Evaluation Evaluation

David Powers

AI Lab, InfoEng, Flinders University

Language Learning

HCSNet

The ARC Network in Human Communication Science

Evaluation in NLP is biased…

- *Consider a POS tagging task*
  - Is this word “water” a Noun in this context?
- *For chance level performance*

  Proportion of +ve predictions correct Proportion of outcomes +ve
  - Precision reflects *Prevalence* (is_Noun 90%)

  Proportion of +ve outcomes correct Proportion of predictions +ve
  - Recall reflects *Bias* (tagged_Noun 100%)

  Proportion of outcomes or predictions correct
  - Accuracy/F-factor reflect both (90%/95%)

  Proportion of +ve outcomes and predictions correct

High Accuracy/F-factor can arise from high bias!

Medicine, Psychology & Gambling

- *Receiver Operating Characteristics (ROC)*
  - Recall vs Fallout tradeoff (*tpr* vs *fpr*)
  - Unbiased default or can allow for *cost / skew*
- *Normative Measure of Contingency (DeltaP)*
  - Predictor of Human Associative Judgements
- *Bookmaker Gambling Edge (Informedness)*
  - Expected profit betting with fair odds

Outcomes

Predictors

Profits

What is the probability of an informed choice?

Correlation and Significance

Recall = Informedness·invPrev+Bias

- *Informedness or DeltaP'* = Recall + invRecall – 1 (0)
  Proportion of time *Predictions* are correct (+ve: **Precision**; all: **Accuracy**)
  – Probability *predictions* are informed vs chance outcome

Precision = Markedness·invBias+Prev

- *Markedness or DeltaP* = Precision + InvPrecision – 1 (0)
  Proportion of time *Real outcomes* are correct (+ve: **Recall**; all: **Accuracy**)
  – Probability *outcomes* are marked vs chance prediction

- *Correlation:* ρ^2 = Markedness·Informedness (0)
  Proportion of time
  – Probability variance is explained reciprocally

Evenness = GeometricMean(Prev·invPrev·Bias·invBias)

- *Significance:* χ^2 = $_k$N·Evenness·Markedness·Informedness (0)
  Proportion of time *Superstitious* contingency ($r=(k-1)^2$ freedoms; k variables; N samples)
  – Probability rejecting true null (Type I Error): $Q(r/2, \chi^2/2)$

http://david.wardpowers.info/~powers/BM David.Powers@flinders.edu.au

Evaluation Evaluation

Over the last decade I have been very concerned by the meaninglessness of evaluation in NLP. In particular, the simple conditional probabilities Recall, Precision and Accuracy are not meaningful as evaluation measures, either individually or in combination, without knowledge of the Bias and Prevalence of the contigency being tested, or equivalently the expectation due to chance.

In 1997 I exemplified this with a number of examples, including in particular the positive effect on all three measures of a POS tagger predetermining that *water* was only ever a Noun.

In 2003 I introduced a measure called Bookmaker Informedness and proved that that was the only unbiased indicator of the probability that an informed decision was being made rather than guessing.

In Psychology, the simple binary (yes/no) dichotomous version of Informedness is known as DeltaP′ and is the complement of DeltaP, both of which are highly predictive of humans forming association between contingent events, although again the complaint surfaces that simple conditional probabilities are being used as measures without proper justification.

In Medicine, ROC analysis is commonly used for dichotomous analysis of variates and parameterization of models, which also turns out to optimize Informedness in this binary case.

Over the last year, I have further developed the Bookmaker measures and introduced a dual measure, Markedness which corresponds to DeltaP in the binary case. The two measures lead to surprisingly simple and insightful characterizations of Recall, Precision, Accuracy, Correlation & Significance.