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。(注： 相对来说，IR 的 ground truth 很多时候是一个 Ordered List, 而不是一个 Bool 类型的 Unordered Collection，在都找到的情况下，排在第三名还是第四名损失并不是很大，而排在第一名和第一百名，虽然都是“找到了”，但是意义是不一样的，因此 更多可能适用于 MAP 之类评估指标。)

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].

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(NonRel**
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 positive 例子中就是分 娘横行,这个错
未被检索到(Not Retrieved)	false negatives(FN 正类判定为负类,"去真",	true negative

	例子中就是,分明是女生,这哥们却判断为男生--梁山伯同学犯的错就是这个)	一个男生被判准儿就会在此

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

,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

,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

调整下也就是

例子中 F1-measure 也就是约为 57.143%(

).

需要说明的是,有人[2]列了这样个公式

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.

话虽然很多,其实实现非常轻松,点击此处可以看到我的一个简单的实现.

References

[1] 李航. 统计学习方法[M]. 北京:清华大学出版社,2012.

[2] 准确率（Precision）、召回率（Recall）以及综合评价指标（F1-Measure ）

假设一个班级有100个学生，其中男生70人，女生30人。如下图，蓝色矩形表示男生，橙色矩形表示女生。

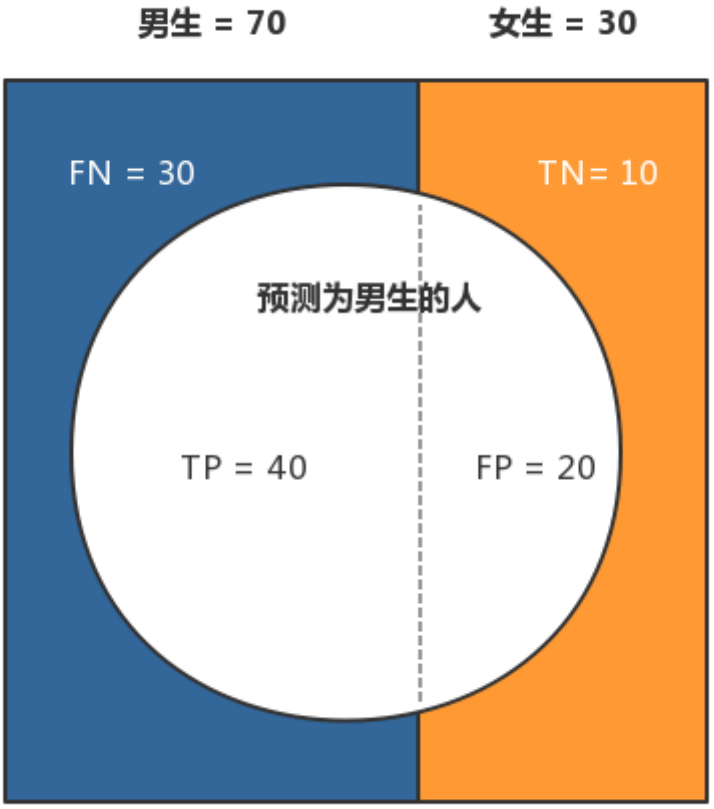
又假设，我们不知道这些学生的性别，只知道他们的身高和体重。我们有一个程序(分类器)，这个程序可以通过分析每个学生的身高和体重，对这100个学生的性别分别进行预测。最后的预测结果为，60人为男生，40人为女生，如下图。

