8.1.1 什么是机器学习

2023 年 9 月 18 日

导论

1. 机器学习的历史

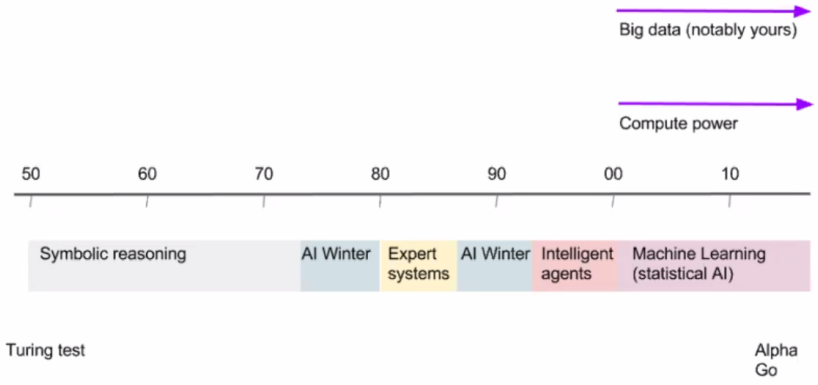

图 1-1 机器学习的历史

从上世纪战后 50 年的**符号推理** (Symbolic reasoning) 开始，经过近 70 年的发展，经历了数次寒冬。终于 21 世纪 10 年后，随着计算机算力的提升，以及互联网大数据存储和计算技术的普及，以**机器学习**（Machine Learning）和**深度学习**（Deep Learning）为代表的 AI 模型开始进入大规模应用。

大多数人听到“机器学习”，往往会在脑海中勾勒出一个机器人：一个可靠的管家，或是一个可怕的终结者，这取决于你问的是谁。但是机器学习并不是未来的幻想，它已经来到我们身边了。

事实上，一些特定领域已经应用机器学习几十年了，比如光学字符识别（Optical Character Recognition，OCR）。但是直到 1990 年代，第一个影响了数亿人的机器学习应用才真正成熟，它就是垃圾邮件过滤器（spam filter）。虽然并不是一个有自我意识的天网系统（Skynet），垃圾邮件过滤器从技术上是符合机器学习的（它可以很好地进行学习，用户几乎不用再标记某个邮件为垃圾邮件）。后来出现了更多的数以百计的机器学习产品，支撑了更多你经常使用的产品和功能，从推荐系统到语音识别。

2. 什么是机器学习?

图 1-2 什么是机器学习

机器学习（Machine Learning）就是教机器自己来完成任务，就这么简单，而复杂性来源于细节。[1]

```
Machine Learning is the science (and art) of programming computers so they can learn from data
```

更一般化的定义:

```
[Machine Learning is the] field of study that gives computers the ability to learn without bei
—Arthur Samuel, 1959
```

或者更工程导向的定义:

```
A computer program is said to learn from experience E with respect to some task T and some per
—Tom Mitchell, 1997
```

想象一下，你手里有很多数据，需要对数据进行识别、分类和预测。但是你又很忙，你想安排机器来完成这些任务。那么机器是你的学生，你就需要扮演老师角色，教一教机器如何自己来完成上述任务。

机器学习的目标就是通过若干示例（怎样做或不做）让机器自主学习完成任务。

机器学习的起点和终点分别是什么呢？确切的讲，机器进行学习是什么意思？如果我下载了一份维

基百科的拷贝，我的电脑就真的学会了什么吗？它马上就变聪明了吗？

3. 它智能在哪儿呢？

假设你每天早上打开电脑，都会做同样的事情：删除垃圾邮件，保留下对你有价值的邮件。过了一段时间，你感到厌烦，开始琢磨是否可以让这种琐事自动完成。一种方式是分析你的大脑，将整理电子邮件时大脑思考过程的规则记录下来。然而，这种方式相当麻烦，而且总不完美。你总会遗漏一些规则，同时又会对另一些规则细致过头。

