机器学习 (/tags/#机器学习)

机器学习模型错误的4个原因（以及如何修复它）

Posted by 永超 on December 22, 2016

阅读：145次

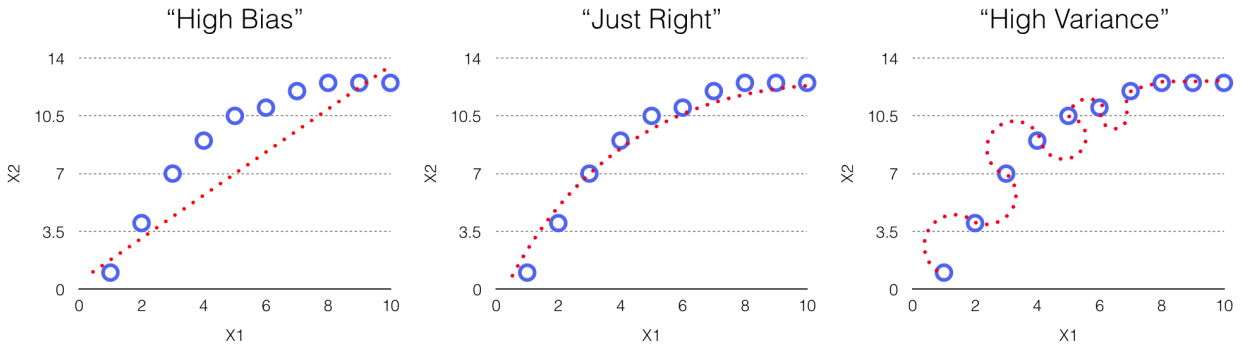
在机器学习中，通常会有很多的模型可以选择。 通常情况下，会使用**线性回归（Linear Regression）**来预测值，**Logistic回归（Logistic Regression）**来归类，以及使用**神经网络（Neural Networks）**来对非线性行为建模。

当构建了模型的时候，通常会使用一些历史数据来帮助机器学习算法学习一组输入特征与预测输出之间的关系。但即使这个模型能够准确的预测历史数据，但是如何知道该模型对新数据有效呢？

更加直白的说，如何评估机器学习模型是否真正的**“好”**呢？

这篇文章中，会介绍一些常见的场景，其中看似良好的机器学习模型可能仍然是错误的。还会展示如何通过评估偏差与方差以及精确度、召回的指标来评估一个模型，并提供改进此类问题的一些解决方案。

高偏差(Bias)或高方差(Variance)



在评估一个模型的时候，首先要评估的是模型是否具有“高偏差”或者“高方差”。

高偏差(High Bias)：指模型“不适合”样本数据（见上图“High Bias”）的情况。这是一种比较坏的情况，因为模型并没有呈现出一个非常准确或者具有代表性的输入与输出之间的关系，并且经常输出很高的误差（例如：模型的预测值和实际值之间的差异）。

