32 | 概率统计篇答疑和总结：为什么会有欠拟合和过拟合？

2019-02-27 黄申

程序员的数学基础课 进入课程 >

讲述：黄申
时长 12:26　大小 11.40M

你好，我是黄申。

在概率统计这个模块中，我们讲了很多监督式机器学习相关的概念。你可能对朴素贝叶斯、决策树、线性回归这类监督式算法中的一些概念还是不太清楚。比如说，为什么要使用大量的文档集合或者语料库来训练一个朴素贝叶斯模型呢？这个过程最后得到的结果是什么？为什么训练后的结果可以用于预测新的数据？这里面其实涉及了很多模型拟合的知识。

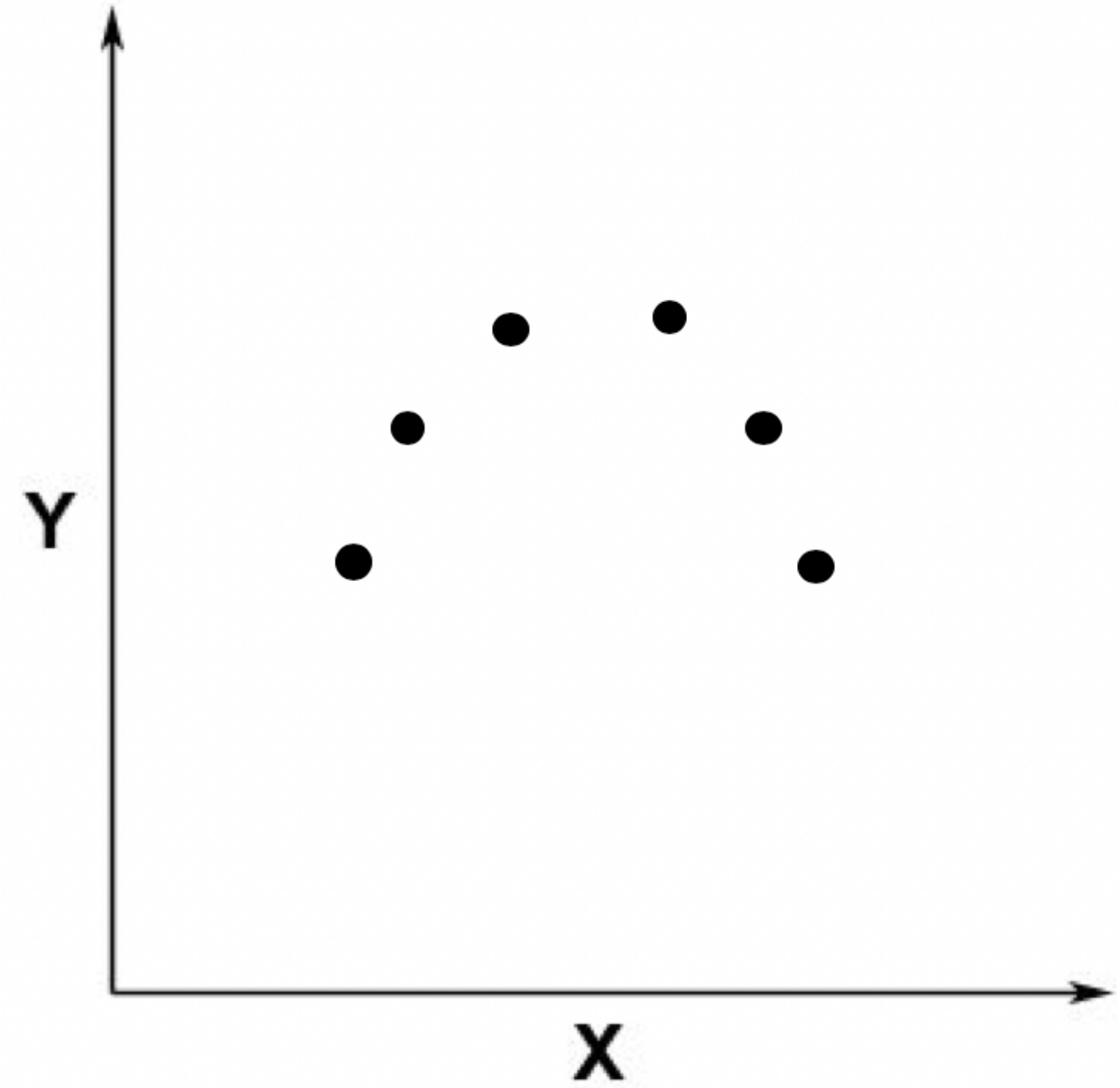
为了帮助你更好地理解这些内容，今天我就来说说监督式学习中几个很重要的概念：拟合、欠拟合和过拟合，以及如何处理欠拟合和过拟合。

