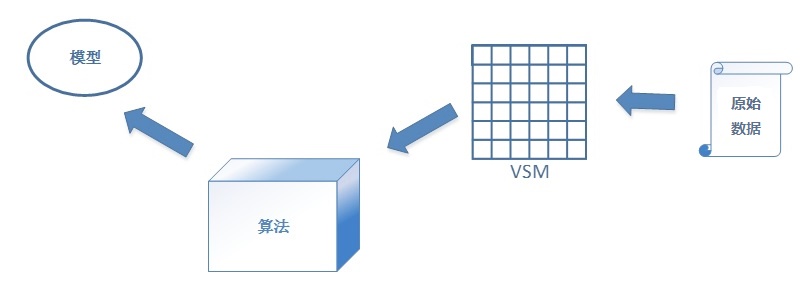


第06课：机器学习三要素之数据、模型、算法

机器学习三要素包括数据、模型、算法。简单来说，这三要素之间的关系，可以用下面这幅图来表示：



总结成一句话：****算法通过在数据上进行运算产生模型****。

下面我们先分别来看三个要素。

### 数据

关于数据，其实我们之前已经给出了例子。

#### 源数据

上一篇中，****图1老鼠和其他动物****和****图2小马宝莉六女主****就是现实中的两份样本集合。如果我们要训练“老鼠分类器”，或者做“小马种族聚类” 分析的话，它们就是原始数据（Raw Data）。

不过，我们之前也说了，计算机能够处理的是数值，而不是图片或者文字。

#### 向量空间模型和无标注数据

那么，我们就需要构建一个**向量空间模型（Vector Space Model，VSM）。**VSM 负责将格式（文字、图片、音频、视频）转化为一个个向量。

然后开发者把这些转换成的向量输入给机器学习程序，数据才能够得到处理。

比如图2小马宝莉中的6为女主角，我们要给她们做聚类，而且已经知道了，要用她们的两个特征来做聚类，这两个特征就是：独角和翅膀。

那么我们就可以定义一个二维的向量 A=[a\_1,a\_2]。a\_1 表示是否有独角，有则 a\_1 = 1, 否则 a\_1 = 0。而 a\_2 表示是否有翅膀。

那么按照这个定义，我们的6匹小马最终就会被转化为下面6个向量：

X\_1 = [1,0]

X\_2 = [0,0]

X\_3 = [0,0]

X\_4 = [0,1]

X\_5 = [0,1]

X\_6 = [1,0]

这样，计算机就可以对数据 X\_1，……，X\_6 进行处理了。这6个向量也就叫做这份数据的****特征向量（Feature Vector）****。

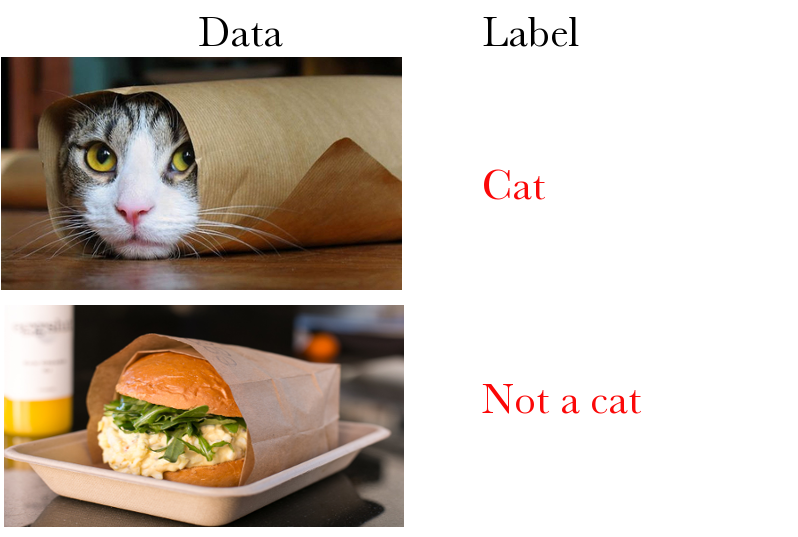
这是****无标注数据****。

#### 有标注数据

和无标注对应的是有标注。

数据标注简单而言就是给训练样本打标签。这个标签是依据我们的具体需要给样本打上的。

比如，我们要给一系列图标做标注，所有图片分为两类：“猫”或者“不是猫”。那么就可以标注成下图这样：



我们把样本的标签用变量 y 表示，一般情况下，y 都是一个离散的标量值。

标注数据当然也要提取出特征向量 X。****每一个标注样本既有无标注样本拥有的 X，同时还比无标注样本多了一个 y。**** 例如：

我们用三维特征向量 X 表示老鼠分类器的源数据，每一维分别对应“耳朵是圆的”、“有细长尾巴”、“是尖鼻子”。同时用一个整型值 y 来表示是否为老鼠，是的话 y=1，否则 y=0。

