

第08课：模型的质量和评判指标

### 衡量模型质量

通过训练得到模型后，我们就可以用这个模型，来进行预测了（也就是把数据输入到模型中让模型吐出一个结果）。

预测肯定能出结果，至于这个预测结果是否是你想要的，就不一定了。

一般来说，没有任何模型能百分百保证尽如人意，但我们总是追求尽量好。

什么样的模型算好呢？当然需要****测试****。

当我们训练出了一个模型以后，为了确定它的质量，我们可以用一些知道预期预测结果的数据来对其进行预测，把实际的预测结果和实际结果进行对比，以此来评判模型的优劣。

由此，我们需要一些评价指标来衡量实际预测结果和预期结果的相似程度。

### 分类模型评判指标： Precision、Recall 和 F1Score

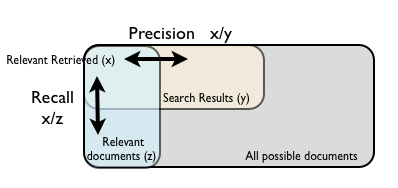
对于分类而言，最简单也是最常见的验证指标：****精准率（Precision）****和****召回率（Recall）****，为了综合这两个指标并得出量化结果，又发明了 ****F1Score****。

对一个分类模型而言，给它一个输入，它就会输出一个标签，这个标签就是它预测的当前输入的类别。

假设数据 data1 被模型预测的类别是 Class\_A。那么，对于 data1 就有两种可能性：data1 本来就是 Class\_A（预测正确），data1 本来不是 Class\_A（预测错误）。

当一个测试集全部被预测完之后，相对于 Class\_A，会有一些实际是 Class\_A的数据被预测为其他类，也会有一些其实不是 Class\_A 的，被预测成Class\_A，这样的话就导致了下面这个结果：





****精准率：Precision=TP/（TP+FP）****，即在所有被预测为 Class\_A 的测试数据中，预测正确的比率。

****召回率：Recall=TP/（TP+FN）****，即在所有实际为 Class\_A 的测试数据中，预测正确的比率。

****F1Score = 2\*(Precision \* Recall)/(Precision + Recall)****

显然上面三个值都是越大越好，但往往在实际当中 P 和 R 是矛盾的，很难保证双高。

此处需要注意，P、R、F1Score 在分类问题中都是对某一个类而言的。

也就是说假设这个模型总共可以分10个类，那么对于每一个类都有一套独立的 P、R、F1Score 的值。衡量模型整体质量，要综合看所有10套指标，而不是只看一套。

同时，这套指标还和测试数据有关。同样的模型，换一套测试数据后，很可能 P、R、F1Score 会有变化，如果这种变化超过了一定幅度，就要考虑是否存在 bias 或者 overfitting 的情况。

NOTE：这几个指标也可以用于 seq2seq 识别模型的评价。

seq2seq 识别实际上可以看作是一种位置相关的分类。每一种实体类型都可以被看作一个类别，因此也就同样适用 P、R、F1Score 指标。

### 指标对应的是模型&数据集

上面我们讲了 P、R 和 F1Score 这一套指标，无论是这套，还是 ROC、PR 或者 AUC 等（这些大家可以自行查询参考），或者是任意的评价指标，都****同时指向一个模型和一个数据集****，两者缺一不可。

同样一套指标，用来衡量同一个模型在不同数据集上的预测成果，最后的分数值可能不同（几乎可以肯定不同，关键是差别大小）。

上面我们一直以测试集为例。其实，在一个模型被训练结束后，它可以先用来预测一遍训练集中所有的样本。

比如，我们训练了一个 Logistic Regression，用来做分类。

一次训练过程完成后，我们可以先用当前结果在训练集合上预测一遍，算出训练集的 P、R 和 F1；再在验证集上跑一下，看看验证集的 P、R 和 F1。

几轮训练后，再在测试集上跑，得出测试集的相应指标。

### 模型的偏差和过拟合

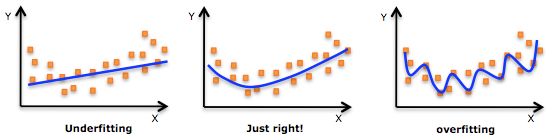
为什么我们明明是用训练集合训练出来的模型，还要再在训练集上跑预测呢？

首先，我们要知道一点，一个模型用来预测训练数据，并不能保证每一个预测结果都和预期结果相符（为什么这样，当我们讲到模型时自然就会清楚）。

一个机器学习模型的质量问题，从对训练集样本拟合程度的角度，可以分为两类：****欠拟合（Underfitting）****和****过拟合 （Overfitting）****。

如何严格定义欠拟合还是过拟合，还要涉及到几个概念：****bias****、****error**** 和****variance****。这些我们在具体讲到模型时再说。

这里先建立一点感性认识：



如果一个模型，在训练集上的预测结果就不佳，指标偏低，那一般是欠拟合的问题。

如果在训练集上指标很好，而在验证/测试集上指标偏低，则很可能是过拟合问题。

对于两种不同的问题，解决方法各不相同。

欠拟合多数情况下是因为选定模型类型太过简单，特征选取不够导致的。而过拟合则相反，可能是模型太过复杂，特征选择不当（过多或组合不当）造成的。

相应的解法，当然是有针对性地选择更复杂/简单的模型类型；增加/减少特征；或者减小/增大正则项比重等。

但有一点，无论哪种问题，增大训练数据量都可能会有所帮助。

[邀请好友一起学，获得 25% 返现奖励](https://gitbook.cn/m/mazi/columns/5bc6ac7442d7d32f50f19a98/topics/5bf2812dfd72950cafdccd82)

[IMG_260](https://gitbook.cn/m/mazi/comp/column?columnId=5bc6ac7442d7d32f50f19a98%26tag=2#catalog)