拟合、欠拟合和过拟合

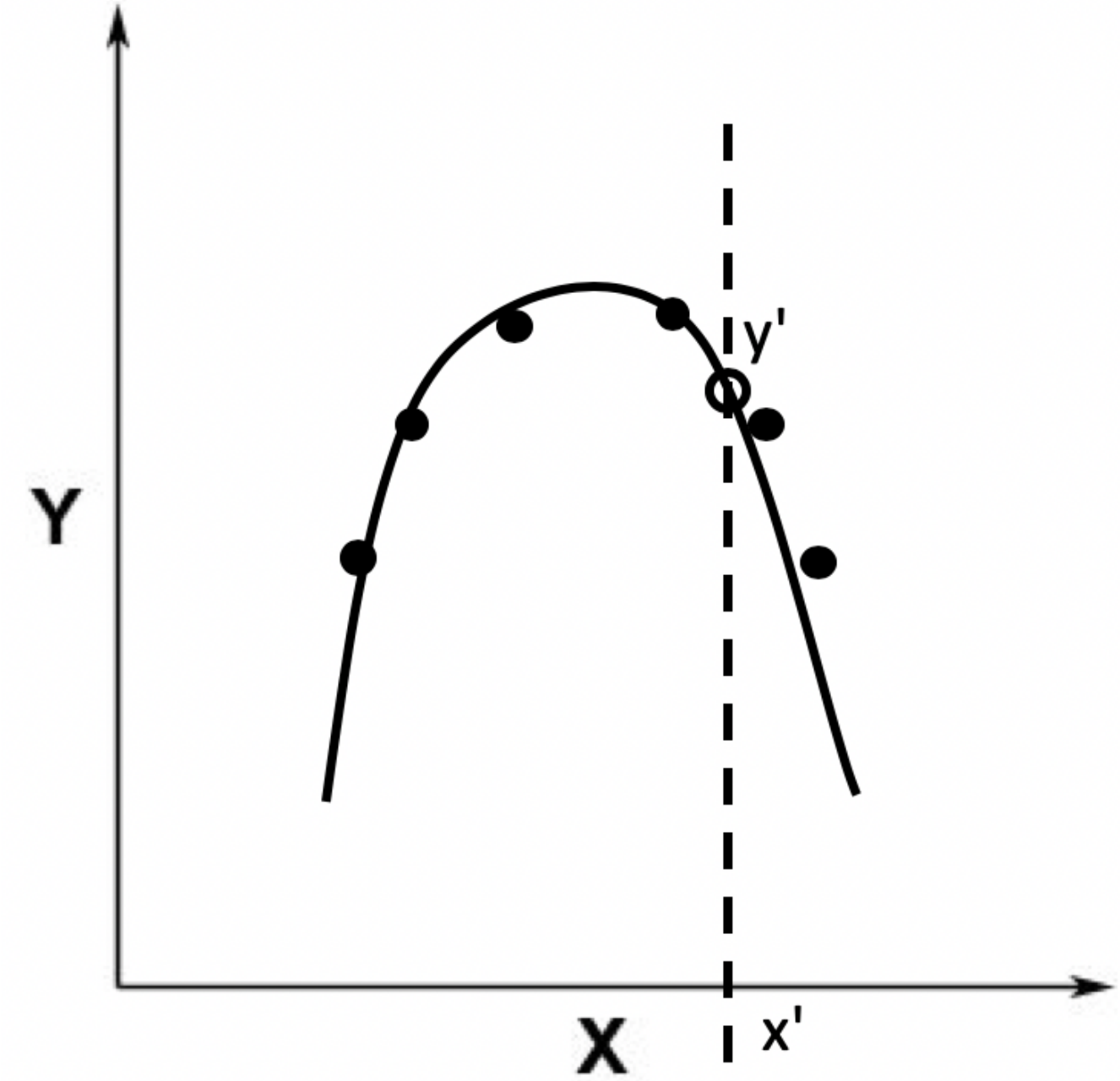
每种学习模型都有自己的假设和参数。虽然朴素贝叶斯和决策树都属于分类算法，但是它们各自的假设和参数都不相同。朴素贝叶斯的假设是贝叶斯定理和变量之间的独立性，而决策树的假设是集合的纯净程度或者混乱程度。我们这里所说的参数，是指根据模型假设和训练样本推导出来的数据，例如朴素贝叶斯中的参数是各种先验概率和条件概率，而决策树的参数是各个树结点以及结点上的决策条件。

了解了什么是模型的假设和参数，我们来看看什么是模型的**拟合**（Model Fitting）。在监督式学习中，我们经常提到“训练一个模型”，其实更学术的说法应该是“拟合一个模型”。

拟合模型其实就是指通过模型的假设和训练样本，推导出具体参数的过程。有了这些参数，我们就能对新的数据进行预测。这样说有些抽象，我画了张一元回归的图来帮助你理解。假设我们的数据点分布在一个二维空间。