TP：实际为男生，预测为男生；

FP：实际为女生，预测为男生；

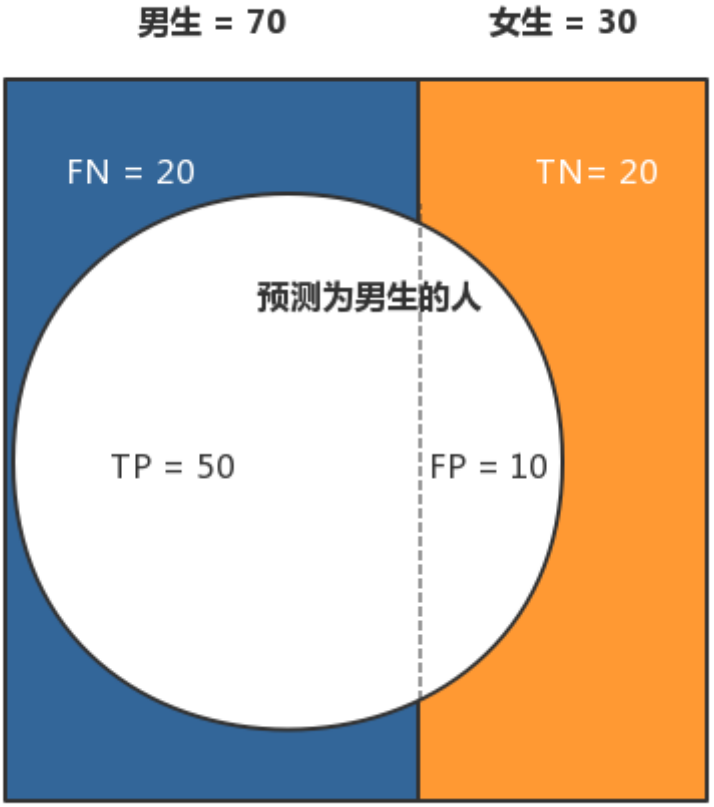
FN：实际为男生，预测为女生；

TN：实际为女生，预测为女生；



- 准确率(Accuracy) = (TP + TN) / 总样本 = (40 + 10)/100 = 50%。 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。
- 精确率(Precision) = TP / (TP + FP) = 40/60 = 66.67%。它表示：预测为正的样本中有多少是真正的正样本，它是针对我们预测结果而言的。Precision又称为查准率。
- 召回率(Recall) = TP / (TP + FN) = 40/70 = 57.14% 。它表示：样本中的正例有多少被预测正确了， 它是针对我们原来的样本而言的。Recall又称为查全率。

可以看到，上面的预测结果并不是很好。假设我们优化了程序后，再次进行预测。预测结果为：



- 准确率(Accuracy) = (TP + TN) / 总样本 = (50 + 20)/100 = 70%
- 精确率(Precision) = TP / (TP + FP) = 50/60 = 83%
- 召回率(Recall) = TP / (TP + FN) = 50/70 = 71.43%

各项指标都比第一次高，说明预测效果更好。从图上也能看出来，预测为男生的范围与实际男生范围更接近。

==========

自己理解 + 我老师的说法就是，准确率就是**找得对**，召回率就是**找得全**。

大概就是你问问一个模型，这堆东西是不是某个类的时候，准确率就是 **它说是，这东西就确实是**的概率吧，召回率就是， **它说是，但它漏说了（1-召回率）这么多。**

==========

在信息检索、分类体系中，有一系列的指标，搞清楚这些指标对于评价检索和分类性能非常重要，因此最近根据网友的博客做了一个汇总。

准确率、召回率、F1

信息检索、分类、识别、翻译等领域两个最基本指标是**召回率(Recall Rate)**和**准确率(Precision Rate)**，召回率也叫查全率，准确率也叫查准率，概念公式:

召回率(**R**ecall) = 系统检索到的相关文件 / 系统所有相关的文件总数

准确率(**P**recision) = 系统检索到的相关文件 / 系统所有检索到的文件总数

图示表示如下：

A： （搜到的也想要的）

B：检索到的，但是不相关的 （搜到的但没用的）

C：未检索到的，但却是相关的 （没搜到，然而实际上想要的）

D：未检索到的，也不相关的 （没搜到也没用的）

注意：准确率和召回率是互相影响的，理想情况下肯定是做到两者都高，但是一般情况下准确率高、召回率就低，召回率低、准确率高，当然如果两者都低，那是什么地方出问题了。

一般情况，用不同的阀值，统计出一组不同阀值下的精确率和召回率，如下图：

如果是做搜索，那就是保证召回的情况下提升准确率；如果做疾病监测、反垃圾，则是保准确率的条件下，提升召回。

所以，在两者都要求高的情况下，可以用F1来衡量。

F1 = * P * R / (P + R)

公式基本上就是这样，但是如何算图1中的A、B、C、D呢？**这需要人工标注，人工标注数据需要较多时间且枯燥，如果仅仅是做实验可以用用现成的语料。当然，还有一个办法，找个一个比较成熟的算法作为基准，用该算法的结果作为样本来进行比照**，这个方法也有点问题，如果有现成的很好的算法，就不用再研究了。

AP和mAP(mean Average Precision)

mAP是为解决P，R，F-measure的单点值局限性的。为了得到 一个能够反映全局性能的指标，可以看考察下图，其中两条曲线(方块点与圆点)分布对应了两个检索系统的准确率-召回率曲线

