写在前面

吴恩达老师(Andrew Ng)的机器学习课程是Coursera上的第一门课程，也是他的经典之作。这门课程从2011年上线到2017年，本门课程已经在全球积累了180万名学员，给许多人工智能入门者提供了全新的学习途径。

这门课可以说是通向人工智能的"必经之路"。我身为一名研究NLP方向的研究生，我从刚开始连论文都看不懂到现在对机器学习的原理和其背后的数学含义有一个初步的了解，离不开这门课程对我的帮助。

在笔记的开始特别要感谢网易云课堂为我提供了一个学习的平台，课程名称为吴恩达机器学习。

同时要感谢黄海广博士提供的字幕以及学习笔记。

在我的笔记中我会参考黄博士的笔记并根据课程内容加以自己的理解，希望能跟大家一起交流，一起学习，共同进步。

> 搜索微信公众号:'AI-ming3526'或者'计算机视觉这件小事' 获取更多人工智能、机器学习干货

> csdn：https://blog.csdn.net/qq_36645271

> github：https://github.com/aimi-cn/AILearners

本笔记仅供学习交流使用！

第一章 绪论：初识机器学习

1.1 欢迎参加机器学习课程

机器学习(Machine Learning, ML)是一门多领域**交叉学科**，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。**专门研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。** 它是人工智能的核心，是使计算机具有智能的根本途径，其应用遍及人工智能的各个领域，它主要使用归纳、综合而不是演绎。在这门课程中，你将学习到机器学习这门技术，并可以自己实现机器学习算法。

其实每天我们都会与许多机器学习算法打交道

- 在浏览器中进行搜索
- 淘宝中对你可能感兴趣的商品的推荐
- 电子邮箱中的垃圾邮件过滤系统

在这门课程中，仅仅了解机器学习算法和数学知识是不足以让你能够解决实际问题的，所以我们要花时间去做一些练习，通过让自己亲自动手去实现这些算法来了解其内部的基理。**由于现在在数据科学领域python语言十分火爆，所以笔记中我会用python语言来完成课后作业而不是课程中所讲的Octave。**

那么为什么机器学习现在如此流行呢？那是因为

- 机器学习是从人工智能即AI中发展出来的一个领域
- 机器学习是为计算机开发出的一项新功能

与传统的编程（类似于求解从A到B的最短距离）不同，一些例如网页搜索，相片标记，过滤垃圾邮件等**工作不能通过人类编写简单的逻辑代码来实现，实现这些功能的唯一方法就是让机器自己"学习"如何去做，这不仅是机器学习这门课程名称的由来，也是它的魅力所在！**

最后再为大家介绍其他几个机器学习的例子：

- 数据挖掘

  机器学习被用于数据挖掘的原因之一是网络和自动化技术的增长，例如网页点击数据、医疗记 录、计算生物学、各类工程领域。

- 人类无法编写的程序

  例如自动驾驶直升飞机、笔迹识别、自然语言处理（NLP）、计算机视觉。

- 个性化订制程序

  比如亚马逊、网飞、淘宝、爱奇艺等的推荐系统。

如果说看到控制台上显示出的'Hello World!'是程序员的第一节课的话，那么我认为打开手机淘宝所看到的精准的产品推送就是算法工程师的第一节课。我相信当我们完整的学习完这门课程之后，这种精确推送背后的原理就会变得不再神秘。

机器学习绝对是目前IT界最受欢迎的计算机技术，就让我们带着对AI梦的追求和对高薪的渴望，一起来开始我们的机器学习吧！

1.2 什么是机器学习

1.2.1 机器学习的定义

Arthur Samuel（1959）：在进行特定编程的情况下，使计算机具有学习能力的研究领域。

Tom Mitchell（1998）：计算机程序从经验**E(Experience)**中学习，解决某一任务**T(Task)**，达到性能度量值**P(Performance)**，通过**P**测定在**T**上的表现因经验**E**而提高。例如在一个跳棋程序中，经验**E**表示程序与自己下的几万盘棋，任务**T**表示玩儿跳棋，性能度量**P**表示与对手玩儿跳棋时赢的概率。

