# Why is the F-Measure a harmonic mean and not an arithmetic mean of the Precision and Recall measures?

Asked 6 years, 4 months ago Active 1 month ago Viewed 19k times

89

When we calculate the F-Measure considering both Precision and Recall, we take the harmonic mean of the two measures instead of a simple arithmetic mean.

What is the intuitive reason behind taking the harmonic mean and not a simple average?

35

machine-learning classification data-mining

Share Improve this question Follow

asked Oct 14 '14 at 8:22

London guy

**24.6k** 40 110 166

1 The intuition is to balance precision and recall (usually the best measurement, but in some case you want to maximize precision or recall, which is a different story). You cannot get a high f-score if either one is very low. – greeness Oct 14 '14 at 11:44

1 cse.unsw.edu.au/~teachadmin/info/harmonic3.html This is a good resource to understanding HM – Sudip Bhandari Jun 26 '17 at 7:20

2 Fix the link above: di.unipi.it/~bozzo/The%20Harmonic%20Mean.htm or the original @archive.org – stason Sep 17 '18 at 4:30

## 5 Answers

Active | Oldest | Votes

17

Here we already have some elaborate answers but I thought some more information about it would be helpful for some guys who want to delve deeper(especially why F measure).

According to the theory of measurement, the composite measure should satisfy the following 6 definitions:

1. Connectedness(two pairs can be ordered) and transitivity(if e1 >= e2 and e2 >= e3 then e1 >= e3)
2. Independence: two components contribute their effects independently to the effectiveness.
3. Thomsen condition: Given that at a constant recall (precision) we find a difference in effectiveness for two values of precision (recall) then this difference cannot be removed or reversed by changing the constant value.
4. Restricted solvability.

5. Each component is essential: Variation in one while leaving the other constant gives a variation in effectiveness.
6. Archimedean property for each component. It merely ensures that the intervals on a component are comparable.

We can then derive and get the function of the effectiveness:

$$E = 1 - \frac{1}{\alpha\left(\frac{1}{P}\right) + (1-\alpha)\frac{1}{R}}$$

And normally we don't use the effectiveness but the much simper F score because:

$$\begin{aligned} E &= 1 - \frac{1}{\frac{1}{\beta^2+1}\frac{1}{P} + \left(1 - \frac{1}{\beta^2+1}\right)\frac{1}{R}}, \\ &= 1 - \frac{PR}{\frac{1}{\beta^2+1}R + \frac{\beta^2+1-1}{\beta^2+1}P}. \\ &= 1 - \frac{(\beta^2+1)PR}{R+\beta^2 P}. \end{aligned}$$

Now you see that

$$E = 1 - F_\beta.$$

Now that we have the general formula of F measure:

$$\frac{(\beta^2+1)PR}{R+\beta^2 P}$$

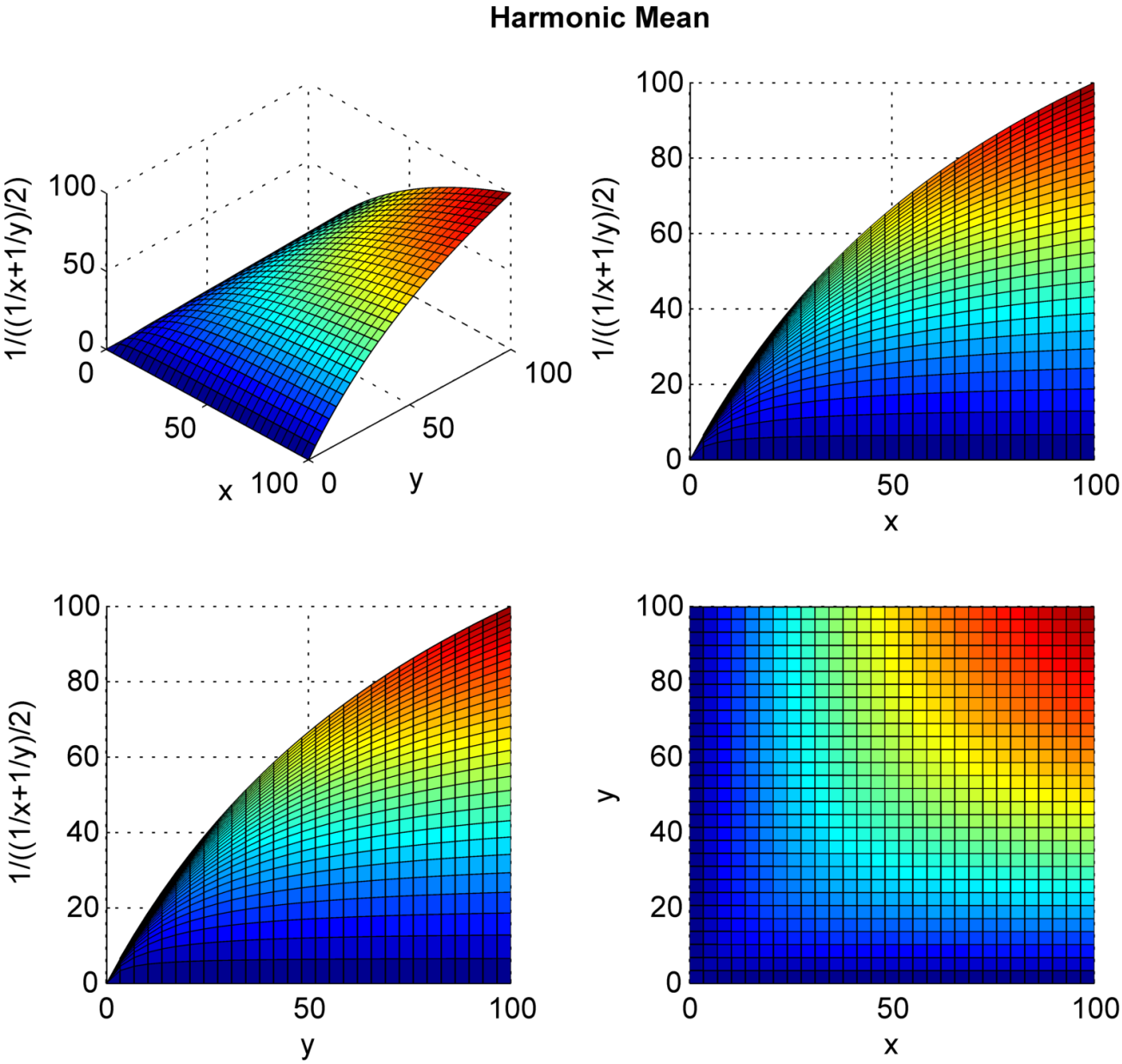
where we can place more emphasis on recall or precision by setting beta, because beta is defined as follows:

$$\beta = R/P, \quad where \quad \frac{\partial E}{\partial P} = \frac{\partial E}{\partial R}.$$

If we weight recall more important than precision(all relevant are selected) we can set beta as 2 and we get the F2 measure. And if we do the reverse and weight precision higher than recall(as many selected elements are relevant as possible, for instance in some grammar error correction scenarios like CoNLL) we just set beta as 0.5 and get the F0.5 measure. And obviously, we can set beta as 1 to get the most used F1 measure(harmonic mean of precision and recall).

I think to some extent I have already answered why we do not use the arithmetic mean.

Let's see the 3D plot of the harmonic mean. We can see that the harmonic mean is sensitive to the lowest value, especially the harmonic mean is 0 when at least one is 0 which doesn't hold for the simple arithmetic mean.




For more visualization of this topic please refer to this article: F1 score explained.

References:

1. https://en.wikipedia.org/wiki/F1_score
2. The truth of the F-measure
3. Information retrival
4. File:Harmonic mean 3D plot from 0 to 100.png

Share Improve this answer Follow edited Jan 18 at 14:18 answered May 14 '19 at 10:33 Lerner Zhang


**4,167** 1 32 45

102

To explain, consider for example, what the average of 30mph and 40mph is? if you drive for 1 hour at each speed, the average speed over the 2 hours is indeed the arithmetic average, 35mph.

However if you drive for the same distance at each speed -- say 10 miles -- then the average speed over 20 miles is the harmonic mean of 30 and 40, about 34.3mph.

