Recall & Precision vs The Bookmaker

The Research Gamble

Research is about trial and error, posing problems, searching for solutions. Frequently we are in the position of testing a highly approximate theory or model and then having to evaluate its predictions against real world results.

This leads to a contingency matrix in which we tabulate the statistics for how these actual results compare with our model. The simplest case has positive and negative outcomes for both the real labels and the predictions. The actual results are shown in columns for ±R whilst the predictions are showing in rows for ±P. Ideally the main diagonal, +P+R and –P–R, has 100% of cases and the off diagonal cells are all 0 – no misclassifications

In the social and empirical sciences, the next step would be to perform a significance test to assess the probability that the results are due to chance and hence estimate whether there is sufficient test data for the conclusions to be valid. Multiplying the figures by some large factor whilst retaining the same proportions is always sufficient to assure significance.

For information retrieval, machine learning and neural nets, the problem is not normally too little data but too much! Large amounts of data are used in training and significant proportions are set aside for validation and testing to avoid overtraining. In information retrieval or web search, the problem is that keyword search returns masses of 'hits' and we need to assess how useful the results are – that is we want to know the 'accuracy'.

With gambling and trading, we are not interested in just any arbitrary definition of accuracy but want a formula that will translate directly to dollars and cents, quantifying to what extent our system gives us an edge. In the final analysis the bookmaker gives us odds, the market gives us prices, and the broker charges us commissions – and these do translate our decisions directly to dollars and cents, whether on paper or in practice.

In this case, if the proportions are the same in our contingency table, the percentage return will be the same too, unlike significance. This paper shows that fair bookmaker odds define the unique formula for determining the probability or proportion of time that we are making an informed correct decision, versus guessing or making a superstitious decision.

This is different from significance in that it is scale independent, assessing how useful the contingency is rather than how rare. In a scientific or machine learning application, some kind of significance test or validation should also be performed to ensure that the results are not due to chance or overtraining.

Trotting out Recall and Precision

The standard accuracy measures used to evaluate neural nets, learning, parsing, tagging, searching, etc. comes from the search application where we have a pool of documents, the corpus, some of which are relevant (+R) but most of which are not (– R). Our search procedure returns a set of documents predicted as useful (+P) and omits others that are predicted as irrelevant (–P). The proportion of relevant documents returned is *recall*, |+P+R|/|+R|, whilst the proportion of returned documents relevant is *precision*, |+P+R|/|+P|.

	+R	–R	
+P	12	28	40
–P	18	42	60
	30	70	100

Correctly labeled cases lie on the main diagonal: *accuracy* is 12+42 correct labels out of 100 or 54%.

+R is class 1, +ve cases, 12 predicted, 18 not predicted; *recall* or *sensitivity* = *tpr* = 12/30 = 40% = *chance* level

–R is class 2, – ve cases, 42 identified, 28 misclassified; so *inverse recall* or *specificity* is 42/70 = 60%, *fallout* = *fpr* = 28/70 = 40%

The model predicted 40 +ve cases and 60 – ve cases, but in this example we were just *guessing* so we expect 40% *recall* just because we gave 40% positive labels

+P is label 1, +ve predictions, 12 right 28 wrong; *precision* is 12/40 or 30% = *chance* level = *guessing*

– P is label 2, – ve predictions, 42 right 18 wrong; *inverse precision* is 42/60 or 70%

Misclassified cases are off diagonal.

30 cases are winners and 70 are irrelevant. If we had no way of predicting and just *guessed* then we'd expect to be right 30% of the time and thus achieve 30% *precision*

The total number, *N*, of labeled cases is 100 in this example, so we can also read these figures as %. The table shows the expected results for mere *guessing*

Recall and Precision are Losers

Recall and precision suffer from a number of disadvantages that make them unsuitable for defining an accuracy measure:

- They assess only a single condition
- There is a tradeoff between them
- Neither can be interpreted alone
- They ignore the cost of errors
- Each is easily inflated:
  - ¤ Recall by labeling more cases +ve
  - ¤ Precision by labeling hard ones –ve

Accuracy can be Biased

These problems can be addressed in part by defining an overall accuracy across all conditions.

The weighted averages of correct cases, of recall and inverse recall, of precision and inverse precision, are all equivalent *accuracy* measures. But this definition does not take into account the cost of errors or the baseline for guessing.

The Bookmaker always Wins

In gambling, the house always wins, and in horseracing, the bookmaker is no exception. The basis of the odds set by a bookie is the assessed likelihood of a horse winning, as influenced by talk and bets from those in the know. The bookie will then add on a percentage as she calculates the odds.

We will work with fair odds based on statistics on past performance. In our example our horse has won 30 out of 100 starts leading to odds of 7:3 for our field. This means if the horse wins I win \$7 on a \$3 bet that I stand to lose if it loses. Note that these fair odds mean there is a 30% chance of winning \$7 and a 70% chance of losing \$3 and they are fair in the sense of being zero sum: the expected gain is 30% of \$7 – 70% of \$3 = 0.

This punter doesn't know anything – he's just guessing! A real bookie would relieve him of his money in no time. But the fair Bookmaker algorithm simply reports that his *edge* is zero!

He won 12 × \$7 + 42 × \$3 = \$210.
He lost 18 × \$7 + 28 × \$3 = \$210.

He was lucky to break even!

