

一、机器学习

一、机器学习

1.1 机器学习算法应用

- **web**搜索
- 照片标记
- 反垃圾邮件
- 数据库挖掘：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
- 医疗记录：随着自动化的出现，我们现在有了电子医疗记录。如果我们可以把医疗记录变成医学知识，我们就可以更好地理解疾病。
- 计算生物学：还是因为自动化技术，生物学家们收集的大量基因数据序列、**DNA**序列和等等，机器运行算法让我们更好地了解人类基因组，大家都知道这对人类意味着什么。再比如，
- 工程方面：工程的所有领域，我们有越来越大、越来越大的数据集，我们试图使用学习算法，来理解这些数据。
- 机械应用：有些人不能直接操作。例如，我已经在无人直升机领域工作了许多年。我们不知道如何写一段程序让直升机自己飞。我们唯一能做的就是让计算机自己学习如何驾驶直升机。
- 手写识别：现在我们能够非常便宜地把信寄到这个美国甚至全世界的原因之一就是当你写一个像这样的信封，一种学习算法已经学会如何读你信封，它可以自动选择路径，所以我们只需要花几个美分把这封信寄到数千英里外。
- 自然语言处理或计算机视觉：这些语言理解或图像理解都是属于AI领域。大部分的自然语言处理和大部分的计算机视觉，都应用了机器学习。
- 定制程序：每次你去亚马逊或**Netflix**或**iTunes Genius**，它都会给出其他电影或产品或音乐的建议，这是一种学习算法。

1.2 机器学习

定义

- 第一个机器学习的定义来自于**Arthur Samuel**。他定义机器学习为，在进行特定编程的情况下，给予计算机学习能力的领域。
- 另一个年代近一点的定义，由**Tom Mitchell**提出，来自卡内基梅隆大学，**Tom**定义的机器学习是，一个好的学习问题定义如下，他说，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。举例，经验**E**就是程序上万次的自我练习的经验，任务**T**就是下棋，性能度量值**P**就是它在与一些新的对手比赛时，赢得比赛的概率

分类

- 监督学习：我们将教计算机如何去完成任务
- 无监督学习中：我们打算让它自己进行学习
- 强化学习和推荐系统等

1.3 监督学习

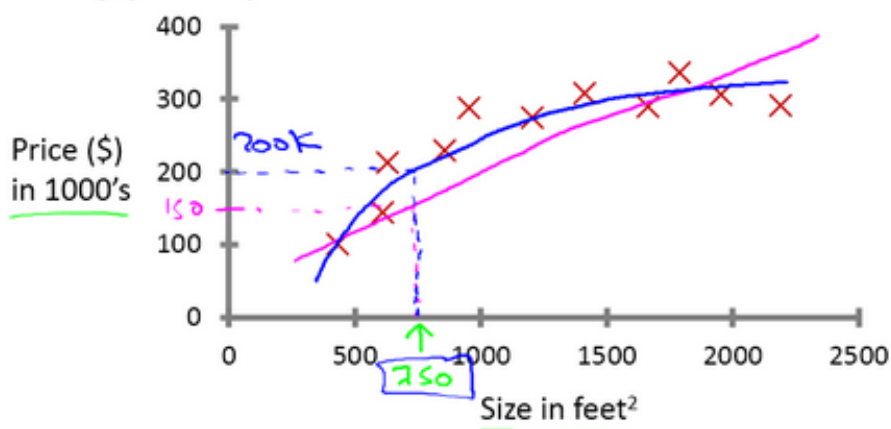
定义

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。

分类

回归：我们在试着推测出这一系列连续值属性

Housing price prediction.



