机器学习综述

## 一、机器学习的基本原理

1、机器学习的定义

机器学习是为了解决数据的分类、回归、标注等任务，基于数据集的特点构建模型，并运用模型解决任务的过程。

这个定义中有三个重要概念：“数据集”、“模型”和“任务”，下面分别解释一下它们：

（1）数据集：

由大量的“样本”（又叫实例）构成，所有样本都有同样的一组“特征”（又叫属性），这组特征构成一个特征向量。不同特征具有不同取值范围；不同样本的特征向量的取值不同。另外，在有些问题（监督学习问题）中，样本集中每个样本还有一个“标签”，样本的标签值代表与样本相关的一个我们感兴趣的指标。

数据集可以表示成一个矩阵的形式（不包含标签），矩阵的行数m就是样本个数，矩阵的列数n就是特征数。矩阵的每一行对应一个样本，每一列对应一个特征。

带标签的数据集一般表示为：

（2）模型：

模型代表了数据集的结构（或者说数据集的分布、规律、信息）。一般来说，它的形式是一个映射关系，或者说映射规则。它输入一个给定的样本（的特征向量），输出样本的一个指标（该指标对于不同的任务具有不同的意义）。

每个模型都有其独特的一个“形式”。

一个模型可以应用于不同的数据集，但在不同的数据集上有不同的具体形式——因为它要适应不同数据的结构特点。模型的不同具体形式通过模型的不同参数来体现。

模型使用数据集进行训练，得到具体的形式（参数值），训练成功的模型能够完美地学到数据集的结构。从而能够对该数据集对应的总体的新的样本进行预测。

用于训练模型的数据集称为训练集。

注意区分模型的“形式”和“具体形式”。模型的形式是包含未知参数的，具体形式是模型通过在训练集上训练确定了参数后得到的。

（3）任务：

机器学习任务主要包括分类、回归、标注和聚类，有时也统一称为“预测”。

分类问题：输入样本（的特征向量），输出为离散的表示类别的整数值。注意，输入可以是连续的也可以是离散的；

回归问题：输入样本（的特征向量），输出为连续的实数值；

标注问题：标注问题是分类问题的一个推广，其输入是一个观测序列，输出是一个标记序列。注意，这里的样本输入是具有顺序意义的序列，而不是简单的特征向量。

聚类问题：根据样本的特征，将样本集划分为不同的子集（也叫“簇”），使簇内的样本之间的相似度尽量高，簇间样本之间的相似度尽量低。

机器学习的任务可以分为“监督学习”和“非监督学习”。监督学习就是使用带标签的训练集进行训练，学到样本特征与标签之间的关系，用这个关系进行预测；非监督学习就是使用不带标签的训练集进行训练，学到样本特征的分布（或者说结构、模式），再根据分布（或结构、模式）进行进一步的分析或预测。分类问题、回归问题、标注问题属于监督学习，聚类问题属于非监督学习。

一般我们讨论机器学习模型，如无特别说明，都默认指监督学习。

2、模型的运作

那么机器学习的模型具体是怎么运作的呢？一般来说需要经历三个步骤：给出模型的形式、构建目标函数最优化问题、对目标函数进行优化。

（1）模型形式（又叫“假设”）：即模型对输入和输出关系的假设。模型的形式包含未知参数，模型在训练集上进行训练，目的就是为了确定参数，当参数确定后，就得到的模型的“具体形式”，就可以拿来作预测了。

比如，线性回归模型的形式为：

其中x是输入的样本特征向量，w是模型参数。

逻辑回归模型的形式为：

另外，有一些模型的形式不具有简洁的解析形式，比如决策树模型（注意，决策树模型也可以形式化，但是它还是要加上语言描述，才能完备地表示其形式）。

（2）目标函数：给定基于数据集的模型性能评价准则，构建目标函数的最优化问题。准则就是告诉人们，模型对样本作出怎样的预测才算是个好模型。

最基本的回归问题的准则是最小均方误差准则（MSE），而分类问题的准则就比较多了。

根据模型的形式和模型性能评价准则，即可构建目标函数，目标函数指示了在该模型形式上的模型性能。目标函数通常是基于训练集构建的，它是模型参数的函数。目标函数的最优化（最大或最小），就对应着模型的最佳性能，相应的模型参数，就是我们认为的最优参数。

