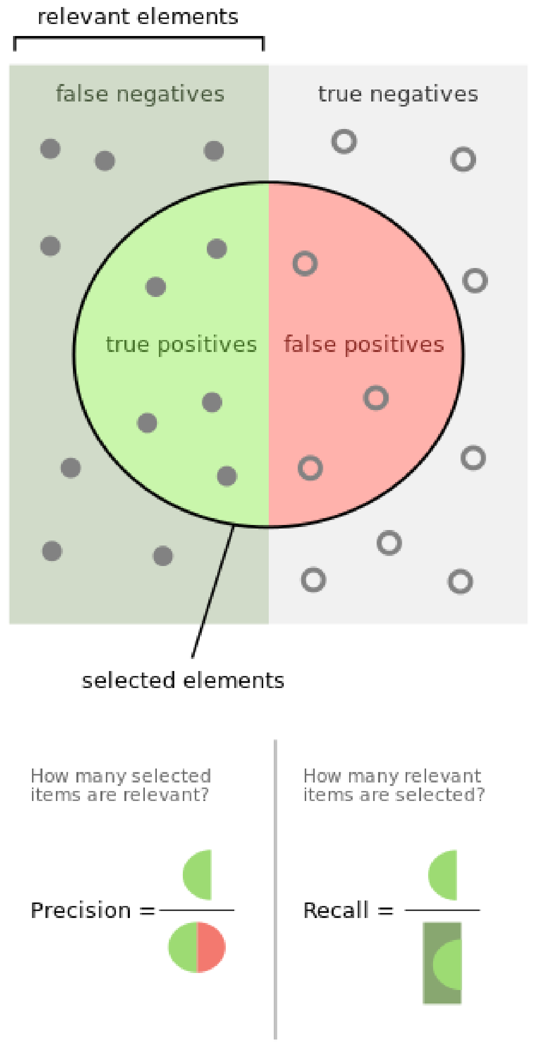
WIKIPEDIA

F1 score

In statistical analysis of binary classification, the $\mathbf{F_1}$ **score** (also **F-score** or **F-measure**) is a measure of a test's accuracy. It considers both the precision *p* and the recall *r* of the test to compute the score: *p* is the number of correct positive results divided by the number of all positive results returned by the classifier, and *r* is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive). The F_1 score is the harmonic average of the precision and recall, where an F_1 score reaches its best value at 1 (perfect precision and recall) and worst at 0.



Precision and recall

Contents

Etymology

The name F-measure is believed to be named after a different F function in Van Rijsbergen's book, when introduced to MUC-4.[1]

Definition

The traditional F-measure or balanced F-score ($\mathbf{F_1}$ **score**) is the harmonic mean of precision and recall:

$$F_1 = \left(\frac{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}}{2}\right)^{-1} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}.$$

The general formula for positive real β is:

$$F_\beta = (1 + \beta^2) \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{(\beta^2 \cdot \mathrm{precision}) + \mathrm{recall}}.$$

The formula in terms of Type I and type II errors

$$F_\beta = \frac{(1 + \beta^2) \cdot \mathrm{true\ positive}}{(1 + \beta^2) \cdot \mathrm{true\ positive} + \beta^2 \cdot \mathrm{false\ negative} + \mathrm{false\ positive}}.$$

Two other commonly used F measures are the F_2 measure, which weighs recall higher than precision (by placing more emphasis on false negatives), and the $F_{0.5}$ measure, which weighs recall lower than precision (by attenuating the influence of false negatives).

The F-measure was derived so that F_β "measures the effectiveness of retrieval with respect to a user who attaches β times as much importance to recall as precision".[2] It is based on Van Rijsbergen's effectiveness measure

$$E = 1 - \left(\frac{\alpha}{p} + \frac{1-\alpha}{r}\right)^{-1}.$$

Their relationship is $F_\beta = 1 - E$ where $\alpha = \dfrac{1}{1+\beta^2}$.

The F_1 score is also known as the Sørensen–Dice coefficient or Dice similarity coefficient (DSC).

