

第05课：机器是如何学习的？

各位同学，从本节开始，我们将进入《21 天入门机器学习》的第二个学习主题：“基本原理”，学时两天，共有 5 篇文章（05 课~09 课），对应一次打卡任务，大家可根据自己的时间来安排阅读和提交笔记。如有任何疑问，也可以在学习笔记中向我提问。下面我们就进入正题：

什么是机器学习？字面理解，就是让机器自己学会某种东西。

更准确一点，机器学习就是：让计算机程序（机器），不是通过人类直接指定的规则，而是通过自身运行，习得（学习）事物的规律，和事物间的关联。

对人类而言，一个概念对应的是具体的事物，一般而言，人类对其有所认知的事物，都不是孤立的，互相之间有着各种各样的关联。

比如说：

当我们对一个人说“苹果”的时候，TA 可能马上就会想到那个圆圆的、香甜的、有皮有核的水果。

也许还会想到香蕉、菠萝等其他水果；或是想到美味的苹果派、伊甸园里的故事、咬了苹果一口的白雪公主等等……

如果我们将 “苹果”这两个字输入给计算机，计算机并不会幻视出一个水果，也不会像人那样“意识到”这个单词的含义。

****计算机程序能够处理的，只有数值和运算。****

计算机程序不过是一段存储在硬盘上的0、1代码，运行时被读取进内存，CPU 根据代码转换成的指令来做一组特定的操作，让这些0、1数字通过逻辑电路进行若干运算后，生成计算结果。

****要让一段程序了解客观世界变化万千的事物，则必须将这些事物数值化，将事物的变化和不同事物之间的关联转化为运算。****

当若干现实世界的事物转换为数值之后，计算机通过在这些数值之上的一系列运算来确定它们之间的关系，再根据一个全集之中个体之间的相互关系，来确定某个个体在整体（全集）中的位置。

我们继续前面的举例——很可能，我说“苹果”的时候，有些人首先想到的不是苹果，而是乔帮主创立的科技公司。

但是，我继续说：“苹果一定要生吃，蒸熟了再吃就不脆了。”——在这句话里，“苹果”一词确定无疑指的是水果，而不是公司。

因为在我们的知识库里，都知道水果可以吃，但是公司不能吃。出现在同一句话中的 “吃”对“苹果”起到了限定作用——这是人类的理解。

对于计算机， “苹果”被输入进去的时候，就被转化为一个数值 Va。

经过计算，这个数值和对应“吃”的数值 Ve 产生了某种直接的关联，而同时和 Ve 产生关联的还有若干数值，它们对应的概念可能是“香蕉”（Vb）、“菠萝”（Vp）、“猕猴桃”（Vc）……

那么据此，计算机就会发现 Va、Vb、Vp、Vc 之间的某些关联（怎么利用这些关联，就要看具体的处理需求了）。

****NOTE 1****：说到数值，大家可能本能地想到 int、double、 float……但实际上，如果将一个语言要素对应成一个标量的话，太容易出现两个原本相差甚远的概念经过简单运算相等的情况。

假设“苹果“被转化为2，而“香蕉”被转化为4，难道说两个苹果等于一个香蕉吗？

因此，一般在处理时会将自然语言转化成 n 维向量。只要转化方式合理，规避向量之间因为简单运算而引起歧义的情况还是比较容易的。

这种现实世界和计算机之间，概念 <-> 数值、关系 <-> 运算的映射，造就了机器可以自主学习获得事物规律的可能。

### 机器学习的基本原理

既然机器是有可能自己学习事物规律的，那么如何才能让它学到规律呢？ 我们先来看一个故事：

猫妈妈让小猫去捉老鼠，小猫问：“老鼠是什么样子啊？”

猫妈妈说：“老鼠长着胡须。”结果小猫找来一头大蒜。

猫妈妈又说：“老鼠四条腿。”结果小猫找来一个板凳。

猫妈妈再说：“老鼠有一条尾巴。”结果小猫找来一个萝卜。

在这个故事里，小猫就是一个****基于规则的（Rule-Based）**** 计算机程序，它完全听命于开发者猫妈妈的指令行事。但是因为三次指令都不够全面，结果，三次都得出了错误的结果。

