**机器学习的本质**

Holbrook

这个问题在[知乎](http://www.zhihu.com/question/19830921" \t "_blank)上有很多讨论，但似乎没有达成共识。我个人的观点如下：

机器学习的目标是让计算机自动获取知识，然后用获取的知识解决问题。

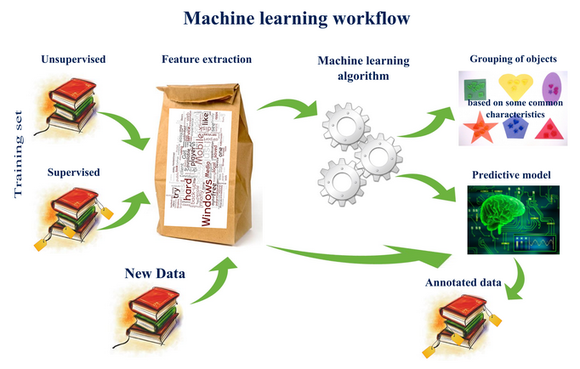
获取知识的方式是通过大量的数据，得到一个函数。

从程序的角度来说，传统的程序在编码完成时，函数就已经确定了，运行程序就是执行这个函数，其结果是确定的；  
而机器学习程序在编码完成后，需要先用数据进行训练/学习，才能确定函数形式，执行这个函数的结果取决于  
训练/学习的数据，是不确定的。

机器学习得到的函数可能是基于概率的(基于统计学习)，也可能是确定的函数关系。

从数学角度，机器学习与生物学习的过程本质上是一样的，都是映射。机器学习得到的函数就是这种映射关系的体现。  
而映射关系从数学上来讲，就是空间搜索和函数的泛化。需要从数据中过滤噪声(noise),提取事物的特征(feature)，  
最后通过特征(feature)确定映射关系。

**机器学习的一般流程**

[[](https://holbrook.github.io/2017/10/13/ML_index/machine_learning_workflow.png)](https://holbrook.github.io/2017/10/13/ML_index/machine_learning_workflow.png" \o "机器学习流程)

[机器学习流程](https://holbrook.github.io/2017/10/13/ML_index/machine_learning_workflow.png" \o "机器学习流程)

机器学习问题，通常可以分为三个步骤:

1. 数据准备与预处理
2. 模型选择与训练
3. 模型验证与参数调优

**输入空间、特征空间和输出空间**

输入空间和输出空间其实就是输入和输出所有可能取值的集合。  
输入和输出空间可以是有限元素的集合，也可以是整个欧式空间。  
输入空间和输出空间可以是一个空间，也可以是不同的空间；通常情况下，输出空间要远远小于输入空间。

特征空间是所有特征向量存在的空间。特征空间的每一维对应于一个特征。  
有时假设输入空间与特征空间为相同的空间，对它们不予区分；  
有时假设输入空间与特征空间为不同的空间，将实例从输入空间映射到特征空间。模型实际上都是定义在特征空间上的。

**机器学习的分类**

**按照输入空间分类**

根据输入空间的数据类型，可以分为：

* 具体特征（Conctrete Feature）

输入X的每一维特征都被人类进行的整理和分析，这种分析常常是与专业领域关联的。

比如对硬币分类需要的特征是（大小，重量）；对信用分级需要的特征是客户的基本信息。  
这些特征中已经蕴含了人的思考。

* 原始特征（Raw Features）

需要人或者机器（深度学习，Deep Learning）进行转化，将原始特征转化成为具体的特征，  
在机器视觉和声音信号的辨识都是属于该类。

例如，数字识别中，原始特征是图片的像素矩阵；声音识别中的声波信号；机器翻译中的每个单词。

其中，深度学习是通过机器自动的进行特征提取的。  
它需要有大量的资料或者非监督式学习的方式去学习如何从中抽取出非常具体的特征。  
深度学习通过组合低层特征形成更加抽象的高层表示属性类别或特征，以发现数据的分布式特征表示。