（3）目标函数的优化：实际上，目标函数的优化跟机器学习问题是独立的，因为它本质上就是一个求函数极值或泛函极值的问题。但是有一些优化算法，是受到机器学习情景下目标函数优化任务的启发而提出的，比如随机梯度下降。

常见的优化方法有：梯度下降法、随机梯度下降法、Adam、共轭梯度法、牛顿法，以及一些启发式算法。

3、“没有免费午餐”定理

没有免费的午餐定理是关于模型对不同数据集适用性的一个结论。它的大致意思是说，任何模型在所有的数据集上的平均性能是相等的，都等价于随机预测。这是因为每个模型都是针对特定的一类问题，或者说针对具有某种特定的特点的一类数据而言的（当然，有可能有多个模型针对具有类似特点的数据）。比如线性回归，它是为了拟合具有线性关系的X和Y的数据的，你用它去拟合具有二次函数关系的X和Y的数据，显然效果不会好。

## 二 机器学习的本质

机器学习本质上是对数据进行建模的科学。从数据中挖掘信息是它的根本目标。比如提取数据中的结构、规律、模式，或者拟合数据的（样本与标签之间的）关系。机器学习针对不同的数据和要挖掘的不同信息，建立不同的模型。模型中就蕴含了我们要提取的信息。

从数据的本质上来看（或者说从统计学的角度来看），我们的离散、有限的数据相当于一个总体分布（或者说整体结构）的一部分。这个总体分布（整体结构）是连续的，我们离散、有限的数据是从这个总体中采出来的一部分数据。而机器学习模型，都是为了根据这部分的数据去拟合总体分布（整体结构）。

从概率统计的视角来看，我们从有限的样本集去拟合总体分布，势必要有先验，即数据+先验=总体分布。我们的机器学习模型就起到先验的作用。

机器学习模型分为有监督和无监督。

有监督就是样本有标签，用模型拟合样本和标签之间的关系。有监督的任务主要有三类：回归、分类、标注。回归的模型比较少，一般就是线性回归、多项式回归，以及加了正则化的岭回归和脊回归，此外，决策树也可以用于回归；分类的模型有很多，比如最简单的感知机模型、单纯基于概率的朴素贝叶斯和最大熵模型、基于距离的KNN和SVM、基于距离但包装成概率的逻辑回归、基于规则的决策树，还有一个模型框架：集成学习（包括bagging和boosting）；标注模型主要就是HMM和CRF。

无监督模型跟有监督模型不同，无监督的任务不像有监督的任务那样，有明确的目标：拟合关系。无监督任务的目标是“提取数据中的结构（或者说模式，或者说规律）”，那么这个“结构”就有很多层意思，比如数据的聚类结构、数据的空间结构、数据的分布等。针对不同的意义，我们使用不同的无监督模型来挖掘数据的结构。比如对于聚类结构，我们使用聚类算法；对于数据的空间结构，我们使用降维算法或者流形学习；对于数据的分布，我们使用无监督生成模型，如VAE、GAN。

总结来说，机器学习就是一门针对数据进行建模的技术。它想方设法挖掘数据中的信息。这是很有意义的事情，因为数据中有很多有价值的信息。但是这也是机器学习的局限性：它过分依赖数据，无法获取数据蕴含的信息以外的知识。

## 三 机器学习模型的分类

1、概率模型与非概率模型

机器学习模型种类繁多，形式各异。但按照是否基于概率思想，可以分为概率模型和非概率模型。概率模型将数据集的每个特征看作是一个随机变量，所有特征构成一个随机向量X，同时所有标签也构成一个随机变量Y。它们的联合分布为，则数据集中的任意样本都是由该联合分布生成的。

