机器学习的性能评估-查准率和查全率_吴恩达_机器学习笔记

1、确定算法优化的步骤优先级

本节主要通过垃圾邮件分类的案例来说明怎么确定一个机器学习的项目优先级：

垃圾邮件分类主要通过文本内容正文、标题、发送人邮箱地址等具体的特征来构建一个10000-50000数量级的特征，然后收集很多垃圾邮件来进行正负样本分类训练，当文本出现打折、促销等关键单词时，这个特征值就等于1，否则等于0。

那么到底怎么来确定选择哪些特征来改进一个机器学习项目呢？到底是增加特征，还是增加数据样本量，还是其他方法？那么误差评估将会是一个很直接的办法。

2、误差分析

当面临一个机器学习项目时，吴恩达建议的步骤是如下：

1、先根据业务特征构建一个机器学习模型，直接粗暴的通过划分训练集、交叉验证集、测试集，并通过交叉验证评估算法的性能值。

2、画出该算法的学习曲线，判断数据到底是需要更多特征、更多数据样本、还是其他方向

3、进行模型的误差分析，这就是本节要讲解的内容。通过观察被算法误分类的样本，总结被误分类的样本的一些共性特征，将非常有助于我们优化算法。

- 例如，做垃圾邮件分类时，误分类的数据中，有100封邮件被误分类为非垃圾邮件，那么观察这些邮件，发现制药类垃圾邮件5封，商品打折广告90封，盗号邮件5封，那么就可以重点总结打折广告的邮件为何总是被误分类。
- 第二个方向是，观察垃圾邮件的特征，比如是否都有特殊的标点符号50封，特殊的邮件标题3封，导致被误分类，然后根据特点来基于特殊标点符合构建新的特征。

上面介绍的是基于业务特征来进行误差分析提升算法性能的一个很好的方法。

另外一个方向就是，对算法的性能提升需要通过数值评估误差。比如，在一个垃圾邮件分类中，文本分析中，对不同单词到底要不要使用不同单词的前几个字母来判断是否归类为同一类单词意思，就需要将该类特征直接进行误差数值评估，以此来看是否有效。比如discount和discounting，universe和university。有的是一个意思，有的意思千差万别。到底这个特征是否有效，直接用数值评估更加直观。

例如，构建一个分类器的时候，选择是否将词干纳入一个特征作为两个不同的算法，纳入词干特征的算法的错误率是3%，没有纳入词干特征的算法错误率是5%。那么这个特征就是有效的。

特别需要注意的是，强烈建议在交叉验证集上评估误差，而不是测试集上评估。

一句话总结，粗暴的上一个算法——评估误差——寻找问题方向——优化算法，这样更有效率

3、样本不均衡问题的误差评估

以是否患癌症为例，使用逻辑回归算法，假设数据中只有2个负样本和198个正样本，那么使用错误率这种评估方法的话，假设交叉验证集算法的评估正确率是99%。错误率是1%。

现在我们观察，测试集中真正患癌症的人比例是0.05%。这样一来，算法的正确率就不太好了，因为仅仅两个人患癌的情况下，都被误判了一个，可以评估实际上从患癌角度来说，算法的错误率是50%。

假设我们设计另外一个不是真正意义算法的算法：

```
#不考虑x,总是将样本预测为0（未患癌症）
```

那么这个算法最终的错误率也是0.05%，因为只评估错误了2个。剩下的198个都被分类正确。实际上，这比我们交叉验证集的错误率1%还低。

这样一个简单的例子说明，只使用简单的错误率，在样本不均衡的情况下，基本上是废材。

假设使用准确率或者错误率这样单调的评估算法的话，那么就算将一个算法从99.2%提升到了99.8%，我们并不清楚这个算法是真正的提升了，还是只是用了一个类似不是算法的算法，总是将样本判断为未患癌症而已。

下面我们来复习一下查准率和召回率。

对于二分类问题，可将样例根据其真实类别与学习器预测类别的组合划分为真正例(true positive)、假正例(false positive)、真反例(true negative)、假反例(false negative)四种情形，令 TP、FP、TN、FN 分别表示其对应的样例数，则显然有 $TP+FP+TN+FN$ =样例总数. 分类结果的“混淆矩阵”(confusion matrix)如表 2.1 所示.