1.2.2 一个小练习

> 假设您的电子邮件程序会观察收到的邮件是否被你标记为垃圾邮件。在这种Email客户端中，你点击"垃圾邮件"按钮，报告某些Email为垃圾邮件，不会影响别的邮件。基于被标记为垃圾的邮件，您的电子邮件程序能更好地学习如何过滤垃圾邮件。请问，在这个设定中，E,P,T分别对应哪几项？
>
> > 给邮件分类
> >
> > 观察你是否把邮件标记为垃圾邮件
> >
> > 正确归类的邮件的比例

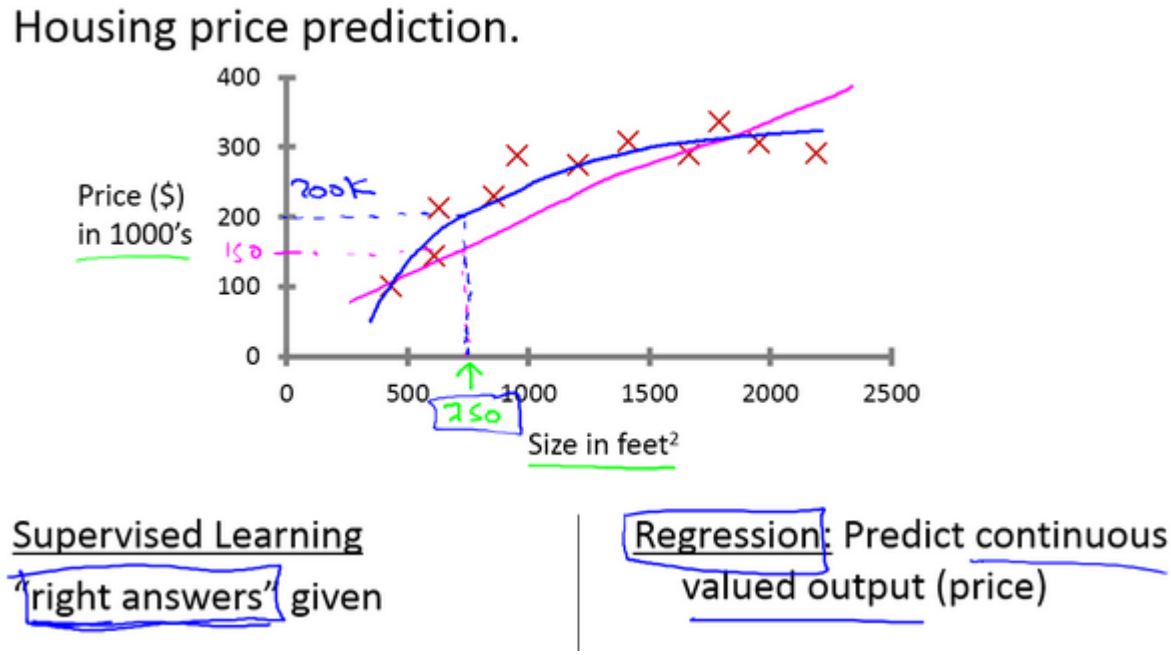
正确答案分别是E,P,T。通过这个例子让我们能够进一步理解机器学习的定义——**我们的系统在任务T上的性能在得到经验E之后会提升性能P。**