为了对给定的x输出其预测值，概率模型需要从训练集中学出联合分布或者条件分布。一般来说，首先给出一个分布的形式，形式中包含未知参数，概率模型利用训练集求取最优的参数，得到分布的具体形式，就能够执行预测的任务了。利用训练集求解最优参数的过程，是不是感到有点熟悉？没错，其实那就是统计学中的参数估计问题。一般使用极大似然法求解分布的最优参数，但对于包含隐变量的概率模型，一般使用EM算法。

可以看出，概率模型的本质就是X与Y的联合分布。

2、生成模型与判别模型

前面说过，监督学习模型试图对每个样本x给出其预测值。

生成模型能够生成新的样本与标签对，当然，这不是随便生成的，因为模型既学到了数据集的样本特征分布，又学到了样本特征与标签的关系，所以它生成的样本与标签对是符合数据集的分布特点的。

判别模型只学到了关系，没学到分布，故不能生成新数据，只能对给定的样本特征x，输出其预测值。

注意，理论上，一个模型是生成模型还是判别模型与它是概率模型还是非概率模型无关，但是实际上，生成模型基本上都是概率模型。因为概率模型可以学到联合分布，然后基于联合分布生成新的样本。

在常用的模型中，属于生成模型的是：朴素贝叶斯和隐马尔科夫模型；属于判别模型的是：KNN、决策树、logistic回归、最大熵模型、SVM、集成模型、条件随机场等。

可见，生成模型比较少，大多数模型都是判别模型。

3、线性模型与非线性模型

线性模型和非线性模型都是针对回归模型而言的。狭义的线性模型就是指线性回归：

线性回归模型的输出y是输入x的线性函数。

而广义上的线性模型是指可以写成如下形式的模型：

即在线性回归外面套一层函数g。

注意，一般来说，多项式回归模型不被看成是非线性模型，因为它可以通过简单的特征转化变成线性模型。

## 四 模型的欠拟合与过拟合

1、欠拟合与过拟合的定义

欠拟合与过拟合是用来分析模型与数据吻合程度以及模型在数据上的表现的。

当在一个训练集上对模型进行训练时，我们会发现，当使用的模型复杂度较低时（也说模型的容量较小），在训练集上预测性能可能会很差；当使用的模型复杂度较高时（也说模型的容量较大），往往在训练集上性能很好，但是在测试集上性能很差。

前面一种情况其原因在于：模型容量太小，相对来说，训练集的分布太复杂（或者说训练集中的信息太多，或者说训练集蕴含的规律太复杂），导致模型应付不来。这称之为欠拟合，这是比较容易理解的。解决该情况的方法就是增加模型的复杂度，使之与数据分布的复杂度相吻合。

而后面一种情况称为过拟合，这是我们着重研究的情况。那么造成该情况的原因又是什么呢？

我们认为，训练集和测试集都是从总体中采集出来的部分样本的集合。这个“总体”，从统计的观点来看，是一个“真实分布”，从模型的观点来看，是一个“完美模型”生成的总体数据集。采样的过程必然存在误差，模型复杂度高的情况下，模型可能会将训练集中的误差的规律当作总体的规律学习到了，而这误差的规律是在训练集上的，当然不能推及测试集，所以在测试集上预测准确度不高。