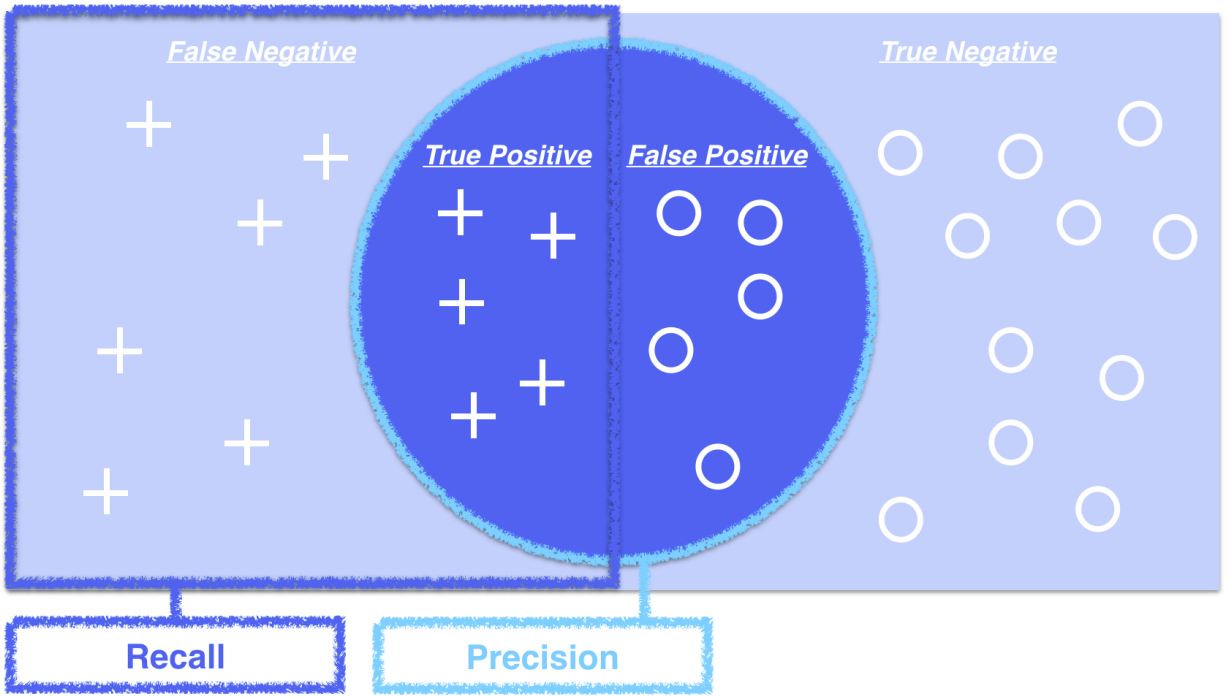
高方差(High Variance)：指相反的情况（见上图“High Variance”）。在高方差或者“高度拟合”的情况下，模型表现的非常准确，以至于完全适合样本数据集。虽然看起来这是一种好的结果，但这也正是担忧的原因，因为这样而模型通常无法推广扩展到未来的数据集中。因此，虽然模型适用于现有的数据，但不知道对于其他样例数据的结果。

但是如何知道模型是否具有“高偏差”或者“高方差”呢？

一种直截了当的方式是对数据进行**训练-测试分割**。举例说明，使用数据集的70%作为训练模型的数据，接下来使用剩下的30%来测量其错误率。如果模型在训练和测试数据集中具有很高的错误，说明模型不适合两个数据集，并且具有“高偏差”（见上图“High Bias”）。如果模型在训练数据集上有很低的错误，但是在测试数据集上有很高的错误，这表示模型具有“高方差”（见上图“High Variance”），因为模型并没有扩展到第二个数据集，也就是测试数据集上。

如果模型在训练数据集和测试数据集上都表现出低误差，那么这个模型就是一个相对“正确”的模型，并平衡了正确的偏差和方差水平（见上图“Just Right”）。

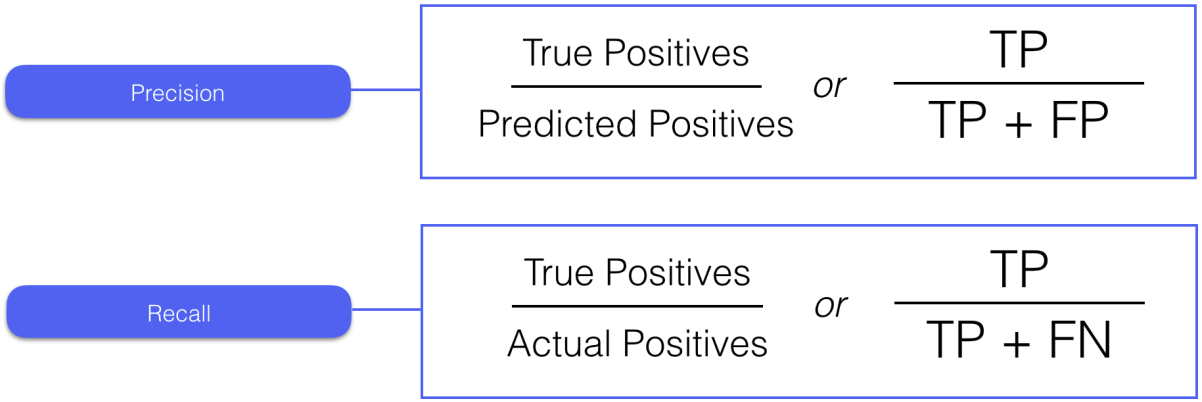
低精度(Precision)或低召回(Recall)



