

# Average precision与PR曲线线下面积

2019年1月20日 17:09

在使用sklearn做简单的多标签分类问题时，注意到sklearn.metrics中有计算average precision (AP) 的方法，而没有计算area under precision-recall curve (AUPRC) 的方法。计算AP的方法是sklearn.metrics.average\_precision\_score，官方文档中有这样的说明：

AP summarizes a precision-recall curve as the weighted mean of precisions achieved at each threshold, with the increase in recall from the previous threshold used as the weight:

$$AP = \sum_n (R_n - R_{n-1}) P_n$$

where  $P_n$  and  $R_n$  are the precision and recall at the  $n$ th threshold [1]. This implementation is not interpolated and is different from computing the area under the precision-recall curve with the trapezoidal rule, which uses linear interpolation and can be too optimistic.

可见AP相当于PR曲线的线下面积，只是区别于插值的方法。上面也提到插值的方式计算出的AUPRC too optimistic，因为常用的插值策略，在特定的recall值点 $r$ ，选择 $> r$ 的最大precision作为 $r$ 点的precision值。下文将首先介绍sklearn中AP的算法，再介绍常用的11点插值AUPRC。

## 一、AP

average\_precision\_score接收的两个重要的参数是y\_true和y\_score，y\_true是测试集的真实labels，y\_score是分类器预测出的概率值，有些分类器通过调用“decision\_function”返回y\_score。我在使用这个AP函数时，有两个疑问：

- (1) **y\_score用来做什么**？在计算AP时，我们已知它相当于PR曲线的线下面积，那么计算时需要一些 (precision, recall) 点，这个函数如何根据y\_true和y\_score计算 (precision, recall) 呢？
- (2) AP名字为average precision，名字中没有出现recall，也就是说计算结果相当于跟recall没关系，但它有时PR曲线的线下面积，**为什么此时AP=AUPRC**呢？

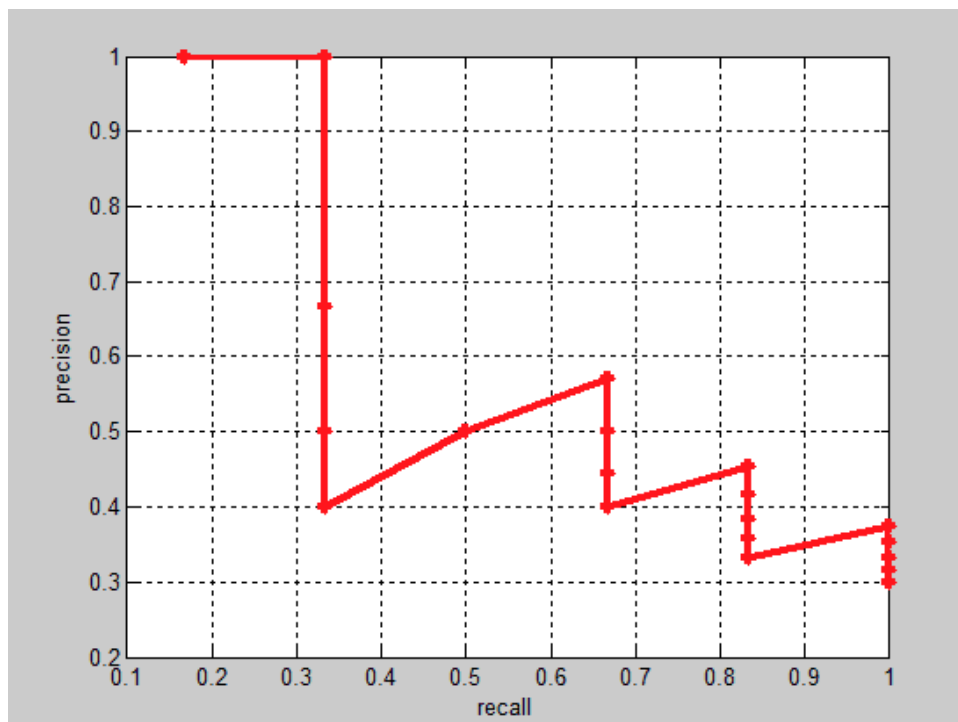
下面以具体的例子来说。

假设测试集中有20个样本，第二列是输入的y\_score，第三列为ground truth label，即y\_true。首先将数据根据y\_score降序排列，如下图所示。

