机器学习 (/tags/#机器学习)

机器学习模型评估指标

Posted by 永超 on December 23, 2016

阅读：105次

机器学习(ML),自然语言处理(NLP),信息检索(IR)等领域,评估(Evaluation)是一个必要的工作,而其评价指标往往有如下几点:准确率(Accuracy),精确率(Precision),召回率(Recall)和F1-Measure。

本文将简单介绍其中几个概念。中文中这几个评价指标翻译各有不同，所以一般情况下推荐使用英文。

现在我先假定一个具体场景作为例子。

> *假如某个班级有男生80人,女生20人,共计100人.目标是找出所有女生. 现在某人挑选出50个人,其中20人是女生,另外还错误的把30个男生也当作女生挑选出来了. 作为评估者的你需要来评估(evaluation)下他的工作。*

首先我们可以计算准确率(accuracy),其定义是: 对于给定的测试数据集，分类器正确分类的样本数与总样本数之比。也就是损失函数是0-1损失时测试数据集上的准确率[1](https://argcv.com/articles/1036.c#ref_1)。

这样说听起来有点抽象，简单说就是，前面的场景中，实际情况是那个班级有男的和女的两类，某人(也就是定义中所说的分类器)他又把班级中的人分为男女两类。accuracy需要得到的是此君分正确的人占总人数的比例。很容易，我们可以得到:他把其中70(20女+50男)人判定正确了,而总人数是100人，所以它的accuracy就是70% (70 / 100)。

由准确率，我们的确可以在一些场合，从某种意义上得到一个分类器是否有效，但它并不总是能有效的评价一个分类器的工作。举个例子,google抓取了argcv 100个页面，而它索引中共有10,000,000个页面,随机抽一个页面，分类下,这是不是argcv的页面呢?如果以accuracy来判断我的工作，那我会把所有的页面都判断为”不是argcv的页面”,因为我这样效率非常高(return false,一句话),而accuracy已经到了99.999%(9,999,900/10,000,000),完爆其它很多分类器辛辛苦苦算的值,而我这个算法显然不是需求期待的,那怎么解决呢?这就是precision,recall和f1-measure出场的时间了.

在说precision,recall和f1-measure之前,我们需要先需要定义TP,FN,FP,TN四种分类情况. 按照前面例子,我们需要从一个班级中的人中寻找所有女生,如果把这个任务当成一个分类器的话,那么女生就是我们需要的,而男生不是,所以我们称女生为”正类”,而男生为”负类”.

	相关(Relevant),正类	**无关(NonRelevant),负类**
被检索到(Retrieved)	true positives(TP 正类判定为正类,例子中就是正确的判定”这位是女生”)	false positives(FP 负类判定为正类,”存伪”,例子中就是分明是男生却判断为女生)
未被检索到(Not Retrieved)	false negatives(FN 正类判定为负类,”去真”,例子中就是,分明是女生,却判断为男生)	true negatives(TN 负类判定为负类,也就是一个男生被判断为男生)

通过这张表,我们可以很容易得到这几个值: TP=20 FP=30 FN=0 TN=50

精确率(precision)的公式是

$$P = \frac{TP}{TP+FP}$$

它计算的是所有”正确被检索的item(TP)”占所有”实际被检索到的(TP+FP)”的比例.

在例子中就是希望知道此君得到的所有人中,正确的人(也就是女生)占有的比例.所以其precision也就是40%(20女生/(20女生+30误判为女生的男生)).

召回率(recall)的公式是

$$R = \frac{TP}{TP+FN}$$

它计算的是所有”正确被检索的item(TP)”占所有”应该检索到的item(TP+FN)”的比例。

在例子中就是希望知道此君得到的女生占本班中所有女生的比例,所以其recall也就是100%(20女生/(20女生+ 0 误判为男生的女生))

F1值就是精确值和召回率的调和均值,也就是

$$\frac{2}{F_1} = \frac{1}{P} + \frac{1}{R}$$

调整下也就是

$$F_1 = \frac{2PR}{2TP + FP + FN}$$

例子中 F1-measure 也就是约为

$$57.143\%(\frac{2*0.4*1}{0.4+1})$$

需要说明的是,有人[2] (https://argcv.com/articles/1036.c#ref_2)列了这样个公式

$$F_a = \frac{(a^2+1)PR}{a^2(P+R)}$$

将F-measure一般化.

F1-measure认为精确率和召回率的权重是一样的,但有些场景下,我们可能认为精确率会更加重要,调整参数a,使用Fa-measure可以帮助我们更好的evaluate结果.

PREVIOUS
机器学习模型错误的4个原因（以及如何修复它）
(/2016/12/22/ML_MODEL1/)

NEXT
在IOS上使用TALKINGDATA实现远程推送消息
(/2017/04/24/IOS_PUSH/)

FEATURED TAGS (/tags/)

调研 (/tags/#调研) iOS (/tags/#iOS) 机器学习 (/tags/#机器学习) 技术 (/tags/#技术)

FRIENDS


(https://zhuanlan.zhihu.com/talkingdata)


(http://weibo.com/TalkingData)


(https://github.com/TalkingData)

Theme by voyagelab (http://leopan.cn/) | Star 5