第1节 人工智能、机器学习的相关概念

机器学习是人工智能研究发展到一定阶段的必然产物。

二十世纪五十年代到七十年代初，人工智能研究处于**推理期**，那时人们认为只要能赋予机器逻辑推理能力，机器就具有智能。

随着研究发展，在七十年代中期开始，人工智能研究进入了**知识期**，要使机器具有智能，就必须设法使机器拥有知识。此期间大量的专家系统面世。

在八十年代，**从样例中学习**（监督和无监督学习等）的一大主流是符号主义学习。其代表包括**决策树(decision tree)**和基于逻辑的学习。

九十年代中期之前，**从样例中学习**的另一主流技术是基于神经网络的**连接主义**学习。与符主义学习能产生明确的概念表示不同，连接主义学习产生的是**黑箱**模型。连接主义最大的局限是**试错性**：学习过程涉及大量参数，而参数的设置缺乏理论指导，主要靠手工调试。

九十年代中期，**统计学习(statistical learning)**登场。代表技术是**支持向量机(support vector machine)**。

二十一世纪初，连接主义通过**深度学习**卷土重来。所谓深度学习，即很多层的神经网络。在涉及语音、图像等复杂对象的应用中，深度学习取得了优越性能。深度学习虽然缺乏严格的理论基础，但是它显著降低了机器学习的门槛，为机器学习的实践带来了便利。

当前时代，互联网和硬件高度发达，人们进入了大数据时代，深度学习取得了大发展。随着物联网、边缘计算、5G网络、IPV6等的发展和普及，相信人工智能会在人类社会发挥更大的作用。

算法

机器学习致力于研究如何通过计算的手段，利用经验来改善系统自身的性能。在计算机系统中，**经验**通常以**数据**形式存在。因此，机器学习所研究的主要内容，是关于在计算机上从数据中产生**模型(model)**的算法。有了算法，我们把经验数据提供给它，它就能基于这些数据产生模型；在面对新的情况时，模型会给我们提供对应的判断。

数据集

要进行机器学习，先要有数据。假定我们收集了一批关于西瓜的数据，例如

```
{色泽=青绿；根蒂=蜷缩；敲声=浊响}
{色泽=乌黑；根蒂=稍蜷；敲声=沉闷}
{色泽=浅白；根蒂=硬挺；敲声=清脆}
```

这样的一组数据称为一个**数据集(data set)**，其中每条记录是关于一个事件或对象的描述，称为一个**示例(instance)或样本(sample)**。如果把每个样本中的色泽、根蒂和敲声作为三个坐标轴，则它们张成一个用于描述西瓜的三维空间，每个西瓜都可以在这个空间中找到自己的位置。当然，一般来说，维数越多，描述就会越精确。空间中每个点对应一个坐标向量，因此我们也把一个样本称为**特征向量(feature vector)**

训练

从数据中学得模型的过程称为**学习(learning)** 或**训练(training)**，这个过程通过执行某个算法来完成。训练过程中使用的数据称为**训练数据(training data)**，其中每个样本称为一个**训练样本(training set)**。学得的模型会对应关于数据的某种规律。

