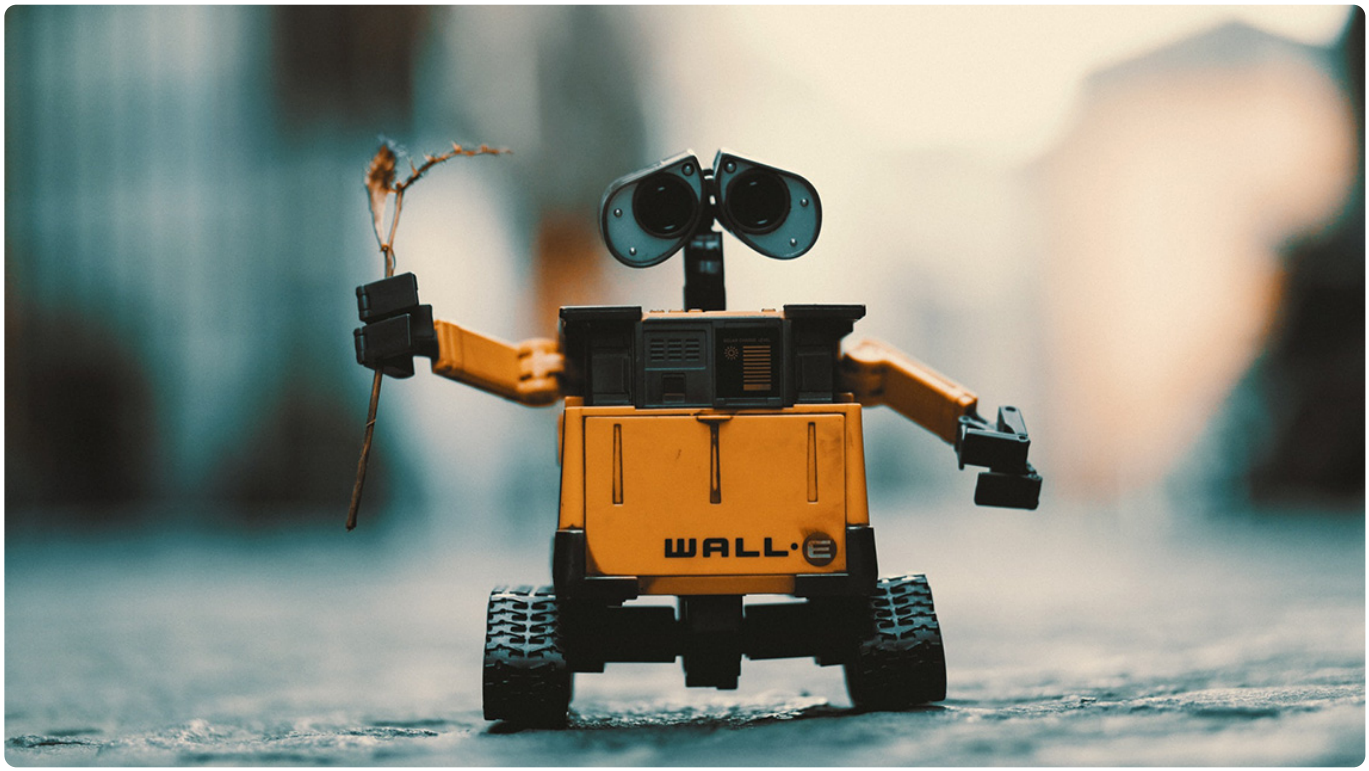
004 | 数据科学家基础能力之机器学习

2017-10-11 洪亮劼

AI技术内参

进入课程 >



讲述：初明明

时长 11:39　大小 5.34M

想要成为合格的，或者更进一步成为优秀的人工智能工程师或数据科学家，机器学习的各种基础知识是必不可少的。然而，机器学习领域浩如烟海，各类教材和入门课程层出不穷。特别是机器学习基础需要不少的数学知识，这对于想进入这一领域的工程师而言，无疑是一个比较高的门槛。

今天，我来和你聊一聊如何学习和掌握机器学习基础知识，又如何通过核心的知识脉络快速掌握更多的机器学习算法和模型。

监督学习和无监督学习

要问机器学习主要能解决什么问题，抛开各式各样的机器学习流派和层出不穷的算法模型不谈，**机器学习主要解决的是两类问题：监督学习和无监督学习。掌握机器学习，主要就是学**

