（了解）机器学习的定义和发展史

机器学习的定义：

* 机器学习是对能通过经验自动改进的计算机算法的研究。

——汤姆·米切尔（Tom Mitchell）[Mitchell, 1997]

* 通俗地讲，机器学习(Machine Learning，ML)就是让计算机从数据中进行自动学习，得到某种知识(或规律)。
* 作为一门学科，机器学习通常指一类问题以及解决这类问题的方法，即如何从观测数据 (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。

机器学习的发展史：

* 神经元模型研究：20世纪50年代中期到60年代初期，也被称为机器学习的热烈时期，最具有代表性的工作是罗森勃拉特1957年提出的感知器模型。
* 符号概念获取：20世纪60年代中期到70年代初期。其主要研究目标是模拟人类的概念学习过程。这一阶段神经学习落入低谷，称为机器学习的冷静时期。
* 知识强化学习：20世纪70年代中期到80年代初期。人们开始把机器学习与各种实际应用相结合，尤其是专家系统在知识获取方面的需求，也有人称这一阶段为机器学习的复兴时期。温斯顿的概念学习和昆兰的决策树ID3算法。
* 连接学习和混合型学习：20世纪80年代中期至今。把符号学习和连接学习结合起来的混合型学习系统研究已成为机器学习研究的一个新的热点。

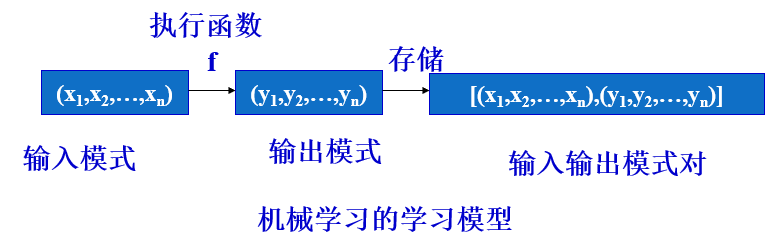
（理解）知识发现

知识发现(Knowledge Discovery in Database, KDD)，是所谓"[数据挖掘](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E6%8C%96%E6%8E%98)"的一种更广义的说法，即**从各种信息中，根据不同的需求获得知识的过程**。知识发现的目的是向使用者屏蔽[原始数据](https://baike.baidu.com/item/%E5%8E%9F%E5%A7%8B%E6%95%B0%E6%8D%AE/10119290)的繁琐细节，从原始数据中提炼出有意义的、简洁的知识，直接向使用者报告。基于数据库的知识发现([KDD](https://baike.baidu.com/item/KDD/1407460))和[数据挖掘](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E6%8C%96%E6%8E%98)还存在着混淆，通常这两个术语替换使用。KDD表示将低层[数据转换](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E8%BD%AC%E6%8D%A2)为高层知识的整个过程。可以将KDD简单定义为：KDD是确定数据中有效的、新颖的、潜在有用的、基本可理解的模式的特定过程。而数据挖掘可认为是观察数据中模式或模型的抽取，这是对数据挖掘的一般解释。虽然数据挖掘是知识发现过程的核心，但它通常仅占KDD的一部分(大约是15% 到25%) 。因此数据挖掘仅仅是整个KDD过程的一个步骤，对于到底有多少步以及哪一步必须包括在KDD过程中没有确切的定义。然而，通用的过程应该接收原始数据输入，选择重要的[数据项](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E9%A1%B9/3227309)，缩减、预处理和浓缩数据组，将数据转换为合适的格式，从数据中找到模式，评价解释发现结果。