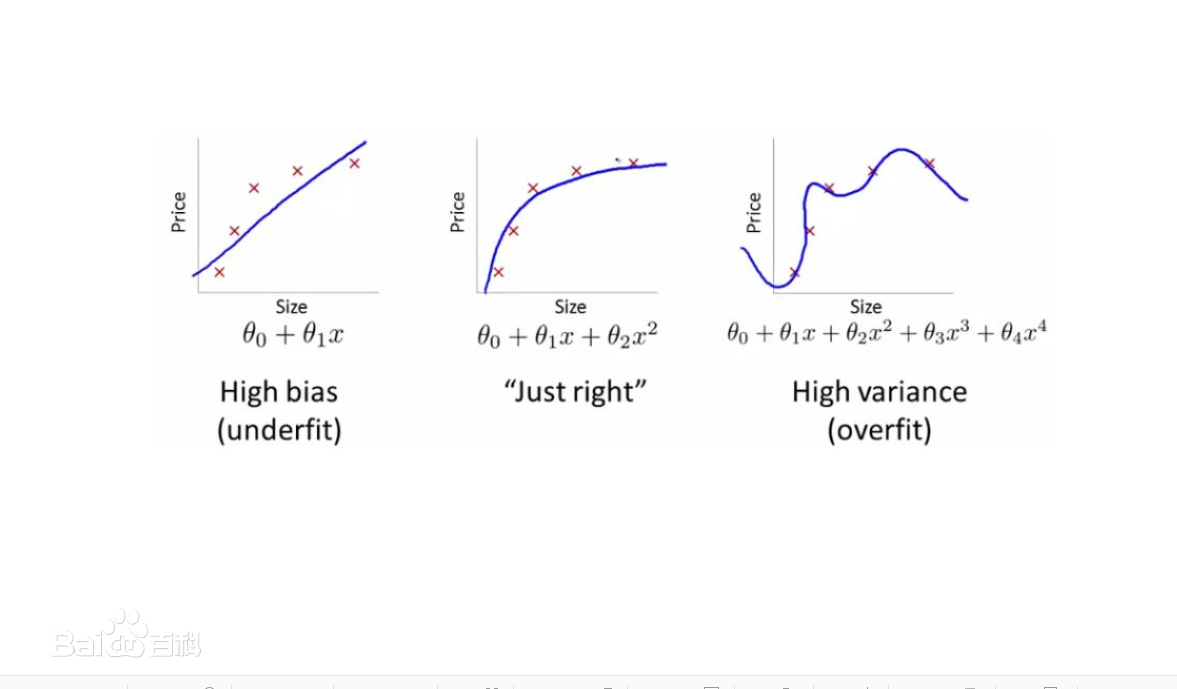


图1 欠拟合与过拟合示意图

注意，训练集是来自于总体的、有限的部分样本集合，我们的目的是用这总体的一部分去推断总体的真实情况，这并不容易——显然，你无法只从总体的一部分中获得总体的全部信息（除非是分形问题^\_^）。如果没有其他（先验）信息，你无法知道训练集中哪部分是误差，哪部分是反映总体的真实规律。比如图1中所示的情况，如果没有其他（先验）信息，你无法判断第三个图中的模型（曲线）是真的过拟合了，还是真实关系本来就是这样子的。

针对这个问题，我们提出了模型的“泛化能力”的概念，希望训练出的模型不仅在训练集上有尽可能高的预测性能，也希望在训练集之外的数据集上有很高的预测性能。注意，这里不只是希望模型在测试集上有好的预测性能，而是在总体数据上有好的预测性能。测试集只是用来测试泛化能力的一个数据集。

模型的泛化能力，本质上由模型与训练集对应的总体在“复杂度”方面的契合程度决定。当训练集给定时，模型越简单，越容易忽略数训练中的采样误差的规律；模型越复杂，越容易学习到训练集中采样误差的规律。

注意，模型的“泛化能力”针对的是模型的形式（包含未知参数）而言的，而不是针对模型的具体形式（参数已确定）而言的。

2、偏差-方差分解

偏差-方差分解理论用“泛化误差”来形式化地表征模型的泛化能力。

其基于回归问题的均方误差作为衡量模型预测性能的指标，使用统计学中的术语（期望、方差）定义了“偏差”和“方差”，并解析了二者如何影响模型的泛化能力。

注意，偏差和方差分解理论说明了偏差和方差是影响模型泛化能力的两个因素，并解释了二者的“含义”（或者说表征了什么），但并未进一步解释这两个因素又是由什么决定的（实际上，是由模型的复杂度决定的，或者说由模型的容量决定的）。