那么图1老鼠和其他动物对应的数据就是这样的：

X\_1 = [1,1,1]; y = 1

X\_2 = [1,1,1]; y = 1

X\_3 = [1,1,1]; y = 1

X\_4 = [1,1,1]; y = 1

X\_5 = [1,1,1]; y = 1

X\_6 = [0,1,1]; y = 0

X\_7 = [0,0,0]; y = 0

X\_8 = [0,1,0]; y = 0

X\_9 = [0,0,1]; y = 0

在数据转换到 VSM 之后，机器学习程序要做的就是把它交给算法，通过运算获得模型。

大家已经看到了，我们之所以能把具体的一系列童话人物转化为2维或者3维的向量，是因为我们已经确定了对某些人物用哪些特征。

这里其实有两步：

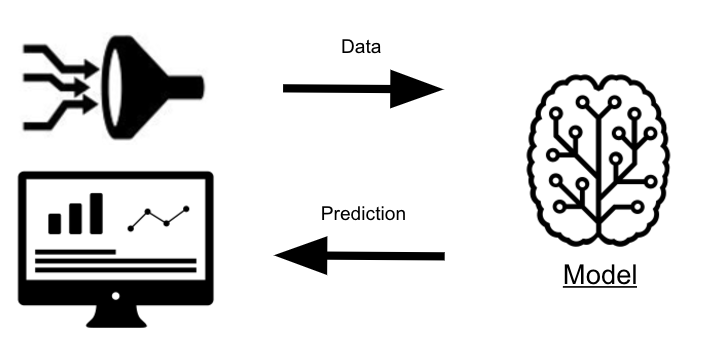
1. 确定用哪些特征来表示数据；
2. 确定用什么方式表达这些特征。

这两步做的事情就叫做****特征工程****。有了特征工程，才有下一步的 VSM 转换。

在机器学习中，特征工程是非常重要的。以后的章节中，我们会单独讲。

### 模型

#### 模型是什么



模型是机器学习的结果，这个学习过程，称为训练（Train）。

一个已经训练好的模型，可以被理解成一个函数： y=f(x)。

我们把数据（对应其中的 x）输入进去，得到输出结果（对应其中的 y）。

这个输出结果可能是一个数值（回归），也可能是一个标签（分类），它会告诉我们一些事情。

比如，我们用老鼠和非老鼠数据训练出了老鼠分类器。这个分类器就是分类模型，它其实是一个函数（具体是什么函数和我们选用的模型类型有关，我们下面再说）。

当这个分类函数被确定了以后，又有一个新数据出现了，比如它：



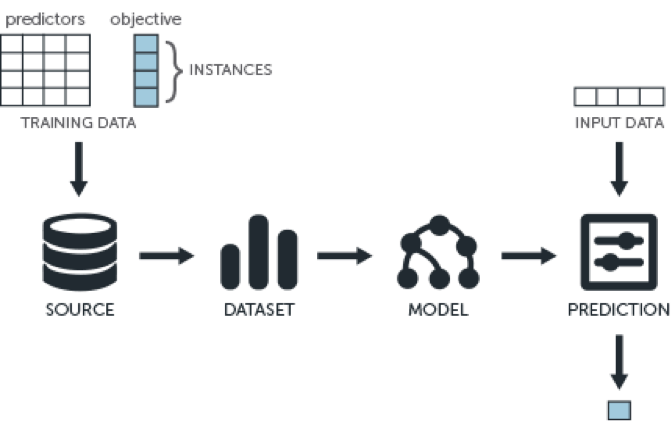
这时候，预测程序（将训练好的模型应用到数据上的过程叫预测） 先将喜羊羊转化到 VSM，变成 X = [0,0,0]，然后将它输入给模型，得到结果 y’= f(X)。

