机器学习

机器学习(Machine Learning, ML)是一门多领域交叉学科，涉及**概率论、统计学、逼近论、凸分析、算法复杂度理论**等多门学科。专门研究计算机**怎样模拟或实现人类的学习行为**，以**获取新的知识或技能**，重新组织已有的知识结构使之不断改善自身的性能。 它是人工智能的核心，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**。【实质上，最主要的内容应该是概率论、统计学、线性代数、微积分等】

**一、概念**

[描述来源： Mitchell, T. (1997). Machine Learning. McGraw Hill. ]

特点：**不需要先在大量的数据中进行人工分析，然后提取规则并建立模型**，而是提供了一种**更有效的方法来捕获数据中的知识**，逐步提高预测模型的性能，以完成**数据驱动**的决策。

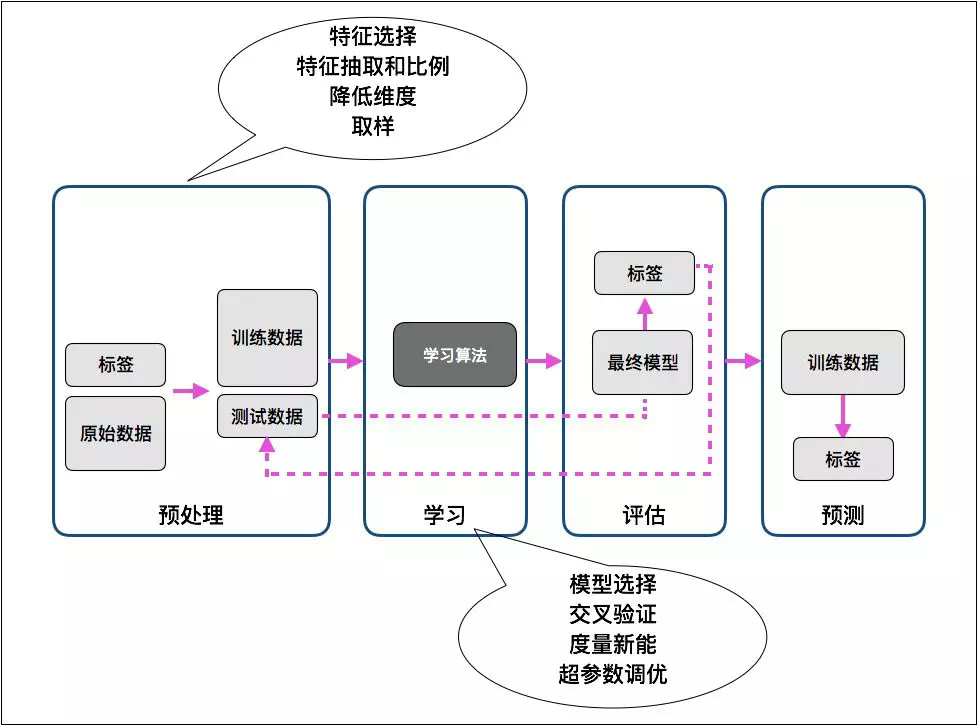
优势：可缩短编程时间的工具，因为不需要提取规则进行建模，只需要进行一定的训练就可以自动建模，因而缩短了编程时间；

优势2：定义自己的产品？；

应用：人脸识别、图文识别、语音识别、信息检索、信息服务、爬虫，这些领域传统编程很难解决，机器学习则可通过特征值的方式来实现。

【趋势：未来可以通过机器学习进行**文学艺术创作，科技预测，产品设计**】。

流程：



**二、分类**

**1.有监督**

**1.1 定义（**Supervised learning**）**

从**有标签**的训练数据中学习模型，以便对未知或未来的数据做出**预测**。

**1.2 分类**

**1.2.1 回归模型**

**1）**概念：回归分析是一种用于估计变量之间的关系（当一个自变量变化而其它变量固定时，因变量会如何变化）的统计过程

**2）**特点：**特征值连续**

**3）常见类型：线性回归**和**逻辑回归**（Logistic Regression）。另外还有**多变量回归、泊松回归、逐步回归、脊回归**（Ridge Regression）、**套索回归**（Lasso Regression）和**多项式回归**等等。**随机梯度下降**（SGD）就是一种起源于回归分析的常用方法，可用于控制复杂度。

4）作用：预测。

**1.2.2 分类模型**

**2）**特点：**结果一般不连续**

**1.3 流程**

1.3.1. 预处理——**整理数据**

**1）特征值的选取——需具备专业知识。有的**要求所选择特征的测量结果具有相同的单位。

**2）转换：**

**a)**把特征数据变换为[0, 1]的取值范围或者均值和单位方差为0的标准正态分布来实现。

**b)** 降维:

**情形一**：某些选定的特征可能是**高度相关**的，因此在某种程度上是多余的。

**情形二：**数据集包含大量不相关的特征或噪声，即数据集**具有较低的信噪比。**

**1.3.2 训练和选择模型预测**

**1.3.3 评估模型和预测新样本**

**2.无监督**

2.1 特点：训练使用的数据是无标签或结构未知的数据。

2.2 常用方法：

2.2.1聚类

【将物理或抽象对象的集合分成由类似的对象组成的多个类的过程被称为聚类。由聚类所生成的簇是一组数据对象的集合，这些对象与同一个簇中的对象彼此相似，与其他簇中的对象相异。“物以类聚，人以群分”，在自然科学和社会科学中，存在着大量的分类问题。**聚类分析又称群分析**，它是研究（样品或指标）分类问题的一种统计分析方法。聚类分析起源于分类学，但是聚类不等于分类。聚类与分类的不同在于，聚类所要求划分的类是未知的。聚类分析内容非常丰富，有**系统聚类法**、**有序样品聚类法**、**动态聚类法**、**模糊聚类法**、**图论聚类法**、**聚类预报法**等。】

2.2.2 降维

**3.强化学习**

**1.概念**

强化学习是一种试错方法，其目标是让软件智能体在特定环境中能够采取回报最大化的行为。强化学习在马尔可夫决策过程环境中主要使用的技术是**动态规划**（Dynamic Programming）。流行的强化学习方法包括**自适应动态规划**（ADP）、**时间差分**（TD）学习、**状态-动作-回报-状态-动作**（SARSA）算法、**Q 学习**、**深度强化学习**（DQN）；其应用包括**下棋**类游戏、**机器人**控制和**工作调度**等。

**2．目标和方式**

强化学习的目标是**开发系统或代理**，通过它们与环境的交互来提高其预测性能。当前环境状态的信息通常包含所谓的**奖励信号**，可以把强化学习看作是与有**监督学习**相关的领域。

【逻辑：人工智能领域用逻辑来理解智能推理问题；它可以提供用于分析编程语言的技术，也可用作分析、表征知识或编程的工具。目前人们常用的逻辑分支有命题逻辑（Propositional Logic ）以及一阶逻辑（FOL）等谓词逻辑。】