* 抽象特征（Abstract Feature）

抽象特征通常没有任何真实意义，更需要认为地进行特征转化、抽取和再组织。

例如，预测某用户对电影的评分，原始数据是(userid, itemid, rating)，  
rating 是训练数据的标签，相当于y。  
这里的(userid, itemid)本身对学习任务是没有任何帮助的，  
我们必须对数据所进一步处理、提炼、再组织。

离散特征具有丰富的自然含义；原始特征有简单的自然含义；抽象特征没有自然含义。

离散特征一般只需要简单选取就够了，  
原始特征、抽象特征都需要再处理，此过程称为特征工程(feature engineering)，  
是机器学习、数据挖掘中及其重要的一步。

**按照输出空间分类**

* 二值分类 Classification）问题

即“是非问题(say YES/NO)”，其输出空间Y={-1，+1}

二类分类应用非常广泛，比如判断是否为垃圾邮件、广告投资是否能盈利、学习系统上在下一题答题是否正确。  
二类分类在机器学习中地位非常重要，是其他算法的基础。

* 多值分类 Classification）问题

输出空间为有限个类别： Y={1,2，…,K}

多类别分类的应用，比如根据一张图片，得出图片中是苹果、橘子还是草莓等；  
还有像Google邮箱，将邮件自动分成垃圾邮件、重要邮件、社交邮件、促销邮件等。  
多类别分类在视觉或听觉的辨识中应用很广泛。

* 回归问题（Regression）

输出空间为实数范围( Y= R ) ，或区间 (Y = [lower, upper] )。 输出是无限种可能。

回归分析在股票价格预测和天气气温的预测上被广泛应用。

* 结构学习问题（Structured Learning）/序列标记（Sequence Tagging）问题

输出是一个结构化数据(structures)，比如句子中每个单词的词性。  
可以成为 hyperclass， 通常难以显示地定义该类。

该学习模型也可以看做是多类别学习的一种，这里可能涉及到数量很大的类别。

在自然语言处理中，自动的词性标注是很典型的结构学习的例子。比如给定机器一个句子，由于词语在不同的句子当中可能会有不同的词性，所以该方法是用来对句子的结构特性的理解。这种学习方法可以被看做是多类别分类，但与多类别分类不同的是，其目标的结构种类可能规模很大，其类别是隐藏在句子的背后的。

这种结构学习的例子还有例如说，生物中蛋白质3D立体结构，自然语言处理方面。

结构学习在有些地方被描述成标注（tagging）问题，标注问题的输入是一个观测序列，输出是一个标记序列或状态序列。标注问题的目标在于学习一个模型，使它能够对观测序列给出标记序列作为预测。

**按照应用类型**

* 关联
* 回归(Regression)
* 密度估计(Density Estimation)

输出是一个概率密度函数或者概率函数。可以近似看做无监督的有界回归问题。

典型例子是根据位置的交通情况报告，预测事故危险多发的区域。

* 分类(Classification)
* 聚类(Clustering)

数据的类别是不知道的，根据某种规则得到不同的分类。可以近似看做无监督的多类别分类。

比如将网络上各式各样的文章分成不同的主题;  
商业公司根据不同的顾客的资料，将顾客分成不同的族群。

* 异常检测(Outlier Detection）

可以看做近似的无监督二类分类问题。

比如根据网络的日志的情况，检测是否有异常入侵行为，这是一个极端的“是非题”，可以用非监督的方法来得出。

* 预测
* 时间序列
* 智能优化

**训练/学习方式**

根据学习的方式，可以分为有监督学习和无监督学习。其中有监督学习，需要训练数据中已经人工做好标签(Label)。  
另外，还有介于二者之间的半监督学习，以及通过结果进行反馈的增强学习。

* 有监督学习(Supervised Learning)

有监督学习需要训练数据，训练数据中每个样本都包含人工做好的标签(Label)。  
在建立预测模型的时候，监督式学习建立一个学习过程，  
将预测结果与“训练数据”的实际结果进行比较，不断的调整预测模型，直到模型的预测结果达到一个预期的准确率。