表 2.1 分类结果混淆矩阵

真实情况	预测结果	
	正例	反例
正例	TP (真正例)	FN (假反例)
反例	FP (假正例)	TN (真反例)

查准率 P 与查全率 R 分别定义为

$$P=\frac{TP}{TP+FP}, \tag{2.8}$$

$$R=\frac{TP}{TP+FN}. \tag{2.9}$$

查准率和查全率是一对矛盾的度量. 一般来说，查准率高时，查全率往往偏低；而查全率高时，查准率往往偏低. 例如，若希望将好瓜尽可能多地选出来，则可通过增加选瓜的数量来实现，如果将所有西瓜都选上，那么所有的好瓜也

查准率：对于我们所有预测为患有癌症的病人当中，有多大比率的病人是真正患有癌症的。

查全率：在真正患有癌症的病人当中，有多大比率我们正确的预测出了他们患有癌症。

样本真实标签，1为患癌，0为未患癌

样本预测标签	1	0
1	**True positive** **TP** **真阳性** **预测为真，实际为真**	**False positive** **FP** **假阳性** **预测为真，实际为假**
0	False negative FN 假阴性 预测为假，实际为真	True negative TN 真阴性 预测为假，实际为假

预测为真实际为真/所有预测为真的总数=查准率

TP/(TP+FP)

预测为真实际为真/实际为真的总数=召回率

TP/(TP+FN)

4、查准度和查全率的权衡

举例使用逻辑回归预测病人是否患癌来说明查准率和召回率的意义。

- y=1 预测患癌
- y=0 预测健康

- 假设为了保证我们的对癌症病人的预测非常的精确。那么我们需要将分类阀值概率从0.5提升到0.8。

这时候，我们预测为癌实际为癌的人，因为阀值的提升会减少，那么整个预测为癌症的人就会减小，这时候查准率的分母整体变小。因此查准率提升。

真正患有癌症的人，有更多的概率被预测患有癌症。

- 假设为了保证尽可能多的样本能够做进一步的癌症前期排查治疗，降低遗漏的未被找到的真正患癌的样本数。那么我们就会将分类阀值概率从0.5降低到0.3。

这时候，我们预测为癌，实际不为癌的人，因为阀值的降低会增多，预测不为癌，实际患癌的会降低。那么查全率的分母整体变小，因此查全率提升。

第一，更多的真正患癌的病人被找到，就是TP会增大，同时，预测为患癌但是实际不患癌的人也会增大。低查准率。

F1值计算公式：

$$F1 = \frac{2 * P * R}{P + R}$$

这个公司的意义是，会考虑查准率和查全率平均值，但是同时会给查准率和查全率中更低的一个值更高的权重。

因此，假设P或者R中任何一个值为0，那么整个F值的也会为0。而不是像平均值那样选择。另外，假设P和R都非常好，都是1，那么F值的结果也会是1。

假设希望得到更好的F值，那么可以多次测试不同的阀值所带来的交叉验证集上的F值的变化情况，取最大的F值所对应的阀值结果，是一个好办法。

5、更大量的数据在什么时候对机器学习有强化作用

本节讲解的问题是，如何在拥有大数据集的情况下，保证数据量大的情况下，能够对算法的优化有很大的帮助？

1、选择一个带有很多参数的，能够拟合近似所有函数的强大算法，这样可以保证我们不会欠拟合，那么整个数据集至少已经保证低偏差。

2、选择的特征与预测结果息息相关，也就是说，完全契合业务真实环境的逻辑。

3、训练集的数量远远大于参数的数量，就不太可能出现过拟合情况，因为参数远远小于数据集的情况下，数据集的所体现的特征越复杂，越不容易很完美的拟合数据。除非参数不断增大。这时候，训练误差约等于交叉验证误差。同时训练误差很小，那么交叉验证误差也会很小。

实际上这时候，几乎可以说明，测试误差所代表的近似泛化误差也会很小。

4、这时候，大量的数据实际上是非常有助于算法提升的。

总结如下：

- 大量样本，保证数据不会出现过拟合问题，排除高方差问题
- 高级参数的算法，保证数据不会出现欠拟合情况，排除高偏差问题

低方差+低偏差= 好的模型