另一种更好的、更智能的方式是该是怎样的呢？

思考一下，你会如何使用传统的编程技术写一个垃圾邮件过滤器：

你先观察下垃圾邮件一般都是什么样子。你可能注意到一些词或短语（比如亲～、免费、机不可失…）在邮件主题中频繁出现，也许还注意到发件人名字、邮件正文的格式，等等。

你为观察到的规律写了一个检测算法，如果检测到了这些规律，程序就会标记邮件为垃圾邮件。

测试程序，重复第 1 步和第 2 步，直到满足要求。

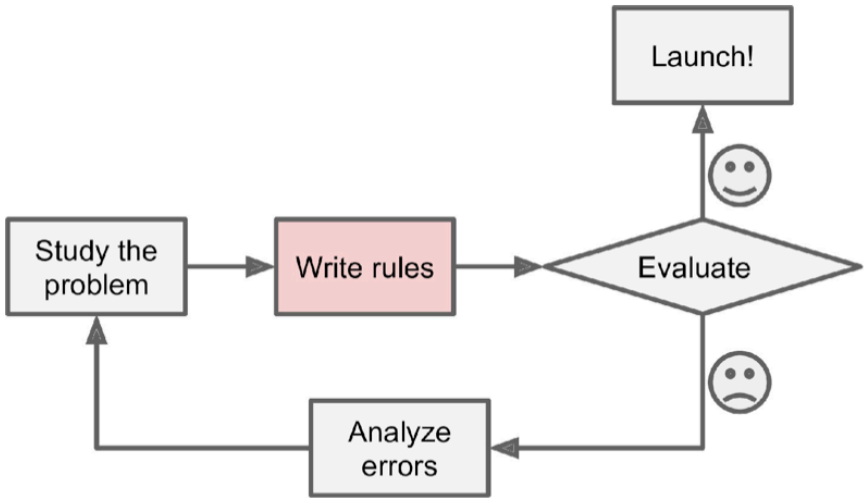

图 1-3 传统方法

这个问题并不简单，你的程序很可能会变成一长串复杂的规则—— 这样就会很难维护。

相反的，基于机器学习技术的垃圾邮件过滤器会自动学习哪个词和短语是垃圾邮件的预测值，通过与普通邮件比较，检测垃圾邮件中反常频次的词语格式（图 1-2）。这个程序短得多，更易维护，也更精确。

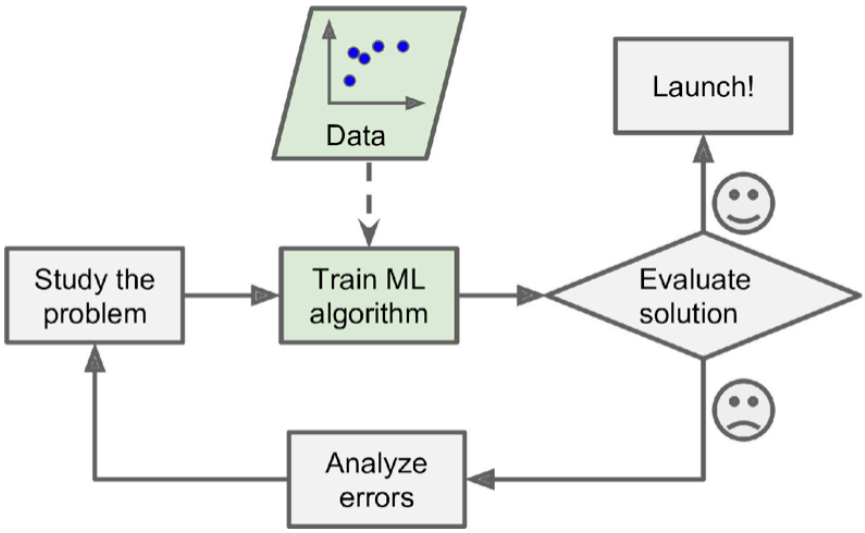

图 1-4 机器学习方法

进而，如果发送垃圾邮件的人发现所有包含 '**免费**' 的邮件都被屏蔽了，可能会转而使用 '**福利**'。使用传统方法的垃圾邮件过滤器需要更新以标记 '**福利**'。如果发送垃圾邮件的人持续更改，你就需要被动地不停地写入新规则。

相反的，基于机器学习的垃圾邮件过滤器会自动注意到 '**福利**' 在用户手动标记垃圾邮件中的反常频繁性，然后就能自动标记垃圾邮件而无需干预了。

参考

1. 《机器学习系统设计》Willi Richert 等著，图灵程序设计丛书
2. 《Hands-On Machine Learning with Scikit-Learn, Keras, and Tensorflow》, Aurelien Geron, OREILLY，中文版在线阅读：https://hands1ml.apachecn.org/#/