9 Jan 2013

eBay/google short course: Problem set 1

1. (Conditional probability; elaboration of case study 1) Consider the HIV screening example we looked at starting in week 1, in which $A = \{$the patient in question is HIV positive$\}$ and $D = \{$ELISA says he's HIV positive$\}$. Let p stand for the prevalence of HIV among people similar to this patient (recall that in our example $p = 0.01$), and let ϵ and π stand for the sensitivity and specificity of the ELISA screening test, respectively (in our case study $\epsilon = 0.95$ and $\pi = 0.98$).

(a) By using either Bayes's Theorem (in probability or odds form) or 2×2 contingency tables (as in the lecture), write down explicit formulas in terms of p, ϵ, and π for the *positive predictive value* (PPV), $P(A|D)$, and *negative predictive value* (NPV), $P(\text{not } A|\text{not } D)$, of screening tests like ELISA (recall that ELISA's PPV and NPV with patients like the one in our case study were 0.32 and 0.99948, respectively). These formulas permit analytic study of the tradeoff between PPV and NPV.

(b) Interest focused in class on why ELISA's PPV is so bad for people, like the man we considered in the case study, for whom HIV is relatively rare ($p = 0.01$). (i) Holding ϵ and π constant at ELISA's values of 0.95 and 0.98, respectively, obtain expressions (from those in (a)) for the PPV and NPV as a function of p, and plot these functions as p goes from 0 to 0.1. (ii) Show (e.g., by means of Taylor series) that in this range the NPV is closely approximated by the simple linear function $(1 - 0.056\, p)$. (iii) How large would p have to be for ELISA's PPV to exceed 0.5? 0.75? (iv) What would ELISA's NPV be for those values of p? (v) Looking at both PPV and NPV, would you regard ELISA as a good screening test for subpopulations with (say) $p = 0.1$? Explain briefly.

(c) Suppose now that p is held constant at 0.01 and we're trying to improve ELISA for use on people with that prevalence of HIV, where "improve" for the sake of this part of the problem means raising the PPV while not suffering too much of a decrease (if any) of the NPV. ELISA is based on the level L of a particular antibody in the blood, and uses a rule of the form $\{$if $L \geq c$ announce that the person is HIV positive$\}$. This means that if you change c the sensitivity and specificity change in a tug-of-war fashion: altering c to make ϵ go up makes π go down, and vice versa. (i) By using the formulas in (a) or 2×2 contingency tables, show that as ϵ approaches 1 with π no larger than 0.98, the NPV will approach 1 but the biggest you can make the PPV is about 0.336. Thus if we want to raise the PPV we would be better off trying to increase π than ϵ. Suppose there were a way to change c that would cause π to go up while holding ϵ arbitrarily close to 0.95. (ii) Show that π would have to climb to about 0.997 to get the PPV up to 0.75. (iii) Is the NPV still at acceptable levels under these conditions? Explain briefly.

2. (Coherence and Dutch book) On 2 Apr 2001 a senior writer for the web page *Sportsline.com*, Mark Soltau, posted an article about the Masters golf tournament that was about to be held on 5–8 Apr 2001. Among other things he identified the 24 players (among the 93 golfers in the field) who were, in his view, most likely to win the tournament, and he posted odds *against* each of them winning (for example, his quoting of 10–1 odds on Phil Mickelson meant that his personal probability that Mickelson would win was $\frac{1}{1+10} \doteq 0.091$), which are summarized in Table 1 below.

(a) If the 24 odds quoted by Mr. Soltau were taken literally, show that the personal probability specification implied by his posted odds was incoherent. (In fact Mr. Soltau may

Table 1: *Odds posted by sports writer Mark Soltau against each of the top 24 golfers competing in the Masters golf tournament, April 2001 (part 1 of table).*

Player	Best Finish	Odds	Comment
Tiger Woods	1st in 1997	3–1	His tournament to lose
Phil Mickelson	3rd in 1996	10–1	Overdue for major breakthrough
Vijay Singh	1st in 2000	10–1	Faldo successfully defended in 1990
Davis Love III	2nd in 1999	15–1	Has come oh-so-close before
Colin Montgomerie	Tied for 8th in 1998	15–1	Sooner or later he'll get it right
José Maria Olazabal	1st in 1994, 1999	20–1	Fearless competitor who never quits
Tom Lehman	2nd in 1994	25–1	Has all the tools to contend again
Nick Price	5th in 1986	25–1	If putter holds up, could be a factor
Ernie Els	2nd in 2000	25–1	Play lacking lately, but ready to rise up
David Duval	Tied for 2nd in 1998	25–1	Wrist, back only question marks
Jesper Parnevik	Tied for 21st in 1997	30–1	A major is next for gritty Swede
Mark Calcavecchia	2nd in 1998	30–1	Streaky player, never backs off
Sergio Garcia	Tied for 38th in 1999	35–1	Doesn't lack game or confidence
Justin Leonard	Tied for 7th in 1997	35–1	Good grinder who won't beat himself
Jim Furyk	4th in 1998	35–1	Will long putter bag a major?
Greg Norman	2nd in 1996	35–1	Everybody's sentimental favorite
Paul Azinger	5th in 1998	40–1	Playing well and knows the layout
Darren Clarke	Tied for 8th in 1998	50–1	Cigar will come in handy at Amen Corner
Loren Roberts	Tied for 3rd in 2000	50–1	Splendid short game comes in handy
Brad Faxon	Tied for 9th in 1993	50–1	Has he ever hit a poor putt?
Fred Couples	Tied for 2nd in 1998	60–1	Never count him out
John Huston	Tied for 3rd in 1990	60–1	The man is a birdie machine
Mike Weir	Tied for 28th in 2000	60–1	Canadian continues to impress
Bernhard Langer	1st in 1993	65–1	Tough, determined and unflappable

well have been quoting un-normalized odds, which is a fairly common practice in sports, but let's take him literally in this part of the problem.)