如果 y’=0，则说明我们的分类模型对喜羊羊的判断是：不是老鼠。若y’= 1，则老鼠分类器把喜羊羊也当成了老鼠。

当然我们都喜欢把喜羊羊分成“不是老鼠”的分类器，而不是相反。这里涉及到模型性能（Performance）的衡量问题，我们后面会单独讲。

我们先来看看——

#### 模型是怎么得到的？



模型是基于数据，经由训练得到的。

训练又是怎么回事？

模型是函数：y=f(x)，x 是其中的自变量，y 是因变量。

从 x 计算出 y 要看 f(x) 的具体形式是什么，它有哪些参数，这些参数的值都是什么。

在开始训练的时候，我们有一些样本数据。如果是标注数据，这些样本本身既有自变量 x（特征）也有因变量 y（预期结果）。否则就只有自变量 x。

对应于 y=f(x) 中的 x 和 y 取值实例。

这个时候，因为已经选定了模型类型，我们已经知道了 f(x) 的形制，比如是一个线性模型 y=f(x)=ax2+bx+c，但却不知道里面的参数 a、b、c 的值。

****训练****就是：根据已经被指定的 f(x) 的具体形式——模型类型，结合训练数据，计算出其中各个参数的具体取值的过程。

训练过程需要依据某种章法进行运算。这个章法，就是算法。

### 算法

有监督和无监督学习的算法差别甚大。因为我们在日常中主要应用的还是有监督学习模型，所以就先以此为重点，进行讲解。

有监督学习的****目标****就是：让训练数据的所有 x 经过 f(x) 计算后，获得的 y’ 与它们原本对应的 y 的差别尽量小。

**我们需要用一个函数来描述 y’ 与 y 之间的差别，这个函数叫做损失函数（Loss Function）L（y, y’）= L(y, f(x))。**

Loss 函数针对一个训练数据，对于所有的训练数据，我们用代价函数（Cost Function）来描述整体的损失。

代价函数一般写作：J（theta）——注意，代价函数的自变量不再是 y 和 f(x)，而是变成了 theta，theta 表示 f(x) 中所有待定的参数（theta 也可以是一个向量，每个维度表示一个具体的参数）！

至此，我们终于得到了一个关于我们真正要求取的变量（theta）的函数。而同时，既然 J（theta） 被称为代价函数，顾名思义，它的取值代表了整个模型付出的代价，这个代价自然是越小越好。

因此，我们也就有了学习的目标（也称为****目标函数****）：argmin J（theta）—— 最小化J（theta）。

能够让 J（theta） 达到最小的 theta，就是最好的 theta。当找到最好的 theta 之后，我们把它带入到原 f(x)，使得 f(x) 成为一个完整的 x 的函数，也就是最终的模型函数。

怎么能够找到让 J（theta） 最小的 theta 呢？这就需要用到优化算法了。

具体的优化算法有很多，比如：**梯度下降法（Gradient Descent）、共轭梯度法（Conjugate Gradient）、牛顿法和拟牛顿法、模拟退火法（Simulated Annealing）** 等等。

其中最常用的是梯度下降法，后面章节会专门讲解梯度下降法。

一般来说，****算法是机器学习和深度学习中最具技术含量的部分****。企业中的“算法工程师”，要求最高，待遇也最好。那些动辄年薪百万的职位，一般都会标明：做算法。

这些算法工程师的职责包括：研发新算法；针对现实问题构造目标函数，选取并优化算法求解；将他人研究的最新算法应用到自己的业务问题上。

在这里需要****强调****一点：要得到高质量的模型，算法很重要，但往往（尤其是在应用经典模型时）更重要的是数据。

有监督学习需要标注数据。因此，在进入训练阶段前必须要经过一个步骤：人工标注。标注的过程繁琐且工作量颇大，却无法避免。

人工标注的过程看似简单，但实际上，标注策略和质量对最终生成模型的质量有直接影响。

往往能够决定有监督模型质量的，不是高深的算法和精密的模型，而是高质量的标注数据。

[邀请好友一起学，获得 25% 返现奖励](https://gitbook.cn/m/mazi/columns/5bc6ac7442d7d32f50f19a98/topics/5bf280e1fd72950cafdccd7c)

[IMG_262](https://gitbook.cn/m/mazi/comp/column?columnId=5bc6ac7442d7d32f50f19a98%26tag=2#catalog)