写给大家看的机器学习书（第一篇）

八汰 · 9 个月前

1. 前言

> 这个世界不缺少专家，我只是期待他们中有一位能把事情说清楚。

机器学习很火。
机器学习专家很贵。
所有大型互联网公司都驾着机器学习的马车朝着人工智能前进。

这里所说的从业者可能是开发工程师，可能是产品经理，也可能是运营，他们与机器学习专家们在同一家公司工作，参与同一个项目，但机器学习算法对他们仍然像黑魔法一样，神秘又疑惑。

这样的局面未免让人沮丧，毕竟如果相对论都可以在高等教育中得到普及，有什么领域是复杂到没办法好好说清楚的呢。据我有限的观察，造成这个局面的原因无非两种：

1. 不少专业人士乐于将机器学习包装得晦涩曲折，以享用他人迷惑眼神中的优越感。
2. 很少人把机器学习以直接的、让人容易理解的方式说出来。有那么几个在这样做的人，面向的也是专业领域学习者而非一般的科普受众。

我鄙视第一种人类。
我希望所有的写作者都能够追逐Richard Stevens的光芒，把复杂的东西变简单，追求简洁明了，追求直接易懂。

这个系列文章，我将试着为开发工程师，产品经理、设计师、所有希望了解学习机器学习的人，介绍机器学习的原理、方法和实战技巧。
我追求它尽可能好理解的同时，也会保持它的准确度和实用度。理论方面，以周志华的《机器学习》西瓜书，林軒田機器學習系列课程(FoundationsTechniques)，Andrew Ng's Machine Learning 为学习资料，结合我个人的理解及日常与朋友同事的讨论。实用实战方面，我将以手机淘宝中第一款**** DAU (Daily Active User) 导购产品——有好货为例子，如果你从事导购或者电商相关工作那么对例子中的场景一定非常熟悉。如果你对导购并不了解也不用担心，讲解的重点仍是机器学习原理和方法的普世应用，理解了原理方法之后可以在任何适合机器学习的场景中进行实践。

这是这个系列的第一篇，看完这篇您将知道

- 什么是机器学习?
- 机器学到的到底是什么?
- 什么样的问题适合用机器学习来解决?

2. 什么是机器学习？机器学到的到底是什么?

2.1 什么是机器学习?