假设f是一个带未知参数的模型的形式，我们使用取自同一总体的多个不同的训练集训练它，得到模型的不同具体形式，然后用这些不同的模型具体形式对新的样本x进行预测。

首先定义模型的“泛化误差”：

其中x是在所有训练集以外的一个新的样本（的特征向量），是f在训练集D上训练好的模型的具体形式对x的输出，y是x对应的真实值（标签）。也就是说，针对多个训练集训练出模型的多个多个具体形式，每个具体形式对新样本x的预测都跟真实值y有误差，这些不同训练集上的误差的期望，就定义为泛化误差。

可以看出，泛化误差不是针对单个训练集，而是考虑了多个训练集上的综合预测误差，因此能衡量模型在多个训练集上的平均性能，或者说在“总体”上的预测性能。这也是泛化能力的本意。显然，泛化误差越低，模型的泛化能力越强。

注意，上式中，不是表示“某个训练集D上的期望”，而是表示“对所有训练集D的期望”。

然后定义模型的“方差”：

设模型在各训练集上的具体形式对新样本x的预测值的期望为：

则方差定义为：

可以看出，模型在每个训练集上训练出的具体形式对x都有一个预测值，对这些预测值统计出的方差，就是模型的方差，它表示模型在各个训练集上形成的具体形式对x的预测值的一致性。方差越大表示不同训练集上形成的具体形式对x的预测值差异较大，也就是各具体形式的性能差异较大；反之，表示不同训练集上形成的具体形式对x的预测值一致性较好，也就是各具体形式的性能一致性较好。

接着定义模型的“偏差”：

可以看出，模型在每个训练集上训练出的具体形式对x都有一个预测值，对这些预测值统计其期望，这个期望跟x的真实标签y的差异，就是模型的偏差。

注意模型偏差跟模型的泛化误差的区别：二者虽然都基于预测值与真实值之间的差异，但偏差是各训练集上具体形式的预测值的平均值与真实值的差异；泛化误差是各训练集上预测值与真实值的差异的平均值。

显然，偏差反映的是模型本身的性能（因为它相当于综合了所有训练集上的具体形式对样本预测的结果（均值），消除了训练集的影响，因此可看作模型本身的性能的衡量）；泛化误差则反映了模型在各训练集上的具体形式的误差的大小。或者可以这样理解：偏差反映模型本身的性能，它通过平均的方式消除了各训练集上具体形式的预测值的误差的波动；泛化误差不仅不消除这些误差，还刻意地统计这些误差的平均值，因为它不是只关注模型本身，而是关注模型在每一个训练集上的性能（因为模型在不同训练集上具体表现不同，这是与训练集有关的）。

可以通过简单的推导，将模型泛化误差分解为方差+偏差+噪声：

其中，最后一项即为噪声项，是样本x的真实标记y与其在各训练集中标记的误差的均值。因为各训练集中可能存在对x错误标记的现象，显然这跟模型无关，是模型无法解决的问题，因此称为噪声。

注意，在这个推导中，我们对泛化误差的定义与原始定义相比稍有改动，将模型（对样本）在各训练集上的预测值跟x的实际标签的对比，改为跟训练集上标签的对比，这是为了考虑噪声的因素。

综上可知，在泛化误差中，偏差反映了模型本身的拟合能力；方差刻画了数据扰动（训练集的不同）的影响；噪声刻画了学习问题本身的难度。

因此，泛化误差并不是单纯的对“某个模型的泛化能力”的度量。而是对“某个模型在某个数据集及其问题上的泛化能力”的度量。前面也已经说过，泛化能力的本质是模型与训练集对应的总体在“复杂度”方面的契合程度，抛开数据集和问题谈模型的泛化能力是没有意义的。