如果要把小猫变成一个****基于机器学习模型的（Model-Based）**** 计算机程序，猫妈妈该怎么做呢？

猫妈妈应该这样做，应该给小猫看一些照片，并告诉小咪咪，有些是老鼠，有些不是，比如下面这些：

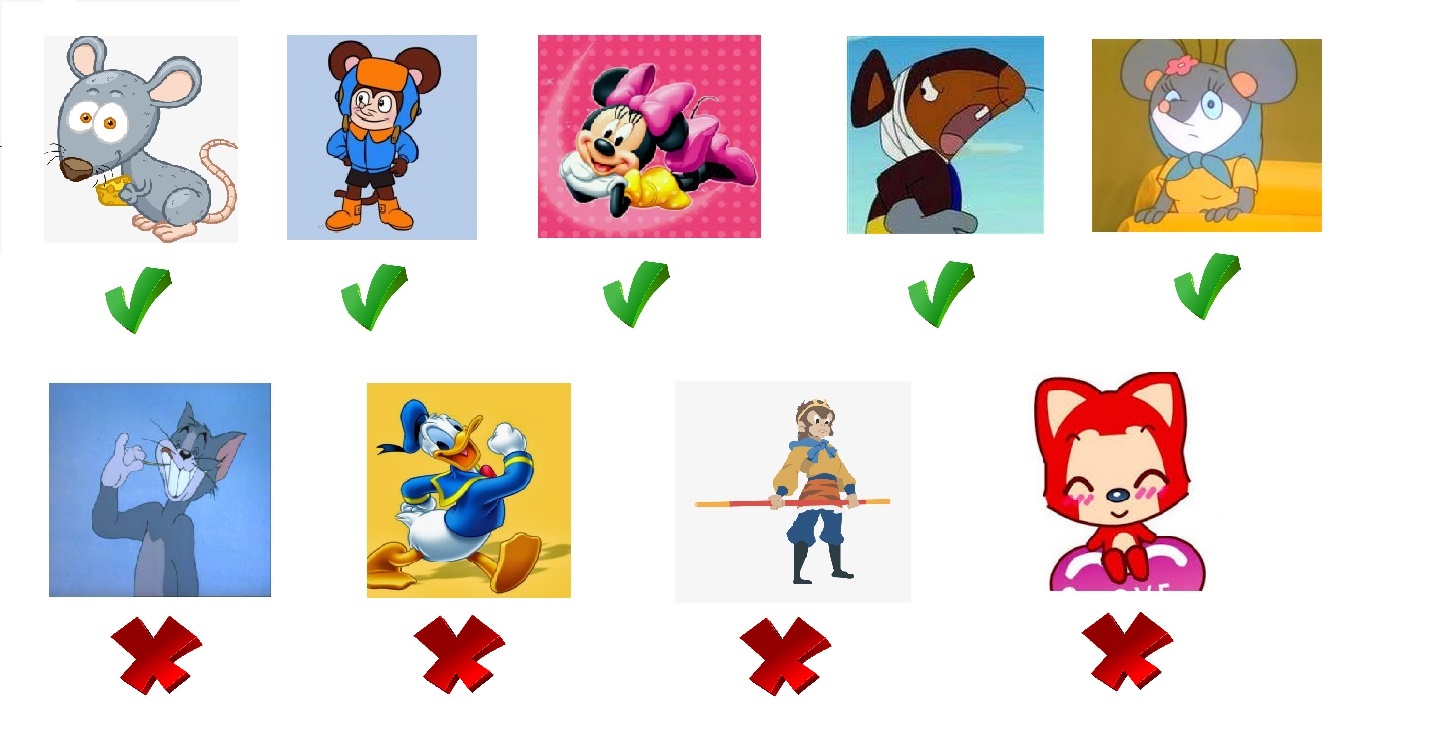


图1 老鼠和其他动物

猫妈妈可以先告诉它：要注意老鼠的耳朵、鼻子和尾巴。

然后小猫通过对比发现：老鼠的耳朵是圆的，别的动物要么没耳朵，要么不是圆形耳朵；老鼠都有尾巴，别的动物有的有，有的没有；老鼠的鼻子是尖的，别的动物不一定是这样。

然后小猫就用自己学习到的：“老鼠是圆耳朵、有尾巴、尖鼻子的动物”的信念去抓老鼠，那么小猫就成了一个“老鼠分类器”。

小猫（在此处类比一个计算机程序）是****机器（Machine）****，它成为“老鼠分类器“的过程，就叫做****学习（Learning）****。

猫妈妈给的那些照片是用于学习的****数据（Data）****。

猫妈妈告知要注意的几点，是这个分类器的****特征（Feature）****。

学习的结果——老鼠分类器——是一个****模型（Model）****。这个模型的类型可能是逻辑回归，或者朴素贝叶斯，或者决策树……总之是一个分类模型。

小猫思考的过程就是****算法（Algorithm）****。

无论有监督学习，还是无监督学习，都离不开这三要素。

什么叫有监督学习？什么又叫无监督学习？我们来解释一下。

### 有监督学习 vs 无监督学习

#### 有监督学习

小猫通过学习成为“老鼠分类器“，就属于典型的有监督学习（Supervised Learning）。

大家请看上面的图1，其中不仅有老鼠、非老鼠的照片，而且在每张老鼠照片的下面还有一个绿色的对勾，标明这是一只老鼠；而非老鼠的照片下是一个红叉子，说明这不是一只老鼠。

每一张照片是一个数据样本（Sample）。下面的对勾或者叉子，就是这个数据样本的标签（Label）。而给样本打上标签的过程，就叫做标注（Labeling）。

标注这件事情，机器学习程序自己是解决不了的，必须依靠外力。这些勾叉都是猫妈妈打上去的，而不是小猫。

小猫通过学习过程获得的，就是给图片打勾打叉的能力。如果小猫已经能够给图片打勾或者打叉了，就说明它已经是一个学习成的模型了。

这种通过标注数据进行学习的方法，就叫做有监督学习或直接叫监督学习（Supervised Learning）。

#### 无监督学习

反过来，如果用于学习的数据只有样本，没有标签，那么通过这种无标注数据进行学习的方法，就叫做无监督学习（Unsupervised Learning）

比如说，我们有这样六个样本：



图2 小马宝莉六女主

要做的事情是，根据她们的体貌区分她们的种族。

明明是六匹马，为什么还要分种族？因为在小马（《Little Pony》）的世界里，小马（Pony）在马这个大类之下，还有细分的种族。

我可以告诉你，要关注的特征（Features）是：独角和翅膀。而她们一共可以被归为3个小马种族。

这样你是不是就能分出来了——两个有独角的一组（她们叫独角兽）；两个有翅膀的一族（她们叫飞马）；另外两个很正常的一组（她们叫陆马）。

OK！聚类完成。我们经历了一次有趣的无监督学习的过程。

[邀请好友一起学，获得 25% 返现奖励](https://gitbook.cn/m/mazi/columns/5bc6ac7442d7d32f50f19a98/topics/5bf280bffd72950cafdccd78)

[IMG_259](https://gitbook.cn/m/mazi/comp/column?columnId=5bc6ac7442d7d32f50f19a98%26tag=2#catalog)