习这两类问题，掌握解决这两类问题的基本思路。

什么是解决这两类问题的基本思路呢？基本思路，简而言之就是“套路”。放在这里的语境，那就是指：

1. 如何把现实场景中的问题抽象成相应的数学模型，并知道在这个抽象过程中，数学模型有怎样的假设。
2. 如何利用数学工具，对相应的数学模型参数进行求解。
3. 如何根据实际问题提出评估方案，对应用的数学模型进行评估，看是否解决了实际问题。

这三步就是我们学习监督学习和无监督学习，乃至所有的机器学习算法的核心思路。机器学习中不同模型、不同算法都是围绕这三步来展开的，我们不妨把这个思路叫作“三步套路”。

那什么是监督学习呢？监督学习是指这么一个过程，我们通过外部的响应变量（Response Variable）来指导模型学习我们关心的任务，并达到我们需要的目的。这也就是“监督学习”中“监督”两字的由来。

也就是说，**监督学习的最终目标，是使模型可以更准确地对我们所需要的响应变量建模**。比如，我们希望通过一系列特征来预测某个地区的房屋销售价格，希望预测电影的票房，或者希望预测用户可能购买的商品。这里的“销售价格”、“电影票房”以及“可能购买的商品”都是监督学习中的响应变量。

那什么是无监督学习呢？通常情况下，无监督学习并没有明显的响应变量。**无监督学习的核心，往往是希望发现数据内部的潜在结构和规律，为我们进行下一步决断提供参考**。典型的无监督学习就是希望能够利用数据特征来把数据分组，机器学习语境下叫作“聚类”。

根据不同的应用场景，聚类又有很多变种，比如认为某一个数据点属于一个类别，或者认为某一个数据点同时属于好几个类别，只是属于每个类别的概率不同等等。

无监督学习的另外一个作用是为监督学习提供更加有力的特征。通常情况下，无监督学习能够挖掘出数据内部的结构，而这些结构可能会比我们提供的数据特征更能抓住数据的本质联系，因此监督学习中往往也需要无监督学习来进行辅助。

我们简要回顾了机器学习中两大类问题的定义。在学习这两大类模型和算法的时候，有这么一个技巧，就是要不断地回归到上面提到的基本思路上去，就是这个“三步套路”，反复用这三个方面来审视当前的模型。另外，我们也可以慢慢地体会到，任何新的模型或者算法的诞生，往往都是基于旧有的模型算法，在以上三个方面中的某一个或几个方向有所创新。

监督学习的基础

监督学习的基础是三类模型：

1. 线性模型
2. 决策树模型
3. 神经网络模型

掌握这三类模型就掌握了监督学习的主干。利用监督学习来解决的问题，占所有机器学习或者人工智能任务的绝大多数。这里面，有 90% 甚至更多的监督学习问题，都可以用这三类模型得到比较好的解决。

这三类监督学习模型又可以细分为处理两类问题：

1. 分类问题
2. 回归问题

分类问题的核心是如何利用模型来判别一个数据点的类别。这个类别一般是离散的，比如两类或者多类。**回归问题的核心则是利用模型来输出一个预测的数值。**这个数值一般是一个实数，是连续的。

有了这个基本的认识以后，我们利用前面的思路来看一下如何梳理监督学习的思路。这里用线性模型的回归问题来做例子。但整个思路可以推广到所有的监督学习模型。

线性回归模型（Linear Regression）是所有回归模型中最简单也是最核心的一个模型。我们依次来看上面所讲的“三步套路”。

首先第一步，我们需要回答的问题是，线性回归对现实场景是如何抽象的。顾名思义，线性回归认为现实场景中的响应变量（比如房价、比如票房）和数据特征之间存在线性关系。而线性回归的数学假设有两个部分：

1. 响应变量的预测值是数据特征的线性变换。这里的参数是一组系数。而预测值是系数和数据特征的线性组合。
2. 响应变量的预测值和真实值之间有一个误差。这个误差服从一个正态（高斯）分布，分布的期望值是 0，方差是σ的平方。

