents' hands based on the cards they have so far played.)

4. (a) From simple enumeration of favorable cases $P_f(\text{favorite wins}|\text{PS} = 8) = \frac{8}{12} \doteq 0.67$, $P_f(\text{favorite wins by at least 8}|\text{PS} = 8) = \frac{5}{12} \doteq 0.42$, and $P_f(\text{favorite wins by at least 8}|\text{PS} = 8 \text{ and favorite wins}) = \frac{5}{8} \doteq 0.63$, where P_f denotes the simple relative-frequency estimate based only on the 12 games with point spread 8 (actually, as somebody recently pointed out to me, these are also classical probabilities based on equipossibility considerations).

(b) The problem suggests that, based on a model for the whole data set, a decent predictive distribution for the actual outcome A of a game, in the form (favorite score − underdog score), would be a Gaussian with mean equal to the point spread PS (assumed to be 8 here) and SD 14. Under this assumption $P_m(\text{favorite wins}|\text{PS} = 8)$ (where P_m stands for the model-based estimate) would be $P(A > 0|\text{PS} = 8)$ where $(A|\text{PS} = 8) \sim N(8, 14^2)$. Denoting by Φ the standard normal CDF, this equals $1 - \Phi\left(\frac{0-8}{14}\right) \doteq 0.72$ (see the sketch below), as can be verified (for example) in R:

```
> 1 - pnorm( ( 0 - 8 ) / 14 )

[1] 0.7161454
```



By the same logic $P_m(\text{favorite wins by at least } 8|\text{PS} = 8) = P(A > 8|\text{PS} = 8) = 0.5$, and then (in notation involving A) $P_m(\text{favorite wins by at least } 8|\text{PS} = 8 \text{ and favorite wins})$ is just $P(A > 8|PS = 8, A > 0)$. Now we know from class that for any propositions B, C, and D, in a kind of generalization of the product rule,

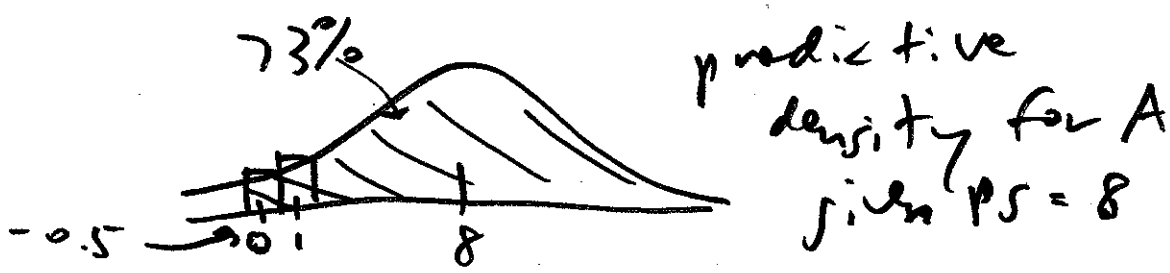
$$P(B, C|D) = P(C|D)P(B|C, D), \quad \text{from which} \quad P(B|C, D) = \frac{P(B, C|D)}{P(C|D)}. \tag{13}$$

An application of (13) yields

$$\begin{aligned} P(A > 8|PS = 8, A > 0) &= \frac{P(A > 8, A > 0|PS = 8)}{P(A > 0|PS = 8)} \\ &= \frac{P(A > 8|PS = 8)}{P(A > 0|PS = 8)} = \frac{0.5}{0.716} \doteq 0.70. \end{aligned} \tag{14}$$

Notice how (14) parallels the relative frequency calculation: the estimate of the third probability ($\frac{5}{8}$) is just the second probability ($\frac{5}{12}$) divided by the first ($\frac{8}{12}$). Notice further that all of the model-based probabilities have (on this occasion, for whatever reason) come out bigger than their relative-frequency counterparts: $(0.72, 0.50, 0.70)$ versus $(0.67, 0.42, 0.63)$.

All of this has ignored the discrete character of A, which really can only take on integer values. As the sketch below indicates, $P(A > 0|\text{PS} = 8)$ might arguably be better approximated by $1 - \Phi\left(\frac{-0.5-8}{14}\right) \doteq 0.73$, with the idea that account should be taken of the edges of the discrete histogram bars. With this adjustment the other two model-based probabilities increase to 0.51 and 0.71, respectively.



(c) There's no way to tell just by looking at the numbers which of the two probability assessments—relative-frequency or model-based—is "better," whatever that word might mean (one possible meaning, which we have no way to test here: in this class we've been judging probability assessments by their calibration, i.e., you're decently calibrated if you make a series of statements of the form "In the future repeatable event B will occur (say) 70% of the time" and in actuality B does occur about 70% of the time). The relative-frequency assessment is only based on the outcome of 12 games; the modeling judgment of a Gaussian predictive distribution was based on 672 games, but not all of them by any means had a point spread of around 8. I get the sense that the "effective sample size" of

the model-based approach was bigger than 12 in this example, however, so I personally would go with the model-based estimates.

5. By Bayes' Theorem, for any D such that $P(D) > 0$,

$$P(A|D) = \frac{P(A)P(D|A)}{P(D)}, \tag{15}$$

and if $P(A) = 0$ the right-hand side of (15) evidently also has to be 0, which proves (a). But if $P(A) = 1$ then $P(D|A)$ has to equal $P(D)$ (any proposition with probability 1 is independent of all other propositions; why?), making the right-hand side of (15) equal to 1 in this case, which demonstrates (b). In practice what this means is that You should try hard not to assign prior probability 0 or 1 to anything, because if You do You won't have left yourself any possibility of learning from whatever data might come along: no matter what the data D says, even if it makes You wish you hadn't assigned prior probability 0 or 1 to A, the posterior probability of A given D has to be the same as the prior probability. Any use of an approach to learning about the world that doesn't permit learning from data can't be a good use of the approach. (We'll see later, however, that Cromwell's Rule is a considerable source of formal embarrassment for Bayesian learning when A is some statement about the *statistical model* that gave rise to the data D.)