id	score	gt_label
1	0.23	0
2	0.76	1
3	0.01	0
4	0.91	1
5	0.13	0
6	0.45	0
7	0.12	1
8	0.03	0
9	0.38	1
10	0.11	0
11	0.03	0
12	0.09	0
13	0.65	0
14	0.07	0
15	0.12	0
16	0.24	1
17	0.1	0
18	0.23	0
19	0.46	0
20	0.08	1

id	score	gt_label
4	0.91	1
2	0.76	1
13	0.65	0
19	0.46	0
6	0.45	0
9	0.38	1
16	0.24	1
1	0.23	0
18	0.23	0
5	0.13	0
7	0.12	1
15	0.12	0
10	0.11	0
17	0.1	0
12	0.09	0
20	0.08	1
14	0.07	0
8	0.03	0
11	0.03	0
3	0.01	0

首先，将score最高的作为正样本，预测其label=1，其他label=0，这样我们得到第一个 (precision, recall) 点，按照本例计算这个点为 (precision=1, recall=1/6)。然后将score排序第一和第二高的作为正样本，预测它们的label=1，其余label=0，这样又得到第二个 (precision, recall) 点。以此类推，直到计算出20个点对，根据这20个点画出PR曲线如下：



再来回顾AP的计算公式： $AP = \sum_n (R_n - R_{n-1}) P_n$ ，当recall值不变化时，对AP的计算没有贡献，上图中有贡献的点是recall发生变化的点，进一步观察可以发现，这些点分别为recall=1/6, 2/6, 3/6, 4/6, 5/6, 6/6。计算AP时加了五部分的矩形面积来近似线下面

积，五个  $(R_n - R_{n-1})$  的值都是1/6，6是测试数据中ground truth为正例的个数。所以实际上AP的计算公式  $AP = \frac{1}{M} \sum_n P_n$ ，M为测试集中label=1的样本个数。因此，AP命名为average precision。

## 二、11点插值AUPRC

理解了AP之后，我们就能很容易理解插值AUPRC。11点插值选取recall=[0, 0.1, 0.2, ..., 1]这11个点，如何选取11点对应的precision值呢？以r=0.1为例，它对应的p值，是  $r > 0.1$  时最大的precision值。因为对于每个点都选用了之后最大的p值，所以说这种方式是optimistic的。

## 三、另一种方法

还有一种计算AP的方法，结合了不插值和选最大的策略。仍以前面的数据为例，计算AP时recall还是原来的那几个点，不过选择precision时，是选之后最大的precision。如下图所示：

top-N	Precision	Recall(r)	Max Precision for Any Recall $r' \geq r$	Average Precision
1	1/1	1/6	1	0.6621
2	2/2	2/6	1	
3	2/3			
4	2/4			
5	2/5			
6	3/6	3/6	4/7	
7	4/7	4/6	4/7	
8	4/8			
9	4/9			
10	4/10			
11	5/11	5/6	5/11	
12	5/12			
13	5/13			
14	5/14			
15	5/15	6/6	6/16	
16	6/16			
17	6/17			
18	6/18			
19	6/19			
20	6/20			

## 四、PR曲线含义

上面提到的三种方法，计算出的结果都可以称为AUPRC。PR曲线能够直观的展示模型的效果，而AUPRC通过计算PR曲线的线下面积值，量化模型的效果。这个值介于0-1之间，越大越好。一个模型较差时，当recall不断增大，它会牺牲precision，因为它增大recall的方式是将较多0预测为1；反之，模型较好时，recall增大的同时，precision也会相对稳定，因此线下面积会较大。

附注：

本文图片摘自博客<https://www.jianshu.com/p/63fc7870d0cb>，这篇博客也对AP有比较好的讲解，推荐阅读。