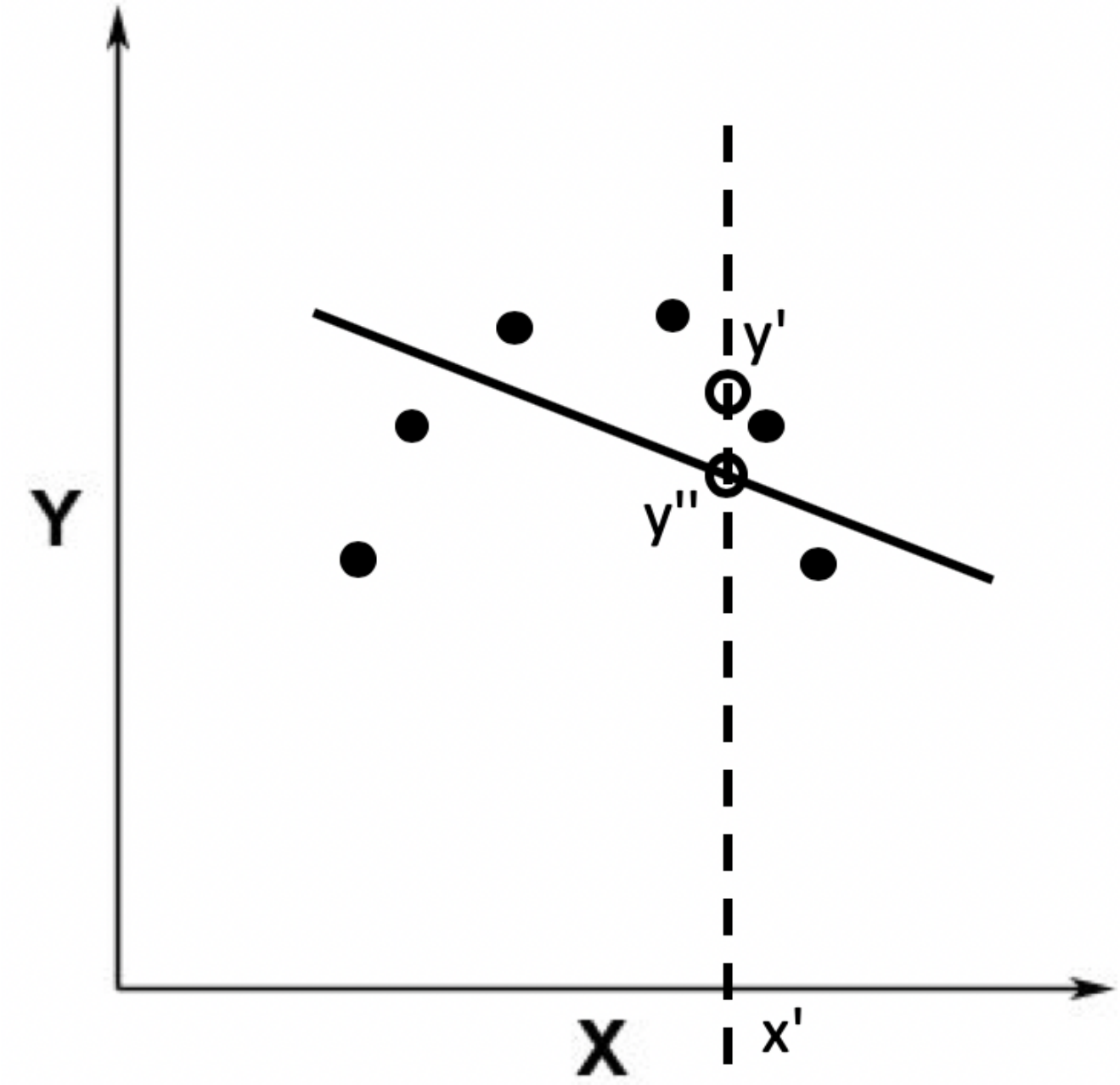


其中黑色的点表示训练数据所对应的点，x 轴表示唯一的自变量，y 轴表示因变量。根据这些训练数据，拟合回归模型之后，所得到的模型结果是一条黑色的曲线。



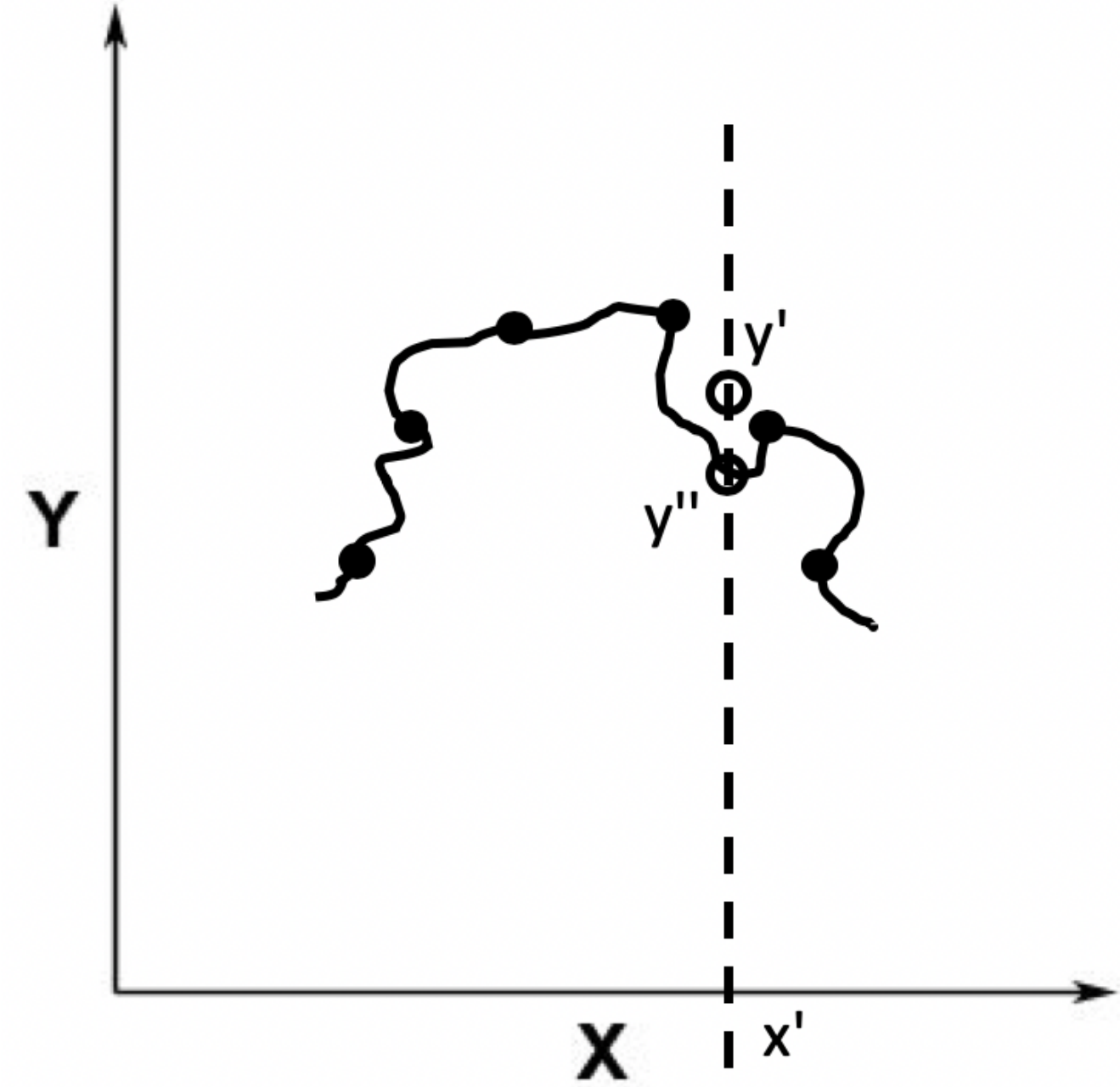
有了这条曲线，我们就能根据测试数据的 x 轴取值（如图中的 x’ ）来获取 y 轴的取值（如图中的 y’ ），也就是根据自变量的值来获取因变量的值，达到预测的效果。这种情况就是**适度拟合**（right fitting）。

