How to assign partial credit on an exam of true-false questions?

1 June, 2016 in math.PR, math.ST | Tags: grading

Note: the following is a record of some whimsical mathematical thoughts and computations I had after doing some grading. It is likely that the sort of problems discussed here are in fact well studied in the appropriate literature; I would appreciate knowing of any links to such.

Suppose one assigns N true-false questions on an examination, with the answers randomised so that each question is equally likely to have "true" as the correct answer as "false", with no correlation between different questions. Suppose that the students taking the examination must answer each question with exactly one of "true" or "false" (they are not allowed to skip any question). Then it is easy to see how to grade the exam: one can simply count how many questions each student answered correctly (i.e. each correct answer scores one point, and each incorrect answer scores zero points), and give that number k as the final grade of the examination. More generally, one could assign some score of A points to each correct answer and some score (possibly negative) of B points to each incorrect answer, giving a total grade of $Ak + B(N-k)$ points. As long as $A > B$, this grade is simply an affine rescaling of the simple grading scheme k and would serve just as well for the purpose of evaluating the students, as well as encouraging each student to answer the questions as correctly as possible.

In practice, though, a student will probably not know the answer to each individual question with absolute certainty. One can adopt a probabilistic model, where for a given student S and a given question n, the student S may think that the answer to question n is true with probability $p_{S,n}$ and false with probability $1 - p_{S,n}$, where $0 \leq p_{S,n} \leq 1$ is some quantity that can be viewed as a measure of confidence S has in the answer (with S being confident that the answer is true if $p_{S,n}$ is close to 1, and confident that the answer is false if $p_{S,n}$ is close to 0); for simplicity let us assume that in S's probabilistic model, the answers to each question are independent random variables. Given this model, and assuming that the student S wishes to maximise his or her expected grade on the exam, it is an easy matter to see that the optimal strategy for S to take is to answer question n true if $p_{S,n} > 1/2$ and false if $p_{S,n} < 1/2$. (If $p_{S,n} = 1/2$, the student S can answer arbitrarily.)

[Important note: here we are *not* using the term "confidence" in the technical sense used in statistics, but rather as an informal term for "subjective probability".]

This is fine as far as it goes, but for the purposes of evaluating how well the student actually knows the material, it provides only a limited amount of information, in particular we do not get to directly see the student's subjective probabilities $p_{S,n}$ for each question. If for instance S answered 7 out of 10 questions correctly, was it because he or she actually knew the right answer for seven of the questions, or was it because he or she was making educated guesses for the ten questions that turned out to be slightly better than random chance? There seems to be no way to discern this if the only input the student is allowed to provide for each question is the single binary choice of true/false.

But what if the student were able to give probabilistic answers to any given question? That is to say, instead of being forced to answer just "true" or "false" for a given question n, the student was allowed to give answers such as "60% confident that the answer is true" (and hence 40% confidence the answer is false). Such answers

would give more insight as to how well the student actually knew the material; in particular, we would theoretically be able to actually see the student's subjective probabilities $p_{S,n}$.

But now it becomes less clear what the right grading scheme to pick is. Suppose for instance we wish to extend the simple grading scheme in which an correct answer given in 100% confidence is awarded one point. How many points should one award a correct answer given in 60% confidence? How about an incorrect answer given in 60% confidence (or equivalently, a correct answer given in 40% confidence)?

Mathematically, one could design a grading scheme by selecting some grading function $f : [0,1] \to \mathbf{R}$ and then awarding a student $f(p)$ points whenever they indicate the correct answer with a confidence of p. For instance, if the student was 60% confident that the answer was "true" (and hence 40% confident that the answer was "false"), then this grading scheme would award the student $f(0.6)$ points if the correct answer actually was "true", and $f(0.4)$ points if the correct answer actually was "false". One can then ask the question of what functions f would be "best" for this scheme?

Intuitively, one would expect that f should be monotone increasing – one should be rewarded more for being correct with high confidence, than correct with low confidence. On the other hand, some sort of "partial credit" should still be assigned in the latter case. One obvious proposal is to just use a linear grading function $f(p) = p$ – thus for instance a correct answer given with 60% confidence might be worth 0.6 points. But is this the "best" option?

To make the problem more mathematically precise, one needs an objective criterion with which to evaluate a given grading scheme. One criterion that one could use here is the avoidance of perverse incentives. If a grading scheme is designed badly, a student may end up overstating or understating his or her confidence in an answer in order to optimise the (expected) grade: the optimal level of confidence $q_{S,n}$ for a student S to report on a question may differ from that student's subjective confidence $p_{S,n}$. So one could ask to design a scheme so that $q_{S,n}$ is always equal to $p_{S,n}$, so that the incentive is for the student to honestly report his or her confidence level in the answer.