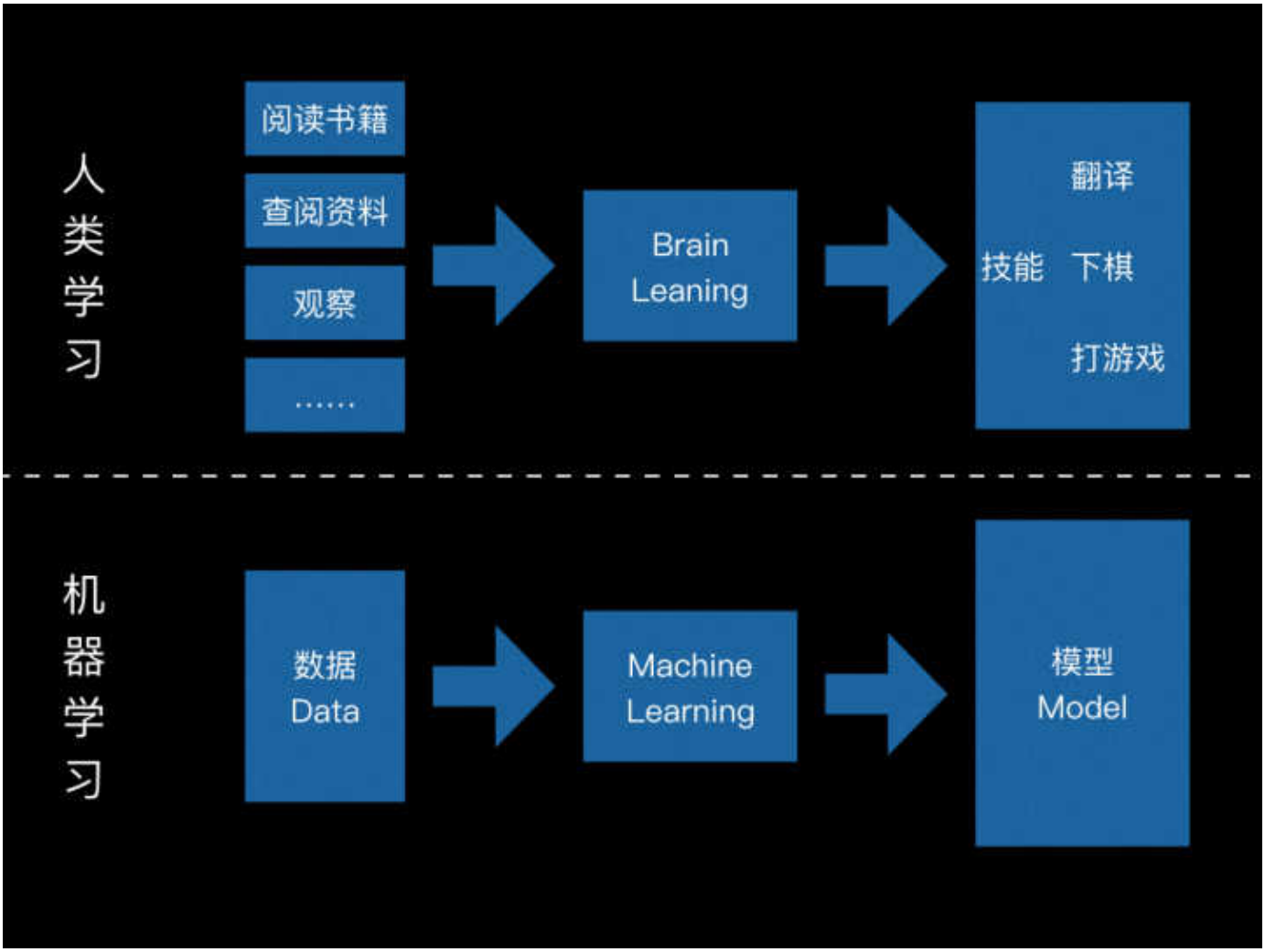
在讨论机器学习之前，我们首先看看人类是如何学习的。如图1上半部分所示，人类通过阅读书籍、查阅资料，观察得到信息，这些信息经过人脑学习，最后习得了某种技能。

机器学习也是类似，只不过机器学习的输入是**数据（Data）**，学到的结果叫**模型（Model）**（各法

- 数据（Data）
- 学习算法（Learning Algorithm）
- 模型（Model）

是机器学习的三个要素。

图1：



当然，上面的类比可能还比较抽象。下面就以“有好货”这个产品为例子，来理解一下机器学习的概念。

首先介绍下有好货（图2所示）。有好货是手机淘宝的一款导购产品，在首页第一屏就能看到产品入口。这款产品在2015-2016我们用了一年的时间，将每日用户数从****做到了****，成为第一个每日用户数破****的导购产品。这里面的一大秘诀就是将个性化推荐技术、机器学习技术与产

这里以有好货的“瀑布流”页面（图2中间）为例。这个页面是一个完全个性化的页面，不同的用户进入到有好货瀑布流页看到的商品推荐是不同的。
有限的屏幕空间，我们希望给每个用户展现他最有可能点击的商品。那什么商品是当前用户最可能点击的呢？这个预测就由机器学习完成。
类比图一的概念，这个场景下的**数据**、**学习算法**和**模型**分别对应着：

- **数据**：输入数据包括：用户对商品的浏览、点击历史行为数据以及相应的用户商品特征数据。这些历史数据记录了什么样的用户点击了什么样的商品，什么样的用户对什么样的商品看了没有点击。我们认为这些历史数据中蕴含了某种规律，希望机器学习能把这种规律挖掘出来，在将来面对新的用户和商品时，就能预测是否会点击。
- **学习算法**：机器学习算法有很多，逻辑回归、随机森林都适用于这个场景，但这里我们先不对具体的算法作展开，暂时读者只需要概念性的知道，基于输入数据执行“学习算法”便可产生模型（模型就代表了学习算法从数据中挖掘出的规律）。
- **模型**：学得模型之后，面对新的用户和商品，模型就能作出相应的判断，用户会点击还是不会点击。利用模型的这个“技能”，我们便可以做到给每个用户推荐他最感兴趣的商品了。

总而言之，机器学习是一门研究“学习算法”的学问，“学习算法”基于历史的经验数据产生模型，进而使计算机有了对新情况进行预判和预测的“技能”（比如预测用户的喜好或股票的涨跌）。

图2：

2.2 机器学到的到底是什么?