可是，有的时候拟合得到的模型过于简单，和训练样本之间的误差非常大，这种情况就是**欠拟合**（Under Fitting）。比如下面这根黑色的曲线，和第一根曲线相比，它离数据点的距离更大。这种拟合模型和训练样本之间的差异，我们就称为**偏差**（Bias）。



欠拟合说明模型还不能很好地表示训练样本，所以在测试样本上的表现通常也不好。例如图中预测的值 y’’ 和测试数据 x’ 对应的真实值 y’ 相差很大。

相对于欠拟合，另一种情况是，拟合得到的模型非常精细和复杂，和训练样本之间的误差非常小，我们称这种情况为**过拟合**（Over Fitting）。比如下面这根黑色的曲线，和第一根曲线相比，离数据点的距离更近，也就是说偏差更小。



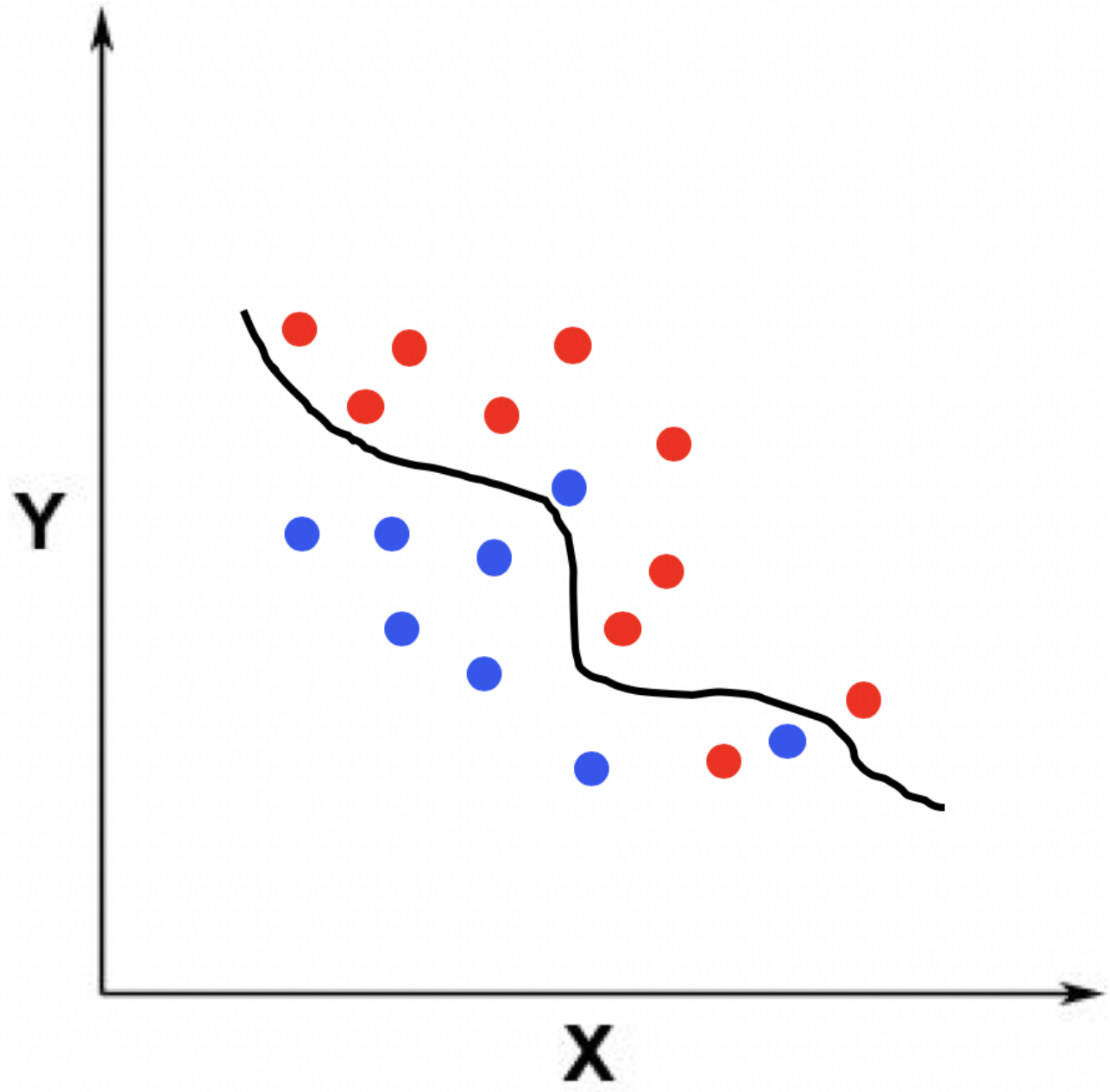
初学者通常都会觉得过拟合很好，其实并不是这样。过拟合的模型虽然在训练样本中表现得非常优越，但是在测试样本中可能表现不理想。为什么会这样呢？这主要是因为，有的时候，训练样本和测试样本不太一致。

