监督学习和无监督学习有什么区别？

机器学习里的基础学习方式有两种：监督学习和无监督学习。从名字就可以看出来，监督学习是有指导的学习，就象小孩在学校里听课，在老师的指导进行学习。相对的，无监督学习是没有指导的学习，类似小孩即使没有听老师讲课，也能通过自己学习和观察获得大量知识。

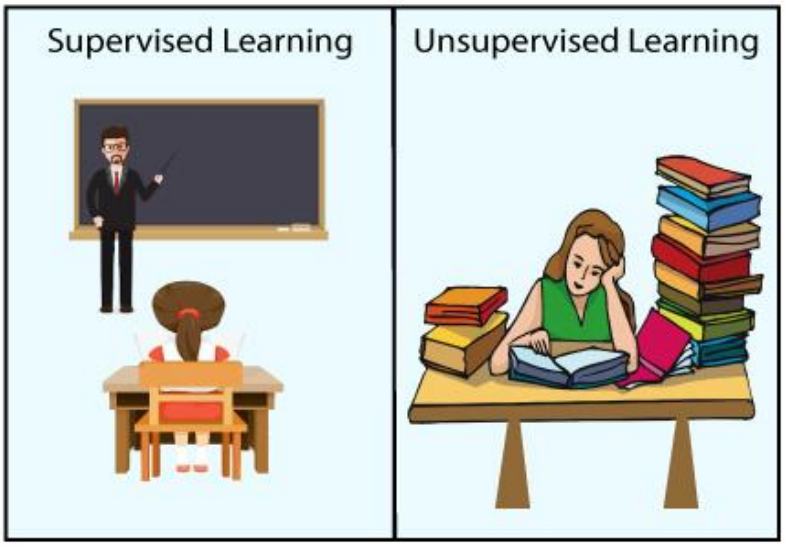

图 1：监督学习与非监督学习[3]

一般来说，监督学习多用于分类或回归任务。在这些任务中，分类和回归的目标需要进行人为标记。这个标记过程就是监督指导的过程。基于这些标记信息，可以学习得到一个分类或回归模型。这些模型以观察变量为输入，输出类别或回归目标值，从而实现分类或回归任务。图 2（a）给出一个基于监督学习的分类任务：首先对训练集中的邮件标注是正常邮件还是垃圾邮件，然后训练一个分类器，其于该分类器即可判断一个未知邮件是否是垃圾邮件。图 2（b）给出一具基于监督学习的回归任务：首先对训练集中的房子标注价格，基于这些标注训练一个回归模型，该模型即可对其他房子的价格进行预测。

无监督学习一般用于聚类和流形学习。聚类是对观察对象进行归类划分，使得相似的对象集中在一类。流形学习是发现数据的高密度低维子空间，从而实现降维和可视化。在这些任务上，不需要提供任何额外标注信息，算法即可以基于观察到的变量进行学习。图 2（c）给出的是一个聚类任务。给定一批图片，不需要对图片进行任何标注，算法通过图片之间的相似性将他们自动归为三个类别。图 2（d）描述的是一个流形学习任务，在一个三维空间中发现一个二维蛋卷形高密度子空间。

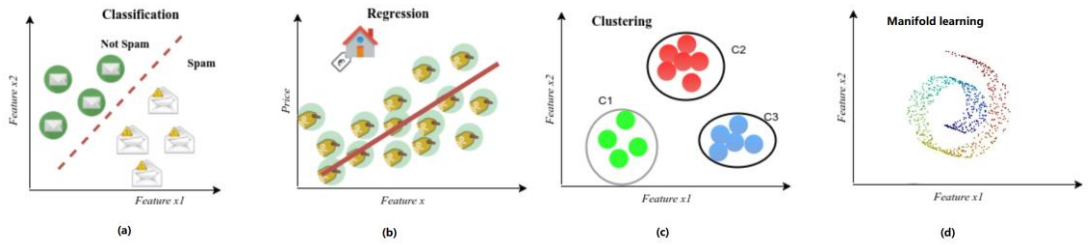

图 2：用于分类（a）和回归（b）的监督学习和用于聚类(c)和流形学习（d）的无监督学习[1]。

总结来说，监督学习和无监督学习的主要区别在于是否需要人为标注，和任务本身并没有特别严格的对应关系。例如，一些分类和预测任务可以天然从数据中得到标签，并不需要人为标注，这种学习方式有时也称为“自监督学习”。
例如通过视频信号学习语音对口唇的预测函数，通过前后文本预测缺失的标点符号等。

在另外一些任务中，监督信息比较弱，人们不对过程中的每一步给以具体监督，而是通过对结果的评价来给出监督信息。这有点类似于教学过程中，老师不告诉孩子如何解题，而是通过对解题结果的判定来对学生进行指导。这种学习方式称为“强化学习”[2]。

不论使用哪种方法，机器学习的根本任务是建立各种变量之间的概率相关性，只要这些相关性建立起来了，就可以完成分类、回归、聚类、降维等各种任务了。

1. https://imgbin.com/png/qbJQykFD/machine-learning-unsupervised-learning-algorithm-png
2. 王东，《机器学习导论》，第 10 章，强化学习，清华大学出版社，2021.2.
3. Difference between Supervised and Unsupervised Learning, https://www.javatpoint.com/difference-between-supervised-and-unsupervised-learning