Introduction: Machine Learning

Shen Hengheng

2017

该笔记是来自 Andrew Ng 的 Machine Learning 课程的第一周的课堂记录，主要讲解了以下几个内容:

- 机器学习的定义
- 机器学习的分类
  - 第一类：监督学习
  - 第二类：学习理论
  - 第三类：非监督学习
  - 第四类：强化学习

0.1 什么是机器学习？

0.1.1 非正式定义

1959 年，Arthur Samuel[1]非正式地定义了机器学习：“在不直接针对问题进行编程的情况下，赋予计算机学习能力的一个研究领域”。

西洋棋就是一个例子，西洋棋自己和自己下棋。由于计算机程序的处理速度非常快，所以 Arthur Samuel 让计算机与计算机自己下了成千上万盘棋，逐渐地，西洋棋意识到了怎样的局势能使自己胜利，什么样的局势导致自己失败，它会反复地自我学习：“如果让对手占据这些地方时，那么我输的概率可能比较大”，或者 “如果我占据这些地方时，我胜利的概率比较大”。在 1959 年，奇迹出现，他的西洋棋程序的棋艺远远超过西洋棋程序的作者！

过去人们的看法是计算机除了做程序明确让其做的事情，除外它什么都不能做，但是 Arthur Samuel 做到了！

0.1.2 现代化的定义

Tom Mitchell 在 1998 年提出现代化的机器学习的定义：“一个合理的学习问题应该是这样定义的：对于一个计算机程序来说，给它一个任务 T 和一个性能评测方法 P，如果在经验 E 的影响下，P 对 T 的测量结果得到了改进，那么就说明程序从中学习到了经验 E.

比如对于西洋棋那个例子来说：

- E - 程序成千上万次的自我练习

[1]发明了西洋棋程序

- T - 下棋
- P - 它与人类棋手对奕的概率

0.2 监督学习

0.2.1 回归问题

下面是给予一组房屋大小与对应的房屋价格的数据进行拟合的例子：

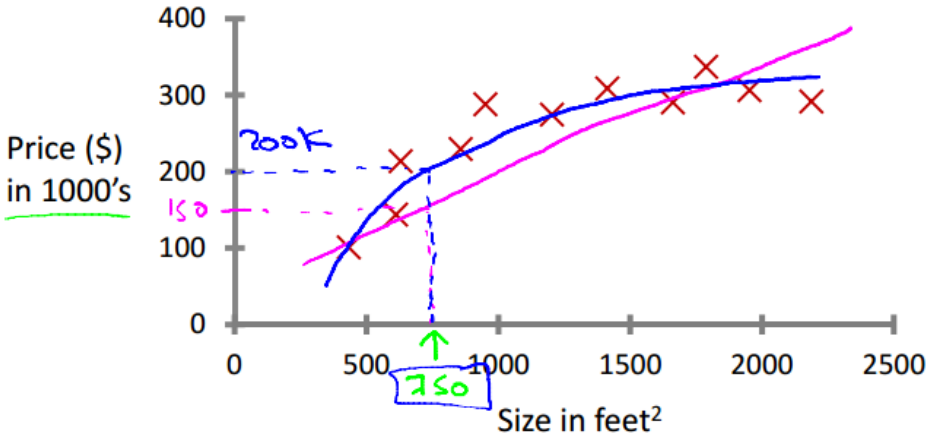


图 1: 房屋价格预测模型

通过学习来预测房屋价格的问题是监督学习问题的一个例子，之所以成为监督学习，是因为我们这个算法提供了一组房屋的大小 $(size)$ 和一组某种程度上可以堪称正确答案的房屋价格 $Price$ 的数据。比如 $(1000, 30)$ 等。

监督问题的学习，给算法提供了一组”正确“的输入和”标准“答案，之后，我们希望算法能够去学习标准输入和标准答案之间的联系，以尝试对于我们提供的其他输入来给我们提供**更为**标准的答案。

0.2.2 分类问题

分类问题是监督学习问题中的另一类问题，它与回归问题最大的区别在于，此时的数据是离散的我而不是连续的.

下面是给予肿瘤的大小和对应的肿瘤是否是恶性（1）或者良性（0），来进行拟合学习最佳的分类决策线。其中

- 数据是一组关于乳腺癌的数据，X 代表肿瘤的大小，Y 代表肿瘤的是否为恶性
- 目标是让一个算法学会预测一个肿瘤是否是恶性或良性

很显然，这是一个 **2 分类**问题。具体如图 2 所示：

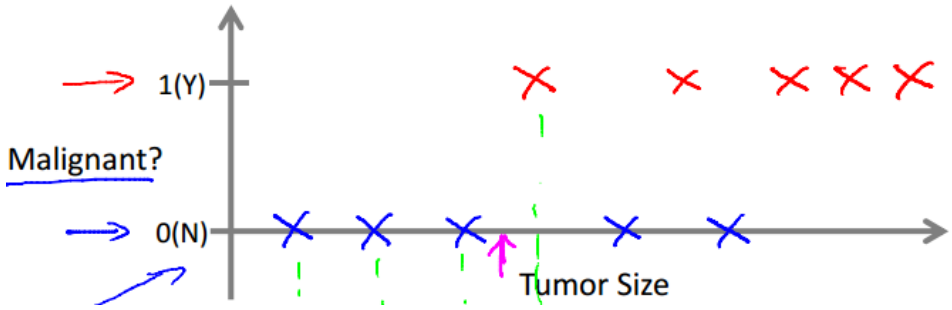


图 2: 肿瘤恶性/良性分类模型

下面考虑一组多个输入变量（多个特征）的数据，下面以两个特征进行举例说明。其中 X 表示肿瘤的大小，Y 表示患者的年龄，对于图中的样本数据的表示进行说明：'o' 表示良性，'x' 表示为恶性，目标是找出最佳的分类决策边界。如图 3 所示：