比如，用于训练的数据都是苹果和甜橙，但是用于测试的数据都是西瓜。在上图中，测试数据 x’ 所对应的 y 值应该是 y’ ，而不是预测的 y’ ’ 。这种训练样本和测试样本之间存在的差异，我们称为**方差**（Variance）。在过拟合的时候，我们认为模型缺乏泛化的能力，无法很好地处理新的数据。

类似地，我以二维空间里的分类为例，展示了适度拟合、欠拟合和过度拟合的情况。仍然假设训练数据的点分布在一个二维空间，我们需要拟合出一个用于区分两个类的分界线。我分

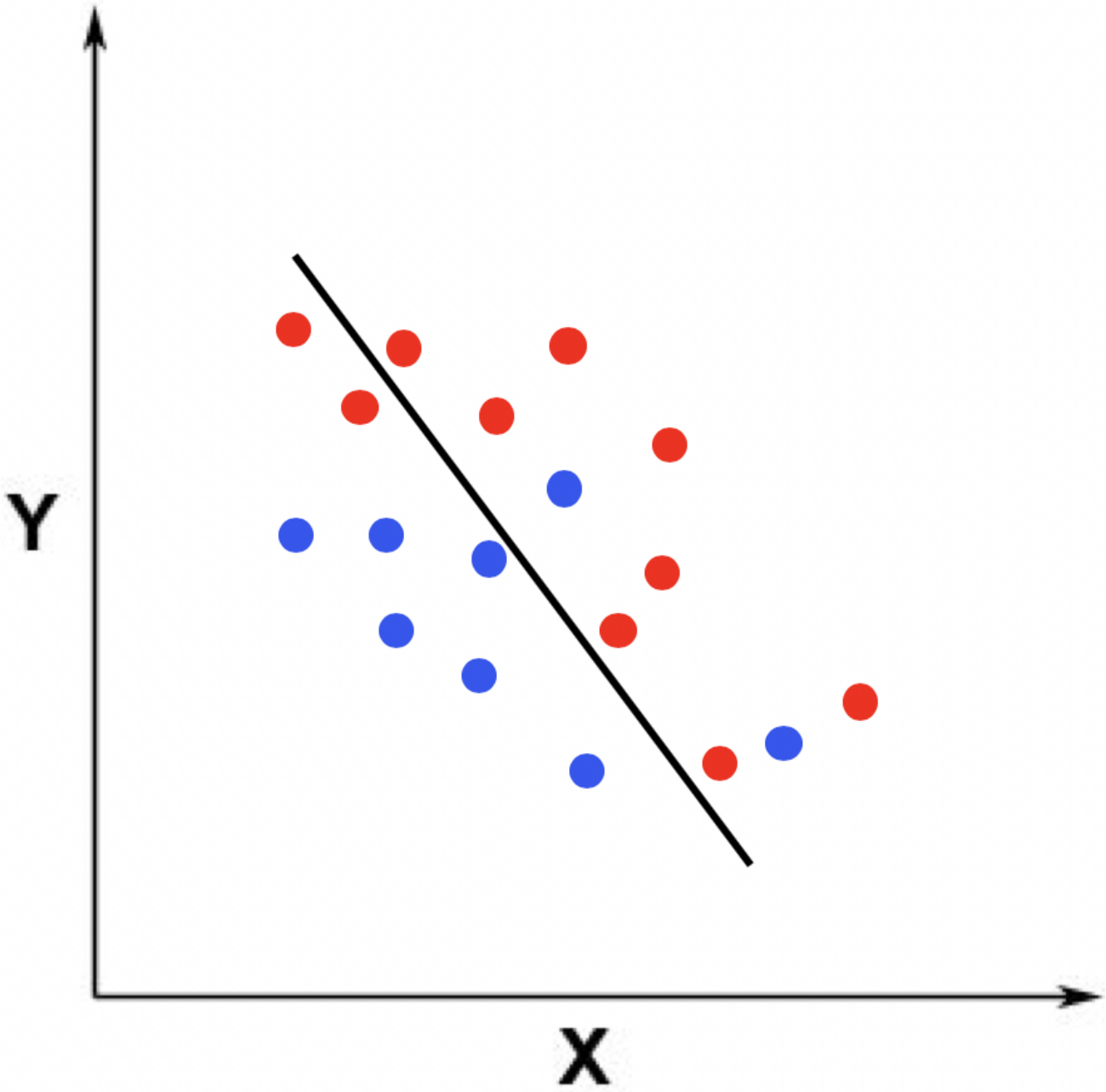
别用三张图展示了这三种情况下的分界线。

首先，第一张是适度拟合的情况。

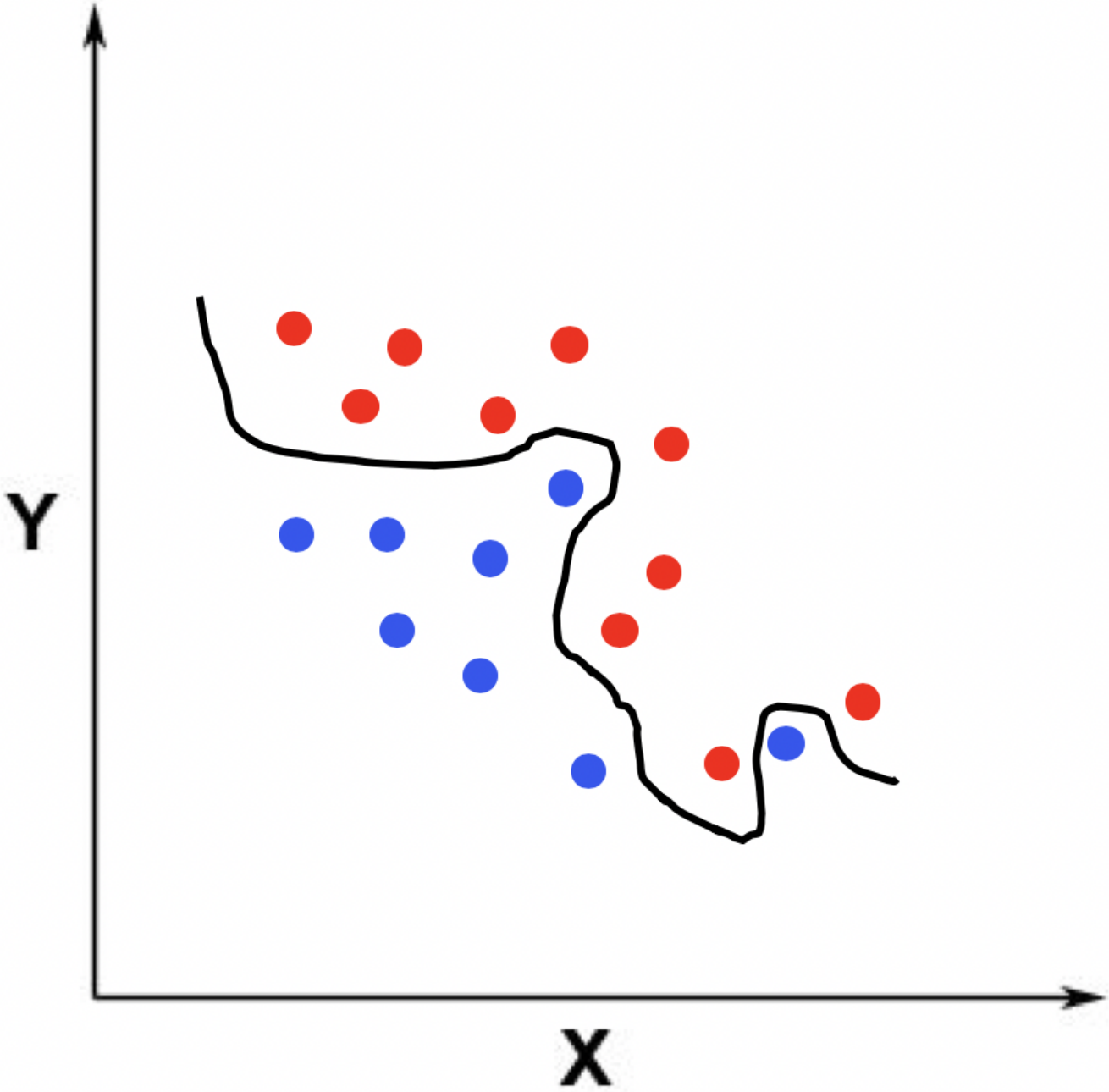


这张图中，蓝色的点表示分类 1 的训练数据点，红色的点表示分类 2 的训练数据点。在适度拟合的时候，分界线比较好的区分了蓝色和红色的点。