有了这样的假设以后。第二步就要看线性回归模型的参数是如何求解的。这里从历史上就衍生出了很多方法。比如在教科书中一般会介绍线性回归的解析解（Closed-form Solution）。线性回归的解析解虽然简单优美，但是在现实计算中一般不直接采用，因为需要对矩阵进行逆运算，而矩阵求逆运算量很大。解析解主要用于各种理论分析中。

线性回归的参数还可以用数值计算的办法，比如梯度下降（Gradient Descent）的方法求得近似结果。然而梯度下降需要对所有的数据点进行扫描。当数据量很多的时候，梯度下降会变得很慢。于是随机梯度下降（Stochastic Gradient Descent）算法就应运而生。随机梯度下降并不需要对所有的数据点扫描后才对参数进行更新，而可以对一部分数据，有时甚至是一个数据点进行更新。

从这里我们也可以看到，**对于同一个模型而言，可以用不同的算法来求解模型的参数。这是机器学习的一个核心特点。**

最后第三步，我们来看如何评估线性回归模型。由于线性回归是对问题的响应变量进行一个实数预测。那么，最简单的评估方式就是看这个预测值和真实值之间的绝对误差。如果对于每一个数据点我们都可以计算这么一个误差，那么对于所有的数据点而言，我们就可以计算一个平均误差。

上述对于线性回归的讨论可以扩展到监督学习的三类基本模型。这样你就可以很快掌握这些模型的特点和这些模型算法之间的联系。

无监督学习的基础

现实中绝大多数的应用场景并不需要无监督学习。然而无监督学习中很多有价值的思想非常值得初学者掌握。另外，**无监督学习，特别是深度学习支持下的无监督学习，是目前机器学习乃至深度学习的前沿研究方向。**所以从长远来看，了解无监督学习是非常必要的。

我们前面说到，无监督学习的主要目的就是挖掘出数据内在的联系。这里的根本问题是，不同的无监督学习方法对数据内部的结构有不同的假设。因此，无监督学习不同模型之间常常

有很大的差别。在众多无监督学习模型中，聚类模型无疑是重要的代表。了解和熟悉聚类模型有助于我们了解数据的一些基本信息。

聚类模型也有很多种类。这里我们就用最常见的、非常重要的**K 均值算法**（K-means），来看看如何通过前面讲过的“三步套路”来掌握其核心思路。

首先，K 均值算法认为数据由 K 个类别组成。每个类别内部的数据相距比较近，而距离所有其他类别中的数据都比较遥远。这里面的数学假设，需要定义数据到一个类别的距离以及距离函数本身。在 K 均值算法中，数据到一个类别的距离被定义为到这个类别的平均点的距离。这也是 K 均值名字的由来。而距离函数则采用了欧几里得距离，来衡量两个数据点之间的远近。

直接求解 K 均值的目标函数是一个 NP 难（NP-hard）的问题。于是大多数现有的方法都是用迭代的贪心算法来求解。

一直以来，对聚类问题、对无监督学习任务的评估都是机器学习的一个难点。无监督学习没有一个真正的目标，或者是我们之前提到的响应变量，因此无法真正客观地衡量模型或者算法的好坏。

对于 K 均值算法而言，比较简单的衡量指标就是，看所有类别内部的数据点的平均距离和类别两两之间的所有点的平均距离的大小。如果聚类成功，则类别内部的数据点会相距较近，而类别两两之间的所有点的平均距离则比较远。

以上我们通过“三步套路”的三个方面讨论了 K 均值算法的核心思路，这种讨论方法也适用所有的聚类模型和算法。

小结

当你可以熟练使用我今天介绍的“三步套路”，去分析更多监督学习和无监督学习的模型算法以后，对于基础的内容，也就是教科书上经常讲到的内容，你就可以去看这些内容究竟是在讲解这三个方面的哪个方面。

对于绝大多数模型来说，第一部分往往是最重要的，也就是说，这个模型究竟和现实问题的联系是什么。第二部分，也就是模型的求解，取决于模型本身的复杂度和成熟度，现在很多模型往往都有现成的软件包提供求解过程。而第三部分，模型的评估则在现实生产中至关重