The reason is that for the average to be valid, you really need the values to be in the same scaled units. Miles per hour need to be compared over the same number of hours; to compare over the same number of miles you need to average hours per mile instead, which is exactly what the harmonic mean does.

Precision and recall both have true positives in the numerator, and different denominators. To average them it really only makes sense to average their reciprocals, thus the harmonic mean.

Share Improve this answer Follow

answered Oct 14 '14 at 14:54


Sean Owen
**63.4k** 22 134 169

7 Thanks, that is a good argument on why this is supported from theory; my answer was more on the pragmatic side. – Has QUIT--Anony-Mousse Oct 14 '14 at 20:51

79

Because it punishes extreme values more.

Consider a *trivial* method (e.g. always returning class A). There are infinite data elements of class B, and a single element of class A:

```
Precision: 0.0
Recall:    1.0
```

When taking the arithmetic mean, it would have 50% correct. Despite being the *worst* possible outcome! With the harmonic mean, the F1-measure is 0.

```
Arithmetic mean: 0.5
Harmonic mean:   0.0
```

In other words, to have a high F1, you need to *both* have a high precision and recall.

Share Improve this answer Follow edited Oct 15 '14 at 13:21

answered Oct 14 '14 at 12:09

Has QUIT--Anony-Mousse
**69.9k** 12 120 181

When the recall is 0.0 the precision has to be greater than 0.0 right? But I get the point in your example. Nicely explained - Thanks. – London guy Oct 14 '14 at 12:29

1 In your example, precision for class A is 0.5 instead of 0 and recall of class A is 1; precision for class B is 0 and recall of class B is 0 as we'll. I assume your balanced class means the true labels are A and B; each applies to 50% of data. – greeness Oct 15 '14 at 8:52

Let's make infinite elements of class B, and a single element of class A. It doesn't change the math behind F1. – Has QUIT--Anony-Mousse Oct 15 '14 at 13:20

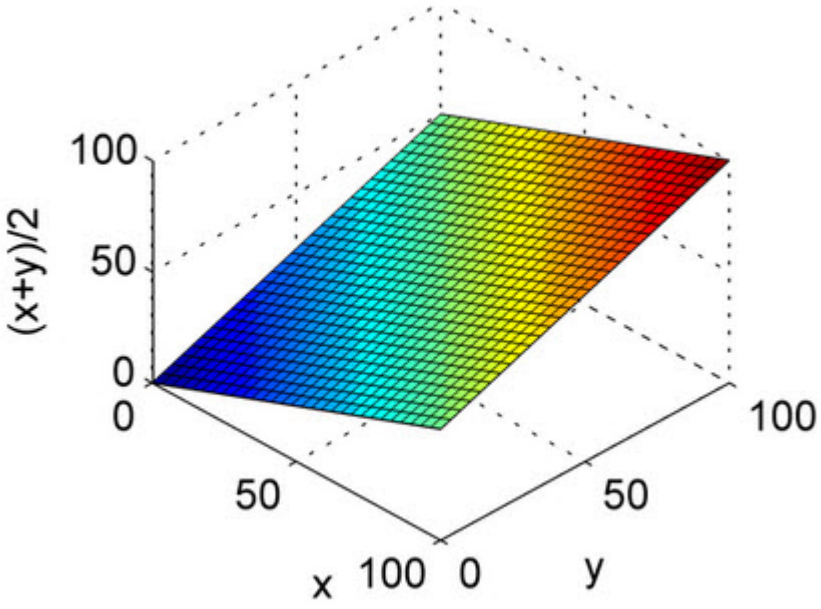
2 It is not just a heuristic to select more balance. Harmonic mean is there only way that makes sense given the units of these ratios. Mean wouldn't have a meaning in comparison – Sean Owen Mar 2 '16 at 9:02

Where does it say "heuristic", and where does your comment differ from my answer? But: F-measure *is* a heuristic in that it assumes precision and recall are equally important. That is why the beta term needs to be chosen - heuristically, one usually uses beta=1. – Has QUIT--Anony-Mousse Mar 2 '16 at 9:06