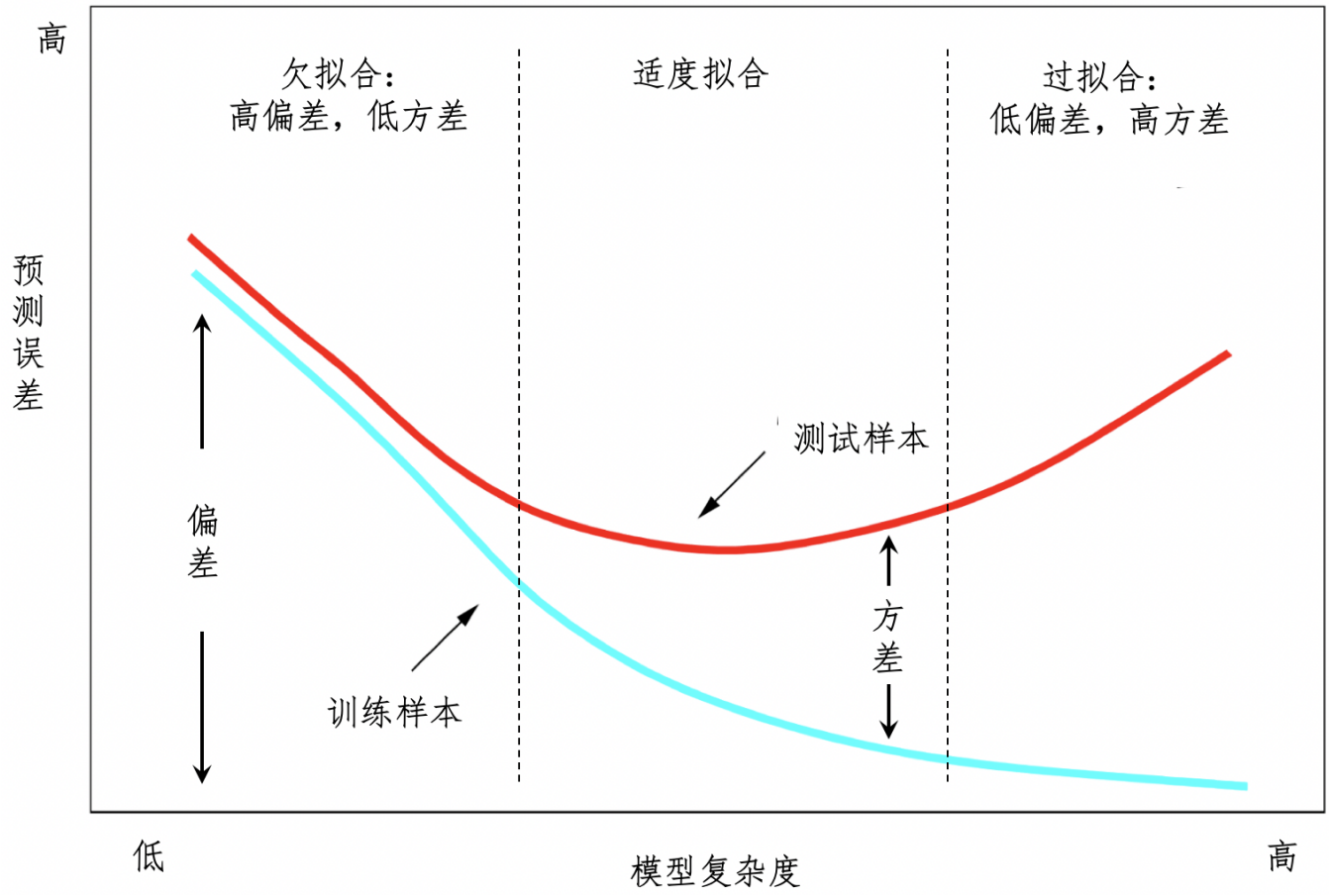
在欠拟合的时候，模型过于简单，分界线区分训练样本中蓝色和红色点的能力比较弱，存在比较多的错误分类。



在过拟合的时候，模型过于复杂，分界线区分训练样本中蓝色和红色点的能力近乎完美，基本上没有错误的分类。但是，如果测试样本和这个训练样本不太一样，那么这个模型就会产生比较大的误差。



在常见的监督式学习过程中，适度拟合、欠拟合和过拟合，这三种状态是逐步演变的。我也用一张图来解释这个过程。



在这个图中，x 轴表示模型的复杂程度，y 轴表示预测的误差。蓝色曲线表示模型在训练样本上的表现，它和 x 轴之间的距离表示了偏差。而红色曲线表示模型在测试样本上的表现，它和蓝色曲线之间的距离表示了方差。

从图的左侧往右侧看，模型的复杂度由简单逐渐复杂。越复杂的模型，越近似训练样本，所以偏差就不断下降。可是，由于过于近似训练样本，模型和测试样本的差距就会加大，因此在模型复杂度达到一定程度之后，在训练样本上的预测误差反而会开始增加，这样就会导致训练和测试样本之间的方差不断增大。

在这个图中，最左边是高偏差、低方差，就是我们所说的欠拟合，最右边是低偏差、高方差，就是我们所说的过拟合。在靠近中间的位置，我们希望能找到一个偏差和方差都比较均衡的区域，也就是适度拟合的情况。

如何处理欠拟合和过拟合？

解释了什么是模型拟合、欠拟合和过拟合，我们下面来说说，有哪些常见的处理过拟合和欠拟合的方法。

想要解决一个问题，我们先要搞清楚产生这个问题的原因。**欠拟合问题，产生的主要原因是特征维度过少，拟合的模型不够复杂，无法满足训练样本，最终导致误差较大。**因此，我们就可以增加特征维度，让输入的训练样本具有更强的表达能力。