要。**牢牢把握这三个方面，来对机器学习模型算法进行讨论，是成长为成熟数据科学家必不可少的过程。**

今天我为你讲了掌握机器学习基础知识的一些核心思路。一起来回顾下要点：第一，机器学习主要的任务有监督学习和无监督学习。这两种机器学习任务的很多模型和算法都可以用一个“三步套路”的思路来进行分析。第二，我们用线性回归作为例子探讨了如何用这个“三步套路”来分析监督学习的模型和算法。第三，我们用 K 均值聚类算法作为例子探讨了如何用“三步套路”来分析无监督学习的模型和算法。

最后，给你留一个思考题，在现实场景中，当你发现一个模型并没有很好地解决你的问题时，从这个“三步套路”的角度来看，究竟哪个方面最容易出问题？

欢迎你给我留言，和我一起讨论。

上一篇　003 | 数据科学家基础能力之概率统计

下一篇　005 | 数据科学家基础能力之系统

精选留言 (14)

写留言

RZ_diversi... 置顶

2017-10-19

9

我认为是第一步，如果针对现实问题的抽象出现了偏差，对抽象设定的假设有问题的话，后续步骤再怎么高效求解参数，评估模型准确性都没办法改正第一步出现的问题。因为模型实际解决的domain已经不一样了。

展开

damonhao 置顶

2017-10-18

5

最容易出问题的是对现实问题的抽象。如果抽象成功，在数学的范围内求解都是比较有保证的。ps：其实我是来抛砖引玉的。。。

橙子

2017-12-14

26

三步套路可以总结为：
1. 提出模型
2. 求解模型
3. 评估模型
如果求解的模型没有很好地解决问题，我觉得应该从两种方面考虑：如果模型在训练集...

展开

作者回复: 你总结得很有道理。

鬼猫猫

2017-11-08

3

这总结得太到位了，作为对机器学习有兴趣的外行人，虽然读了很多书，教程，但还是在云里雾里，看了本篇文章之后，对机器学习有了个整体上的认识。这个专栏订的太值了。

展开

作者回复: 谢谢。

JIA
2017-10-18

2

针对今天的思考题，我觉得最容易出问题的地方就是最重要的第一步，弄清楚模型和现实的联系。如果这一步有问题，那后面做得再好也是白费，方向就错了，当然没办法解决问题。

展开

mortimer
2018-10-04

1

虽然直觉上我也认为是第一步模型容易出问题,但是我在做人脸聚类的经验恰恰是困在第三部----我们花了大量的时间和精力,来设计评估数据模型的准确性,中间可能有硬编码导致的异常结果;也可能是最初设想的模型不够充分导致结果;更有我们设计出一些组合性的数学模型,就连显而易见的数学意义都找不到了,也就更加不好评估.
所以啊,第三步,如果评价模型和算法反而是最容易出问题,也需要反复检查,验证的.

展开

孤帆
2018-03-04

1

老师没有提标注，在《统计机器学习方法》中，周航老师将监督学习分为分类、回归、标注。而老师没有提“标注”，请问是什么原因呢?

套码的汉子
2017-12-07

1

实际应用中，应该是第一步来背锅的几率较大。第三步评估的标准，往往生产环境已经决定。而作者也说，第二步已经有许多现成算法，在实际开发中改进的空间不大。所以，我以前参与开发的测量软件都提供几个算法，一个算法测不准技术支持就会让客户换一个，直到测得准为止。。。

作者回复: 的确需要从第一步多找原因。

吴文敏
2017-10-19

1

最容易出问题应该是假设，也就是说现实的问题与所用模型的假设不一致。

展开

黄德平

2018-11-26

好的开始是成功的一半，第一步至关重要。
提出合理的模型，对问题的本质做合理的抽象是最关键的一步，结果的好坏往往从最开始就决定了

展开

帅帅

2018-10-20

如果模型效果不好，数据的问题往往会比较大；

如果欠拟合，那一般是模型的容量问题，这个比较简单，换用更大容量的算法即可；
如果过拟合，那很可能是数据量太小了，需要去找寻提取更多的特征输入；

展开

海滨

2018-03-01

刚订这个专栏，才读了几篇文章，就觉得已经值回票价，很赞～

展开

udisyue

2017-10-21

最重要的是第一步定义模型，最初定义的模型不够准确，验证结果也就要有偏差，所以才需要很多手段例如正则化等来对模型修正

gg

2017-10-17

厉害厉害

展开