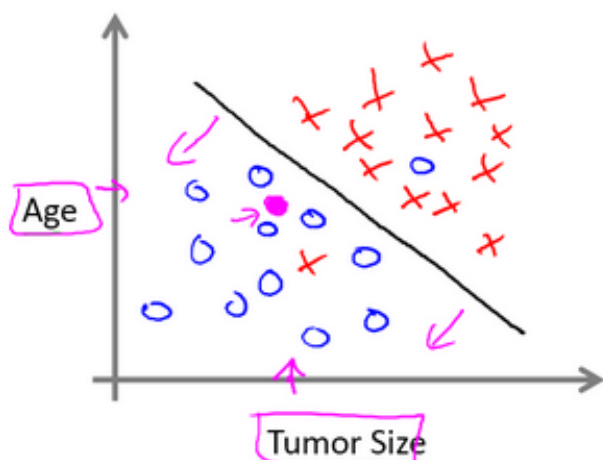
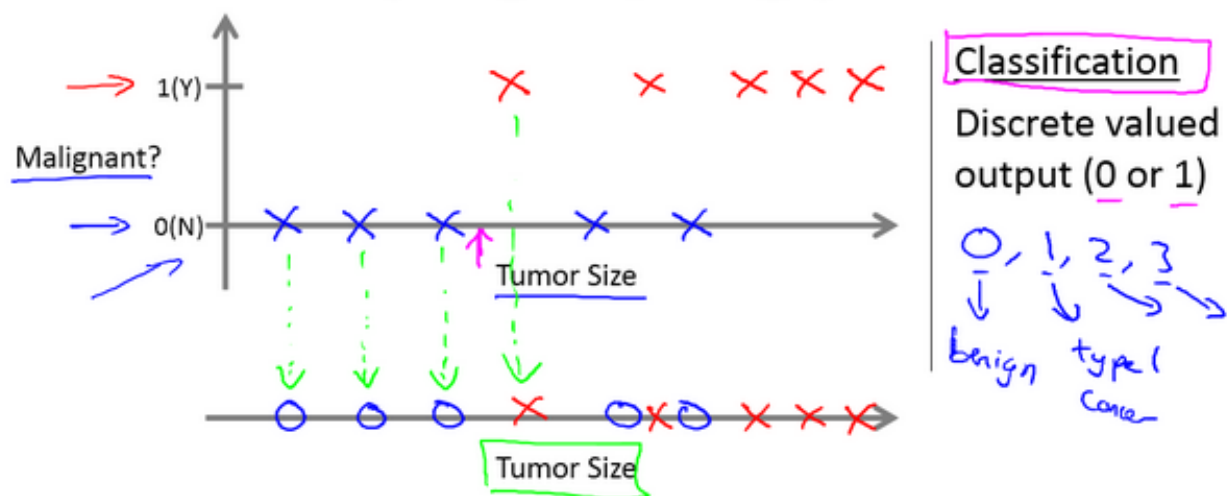
Supervised Learning

"right answers" given

Regression: Predict continuous valued output (price)

分类：我们试着推测出离散的输出值

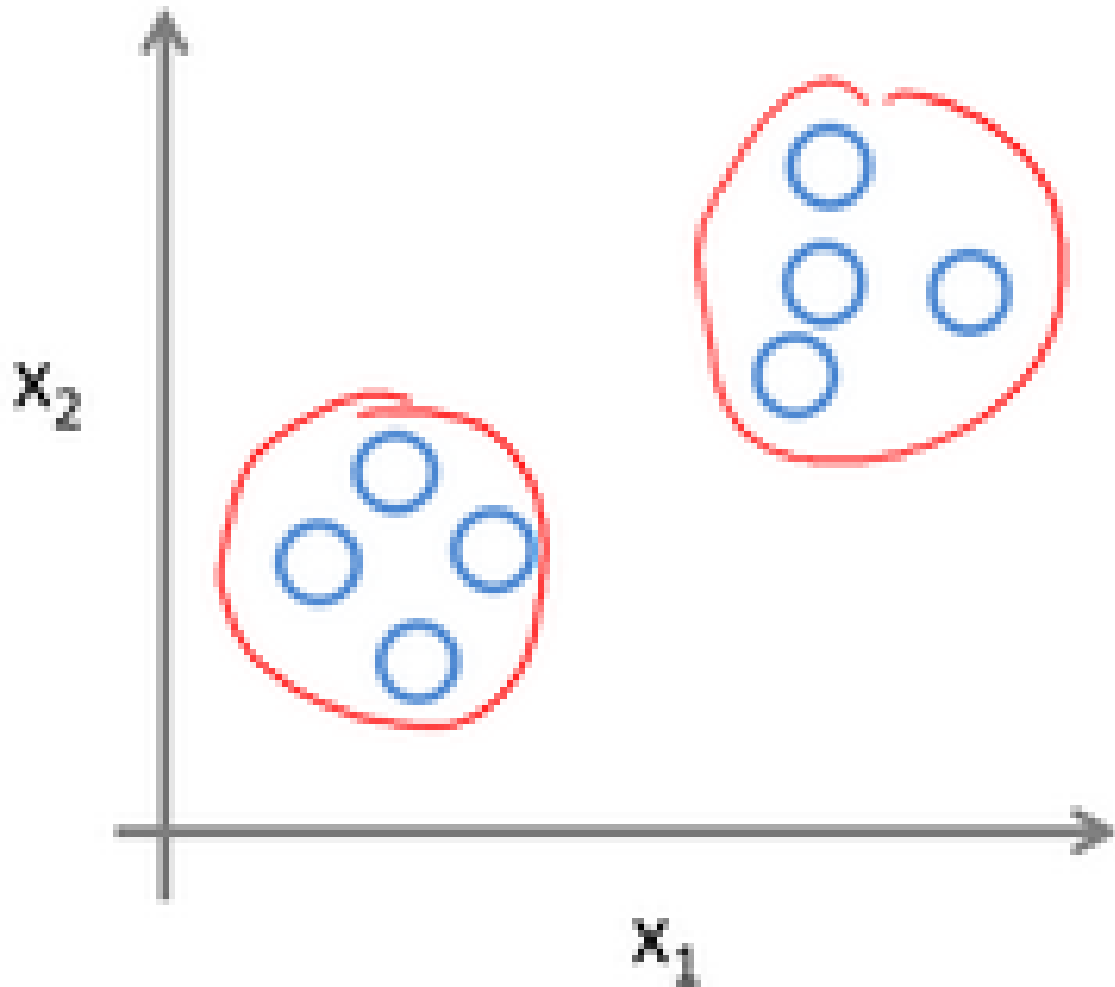
Breast cancer (malignant, benign)