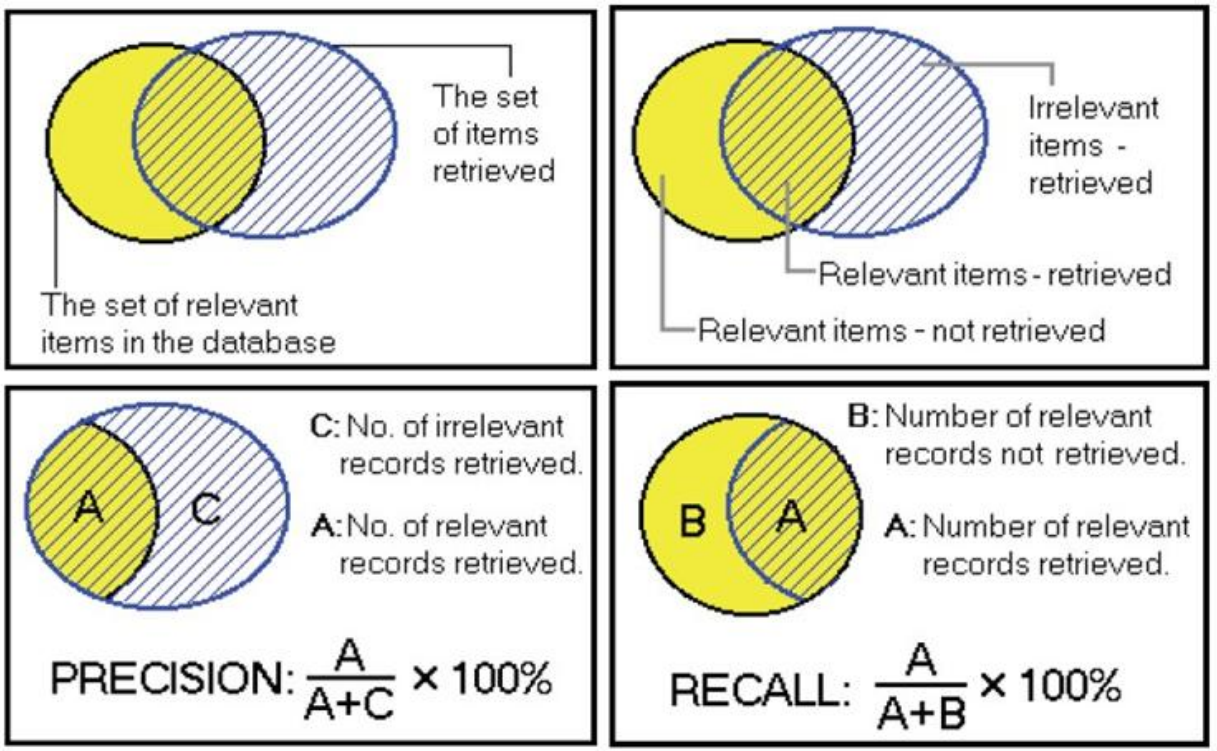
即使机器学习模型具有很高的正确率，模型也可能有其他类型的错误。

例如将邮件分为垃圾邮件（正类）或非垃圾邮件（负类）。99%的情况下，收到的邮件不是垃圾邮件，也就是说大多数情况下都是负类，但是还有1%的情况是垃圾邮件，也就是正类。如果训练得到了一个模型，并且已经学会预测一个邮件是非垃圾邮件，那么99%的情况下模型的预测是正确的，但是还有1%的情况下却是不正确的。