理解了机器学习的概念，我们知道机器学习无非三个要素(1)数据，(2)学习算法，(3)模型

1. **数据**很好理解，当我们希望预测用户是否会点击某个商品，就把历史上用户对商品的点击浏览行为喂给机器学习算法，希望从历史数据中中挖掘出**某种规律**。
2. **学习算法**有很多，上面提到过的逻辑回归、随机森林只是众多算法中的两种。事实上对各种不同学习算法的讨论是机器学习书籍的重点，一章介绍一种，就厚厚一本书了。读者不必着急，这部分我们将慢慢展开。
3. 在这一小节，我想重点讨论的是，我们说机器学习学得的模型可以预测用户是否会点击某个商品，可是**模型在机器内部到底是怎么表示的呢？机器学到的模型到底是什么？**许多人觉得机器学习非常神秘，是因为人类习得的技能并没有一个直观的展示形式，因此很难想象机器学到的模型到底是什么，其实答案非常简单：

映射，在数学的许多分支就等价于函数（*备注2*）。而函数，我们再熟悉不过了，给定一个（输入集合中的）元素，函数唯一对应（输出集合的）一个输出值。

比如函数 $f(x) = x^2$ ，给定任意实数 x ， x 的平方就是函数的输出。

比如函数 $f(x) = w_1 * x_1 + w_2 * x_2 + \ldots + w_d * x_d + b$ ，当 $w_1, w_2, \ldots, w_d$ 是确定的，那么给定一组 $x_1, x_2, \ldots, x_d$ ，就能唯一确定一个输出值 $f(x)$ 。（事实上这个就是最简单的一种机器学习模型——线性模型）。

而在有好货的例子中，机器学习学到的模型就是这样一个函数：

> 给定一个用户和商品，这个函数就能够唯一输出一个分数，表示用户点击该商品的可能性。

这就是机器学习的秘密。

3. 什么样的问题适合用机器学习来解决?

不少计算机科学专业的同学可能会有些疑惑，计算机科学在本科阶段教授了大量的算法——字符串匹配算法、排序算法、贪心、动态规划，算法导论厚厚一千多页，可这些都不属于机器学习的范畴，机器学习也不是计算机科学本科的必修课。

那到底算法导论中的算法跟机器学习算法有什么区别呢?
什么样的问题适合用机器学习来解决？什么情况需要使用机器学习呢?
答案是：

> 难以用规则解决的问题，可以尝试用机器学习来解决。（*备注3*）

算法导论中经典的排序问题，无论解法是快排还是归并排序，解法已经是一个确定的规则。但是机器学习问题，比如垃圾邮件识别，比如辨识一张图片中的物体是不是树叶，就很难用规则来解决。前者的规则难以穷举，后者则根本很难描述辨别树叶的规则。

因此，仍然以规则堆砌的观念来看待算法的朋友们注意了，
永远不要跟机器学习专家说：“加条规则呗”
永远不要跟机器学习专家说：“加条规则呗”
永远不要跟机器学习专家说：“加条规则呗”
God Bless You~

3.1 适合用机器学习解决的问题的必要条件

这其实道出了能用机器学习解决的问题需要具备这样的必要条件

> 有大量数据，并且数据中有隐藏的某种规律或模式

如果某些问题没有任何的规律，比如抛硬币，那么无论有多少数据也是不行的。

3.2 小测试

读到这里，不如试试看你对机器学习理解的怎么样了。
判断下面这些问题适不适合用机器学习解决。能不能用机器学习解决。

问题：预测下一次六合彩的中奖号码。
答案：不能用机器学习解决，因为跟投硬币一样摇奖是随机的，并没有规律。

问题：判断一个图形是否是圆。
答案：无需用机器学习解决，因为有明显的规则。

问题：预测股票的涨跌。
答案：可以用机器学习辅助交易并盈利。要是你发现自己能很好的解这个问题，请跟我做朋友吧：）

问题：预测一个10岁的小朋友长大了会喜欢的女孩子的类型。
答案：可能不能用机器学习解决，因为缺少“大量数据”这点必要条件。

4. 小结和预告

这是系列文章的第一篇，我们首先介绍了机器学习的基本概念，机器学习的三个要素：数据，学习算法，还有模型。

然后我们揭示了机器学到的模型，本质上就是一个映射，或者函数。

最后我们总结了机器学习适合解决的问题，是那些难以用规则解决的问题。并且机器学习的必要前提不仅是有大量的数据，而且需要数据中确实存在隐藏的某种规律，机器学习才能帮的上忙。