之前讲解朴素贝叶斯的时候，我提到“任何两个变量是相互独立的假设”，这种假设和马尔科夫假设中的一元文法的作用一致，是为了降低数据稀疏程度、节省计算资源所采取的措施。可是，这种假设在现实中往往不成立，所以朴素贝叶斯模型的表达能力是非常有限的。当我们拥有足够的计算资源，而且希望建模效果更好的时候，我们就需要更加精细、更加复杂的模型，朴素贝叶斯可能就不再适用了。

比如，在最近非常火的电影《流浪地球》中，计算机系统莫斯拥有全人类文明的数字资料库。假设我们手头也有一个庞大的资料库，也有莫斯那么强大的计算能力，那么使用一元文法来处理数据就有点大材小用了。我们完全可以放弃朴素贝叶斯中关于变量独立性的假设，而使用二元、三元甚至更大的 N 元文法来处理这些数据。这就是典型的通过增加更多的特征，来提升模型的复杂度，让它从欠拟合阶段往适度拟合阶段靠拢。

相对应的，**过拟合问题产生的主要原因则是特征维度过多，导致拟合的模型过于完美地符合训练样本，但是无法适应测试样本或者说新的数据。**所以我们可以减少特征的维度。之前在介绍决策树的时候，我提到了这类算法比较容易过拟合，可以使用剪枝和随机森林来缓解这个问题。

剪枝，顾名思义，就是删掉决策树中一些不是很重要的结点及对应的边，这其实就是在减少特征对模型的影响。虽然去掉一些结点和边之后，决策树对训练样本的区分能力变弱，但是可以更好地应对新数据的变化，具有更好的泛化能力。至于去掉哪些结点和边，我们可以使用前面介绍的特征选择方法来进行。

随机森林的构建过程更为复杂一些。“森林”表示有很多决策树，可是训练样本就一套，那这些树都是怎么来的呢？随机森林算法采用了统计里常用的可重复采样法，每次从全部 n 个样本中取出 m 个（m<n），然后构建一个决策树。重复这种采样并构建决策树的过程若干次，我们就能获得多个决策树。对于新的数据，每个决策树都会有自己的判断结果，我们取大多数决策树的意见作为最终结果。由于每次采样都是不完整的训练集合，而且有一定的随机性，所以每个决策树的过拟合程度都会降低。

从另一个角度来看，过拟合表示模型太复杂，而相对的训练数据量太少。因此我们也可以增加训练样本的数据量，并尽量保持训练数据和测试数据分布的一致性。如果我们手头上有大

量的训练数据，则可以使用交叉验证（Cross Validation）的划分方式来保持训练数据和测试数据的一致性。其核心思想是在每一轮中，拿出大部分数据实例进行建模，然后用建立的模型对留下的小部分实例进行预测，最终对本次预测结果进行评估。这个过程反复进行若干轮，直到所有的标注样本都被预测了一次而且仅一次。如果模型所接受的数据总是在变化，那么我们就需要定期更新训练样本，重新拟合模型。

总结

第二模块中，我介绍了很多概率统计中常用的概念。随机变量和它的概率分布体现了事物发生的不确定性。而条件概率、联合概率和边缘概率体现了多个随机变量之间的关系以及是不是相互独立，通过这三者的关系，我们可以推导出贝叶斯定理。在贝叶斯定理和变量独立性假设的基础之上，我讲了朴素贝叶斯算法中的公式推导，以及如何使用先验概率来预测后验概率。由于朴素贝叶斯假定多个变量之间相互独立，因此特别适合特征维度很多、特征向量和矩阵很稀疏的场景。基于词包方法的文本分类就是个非常典型的例子。

文本分类涉及了词与词之间相互独立的假设，然后延伸出多元文法，而多元文法也广泛应用在概率语言模型中。语言模型是马尔科夫模型的一种，而隐马尔科夫模型在马尔科夫模型的基础之上，提出了两层的结构，解决了我们无法直接观测到转移状态的问题。

由概率知识派生而来的信息论，也能帮助我们设计机器学习的算法，比如决策树和特征选择。统计中的数据分布可以让特征值转换更合理，而假设检验可以告诉我们哪些结论是更可靠的。

由于很多监督式学习都是基于概率统计的，所以我使用了一些学习算法来进行讲解。你会发现，概率和统计可以帮助这些算法学习训练样本中的特征，并可以对新的数据进行预测，这就是模型拟合的过程。

思考题

学习完了概率和统计模块，你觉得自己最大的收获和感触是什么？

欢迎留言和我分享。你可以把今天的内容分享给你的好友，和他一起精进。

上一篇　31 | 统计意义（下）：如何通过显著性检验，判断你的A/B测试结果是不是巧合？

下一篇　33 | 线性代数：线性代数到底都讲了些什么？

精选留言 (1)

写留言

冰冷的梦
2019-03-13

老师，贝叶斯以后我已经基本看不懂了。。。我应该是缺少概率统计相关知识的基础吧？

作者回复: 你可以先把我之前介绍的概率基础，包括联合概率、条件概率、边缘概率等概念弄明白，然后再慢慢看贝叶斯这一块就不难理解了