- Clump Thickness
- Uniformity of Cell Size
- Uniformity of Cell Shape
- ...

分类算法，叫支持向量机，里面有一个巧妙的数学技巧，能让计算机处理无限多个特征。

Unsupervised Learning



定义

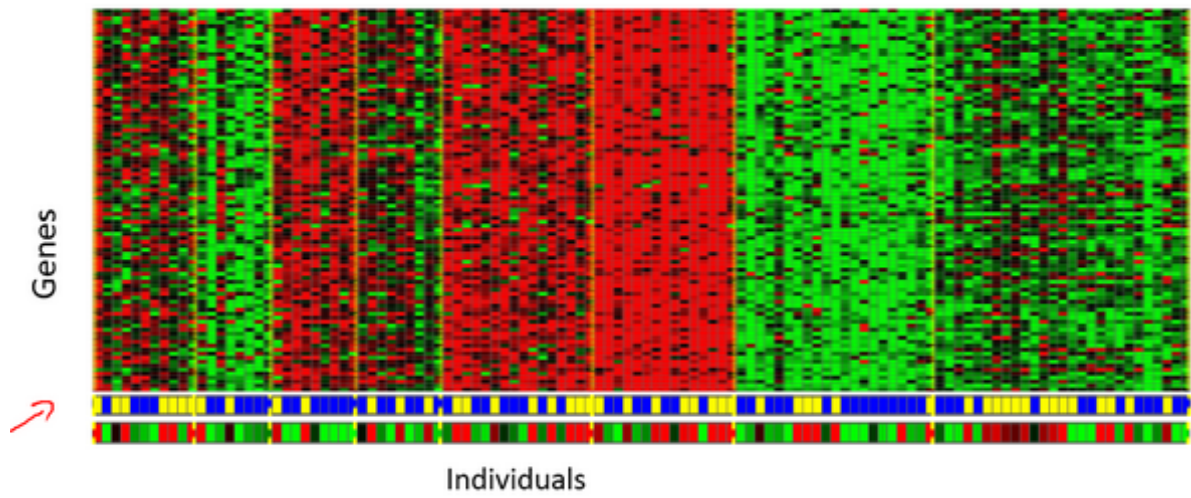
- 无监督学习中没有任何的标签或者是有相同的标签或者就是没标签。
- 无监督学习 (Unsupervised Learning) 指的是：在**没有标签 (y)** 的数据集上，算法通过数据本身的结构、分布、相似性等来发现模式或规律。

分类

- 无监督学习算法可能会把这些数据分成两个不同的簇。所以叫做聚类算法。
- 降维

例子

- 基因学的理解应用，DNA微观数据的例子



- 鸡尾酒宴问题
- 新闻事件分类
- 细分市场