希望我有把事情说清楚，有任何疑惑或者问题，欢迎留言。我回答后会把FAQ附在每篇文章的后面。下一篇将细化具体的机器学习原理，可能会引出一个入门级的机器学习模型。您有什么希望了解学习的内容，也可以留言。

祝开心。

2. en.wikipedia.org/wiki/M... ↩

3. 另一种说法是，机器学习求解的问题，都难以用程序控制结构求解——程序控制结构包括顺序、分支、循环、跳转。 ↩

机器学习 大规模机器学习 人工智能

阿萨姆

“问题：预测一个10岁的小朋友长大了会喜欢的女孩子的类型”
...抖个激灵...把问题改成“小朋友长大是否喜欢女生？”马上把multi class 先简化成了一个binary的预测哈哈

9 个月前 3 赞

Jack

mark

9 个月前

八汰（作者） 回复 王旸 查看对话

坚持写下去这些内容应该都会涉及到。遗传算法我没有接触过，搜索了一些资料，如果用一个字回答，我想答案是：是。毕竟wikipedia机器学习方法下有遗传算法的条目：

Machine learning
and

器学习可能更倾向于从数据中挖掘模型或模式。

这个stackoverflow问题

里有人提到：“you can use Genetic Algorithms as an alternative to the Backpropagation algorithm to update weights in Neural Nets”。希望有全面了解的同学来科普遗传算法和机器学习的关系。

9 个月前

八汰（作者） 回复 阿萨姆 查看对话

哈，是否喜欢女生，这是个问题

9 个月前 1 赞

莫里007

很好的文章

9 个月前

从今以后我

好好，以关注

9 个月前

吴东知识产权律师

看了你这篇浅显易懂的文章，我终于明白那个Ross人工智能律师是怎么一回事了。我现在更确定，中国很快就会出现人工智能的产品，我想动手做一个，先学习写代码。

9 个月前 3 赞

HTML

谢谢。非常感谢你的分享。希望尽快更新出后续的文章。后面的什么时候能出来大概?

9 个月前

八汰（作者） 回复 HTML 查看对话

我写东西比较慢，争取一周有一篇。谢谢您的阅读~

9 个月前

同一秒钟88

9 个月前 1 页

下一页

文章被以下专栏收录

写给大家看的机器学习书

这世界不缺少专家，我只是期待有一位能把事情说清楚

进入专栏

推荐阅读

写给大家看的机器学习书（第三篇）

题记 —— 我们为何出发在开始这个系列文章的第三篇之前，为了对初次见面的朋友更友好，将这... 查看全文 >

八汰 · 9 个月前 · 编辑精选 · 发表于 写给大家看的机器学习书

写给大家看的机器学习书（第四篇）—— 机器学习为什么是可行的（上）

1. 你敢跟着机器学习投资吗？系列文章学到这里，我们已经理解了机器学习的概念，也掌握了一... 查看全文 >

八汰 · 8 个月前 · 编辑精选

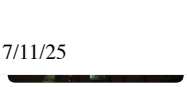
写给大家看的机器学习书（第五篇）—— 机器学习为什么是可行的（中）

1. 提要从这个系列文章的第四篇开始，我们开启了机器学习可行性的讨论。在第四篇中，我们经... 查看全文 >

八汰 · 8 个月前 · 编辑精选 · 发表于 写给大家看的机器学习书

写给大家看的机器学习书（第六篇）—— 机器学习为什么是可行

知 首发于 写给大家看的机器学习书 写文章 ooo



1. 提要从这个系列文章的第四篇开始，我们开启了机器学习句柄性的讨论。这是所有机器学习者... 查看全文 >

八汰 · 8 个月前 · 编辑精选