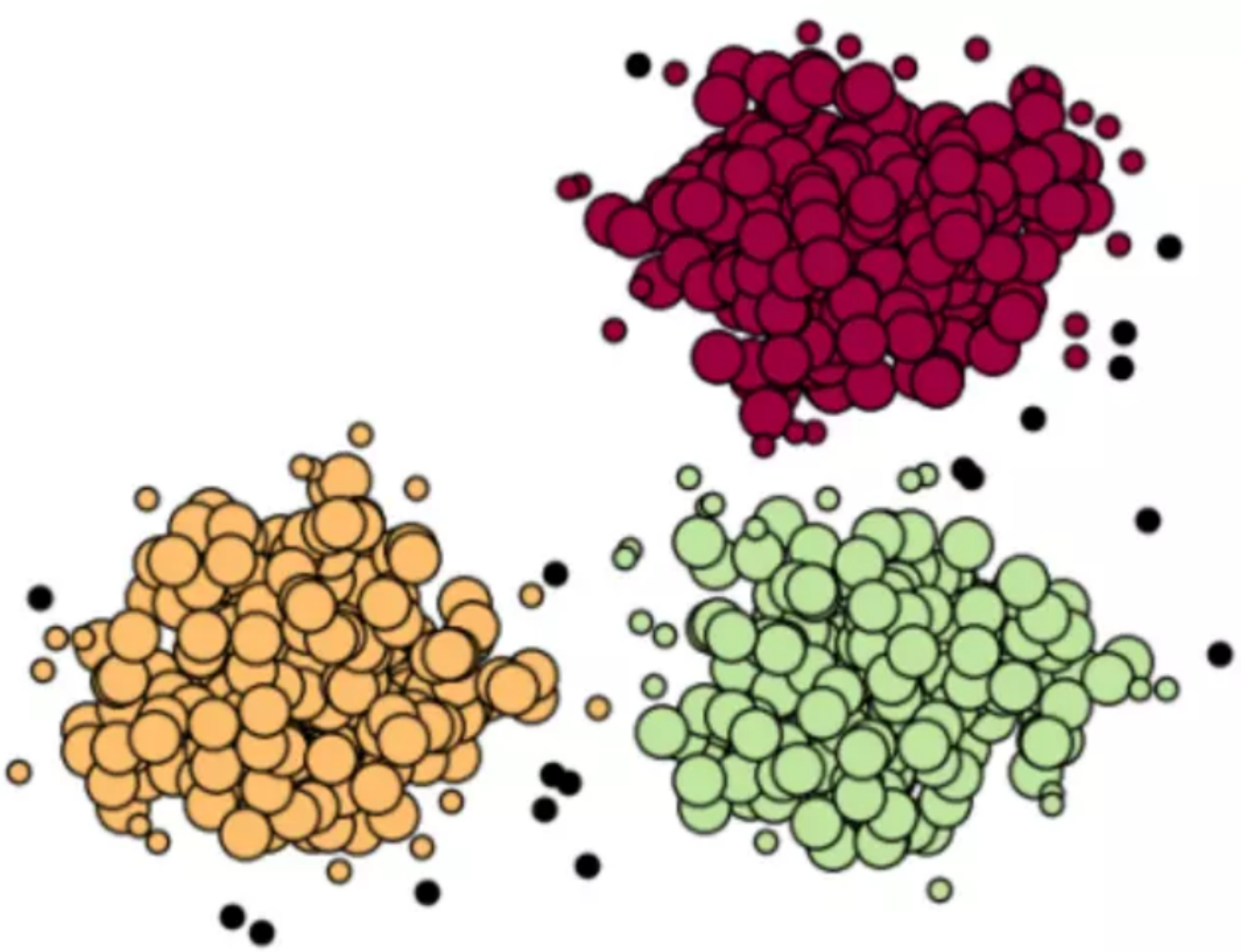
例如，如果希望学得一个能帮助我们判断一个西瓜是不是好瓜的模型，仅仅有前面的数据集是不够的。要建立关于**预测(prediction)的模型，我们需要过得训练样本的结果**信息:
例如`{{色泽=青绿；根蒂=蜷缩；敲声=浊响}，好瓜}`
这个关于结果的信息（好瓜）称为**标记(label)**。

分类、回归、聚类、监督与无监督学习

若我们预测的是离散值，例如**好瓜**、**坏瓜**，此类学习任务称为**分类(classification)**

若预测的是连续值，如西瓜成熟度0.95，0.37，此类学习任务称为**回归(regression)**

我们可以对西瓜做**聚类(cluistering)**。即将训练集中的西瓜分为若干做，每组称为一个**簇(cluster)**
例如，算法自动将数据集分成了3簇，用三种颜色代表。每一簇内较大的点代表核心对象，较小的点代表边界点。黑色的点代表离群点或者叫噪声点。



根据训练数据是否拥有标记信息（好瓜），学习任务可划分为两大类：**监督学习(supervised learning)**和**无监督学习(unsuperviserd learning)**

分类和回归是**监督学习**的代表。聚类是**无监督学习**的代表。