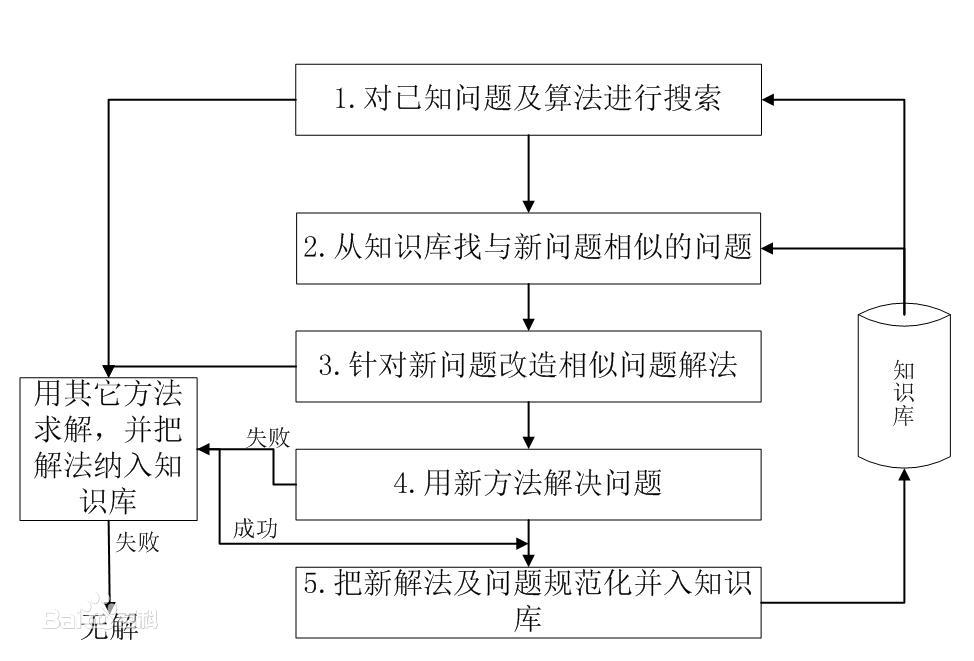
（理解）机器学习的主要策略与基本结构

按照学习中使用推理的多少，机器学习所采用的策略大体上可分为4种：

* 机械学习（记忆学习）
* 示教学习
* 类比学习
* 示例学习



示教学习：学生从环境（教师或其它信息源如教科书等）获取信息，把知识转换成内部可用的表示形式，并将新的知识和原有知识有机地结合。

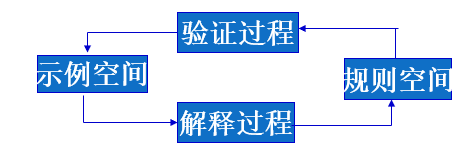


**示例空间**：是我们向系统提供的示教例子的集合。研究问题：例子质量，搜索方法。

**解释过程**：是从搜索到的示例中抽象出一般性的知识的归纳过程。解释方法：常量转换为变量，去掉条件，增加选择，曲线拟合等。

**规则空间**：是事务所具有的各种规律的集合。研究问题：对空间的要求，搜索方法

**验证过程**：是要从示例空间中选择新的示例，对刚刚归纳出的规则做进一步的验证和修改。



基于学习方式，机器学习所采用的策略大体上可分为3种：

* **有导师学习（监督学习）**：从标记的训练数据集学习一个模型，再用此模型对测试样本集进行预测。输入数据中有导师信号，以概率函数、代数函数或人工神经网络为基函数模型，采用迭代计算方法，学习结果为函数。
* **无导师学习（非监督学习）**：直接对输入数据进行建模，提取数据有效信息探索数据的整体结构。输入数据中无导师信号，采用聚类方法，学习结果为类别。典型的无导师学习有发现学习、聚类、竞争学习等。
* **强化学习（增强学习）**：智能体以“试错”的方式进行学习，通过与环境进行交互获得的奖赏指导行为，目标是使智能体获得最大的奖赏。以环境反馈（奖/惩信号）作为输入，以统计和动态规划技术为指导的一种学习方法。

基于学习目标，机器学习所采用的策略大体上可分为5种：

* **概念学习**：即学习的目标和结果为概念，或者说是为了获得概念的一种学习。典型的概念学习有示例学习。
* **规则学习**：即学习的目标和结果为规则，或者说是为了获得规则的一种学习。典型的规则学习有决策树学习。
* **函数学习**：即学习的目标和结果为规则，或者说是为了获得函数的一种学习。典型的函数学习有神经网络学习。
* **类别学习**：即学习的目标和结果为对象类，或者说是为了获得类别的一种学习。典型的类别学习有聚类分析。
* **贝叶斯网络学习**：即学习的目标和结果是贝叶斯网络，或者说是为了获得贝叶斯网络的一种学习。其又可分为结构学习和参数学习。

（掌握）归纳学习

归纳学习是指以归纳推理为基础的学习，其任务是要从关于某个概念的一系列已知正例和反例中归纳出一个一般的概念描述。

1. 示例学习

归纳学习的一种特例。它给学习者提供某一概念的一组正例和反例，学习者归纳出一个总的概念描述，并使这个描述适合于所有的正例，排除所有的反例。

1. 决策树学习

一种以示例为基础的归纳学习方法，目前最流行的归纳学习方法之一。在现有的各种决策树学习算法中，影响较大的是ID3算法。本节主要讨论决策树的概念和决策树学习的ID3算法。

（掌握）决策树（ID3算法）

决策树是一种由节点和边构成的用来描述分类过程的层次数据结构。树的根节点表示分类的开始，叶节点表示一个实例的结束，中间节点表示相应实例中的某一属性，而边则代表某一属性可能的属性值。