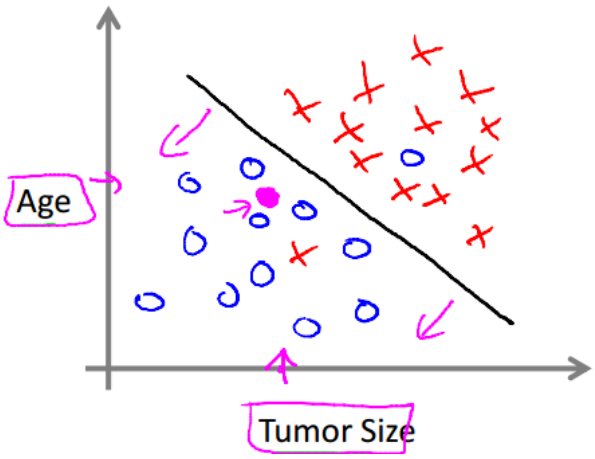


图 3: 多特征的肿瘤恶性/良性分类模型

如果你的数据不能在二维，三维甚至任何有限维空间表示出来，如果你觉得数据实际上存在于无线维空间中该怎么办?

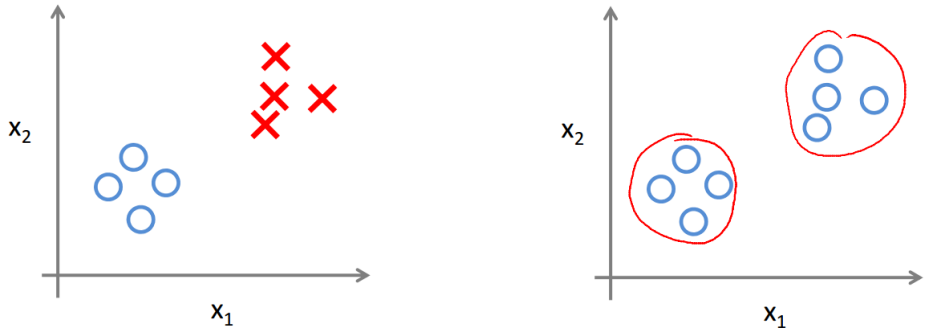
solution: 支持向量机算法[2]，可以将数据映射到无限维空间，所以他不仅能处理像之前例子中的两个特征所表示的数据，而且还可以处理无限种特征.

0.3 学习理论

学习理论主要主要试图了解一下什么算法能够很好地近似不同的函数，并且试图了解一些诸如需要多少训练集数据，测试集数据这样的问题，还比如算法优化，欠拟合和过拟合等问题。

[2]kernel

0.4 无监督学习

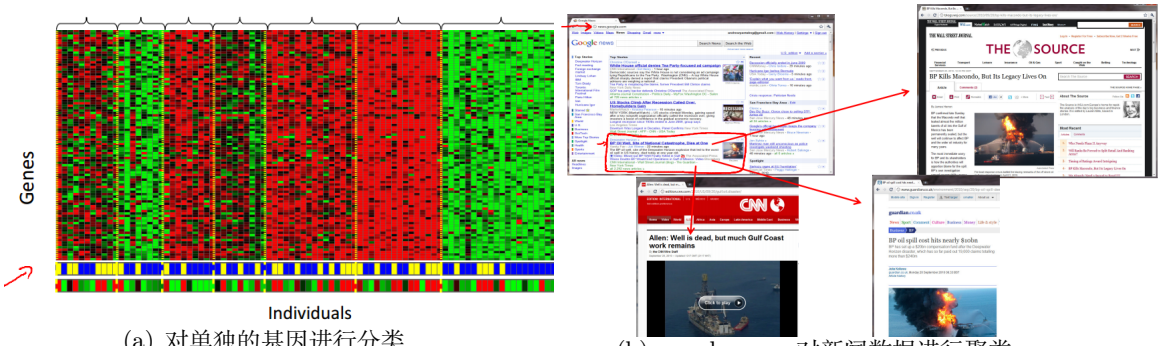


(a) 一些样本数据　　(b) 聚类之后

图 4: 你能在这组数据中寻找一些有趣的结构吗?

一个算法会寻找这样的结构（如图4），它会将里面的数据聚成两类，”聚类问题“就是无监督学习的一个典型问题。主要应用：

1. 无监督学习尝试理解基因数据，会按照基因在实验中体现出的形状的规律，来对单独的基因数据进行分类。
2. google news 尝试对新闻进行聚类。
3. 聚类算法处理图像问题
   特定的无监督学习算法，它会学习对这些像素进行聚类，就是说，这些像素可能是在一起的，那些像素可能是在一起的。对像素进行的分组，他们对于计算机视觉和图像处理领域都很有用。



(a) 对单独的基因进行分类　　(b) google news 对新闻数据进行聚类

图 5: 应用

还有一些例子，比如计算机集群，社会网络分析，市场划分，航天数据分析等等都会用到无监督学习。另外还有在语音识别方面的经典问题：鸡尾酒会问题.

0.5 强化学习

他可以用在一些你不需要进行一次决策的情形中，比如，利用监督学习对癌症进行预测的例子中，对于一个病人，你要预测他的肿瘤是否为恶性，那么你的预测将会决定病人的生死，也就是说，你的决策要么对要么错。那么对于强化学习来说，他主要使用了一种叫做**回报函数**的概念来进行决策，不像监督学习那样一次性决策，它是探索性算法它常常用于机器人领域，网页爬取等领域。