The Bookmaker tests your Edge

Traders and gamblers know that they have to have an edge to win – and this edge has to be bigger than the house percentage or the costs of trading to make a profit. Markets are fairly efficient, and brokers and bookmakers and your fellow speculators aren't all fools – there has to be real information available and correctly used in order to be able to win consistently. This is your edge. This small bias in your favour eventually adds up to a proportionate profit.

General Bookmaker

We now show how to apply Bookmaker to the general classification case where there are *K* classes we are trying to identify. In this case we simply calculate and use the odds separately for each of the *K* horses. Note that once you bet on a horse, your system specifies a label, the value of the bet, the penalty incurred if you lose, is specified independent of which other horse wins. We furthermore present it in a normalized form such that the expected gain directly gives the probability that you are making an informed decision as opposed to guessing.

In defining Bookmaker formally we make use of sample probabilities from the contingency matrix: $\mathrm{p}_{\mathbf{R}}(c) = |\mathbf{R} = c| / N$, $\mathrm{p}_{\mathbf{P}}(l) = |\mathbf{P} = l| / N$ and also $\mathrm{p}_{\mathbf{PR}}(l, c) = |\mathbf{P} = l \;\&\; \mathbf{R} = c| / N$. Our example has two classes using {+, –} to denote *predicted labels* **P** and *real classes* **R**, but in general we use the set $\mathbf{P} = \mathbf{R} = \{1 .. K\}$. In this formulation, precision and recall correspond to conditional probabilities and are respectively $\mathrm{p}_{\mathbf{R}}(c{=}l|l) = \mathrm{p}_{\mathbf{PR}}(l, c) / \mathrm{p}_{\mathbf{P}}(l)$ and $\mathrm{p}_{\mathbf{P}}(l{=}c|c) = \mathrm{p}_{\mathbf{PR}}(l, c) / \mathrm{p}_{\mathbf{R}}(c)$.

The generalized Bookmaker payoff formula is then

$$B = \sum_{l\in \mathbf{P}} \mathrm{p}_{\mathbf{P}}(l) \sum_{c\in \mathbf{R}} \mathrm{p}_{\mathbf{PR}}(l,c)\, \mathrm{w}(c|l),$$

where

$$\begin{aligned} \mathrm{w}(c|l) &= +1 / \mathrm{p}_{\mathbf{R}}(l) && (c = l), \\ &= -1 / (1 - \mathrm{p}_{\mathbf{R}}(l)) && (c \neq l). \end{aligned}$$

In our example, the normalization corresponds to dividing the payoffs by $N\, \mathrm{p}_{\mathbf{R}}(l)\, (1 - \mathrm{p}_{\mathbf{R}}(l)) = 100 \times 0.30 \times 0.70 = 21$ (independent of *l*). This guesswork matrix gives $B = .4(.12/.3 - .28/.7) + .6(.42/.7 - .18/.3) = 0$. The perfect decision matrix has $\mathrm{p}_{\mathbf{P}}(l) = \mathrm{p}_{\mathbf{R}}(l) = \mathrm{p}_{\mathbf{PR}}(l, l)$ so |+P+R| = 30 and |–P–R| = 70 whence $B = .3(.3/.3 - 0/.7) + .7(.7/.7 - 0/.3) = 1$.

The Bookmaker measures Informedness

The Bookmaker formula measures the 'informedness' of our decisions. Suppose that we guess 50% of the time and make a correct informed decision 50% of the time. Our contingency matrix will then be the average of the guessing matrix and the perfect decision matrix giving the calculation $B = .35((.06+.15)/.3 - (.14+0)/.7) + .65((.21+.35)/.7 - (.09+0)/.3) = 0.5$.

The Bookmaker is Unique

The ability to recover informedness is unique to the bookmaker measure, and indeed it also detects informed incorrect decisions, whether deliberate, due to overtraining or a function of atypical data. In this case Bookmaker will return a negative value. Its uniqueness follows from the linearity of the equation combined with the linearity of our assumed mix of guessing and informed decision. Note that in the binary (yes/no) case $B = tpr - fpr$.

© 2003 David M. W. Powers

School of Informatics and Engineering
Flinders University of South Australia
G.P.O. Box 2100, ADELAIDE 5001
SOUTH AUSTRALIA

Contact Details

http://www.infoeng.flinders.edu.au/people/pages/powers_david/

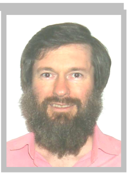
Copies of this poster, paper, spreadsheets, scripts may be obtained from:
David.Powers@flinders.edu.au or powers@acm.org or
http://members.dodo.com.au/~powers/BM/

References

Entwisle, J., & Powers, D. M. W. (1998). The Present Use of Statistics in the Evaluation of NLP Parsers, *Proceedings of NeMLaP3/CoNLL98* (pp. 215-224). Somerset NJ: ACL.

Manning, C. D., & Schutze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

David M. W. Powers is Associate Professor of Computer Science and Head of the AI Lab at Flinders University. He specializes in applications of unsupervised learning to language and speech processing. Dr Powers undertook his PhD in this area, as well as co-founding ACL's SIGNLL and CoNLL. He is also a trader and has a Diploma in Technical Analysis, being the study of how to find and exploit 'edges' in the financial markets.



Joint International Conference on Cognitive Science, July 2003, Sydney Australia, pp529-534 (revised Jan 2006)