在这种情况下，可以看看模型预测正类的百分比，由精度和召回两个指标给出。



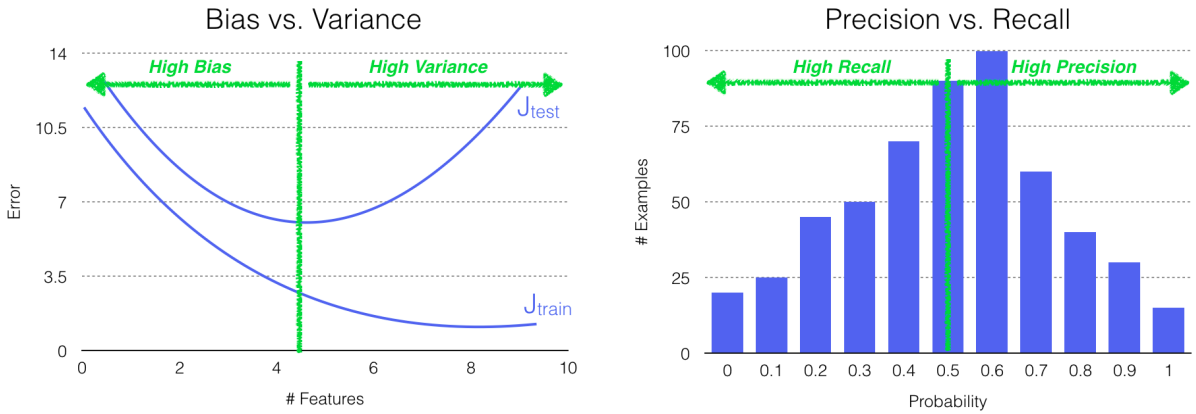
精确率是指模型预测正确的正样本的个数占该分类器所有分类为正样本个数的比例。召回率是指模型预测正确的正样本个数占所有的正样本个数的比例。可参考下图理解：



http://www.hsl.creighton.edu/hsl/Searching/Recall-Precision.html

一个良好的机器学习模型的目标是通过尝试最大化正类数量同时最小化假负类和假正类的数量，从而获取精确率和召回率之间的平衡。

5种改进模型的方法



如果在模型中遇到”高偏差”与”高方差”的问题，或者无法平衡“精确”与“召回”，那么可以使用以下多种策略。

对于模型中“高偏差”的问题，你可以尝试**增加输入特征的数量**。正如所讨论的，“高偏差”出现时，模型缺少底层数据，导致模型在训练和测试数据集上有很高的错误。可以绘制模型的误差图（见上图）直观的观察不同数据集在模型预测中的表现，可以看到更多的特征会提高模型的拟合程度。

然而在“高方差”的情况下，可以**减少特征的数量**。如果模型过度的拟合训练数据，可能是使用了太多的特征，**减少输入特征的数量**将使模型对测试数据和未来数据的处理更加的灵活。类似地，增加训练数据特征数量可以帮助机器学习算法在“高方差”的情况下，构建更可概括的模型。

对于低精度和低召回的情况，可以更改预测正负类的概率阈值（见上图）。对于低精度的情况，可以**适当增加概率阈值**，从而使得模型在正类的预测中正价的保守，另一方面，如果看到低召回，可以**适当的减少概率阈值**，使得模型更加积极的预测正类。

由于有足够的迭代次数，通常可以找到一个适当的机器学习模型与偏差、方差、精度和召回之间的平衡。

分享到：

- 微信
- 微博
- 豆瓣

PREVIOUS
TALKINGDATA开源智能设备情景感知框架“MYNA”
(/2016/12/07/OPENSOURCE_MYNA/)

NEXT
机器学习模型评估指标
(/2016/12/23/ML_ACCURACY/)

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)