This turns out to give a precise constraint on the grading function f. If a student S thinks that the answer to a question n is true with probability $p_{S,n}$ and false with probability $1 - p_{S,n}$, and enters in an answer of "true" with confidence $q_{S,n}$ (and thus "false" with confidence $1 - q_{S,n}$), then student would expect a grade of

$$p_{S,n} f(q_{S,n}) + (1 - p_{S,n}) f(1 - q_{S,n})$$

on average for this question. To maximise this expected grade (assuming differentiability of f, which is a reasonable hypothesis for a partial credit grading scheme), one performs the usual maneuvre of differentiating in the independent variable $q_{S,n}$ and setting the result to zero, thus obtaining

$$p_{S,n} f'(q_{S,n}) - (1 - p_{S,n}) f'(1 - q_{S,n}) = 0.$$

In order to avoid perverse incentives, the maximum should occur at $q_{S,n} = p_{S,n}$, thus we should have

$$p f'(p) - (1-p) f'(1-p) = 0$$

for all $0 \leq p \leq 1$. This suggests that the function $p \mapsto p f'(p)$ should be constant. (Strictly speaking, it only gives the weaker constraint that $p \mapsto p f'(p)$ is symmetric around $p = 1/2$; but if one generalised the problem to allow for multiple-choice questions with more than two possible answers, with a grading scheme that depended only on the confidence assigned to the correct answer, the same analysis would in fact force $p f'(p)$ to

be constant in p; we leave this computation to the interested reader.) In other words, $f(p)$ should be of the form $A \log p + B$ for some A, B; by monotonicity we expect A to be positive. If we make the normalisation $f(1/2) = 0$ (so that no points are awarded for a $50-50$ split in confidence between true and false) and $f(1) = 1$, one arrives at the grading scheme

$$f(p) := \log_2(2p).$$

Thus, if a student believes that an answer is "true" with confidence p and "false" with confidence $1-p$, he or she will be awarded $\log_2(2p)$ points when the correct answer is "true", and $\log_2(2(1-p))$ points if the correct answer is "false". The following table gives some illustrative values for this scheme:

Confidence that answer is "true"	Points awarded if answer is "true"	Points awarded if answer is "false"
0%	$-\infty$	1.000
1%	-5.644	0.9855
2%	-4.644	0.9709
5%	-3.322	0.9260
10%	-2.322	0.8480
20%	-1.322	0.6781
30%	-0.737	0.4854
40%	-0.322	0.2630
50%	0.000	0.000
60%	0.2630	-0.322
70%	0.4854	-0.737
80%	0.6781	-1.322
90%	0.8480	-2.322
95%	0.9260	-3.322
98%	0.9709	-4.644
99%	0.9855	-5.644
100%	1.000	$-\infty$

Note the large penalties for being extremely confident of an answer that ultimately turns out to be incorrect; in particular, answers of 100% confidence should be avoided unless one really is absolutely certain as to the correctness of one's answer.

The total grade given under such a scheme to a student S who answers each question n to be "true" with confidence $p_{S,n}$, and "false" with confidence $1-p_{S,n}$, is

$$\sum_{n:\ \text{ans is true}} \log_2(2p_{S,n}) + \sum_{n:\ \text{ans is false}} \log_2(2(1-p_{S,n})).$$

This grade can also be written as

$$N + \frac{1}{\log 2} \log \mathcal{L}$$

where

$$\mathcal{L} := \prod_{n:\ \text{ans is true}} p_{S,n} \times \prod_{n:\ \text{ans is false}} (1-p_{S,n})$$

is the likelihood of the student S's subjective probability model, given the outcome of the correct answers. Thus the grade system here has another natural interpretation, as being an affine rescaling of the log-likelihood. The incentive is thus for the student to maximise the likelihood of his or her own subjective model, which aligns well with standard practices in statistics. From the perspective of Bayesian probability, the grade given to a student can then be viewed as a measurement (in logarithmic scale) of how much the posterior probability that the student's model was correct has improved over the prior probability.

One could propose using the above grading scheme to evaluate predictions to binary events, such as an upcoming election with only two viable candidates, to see in hindsight just how effective each predictor was in calling these events. One difficulty in doing so is that many predictions do not come with explicit probabilities attached to them, and attaching a default confidence level of 100% to any prediction made without any such qualification would result in an automatic grade of $-\infty$ if even one of these predictions turned out to be incorrect. But perhaps if a predictor refuses to attach confidence level to his or her predictions, one can assign some default level p of confidence to these predictions, and then (using some suitable set of predictions from this predictor as "training data") find the value of p that maximises this predictor's grade. This level can then be used going forward as the default level of confidence to apply to any future predictions from this predictor.

The above grading scheme extends easily enough to multiple-choice questions. But one question I had trouble with was how to deal with *uncertainty*, in which the student does not know enough about a question to venture even a probability of being true or false. Here, it is natural to allow a student to leave a question blank (i.e. to answer "I don't know"); a more advanced option would be to allow the student to enter his or her confidence level as an interval range (e.g. "I am between 50% and 70% confident that the answer is "true""). But now I do not have a good proposal for a grading scheme; once there is uncertainty in the student's subjective model, the problem of that student maximising his or her expected grade becomes ill-posed due to the "unknown unknowns", and so the previous criterion of avoiding perverse incentives becomes far less useful.