Diagnostic testing

This is related to the field of binary classification where recall is often termed as Sensitivity. There are several reasons that the F_1 score can be criticized in particular circumstances.[3]

<table>
<tr><td></td><td></td><td colspan="2">True condition</td><td></td><td colspan="2"></td></tr>
<tr><td></td><td>Total population</td><td>Condition positive</td><td>Condition negative</td><td>Prevalence $= \frac{\Sigma\ \text{Condition positive}}{\Sigma\ \text{Total population}}$</td><td colspan="2">Accuracy (ACC) $= \frac{\Sigma\ \text{True positive} + \Sigma\ \text{True negative}}{\Sigma\ \text{Total population}}$</td></tr>
<tr><td rowspan="2">Predicted condition</td><td>Predicted condition positive</td><td>True positive</td><td>False positive, Type I error</td><td>Positive predictive value (PPV), Precision $= \frac{\Sigma\ \text{True positive}}{\Sigma\ \text{Predicted condition positive}}$</td><td colspan="2">False discovery rate (FDR) $= \frac{\Sigma\ \text{False positive}}{\Sigma\ \text{Predicted condition positive}}$</td></tr>
<tr><td>Predicted condition negative</td><td>False negative, Type II error</td><td>True negative</td><td>False omission rate (FOR) $= \frac{\Sigma\ \text{False negative}}{\Sigma\ \text{Predicted condition negative}}$</td><td colspan="2">Negative predictive value (NPV) $= \frac{\Sigma\ \text{True negative}}{\Sigma\ \text{Predicted condition negative}}$</td></tr>
<tr><td></td><td></td><td>True positive rate (TPR), Recall, Sensitivity, probability of detection, Power $= \frac{\Sigma\ \text{True positive}}{\Sigma\ \text{Condition positive}}$</td><td>False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma\ \text{False positive}}{\Sigma\ \text{Condition negative}}$</td><td>Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$</td><td rowspan="2">Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$</td><td rowspan="2">$F_1$ score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$</td></tr>
<tr><td></td><td></td><td>False negative rate (FNR), Miss rate $= \frac{\Sigma\ \text{False negative}}{\Sigma\ \text{Condition positive}}$</td><td>Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma\ \text{True negative}}{\Sigma\ \text{Condition negative}}$</td><td>Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$</td></tr>
</table>

Applications

The F-score is often used in the field of information retrieval for measuring search, document classification, and query classification performance.[4] Earlier works focused primarily on the F_1 score, but with the proliferation of large scale search engines, performance goals changed to place more emphasis on either precision or recall[5] and so F_β is seen in wide application.

The F-score is also used in machine learning.[6] Note, however, that the F-measures do not take the true negatives into account, and that measures such as the Matthews correlation coefficient, Informedness or Cohen's kappa may be preferable to assess the performance of a binary classifier.[3]

The F-score has been widely used in the natural language processing literature, such as the evaluation of named entity recognition and word segmentation.

Criticism

David Hand and others criticize the widespread use of the F-score since it gives equal importance to precision and recall. In practice, different types of mis-classifications incur different costs. In other words, the relative importance of precision and recall is an aspect of the problem.[7]

Difference from G-measure

While the F-measure is the harmonic mean of recall and precision, the G-measure is the geometric mean.[3]

See also

- BLEU
- Matthews correlation coefficient
- METEOR
- NIST (metric)
- Precision and recall

- Receiver operating characteristic
- ROUGE (metric)
- Sørensen–Dice coefficient
- Uncertainty coefficient, aka Proficiency
- Word error rate (WER)

References

1. Sasaki, Y. (2007). "The truth of the F-measure" (https://www.toyota-ti.ac.jp/Lab/Denshi/COIN/people/yutaka.sasaki/F-measure-YS-26Oct07.pdf) (PDF).
2. Van Rijsbergen, C. J. (1979). *Information Retrieval* (http://www.dcs.gla.ac.uk/Keith/Preface.html) (2nd ed.). Butterworth-Heinemann.
3. Powers, David M W (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation" (http://www.bioinfopublication.org/files/articles/2_1_1_JMLT.pdf) (PDF). *Journal of Machine Learning Technologies*. **2** (1): 37–63.
4. Beitzel., Steven M. (2006). *On Understanding and Classifying Web Queries* (Ph.D. thesis). IIT. CiteSeerX 10.1.1.127.634 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.127.634).
5. X. Li; Y.-Y. Wang; A. Acero (July 2008). *Learning query intent from regularized click graphs* (https://pdfs.semanticscholar.org/6718/f8e95461456023196fe6409073151ab0513d.pdf) (PDF). *Proceedings of the 31st SIGIR Conference*.
6. See, e.g., the evaluation of the [1] (https://dl.acm.org/citation.cfm?id=1119195).
7. Hand, David. "A note on using the F-measure for evaluating record linkage algorithms - Dimensions" (https://app.dimensions.ai/details/publication/pub.1084928040). *app.dimensions.ai*. Retrieved 2018-12-08.

Retrieved from "https://en.wikipedia.org/w/index.php?title=F1_score&oldid=874064435"

This page was last edited on 16 December 2018, at 22:23 (UTC).

Text is available under the Creative Commons Attribution-ShareAlike License; additional terms may apply. By using this site, you agree to the Terms of Use and Privacy Policy. Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a non-profit organization.