可以看出，虽然两个系统的性能曲线有所交叠但是以圆点标示的系统的性能在绝大多数情况下要远好于用方块标示的系统。

从中我们可以 发现一点，如果一个系统的性能较好，其曲线应当尽可能的向上突出。

更加具体的，曲线与坐标轴之间的面积应当越大。
最理想的系统， 其包含的面积应当是1，而所有系统的包含的面积都应当大于0。这就是用以评价信息检索系统的最常用性能指标，平均准确率mAP其规范的定义如下:(其中P，R分别为准确率与召回率)

ROC和AUC

ROC和AUC是评价分类器的指标，上面第一个图的ABCD仍然使用，只是需要稍微变换。

回到ROC上来，ROC的全名叫做Receiver Operating Characteristic。
ROC关注两个指标

True Positive Rate (TPR) = TP / [TP + FN] ， TPR代表能将正例分对的概率

False Positive Rate(FPR) = FP / [FP + TN] ， FPR代表将负例错分为正例的概率

在ROC 空间中，每个点的横坐标是FPR，纵坐标是TPR，这也就描绘了分类器在TP（真正的正例）和FP（错误的正例）间的trade-off。ROC的主要分 析工具是一个画在ROC空间的曲线——ROC curve。我们知道，对于二值分类问题，实例的值往往是连续值，我们通过设定一个阈值，将实例分类到正类或者负类（比如大于阈值划分为正类）。因此我们 可以变化阈值，根据不同的阈值进行分类，根据分类结果计算得到ROC空间中相应的点，连接这些点就形成ROC curve。ROC curve经过（0,0） （1,1），实际上(0, 0)和(1, 1)连线形成的ROC curve实际上代表的是一个随机分类器。一般情况下，这个曲线都应该处于(0, 0)和(1, 1)连线的上方。如图所示。

用ROC curve来表示分类器的performance很直观好用。可是，人们总是希望能有一个数值来标志分类器的好坏。
于是**Area Under roc Curve(AUC)**就出现了。顾名思义，AUC的值就是处于ROC curve下方的那部分面积的大小。通常，AUC的值介于0.5到1.0之间，较大的AUC代表了较好的Performance。
AUC计算工具:
http://mark.goadrich.com/programs/AUC/
P/R和ROC是两个不同的评价指标和计算方式，一般情况下，检索用前者，分类、识别等用后者。
参考链接:

http://www.vanjor.org/blog/2010/11/recall-precision/

http://bubblexc.com/y2011/148/

http://wenku.baidu.com/view/ef91f011cc7931b765ce15ec.html

：Recall，又称“查全率”——还是查全率好记，也更能体现其实质意义。

准确率

“召回率”与“准确率”虽然没有必然的关系（从上面公式中可以看到），在实际应用中，是相互制约的。要根据实际需求，找到一个平衡点。

当 我们问检索系统某一件事的所有细节时（输入检索query查询词），Recall指：检索系统能“回忆”起那些事的多少细节，通俗来讲就是“回忆的能 力”。“能回忆起来的细节数” 除以 “系统知道这件事的所有细节”，就是“记忆率”，也就是recall——召回率。简单的，也可以理解为查全率。

在人工智能中，混淆矩阵（confusion matrix）是可视化工具，特别用于监督学习，在无监督学习一般叫做匹配矩阵。

如有150个样本数据，这些数据分成3类，每类50个。分类结束后得到的混淆矩阵为：

		预测		
		类1	类2	类3
实际	类1	43	5	2
	类2	2	45	3
	类3	0	1	49

每一行之和为50，表示50个样本，

第一行说明类1的50个样本有43个分类正确，5个错分为类2，2个错分为类3

Precision表示被分为正例的示例中实际为正例的比例，precision=TP/（TP+FP）。即，一个二分类，类别分别命名为1和2，Precision就表示在类别1中，分对了的数量占了类别1总数量的多少；同理，也表示在类别2中，分对了的数量占类别2总数量的多少。那么这个指标越高，就表示越整齐不混乱。

而Accuracy是我们最常见的评价指标，accuracy = （TP+TN）/(P+N)，这个很容易理解，就是被分对的样本数除以所有的样本数，通常来说，正确率越高，分类器越好。我们最常说的

就是这个准确率