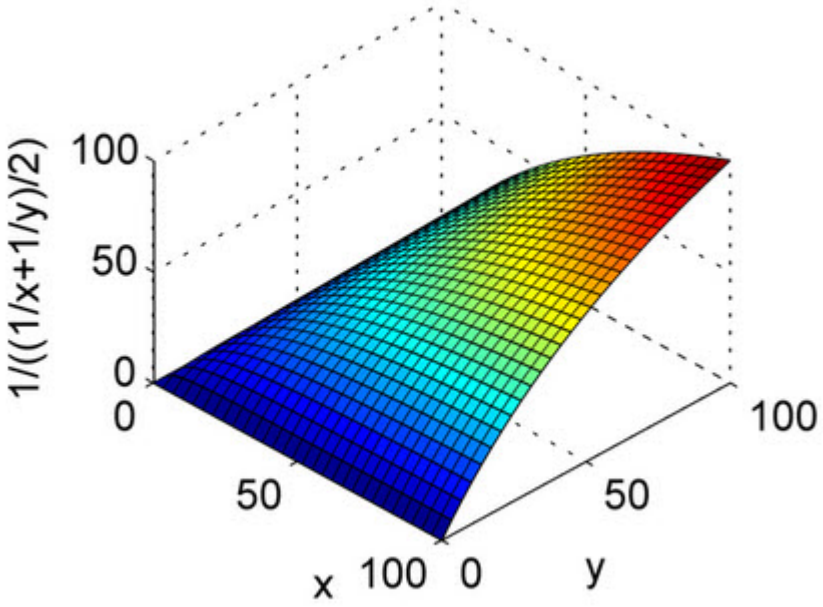
30

The above answers are well explained. This is just for a quick reference to understand the nature of the arithmetic mean and the harmonic mean with plots. As you can see from the plot, consider the X axis and Y axis as precision and recall, and the Z axis as the F1 Score. So, from the plot of the harmonic mean, both the precision and recall should contribute evenly for the F1 score to rise up unlike the Arithmetic mean.

This is for the arithmetic mean.




This is for the Harmonic mean.




Share Improve this answer Follow

edited Sep 17 '18 at 9:55

stason
**3,155** 2 21 36

answered Mar 28 '18 at 13:10

gadde saikumar
**420** 5 6

Please use formatting tools to properly edit and format your answer. Image should be displayed here , its not a hyperlink. – Morse Mar 28 '18 at 13:30

26

The harmonic mean is the equivalent of the arithmetic mean for reciprocals of quantities that should be averaged by the arithmetic mean. More precisely, with the harmonic mean, you transform all your numbers to the "averageable" form (by taking the reciprocal), you take their arithmetic mean and then transform the result back to the original representation (by taking the reciprocal again).

Precision and the recall are "naturally" reciprocals because their numerator is the same and their denominators are different. Fractions are more sensible to average by arithmetic mean when they have the same denominator.

For more intuition, suppose that we keep the number of true positive items constant. Then by taking the harmonic mean of the precision and the recall, you implicitly take the arithmetic mean of the false positives and the false negatives. It basically means that false positives and false negatives are equally important to you when the true positives stay the same. If an algorithm has N more false positive items but N less false negatives (while having the same true positives), the F-measure stays the same.

In other words, the F-measure is suitable when:

1. mistakes are equally bad, whether they are false positives or false negatives

2. the number of mistakes is measured relative to the number of true positives
3. true negatives are uninteresting

Point 1 may or may not be true, there are weighted variants of the F-measure that can be used if this assumption isn't true. Point 2 is quite natural since we can expect the results to scale if we just classify more and more points. The relative numbers should stay the same.

Point 3 is quite interesting. In many applications negatives are the natural default and it may even be hard or arbitrary to specify what really counts as a true negative. For example a fire alarm is having a true negative event every second, every nanosecond, every time a Planck time has passed etc. Even a piece of rock has these true negative fire-detection events all the time.

Or in a face detection case, most of the time you "*correctly don't return*" billions of possible areas in the image but this is not interesting. The interesting cases are when you *do* return a proposed detection or when you *should* return it.

By contrast the classification accuracy cares equally about true positives and true negatives and is more suitable if the total number of samples (classification events) is well-defined and rather small.



answered Mar 31 '15 at 12:34

isarandi
**2,468** 21 32