

# 定义

机器学习，是研究利用计算机，模仿人类认知和学习的经验，期望通过给计算机“灌输”经验数据，让计算机总结出认知经验（学习算法或模型），最后计算机使用这些经验去认知新的未知事物。

## 基本术语

1. 数据集：对一个事件或者对象的描述的集合。
2. 示例（样本）：数据集中每一个事件或者对象的描述。
3. 属性（特征）：描述事件或者对象的表现或性质的指标。
4. 属性值：每个样本关于属性对应的具体取值。
5. 属性空间（样本空间）：由属性张成的空间。
6. 特征向量：每一个属性（特征）再空间中的所有取值，一个一维数列空间。
7. 维数：每个特征向量的长度。
8. 学习（训练）：计算机学习数据得到模型或经验的过程，用来“灌输”给计算机学习的数据就叫做“训练集”，其中每一条数据就是一组“训练样本”。
9. 假设：计算机从数据中学习到了某种潜在规律。
10. 标记：训练数据中对训练样本的真实判断，其组成的空间就是“标记空间”（输出空间）。
11. 测试：计算机通过总结的模型和经验，用来判断未被标记的数据，来给出的判断，就是测试，而测试的数据就是测试集。
12. 分类&回归：分类任务是为了区分数据中存在多少个不同的分类集合，例如通过图片去分析图片中由多少个人，而回归任务是为了通过学习去预测未知数据的具体结果，例如通过模型去预测下一年的具体收入数据。
13. 聚类：通过对数据的学习，来自动的对数据进行分类，但是计算机并不知道每一类的具体含义，就像把红豆和绿豆在混合的豆子中分离出来，但是没必要知道他们具体的品种。
14. 监督学习&无监督学习：监督学习是数据由现实结论标记的训练方式，回归和分类就是监督学习的代表，而无监督学习就是并不存在现实结论标记的训练方式，并不对数据结果做判断，仅仅是区分，聚类任务就是无监督学习的代表。
15. 泛化：泛化能力才是机器学习的最终目的，就是机器通过训练之后，去识别未知事物的能力。

## 假设空间

1. 假设空间：机器学习中可能的函数构成的空间称为“假设空间”。
2. 版本空间：存在着的一个与训练集一致的“假设集合”。

## 归纳偏好

1. 归纳偏好（偏好）：机器学习算法在学习过程中对某种类型假设的偏好。
2. 奥卡姆剃刀：若有多个假设与观察一致，则选最简单的那个