The traditional measure for Machine Learning/Natural Language experiments has been borrowed from Information Retrieval, where Precision reflects the accuracy of positive predictions, and Recall reflects the rate of return of real positives. These measures are totally independent of the number of negative cases that are truly predicted as negative (TN=0). The F-factor is a harmonic mean of Recall and Precision and thus also ignores these cases. The Rand Accuracy can be regarded as a weighted average of Recall and Inverse Recall, or of Precision and Inverse Precision (where the Inverse problem reverses positive and negative), and thus does reflect both. Often however, we do not know all the positive or negative cases, so we cannot calculate anything but Precision and derivatives of Recall based on the known subsets.

A second problem with all of these measures is that they are heavily influenced by two kinds of bias – Prevalence, the proportion of real positives, and prediction Bias, the proportion of positive predictions, pp=α+β/8. A third issue is the confidence we have in these values, or the significance of the purported results.

Powers [1] derived an unbiased measure, Informedness, based on the idea of an “edge” in gambling, given knowledge of the underlying base probabilities and rewarding success and penalizing failures according fair odds (Bookmaker) for an arbitrary number of classes.

**Dichotomous cases – Informedness, Markedness & Correlation**

In the case of just two classes (+ and –), Informedness can be understood in terms of ROC analysis either as the distance from a specific prediction system to the chance line, as the Unbiased Weighted Relative Accuracy (WRAcc for Skew=0), or as twice the area between the curve and the chance line, or in terms of the Area Under the Curve (AUC) as 2AUC-1. In Psychology, Dichotomous Informedness corresponds to Deltap^2. Deltap is identified empirically as the normative predictor of human associative judgements, as one concept primes or marks another. Whereas Informedness corresponds and Deltap^2 and is the unbiased form Recall, Deltap^2 corresponds to Markedness, how much the actual class influences or marks the selected predictor, and relates to Precision, being dual regression coefficients whose product is the Matthews Correlation, and associating these with Cramer’s V gives us the corresponding (p^2) significance estimates. Similarly we can directly set confidence intervals [2].

**Definition 1**

Informedness quantifies how informed a predictor is for the specified condition, and specifies the probability that a prediction is informed in relation to the condition.

\[ \text{Informedness} = \frac{\text{Recall} + \text{Inverse Recall} - 1}{\text{Recall} + \text{Inverse Recall}} = \frac{\text{tp} - \text{fp}}{\text{tp} + \text{fp}} \]

Confidence Interval CI = N(1-Informedness)/N(1) where \( N > 0.05 \).

**Definition 2**

Markedness quantifies how marked a condition is for the specified predictor, and specifies the probability that a condition is marked by the predictor.

\[ \text{Markedness} = \frac{\text{Precision} + \text{Inverse Precision} - 1}{\text{Precision} + \text{Inverse Precision}} = \frac{\text{tp} + \text{tn}}{\text{tp} + \text{tn}} \]

Confidence Interval CM = N(1-Markedness)/N(1) where \( N > 0.05 \).

References


Illustration of Significance and Confidence. 110 Monte Carlo simulations with 11 stepped expected Informedness levels (red line) with Bookmaker-estimated Informedness (blue dots), Markedness (green dot) and Correlation (blue dot), with significance (p=1) calculated using G^2 , Fisher estimates, and confidence bands shown for both the theoretical Informedness and the B=0 and B=1 levels (parallel at p=0.18,0.82). The lower theoretical band is calculated twice, using both CI\(_1\) and CI\(_2\). Here K=5, N=128, \( \chi^2 = 1.96 \) for two-tailed m=0.05. The upper theoretical band is calculated twice, using both CI\(_1\) and CI\(_2\). Here K=5, N=128, \( \chi^2 = 1.96 \) for two-tailed m=0.05.