在实际的问题中，我们通常只有一个训练集用来训练模型。那么我们怎么用基于多个训练集定义的偏差-方差的概念分析模型的泛化能力呢？

其实，一个训练集上的表现跟多个训练集上的表现具有一致性，因为若模型在一个训练集上预测的准确度较高，说明该模型复杂度较高，那么在多个训练集上预测的准确度较高，也就是偏差越低；同理，若一个模型在训练集上表现特别好，而在测试集上表现不佳，说明模型复杂度过高，学到了训练集上的噪声的规律，可以想象，这样的模型在多个训练集上预测结果的差异一定较大，也就是方差较大。

显然，要想降低泛化误差，就应同时降低偏差和方差，但是实际上，偏差和方差是一对矛盾，降低偏差势必会增加方差（这也是不能用偏差-方差理论本身来解释的）。因为降低偏差就要增加模型的复杂度，用复杂度较高的模型在各训练集上学习，就会学到各个训练集中的噪声规律，那么不同训练集上训练的模型对样本x的预测肯结果定差异较大，也就是方差较大；同理，若降低模型复杂度，可能会降低方差，但偏差就会增大。

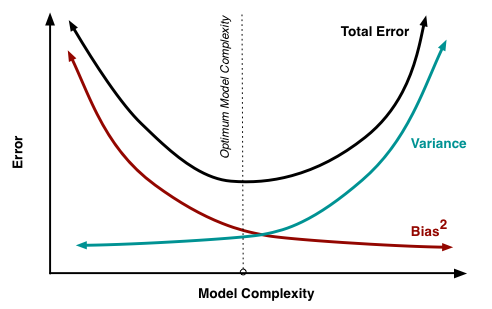


图2 方差-偏差平衡示意图

图2展示了偏差-方差平衡。该图表明，模型复杂度的把控需要考虑偏差-方差这一对因素，使整体泛化误差最小。

当然，样本集容量也影响模型泛化能力。样本集容量越大，其蕴含的关于总体的信息越多。通过增加样本集容量，可以同时降低偏差和方差。

3、提高模型泛化能力的方法

从上面的分析可以看出，提高模型泛化能力的基本思路有两个：一是增加样本容量，二是选取复杂度适中的模型。此外，还可以使用集成方法构建模型，以及使用交叉验证。

（1）正则化

正则化是一种给模型添加先验信息的有效手段，它也可以起到减小模型复杂度的作用。它并不改变模型的形式，而是通过改造模型的目标函数实现其目的：它在一般的损失函数后面加上一个正则化项，形成新的目标函数。比如：基本的线性回归模型的目标函数是MSE损失函数，它加上正则项，就变成了Lasso回归回归；如果加上正则项，则变成岭回归。

（2）提前终止

对于决策树等模型，提前终止是减小模型复杂度的有效方法（剪枝也是一种有效方法）。

注意，虽然减小模型复杂度能提高模型的泛化能力，但不能毫无依据地随便乱减，要根据合理的先验信息（对数据的认识）。

（3）集成方法

如Bagging方法、Boosting方法。

（4）交叉验证

交叉验证是一种选择模型的方法。一般来说，我们会给出一个带有超参数的模型，它相当于是一族模型，交叉验证的目的就是选取它认为的最优的超参数（对应着它认为的泛化能力最强的模型）。

交叉验证是将训练集划分成训练集和验证集两部分。模型使用不同的超参数，在训练集上进行训练，然后在验证集上验证其预测效果，通过效预测果的好坏调整超参数，选出在验证集上预测效果最好的一组超参数作为最优超参数，得到最优的模型。

在实际中应用的较多的是上述简单交叉验证的一个变种：k折交叉验证。

k折交叉验证的运行过程描述如下：

对于每一组超参数，将训练集等比例划分成k份，以其中的一份作为验证集，其他的k-1份数据合起来作为训练集，然后在训练集上训练，在验证集上对预测效果评分，这样算是一次实验。交叉验证把实验重复做了k次，每次实验从k个部分选取一份不同的数据部分作为验证集，剩下的k-1个当作训练数据（最终保证k个部分的数据都作过验证集）。最后把得到的k个实验的预测效果评分进行平均，作为该组超参数的评分。最后选择评分最高的超参数，作为最优模型的超参数。