监督学习常用于分类问题和回归问题。

* 无监督学习（Unsupervised Learning）

训练数据中没有特别的标识，通过学习推断出数据的一些内在结构。

无监督学习常用于关联，聚类，密度估计(density estimation),异常检测。

* 半监督学习（Semi-Supervised Learning)

半监督学习是监督学习与无监督学习相结合的一种学习方法，  
主要考虑如何利用少量的标注样本和大量的未标注样本进行训练和分类的问题。

半监督学习的输入数据中，只有少量标注数据，算法利用未标记的大量数据提升机器学习算法的表现效果。

这类学习模型可以用来进行预测，  
但是模型首先需要学习数据的内在结构以便合理的组织数据来进行预测。

半监督学习的算法一般是对监督式学习算法的延伸，也常用于分类问题和回归问题。主要有五类：

* 1. 基于概率的算法；
  2. 在现有监督算法基础上作修改的方法；
  3. 直接依赖于聚类假设的方法；
  4. 基于多视图的方法；
  5. 基于图的方法。

这些算法首先试图对未标识数据进行建模，在此基础上再对标识的数据进行预测。  
如图论推理算法（Graph Inference）或者拉普拉斯支持向量机（Laplacian SVM.）等。

半监督学习的例子，比如人脸识别，可能只有一小部分人脸是被标记的，大部分是没有被标记的。

* 增强学习（Reinforcement Learning）

增强学习，通常知识是隐含的，无法直接表示什么是正确的。  
学习的方法是通过”惩罚“不好的结果，”奖励“好的结果来优化学习效果。

增强学习是一种以环境反馈作为输入的、特殊的、适应环境的机器学习方法。  
所谓强化学习是指从环境状态到行为映射的学习，以使系统行为从环境中获得的累积奖赏值最大。  
该方法不同与监督学习技术那样通过正例、反例来告知采取何种行为，  
而是通过试错（trial-and-error）的方法来发现最优行为策略。

在这种学习模式下，输入数据不是像监督模型那样作为对模型的反馈，  
而是作为一个检查模型对错的方式。输入数据直接反馈到模型，模型必须对此立刻作出调整。

增强学习的的输出不一定是真正想要得到的输出，而是用过奖励或者惩罚的方式来告诉这个系统做的好还是不好。

增强学习是一种”困难“但是非常自然的方式进行学习。常用于广告系统，扑克、棋类游戏，动态系统以及机器人控制等。

比如一个线上广告系统，可以看做是顾客在训练这个广告系统。这个系统给顾客投放一个广告，  
即可能的输出，而顾客有没有点或者有没有因为这个广告赚钱，这评定了这个广告投放的好坏。  
这就让该广告系统学习到怎么样去放更适合的广告。

常见算法包括Q-Learning以及时间差学习（Temporal difference learning）

在企业数据应用的场景下， 人们最常用的可能就是监督式学习和非监督式学习的模型。  
在图像识别等领域，由于存在大量的非标识的数据和少量的可标识数据，目前半监督式学习是一个很热的话题。 而强化学习更多的应用在机器人控制及其他需要进行系统控制的领域。

**根据学习协议（Protocol）分类**

* 批量学习(Batch learning)

”填鸭式“（duck feeding）, 将所有已知训练数据一次性批量输入给学习算法。

* 在线学习(online learning)

”循序渐进式“（passive sequential）,通过序列化地接收数据，循序渐进地学习，逐渐提高性能。

应用：垃圾邮件， 增强学习。

* 主动学习(active learning/learning by asking)

”提问式“（question asking）。可以看做是机器有问问题的能力。  
开始只有少量label, 通过有策略地”问问题“ 来提高性能。  
比如遇到xi, 不确定输出是否正确，则主动去确认yi 是什么，依次来提高性能。

当label的获取成本非常昂贵时，会利用此法。

比如电子相册中好友照片的标注，需要机器向人提问。