1.2.3 常见的机器学习算法

- 监督学习：人类"教"计算机去做某一件事。
- 无监督学习：人类让计算机自己"学习"。

1.3 监督学习

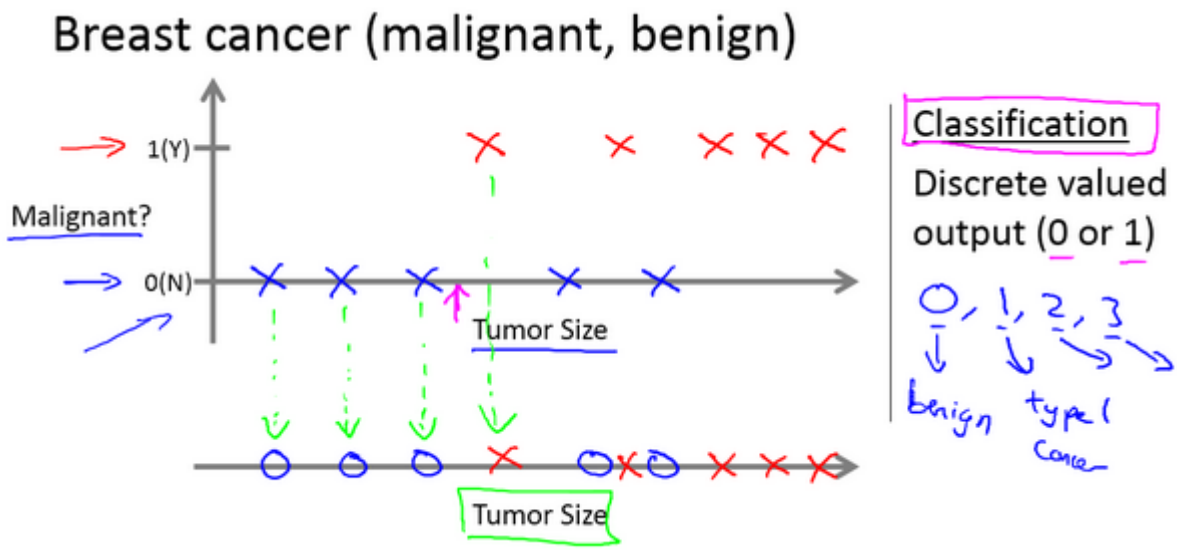
1.3.1 案例1：房价预测



通过一条直线或二次方程去拟合已有的数据，进而预测大小为750平方英尺的房子的价格。这是一个监督学习的例子。**可以看出，监督学习指的就是我们给学习算法一个数据集，这个数据集由"正确答案"组成。**在房价的例子中，我们给了一系列房子的数据，在这个数据集中对于每个样本我们都给出正确价格，即它们实际的售价，监督学习算法的目的就是给出更多的正确答案，即750平方英尺房子的应售价。

同时，这也是一个回归问题，即预测一个连续值的输出。

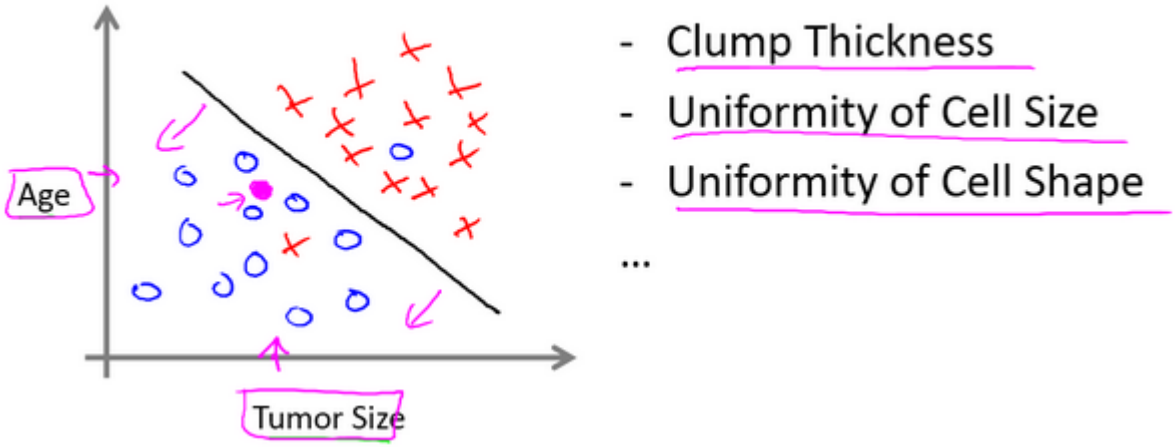
1.3.2 案例二：乳腺癌（恶性，良性）



在这个例子中，机器学习要解决的问题就是估计出肿瘤是良性的还是恶性的概率，数字1代表是肿瘤，数字0代表不是肿瘤。

同时，这也是一个分类问题，即预测一个离散值的输出。

但是在实际中，机器学习要处理的问题可能不仅只有一种特征，除了上面例子中肿瘤的大小，可能还有年龄等特征，如下例子所示。**所有的机器学习算法都是一个不仅仅能处理两到三个特征，而是可以处理无穷多特征的算法。**



