机器学习(Machine Learning)？

hey，这是机器学习系列文章的第一篇哦，主要想以普通的视角来给您普及机器学习的知识(简单的说就是通俗易懂)，希望能给你解决问题提供一种新的思路。资料主要依托于Andrew Ng 的斯坦福课程(CS229-----2008)

2019已经到来，相信你已经在过去的N久时间里，听过机器学习或者AI这个词已经很多次了，那么，机器学习到底是什么呢？

在回答到底是什么之前，我想先用几个例子来告诉你机器学习能做什么。

不过在这之前，我们先把机器学习拆开，拆成 机器 (Machine)、学习 (Learning)

首先，我们简单阐述一下学习

> 学习方式有很多种，这里主要介绍的是归纳学习
>
> 除了归纳学习，还有一种是演绎学习(Deduction Learning),通过一般情况，得到特殊情况。

那么 归纳学习(*Inductive Learning*)是怎么回事呢？

有一句古话叫做："瑞雪兆丰年"(PS：今年的雪额外的多)，为什么古人能得出这个结论呢(学习到这个知识点)？

当然，就在于不断的统计了，

- 第一年下雪，第二年丰收
- 第二年下雪，第三年丰收
- 第三年下雪，第四年丰收
- ...

不下雪的时候：

- 第二年不丰收
- 第三年不丰收
- 第四年不丰收

所以，古人们得出了这个结论，当然，还有很多很多。

- 李时珍的《本草纲目》，在不断的尝试药物之后，将实验结果记录在了书上
- 牛顿的万有引力(观察一系列物体)
- 你背的单词(不断的记忆，记忆之前肯定有一个参考，比如:apple 是苹果，不是梨子)
- 你的用户喜欢的东西，根据他/她历史购买的记录，帮他/她进行抉择
- 对你的产品进行A/B测试（选择更优的方案）
- ....

这一切的学习过程或者方法，我们从中获得了经验，最后选择了理想的结果。

到这里，我们简单的将归纳学习的过程整理复述一下:

通过对事物的观察,得到一定的经验，通过经验，我们就能解决某些问题或者说知道了新的东西。

我们把这些经验叫做**E(experience)**，需要解决的任务叫做**T(target)**

用一个常见的例子：

> 啤酒与尿布
>
> 在美国，在一次销量统计中，某超市发现，尿布和啤酒的销量都很高，超市的方感觉很奇怪。经过调查后发现，是因为男人们在买尿布的同时，会顺手也买啤酒。所以最后，超市把啤酒和尿布放在一起进行售卖，结果超市的尿布和啤酒的销售量又上了一个台阶。于是，“啤酒与尿布”的故事，广为流传，尤其在营销界。

在这个故事里，**E**就是各产品销售的情况(水果、厨具、生活用品、食品...)，**T**就是销售额的增加。

那么，我们简述一下经验学习的过程

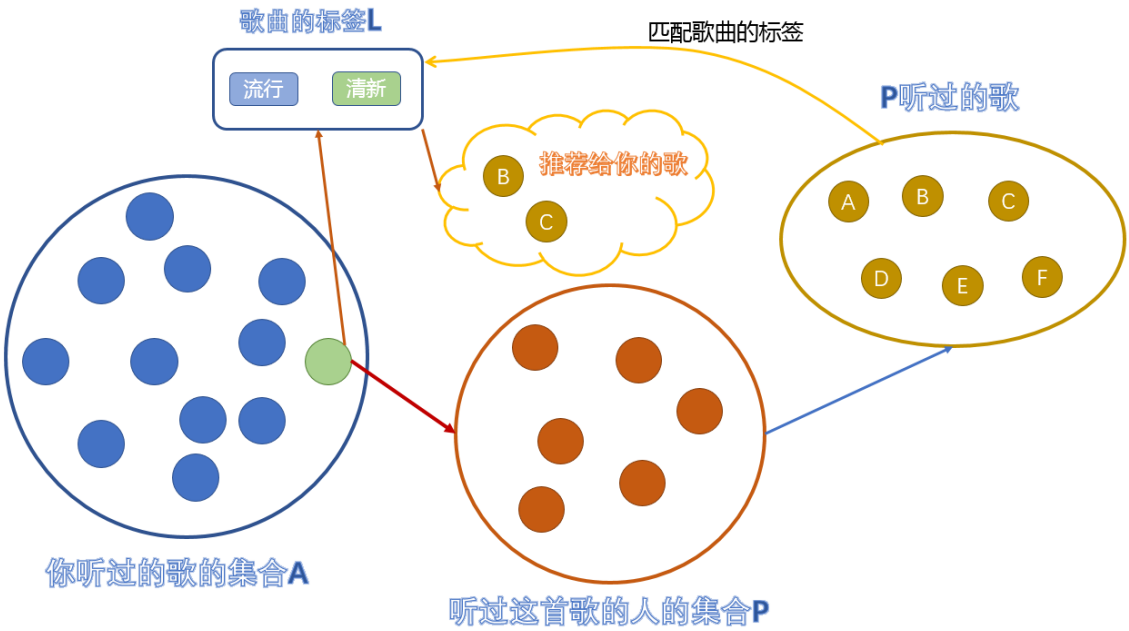
- 通过观察地壳的活动情况，我们可以预测地震的发生。 此处, **E**为地壳的活动情况，**T**为是否发生地震。
- 通过分析邮件的题目或者内容，我们可以区分一封邮件是否为垃圾邮件。此处，**E**为邮件的观察情况，**T**为是否为垃圾邮件。
- 通过分析用户的购买记录，我们可以为用户推荐他可能需要的东西。此处，**E**为用户的购买记录，**T**为用户可能喜欢的东西。