(b) It would be nice to demonstrate Mr. Soltau's incoherence by explicitly providing a set of bets that would be guaranteed to lose him money, but that's actually fairly complicated (hint for the previous part of this question: that's *not* what I had in mind for you to do in (a)). To take a simpler example that has the same flavor as Mr. Soltau's mistake (if his odds are taken literally), pretend that he's handicapping (setting odds for) a tournament in which only Tiger Woods, Phil Mickelson, and some other unnamed golfers are playing, and he announces 3 to 1 odds in *favor* of Woods winning and 1 to 1 odds in favor of Mickelson (again without specifying any odds for the other golfers). (To be clear on the relationship between odds and money, here's how it works in horse-racing (and Mr. Soltau would have to play by the same rules): suppose that a bookie at the horse track offers odds of 4 to 1 against horse A, and I bet (say) \$1 on that horse to win; if horse A wins I enjoy a net gain of \$4, otherwise I suffer a net loss of \$1.) Work out an explicit set of bets to offer Mr. Soltau that would constitute a Dutch book against him. If Mr. Soltau were willing to accept arbitrarily large bets, is there any theoretical limit to the amount of money you would be guaranteed to win from him? Explain briefly.

(c) In practice sports bookies only allow people to make bets *for* individual golfers, so that in reality you're not allowed to construct a wager like {\$$x$ on Woods to win and \$$y$ on Mickelson to lose}. Can you make Dutch book against Mr. Soltau under these conditions? Explain briefly.

3. (Bayes's Theorem; based on problem 7 in chapter 1 of Gelman A, Carlin JB, Stern HS, Rubin DB (2004). *Bayesian Data Analysis*, second edition. New York: Chapman & Hall/CRC.) In the old television game show *Let's Make a Deal*, there are three doors; behind one of the doors is a car, and behind the other two are goats, with the assignment of prizes to doors made at random. You — the contestant, who prefers cars to goats — are asked to pick a door. After you choose (of course you can do no better than picking at random), the emcee, Monte Hall, who knows where the car is, opens one of the other doors to reveal a goat, and he offers you the choice of staying with the door you originally picked or switching to the other unopened door. Suppose that Monte Hall uses the following algorithm to decide which door to reveal to you after you've chosen (say) door 1. If the car is behind door 2 he shows you door 3; if it's behind door 3 he shows you door 2; and if it's behind door 1 he randomizes between showing you doors 2 and 3 with equal probability. Should you switch or stay with your original choice?

(a) Explicitly use Bayes's Theorem to work out the chance of winning the car under each strategy.

(b) How would you explain intuitively to someone who favors the inferior strategy why the other one is better?

4. (Conditional probability, and review of the normal distribution; based on problem 4 in chapter 1 of Gelman et al.) (American) football (not soccer) experts provide a *point spread* (PS) for every football game as a measure of the difference in ability between the two teams. For example, team A might be a 3.5–point favorite over team B. This means that the proposition that A (the favorite) defeats B (the underdog) by 4 or more points is considered a fair bet, i.e., $P(A \text{ wins by more than 3.5 points}) = \frac{1}{2}$. If the PS is an integer, the implication is that A is as likely to win by more points than the PS as it is to win by fewer points than the PS (or to lose); there is a positive probability that A will win by exactly the PS, in which case neither side is paid off. In Chapter 1 Gelman et al. present

data on the PS and actual game outcome for 672 professional football games played during the 1981 and 1983–84 seasons, and they show that the histogram of the quantity (actual outcome – PS) is well approximated by a normal distribution with mean 0.07 and standard deviation (SD) 13.86, suggesting that a good predictive distribution for the actual result of an NFL football game would be normal with mean equal to the PS and SD 14 points (two touchdowns). (If you're in the habit of betting on NFL games this should give you pause, e.g., if a team is favored by a touchdown the chance it will win, according to this uncertainty assessment, is only about 69%.) It turns out that there were 12 games in this data base with PS values of 8 points, and the actual outcomes in those games were –7, –5, –3, –3, 1, 6, 7, 13, 15, 16, 20, and 21, with positive (negative) values indicating wins by the favorite (underdog). Consider the following conditional probabilities:

$$P(\text{favorite wins}|\text{PS} = 8)$$
$$P(\text{favorite wins by at least } 8|\text{PS} = 8)$$
$$P(\text{favorite wins by at least } 8|\text{PS} = 8 \text{ and favorite wins})$$

(a) Estimate each of these using the relative frequencies of the games with an 8–point PS.

(b) Estimate each using the normal approximation to the distribution of (actual outcome – PS). (You can use a normal table from any statistics book, or the error function `erf` in `Maple`, or the `pnorm` function in `R`.)

(c) Which of these approaches to uncertainty assessment seems to have produced better answers here? How should we define "better"? Explain briefly.

5. (Cromwell's Rule and its implications for Bayesian learning) Prove the following two facts: for any D such that $P(D) > 0$,

(a) If $P(A) = 0$ then $P(A|D) = 0$.

(b) If $P(A) = 1$ then $P(A|D) = 1$.

In the usual application of these facts (as in the HIV case study in class), A is a proposition whose truth value is unknown to You (such as the HIV status of the patient) and D represents some data relevant to A (such as the result of a screening test like *ELISA*); in this setting (a) and (b) together are referred to as *Cromwell's Rule* (I'll give the history behind this in class). What are the implications of Cromwell's Rule for the use of Bayes's Theorem as a formal model for learning from data? Explain briefly.