# 1.1引言

Q：如何理解机器学习？

A：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。“经验”通常以数据的形式存在，因此机器学习主要的研究内容是关于**在计算机上从数据中产生“模型”(model)的算法**，即“学习算法”。有了学习算法，我们把数据经验提供给他，他就能基于这些数据产生模型；在面对新的情况时，模型会给我们提供相应的判断。

# 1.2 基本术语

**数据集(data set)**：记录的组合。

**示例/样本(sample)**：数据集中的每一条记录，是关于一个事件或对象的描述。

**属性(attribute)/特征(feature)、属性值(attribute value)、属性空间(attribute space)/输入空间/样本空间(sample space)、特征向量(feature vector)**：反映事件或对象在某方面的表现或性质的事项称为属性或特征，属性上的取值称为属性值，属性张成的空间称为属性空间/输入空间/样本空间。由于空间中每个点对应一个坐标向量，因此把一个示例称为一个特征向量。

**维数(dimensionality)**：每个样本中属性的数量。

**分类(classification)与回归(regression)**：想预测的是离散值（如“好瓜/坏瓜”），此类学习任务称为“分类”(classification)；若为连续值（如西瓜的成熟度为0.96、0.84等），此类学习任务称为“回归”(regression)。

**监督学习和无监督学习**：根据训练数据是否有标记信息划分，分类和回归是监督学习的代表；聚类是无监督学习的代表。

# 1.3 假设空间

“归纳”是从**特殊到一般**的“**泛化**”过程，“从样例中学习”是一个归纳过程，因此亦称为“归纳学习”。归纳学习有**广义**和**狭义**之分。广义的归纳学习相当于从样例中学习；狭义的归纳学习要求从训练数据中学得概念，因此也叫做“概念学习”或“概念生成”，由于难度较高，当前的研究和应用均较少。

**假设空间**：学习过程可以看作是一个在所有假设组成的空间中进行搜索的过程，搜索目标是找到与训练集“匹配”的假设。比如有三种属性，每种属性各有三个可取的属性值，因此假设空间规模为：

**版本空间**：需要注意的是，现实问题中假设空间往往很大，学习过程使用的训练样本数量有限，因此，可能有**多个假设**与训练集一致，称该**假设集合**为“版本空间”。

# 1.4 归纳偏好

**版本偏好**用于在**版本空间**中选择**好的假设**作为输出模型。

# 1.5 发展历程

**20世纪50年代到70年代初**，代表工作有“逻辑推理家”程序以及“通用问题求解”程序，机器**具有推理能力**，**能够推导定理**。从**70年代中期开始**，人工智能研究进入“**知识期**”，大量专家系统问世。

**50年代中后期**，基于**神经网络**的“连接主义”学习开始出现，代表工作有感知机、Adaline等。**60到70年代**基于逻辑表示的“符号主义”学习技术蓬勃发展，代表工作有“结构学习系统”、“基于逻辑的归纳学习系统”等；同时以决策理论为基础的学习技术以及强化学习技术也得到发展，代表工作如“学习机器”；之后的统计学习理论的一些奠基性成果也在此时期取得。

**20世纪80年代**机器学习成为一个**独立的学科领域**，各种机器学习技术百花齐放。机器学习研究被划分为多个种类：“**从样例中学习**”、“**在问题求解和规划中学习**”、“**通过观察和发现学习**”、“**从指令中学习**”等等。其中“从样例中学习”是20世纪80年代以来**研究最多**，**应用最广**的机器学习方法，涵盖了监督学习、无监督学习等。

**在20世纪80年代**，“从样例中学习”的一大主流是**符号主义学习**，其代表包括**决策树**和**基于逻辑的学习**，典型的决策学习以信息论为基础，以信息熵的最小化为目标，直接模**拟了人类对概念进行判定的树形流程**。

**20世纪90年代中期之前**，“从样例中学习”的另一个主流是基于**神经网络**的**连接主义学习**。

**20世纪90年代中期**，“**统计学习**”成为主流，代表工作是**支持向量机(SVM)**以及更一般的“**核方法**”。

**21世纪初**，连接主义学习再次兴起，**深度学习**大流行。深度学习虽然缺乏严格的理论基础，但是性能强大，易上手，因此深受喜爱。