Macro VS Micro VS Weighted VS Samples F1 Score

Asked 3 years, 6 months ago Modified 1 year, 3 months ago Viewed 35k times

33

In sklearn.metrics.f1_score, the f1 score has a parameter called "average". What does macro, micro, weighted, and samples mean? Please elaborate, because in the documentation, it was not explained properly. Or simply answer the following:

1. Why is "samples" best parameter for multilabel classification?
2. Why is micro best for an imbalanced dataset?
3. what's the difference between weighted and macro?

python python-3.x machine-learning scikit-learn metrics Edit tags

Share Edit Follow Close Flag

edited Dec 16, 2019 at 11:57
sentence
7,497 4 31 37

asked Apr 18, 2019 at 6:26
Code Geek
675 2 6 15

I've tried, nothing comes out. – Code Geek Apr 18, 2019 at 6:37

1 Read the documentation of the sklearn.metrics.f1_score function properly and you will get your answer. – abunickabhi Apr 18, 2019 at 6:56

2 Sorry but I did. "because in the documentation, it was not explained properly" – Code Geek Apr 18, 2019 at 7:29

where did you see that "micro is best for imbalanced data" and "samples best for multilabel classification"? – Thibault D. Apr 18, 2019 at 9:02

Answers to your questions here: datascience.stackexchange.com/a/24051/17844 – Kirill Dolmatov Aug 4, 2021 at 14:49

2 Answers

Sorted by:
Reset to default

Date modified (newest first)

5

I found a really helpful article explaining the differences more thoroughly and with examples: https://towardsdatascience.com/multi-class-metrics-made-simple-part-ii-the-f1-score-ebe8b2c2ca1

Unfortunately, it doesn't tackle the 'samples' parameter and I did not experiment with multi-label classification yet, so I'm not able to answer question number 1. As for the others:

2. Where does this information come from? If I understood the differences correctly, micro is not the best indicator for an imbalanced dataset, but one of the worst since it does not

include the proportions. As described in the article, micro-f1 equals accuracy which is a flawed indicator for imbalanced data. For example: The classifier is supposed to identify cat pictures among thousands of random pictures, only 1% of the data set consists of cat pictures (imbalanced data set). Even if it does not identify a single cat picture, it has an accuracy / micro-f1-score of 99%, since 99% of the data was correctly identified as not cat pictures.

3. Trying to put it in a nutshell: Macro is simply the arithmetic mean of the individual scores, while weighted includes the individual sample sizes. I recommend the article for details, I can provide more examples if needed.

I know that the question is quite old, but I hope this helps someone. Please correct me if I'm wrong. I've done some research, but am not an expert.

Share Edit Follow Flag

answered Jul 13, 2021 at 18:54

3scapeX

51 1 1

"micro is not the best indicator for an imbalanced dataset", this is not always true. You can keep the negative labels out of micro-average. E.g. sklearn f1_score function provided labels/pos_label parameters to control this. In many NLP tasks, like NER, micro-average f1 is always the best metrics to use. – Bugface Mar 13 at 17:44

55

The question is about the meaning of the `average` parameter in `sklearn.metrics.f1_score`.

As you can see from the code:

- `average=micro` says the function to compute f1 by considering total true positives, false negatives and false positives (no matter of the prediction for each label in the dataset)
- `average=macro` says the function to compute f1 for each label, and returns the average without considering the proportion for each label in the dataset.
- `average=weighted` says the function to compute f1 for each label, and returns the average considering the proportion for each label in the dataset.
- `average=samples` says the function to compute f1 for each instance, and returns the average. Use it for multilabel classification.

Share Edit Follow Flag

answered Apr 19, 2019 at 8:43

sentence

7,497 4 31 37