1.3.3 一个小练习

假设你经营着一家公司，你想开发学习算法来处理两个问题：

1. 你有一大批同样的货物，假设你有上千件一模一样的货物等待出售，你想预测接下来的三个月能卖多少件？
2. 你有许多用户，这时你想写一个软件来检验每一个用户的账户。对于每一个账户，你想要判断它们是否曾经被盗过？

那这两个问题，它们属于分类问题、还是回归问题？

问题一是一个回归问题，因为如果我有数千件货物，我会把它看成一个实数，即一个连续的值。因此卖出的物品数，也是一个连续的值。

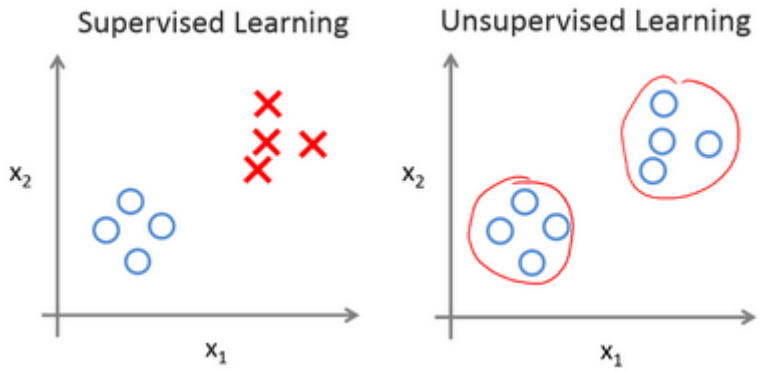
问题二是一个分类问题，因为我可能会用 0 来表示账户未被盗，用 1 表示账户曾经被盗过。所以我们根据账号是否被盗过，把它们定为0 或 1，然后用算法推测一个账号是 0 还是 1，因为只有少数的离散值，所以我把它归为分类问题。

1.3.4 监督学习总结

监督学习，核心在“监督”二字，基本思想是，我们数据集中的每个样本都有相应的“正确答案”，即都被标记。再根据这些“正确答案”对新样本进行“监督”，最后得到预测结果。就像房子和肿瘤的例子那样。同时还介绍了回归问题，即通过回归来推出一个连续的输出，这里还介绍了分类问题，其目标是推出一组离散的结果。

1.4 无监督学习

1.4.1 无监督学习与监督学习的区别及其概念



监督学习的数据集，如上图左1表所示，其中每条数据都已经被标注，例如一个肿瘤是良性或恶性。所以，对于监督学习里的每条数据，我们已经清楚地知道，训练集对应的正确答案。

无监督学习，核心在“无监督”这三个字，我们数据集中的每个样本都没有相应的“正确答案”，即都未被标记，我们将这些数据交给算法，并让算法为我们从中找出某种结构。无监督学习的数据集与我们之前看到的不一样，如上图右1所示，所有数据都没有被标注，即无监督学习中没有任何的标签或者是都具有相同的标签。所以对于一个没有任何标签的数据集，无监督学习算法可以判定该数据集包含两个不同的簇（Cluster）。

同时，这是一个聚类问题。

1.4.2 无监督学习算法的应用

- 新闻网站。新闻网站会用聚类算法将每天收集到的成千上万条的**没有任何标记**的新闻组合成一个个新闻专题。
- 基因组学。通过聚类算法，在没有提前告诉这个算法什么样的个体是什么种类的情况下，把不同的个体归为不同的类。
- 组织大型计算机集群。
- 社交网络分析。
- 市场细分。
- 天文数据分析。
- 鸡尾酒会问题（感兴趣的同学可以到课时4无监督学习中去详细了解）。

1.4.3 一个小练习

在下述例子中，哪一个适合用无监督学习算法去解决?

1. 给定电子邮件数据是/不是垃圾邮件的标签，训练一个电子邮件过滤系统。
2. 给出一系列在网上找到的新文章，将他们分到内容相同的文章的组里。
3. 给出一个顾客数据集，自动进行市场细分，并将不同的顾客分到不同的组中。
4. 给出一组是否患有糖尿病病人的数据集，进而预测一个新病人是否患有糖尿病。

很显然，结合之前的例子我们不难得出2,3适合用无监督学习算法去解决，而1,4适合用监督学习算法去解决。