当然，到这里，你可能觉得太简单了，那么我们来一点复杂的

你知道音乐软件如何知道你喜欢哪种音乐的吗？

我们这里举一些方法：

1. 通过人工编排播单 - 通过人为的添加歌曲，形成歌单
    - 这种是经常见到的方式，你自己创建了多少歌单呢？你的播放列表有多少歌单呢？
2. 通过相似性来选择歌曲
    - 人以类聚，物以群分。歌，就是物品。我们把歌和人分离开。然后选一首你听过的歌，叫做**A**，当然我们也能得到哪些人听过这些歌,这些人叫做**P**，你听的歌也有一些标签(比如：流行、清新、热血...)，叫做**L**。
    - 然后我们把**P**听过的歌找出来，当然，这时候歌很多，我们再通过这些歌的标签是否等于**L**,去除掉一部分标签，然后就缩小了选择范围，也可以产出你可能喜欢的歌曲了。
    - 如下图所示:



- 当然我们可以再筛选一次，把**P**中，和你最像的人挑选出来（通过：听过哪些歌，哪些时间听的歌....），再从挑选出来的人进行一次歌曲匹配。

So,相信你也清楚了人的经验学习过程，那么机器学习又是怎么定义的呢？

我们引用 *Tom Mitchell* 提出的:

> 一个程序被认为能从经验 **E** 中学习，解决任务 **T**，达到性能度量值 **P**，当且仅当，有了经验 **E** 后，经过 **P** 评判，程序在处理 **T** 时的性能有所提升。

那么:

> 经验**E** 就是程序上万次的自我练习的经验而任务 **T** 就是下棋。性能度量值 **P** 呢，就是它在与一 些新的对手比赛时，赢得比赛的概率**P**。

是不是和人的经验学习过程很像呢，不过程序在经过机器学习的训练之后，得到的结果都是概率性的。

为什么结果都是概率性的呢，因为事件都是有不可确定性的。

有兴趣可以看看信息熵的知识。

如何让程序学习呢？

在知道如何学习之前，我们先看一下机器学习主要的分类(当然还有**半监督学习(*semi-supervised learning*)**这里就不举例了，有兴趣可以单独看看或者跟我讨论(・ω・)✧)。

监督学习(Supervised)和**非监督学习(Unsupervised)**

我们举一个例子:

对人进行分类

我们有一些人的数据(也叫做*样本空间*)，一个人(也叫做*样本*)包含了数据有身高、体重、肤色、母语...(也叫做*特征*)，分类的结果叫做目标(Target)或者标签(Label)

姓名	身高	体重	肤色	母语
A	100	40	Y	C
B	150	50	W	E
C	200	60	Y	J
D	250	70	W	C

无监督学习:

- 怎么分类呢，无监督学习也不知道，按照高矮？按照BMI？按照肤色和身高？按照性别(通过分析各个特征，得到该结果，比如：身高，或者BMI)？
- 但是他也能根据特征得出一个分类结果，按照一种分类方式来进行分类，如下表

姓名	类别I(高、矮)	类别II（BMI）	类别III(肤色和身高)	类别IV(性别)
A	矮	偏低	X洲人	M
B	高	正常	Y洲人	F
C	高	偏低	Z洲人	M
D	高	偏低	Z洲人	F

- 所以无监督学习，**不需要对数据进行标注**，也能得到分类结果，但是不是你想要的就不知道了。

监督学习:

- 怎么分类呢？ 显而易见的啦。因为监督学习会给这些人中的部分数据贴上标签，比如直接贴上性别标签。
- 那么很明显，分类目的就是让没有标签样本得到性别的结果，即按照性别进行分类。
- 我们添加两个带标签的人，如下表:

姓名	身高	体重	肤色	母语	性别(Target)
E	150	50	W	C	F
F	200	60	W	F	M

- 没有标签的样本，就通过已有标签的样本的特征信息，进行相似度的匹配，最后得到性别的结果。
- 那我们利用该表就可以得出我们的分类结果，如下表

姓名	类别(性别)
A	M
B	F
C	M
D	F

- 所以监督学习，一般都**需要对数据进行标注(也就是打标签)**，我们就能得到更靠近目的的答案哦。

是不是很简单呢？

我们接下来说一下机器学习完成的任务的分类，

分为**回归(regression)**和**分类(classification)**

其实这两种很简单

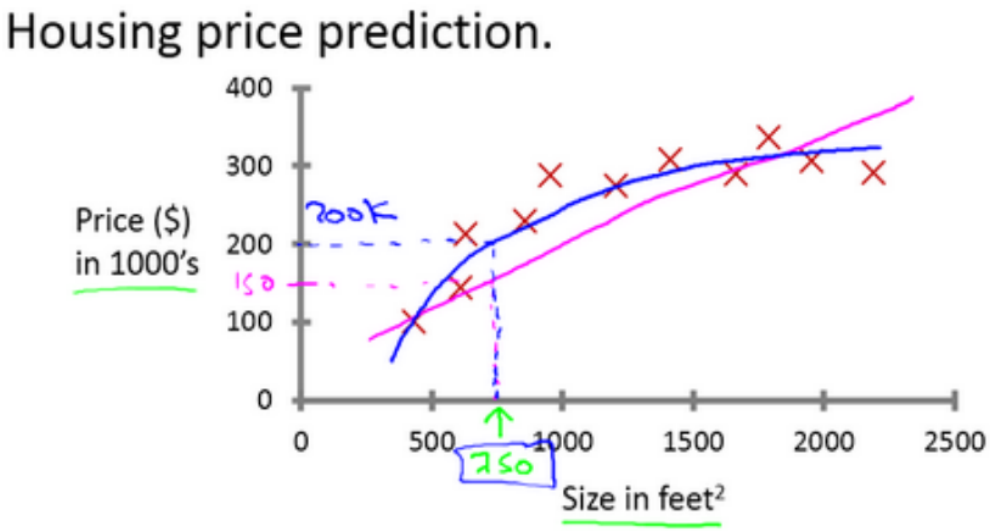
回归

- 一般用来预测，**目标(Target)**是**连续性**数值，比如：身高，体重，长度...
- 如下图:

  房价推测，x轴是房子大小，y轴为价格

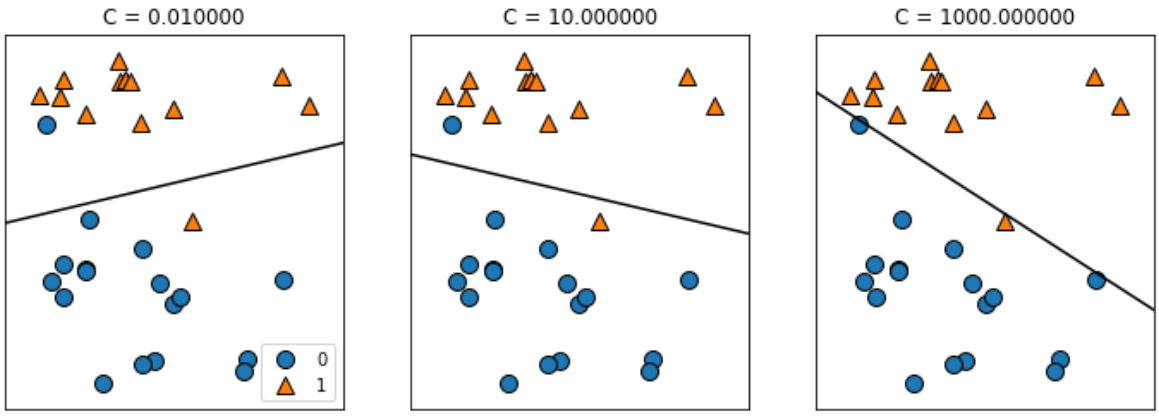
  用一条线进行拟合数据，做出预测结果，像得到一条线，比如:

$$y = ax + b$$



分类

- 如其名，就是用来分类的，目标**(Target)**是**离散型**数值比如：性别，爱好，动物...
- 如下图:
  - 分成两种类别(0,1)，上面的图形都是特征组合的呈现

- 

是不是感觉很简单呢?

以上就是初篇的基本知识啦！我们再简单回顾一下有哪些知识点。

- 经验性学习 (Experience Learning)
- 机器学习(Machine Learning)
- 监督学习和非监督学习 (Supervised and Unsupervised)
- 分类与回归(Classification and Regression)

第一篇文章就结束啦，希望对你有所帮助。是不是感觉意犹未尽呢？之后我们会继续讲解哦。如果有遇到问题的话，欢迎和我一起探讨！

Thanks for your read!

参考资料:

> *Andrew Ng - Stanford University CS229 《Machine Learning》*
>
> *邢无刀 - 极客时间《推荐系统36式》*
>
> *package scikit-learn*