# 1.机器学习概述

## 1.1.人工智能与机器学习

人工智能（Artificial Intelligence，Al）是一种以智能人类思维的类似方式使计算机，计算机控制的机器人或软件智能地思考的方法。

人工智能的核心，是使计算机具有智能的根本途径，其应用遍及人工智能的各个领域，它主要使用归纳、综合而不是演绎。

人工智能是基于计算机科学，生物学，心理学，语言学，数学和工程学等学科的科学和技术。人工智能的主要推动力是开发与人类智能相关的计算机功能。

IMG_256 

机器学习主要是研究如何使计算机从给定的数据中学习规律，即从观测数据（样本）中寻找规律，并利用学习到的规律（模型）对未知或无法观测的数据进行预测。

针对经验E（experience）和一系列的任务T（tasks）和一定表现的衡量P，如果随之经验E的积累，针对定义好的任务T可以提高表现P，就说计算机具有学习能力。

## 1.2.机器学习分类

**1.按照算法分类**

分为4大类：

分类（Classification）、回归（Regression）、聚类（Clustering）、关联（Relation）

分类（classification）：预测是离散值

比如把人分为好人和坏人之类的学习任务

二分类（binary classification）：只涉及两个类别的分类任务

正类（positive class）：二分类里的一个

反类（negative class）：二分类里的另外一个

多分类（multi-class classification）：涉及多个类别的分类

回归（regression）：预测值是连续值

比如预测下个月的收入是3860元

聚类（clustering）：把训练集中的对象分为若干组

比如将客户分为5类

**2.按照学习任务分类**

分为监督学习和无监督学习。

监督学习（supervised learning）：

从标注数据中学习预测模型的机器学习问题

常见的有监督学习算法有：线性回归、逻辑回归、K-近邻、朴素贝叶斯、决策树、随机森林、支持向量机等。

无监督学习（unsupervised learning）：

从无标注数据中学习预测模型的机器学习问题

常见的无监督学习算法有：聚类、EM算法等。

## 1.3.机器学习应用



计算机视觉、语音识别以及自然语言处理（这里特指文本处理）目前是机器学习领域最常见的几类应用领域。

计算机视觉是一门研究如何让机器能够替代人的眼睛，把看到的图片进行分析、处理的一门科学。在图像分类、人脸识别、车牌识别、自动驾驶中的街景识别等场景均有十分广泛的应用。

语音识别是把语音处理、语义理解等技术和机器学习结合起来。常见的应用有：siri、小冰等语音助手。此外，语音识别经常还会和[自然语言处理](https://zh.wikipedia.org/wiki/%E8%87%AA%E7%84%B6%E8%AF%AD%E8%A8%80%E5%A4%84%E7%90%86" \t "_blank)技术中的机器翻译、语音合成等技术构建出更加复杂的应用，如：语音翻译器。

自然语言处理旨在使用自然语言处理技术使计算机能够“读懂”人类的语言。具体的应用有：谷歌翻译、垃圾邮件的识别、知识图谱等。

## 1.4.机器学习常用术语解释

**数据集：**所有数据的集合成为一个数据集 data set

**示例/样本:**一条记录，即一行数据

**属性/特征：**描述事物的某一性质，比如西瓜的颜色属性，其实就是列名

**属性值：**一个属性可能的取值范围，比如西瓜的颜色属性值有 青绿、乌黑

**属性空间/样本空间：**所有属性构成的空间，比如一个西瓜有颜色、大小、敲声三个属性，就构成一个三维空间

**特征向量：**一个示例用向量表示

**维数：**样本的属性个数

**标记：**示例结果的信息，即要求解的y

**假设：**模型对应了数据的某种规律 ，所以模型也称为假设

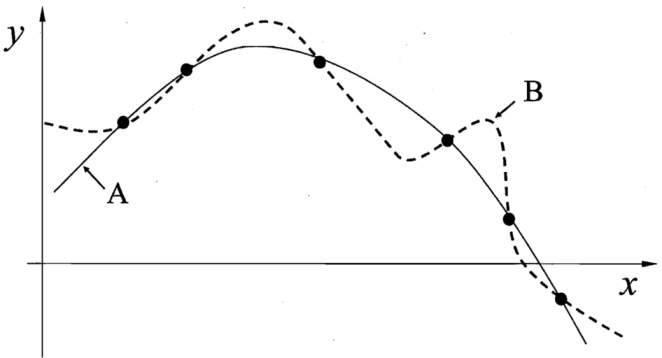
**训练集：**训练模型使用的数据叫训练集

**测试集：**用训练集训练处模型后，被预测的样本叫测试集/测试样本

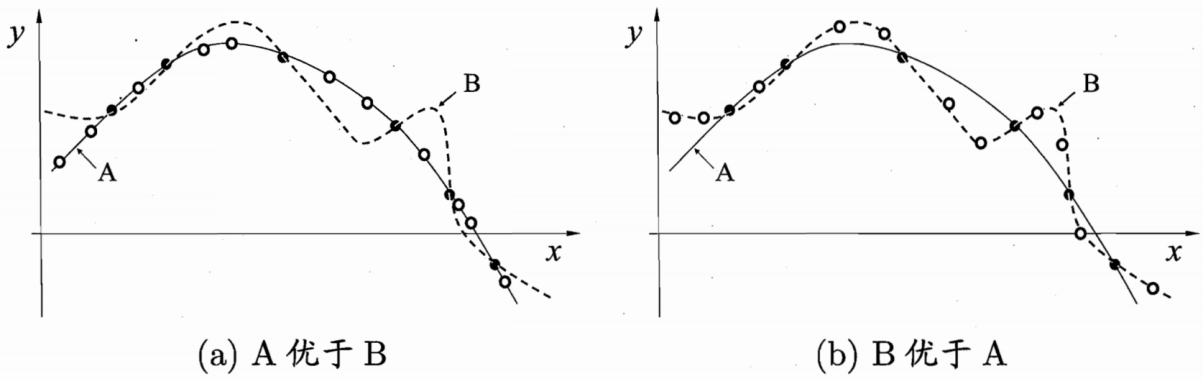
**泛化能力：**最终得到的模型适用于新样本的能力。模型是基于已有数据训练并测试的，但是最后还是要用来预测今后实际的未知的数据

两个原则：

1.奥卡姆剃刀原则：若多个假设与观察一致，则选择最简单的一个，比如下图选择较为平滑的那个。



2.没有免费的午餐定理NFL：在不考虑具体问题的情况下，没有任何一个算法比另一个算法更优，即没有一种机器学习算法是适用于所有情况的。



如上图，对于某些问题，A可能优于B，但一定会相应存在一些情况B优于A。所以，具体问题具体分析。