在决策树中，从根节点到叶节点的每一条路径都代表一个具体的实例，并且同一路径上的所有属性之间为合取关系，不同路径（即一个属性的不同属性值）之间为析取关系。

决策树的分类过程就是从这棵树的根节点开始，按照给定的实例的属性值去测试对应的树枝，并依次下移，直至到达某个叶节点为止。

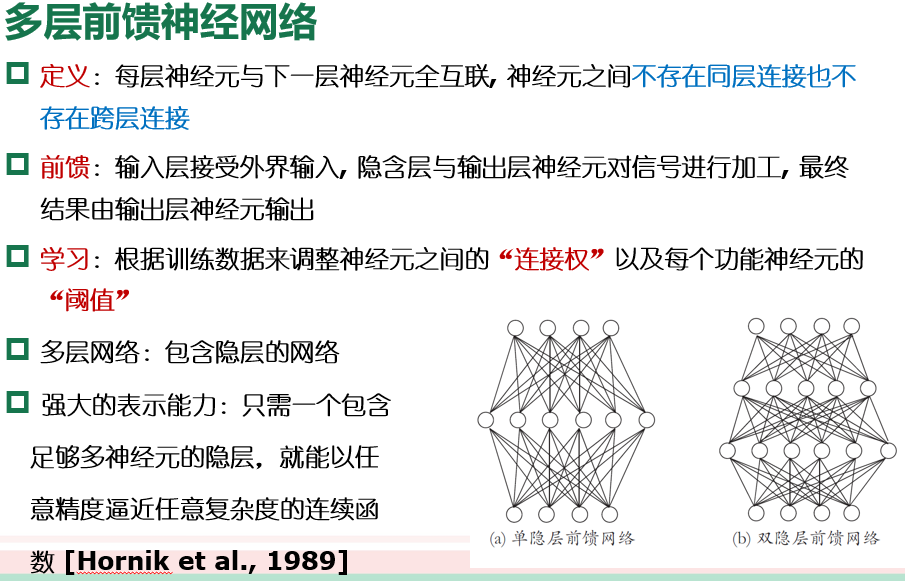
ID3算法是昆兰(J.R.Quinlan)于1979年提出的一种以信息熵(entropy)的下降速度作为属性选择标准的一种学习算法。其输入是一个用来描述各种已知类别的例子集。学习结果是一棵用于进行分类的决策树。

ID3算法的学习过程：

1. 首先以整个例子集作为决策树的根节点S，并计算S关于每个属性的条件熵；
2. 然后选择能使S的条件熵为最小的一个属性对根节点进行分裂，得到根节点的一层子节点；
3. 接着再用同样的方法对这些子节点进行分裂，直至所有叶节点的熵值都下降为0为止。
4. 这时，就可得到一棵与训练例子集对应的熵为0的决策树，即ID3算法学习过程所得到的最终决策树。该树中每一条从根节点到叶节点的路径，都代表了一个分类过程，即决策过程。

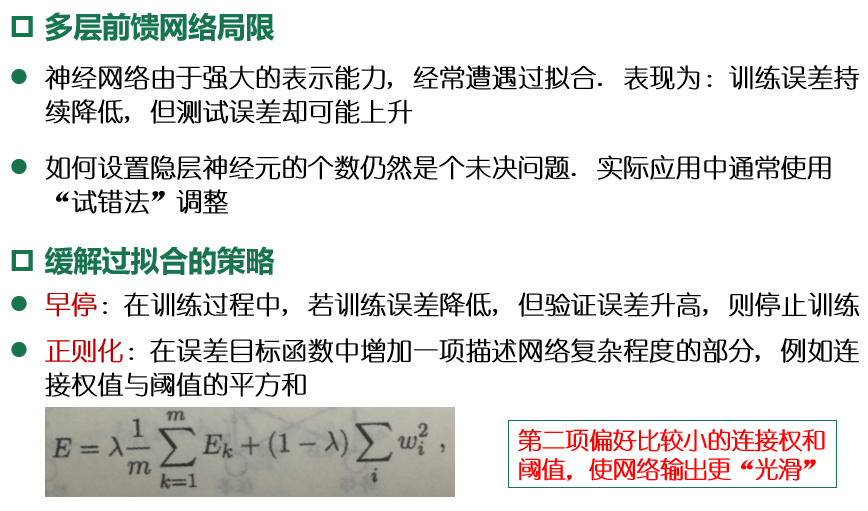
ID3 算法的核心是在决策树各个结点上应用信息增益准则选择特征，递归地构建决策树。

（掌握）BP神经网络



误差逆传播算法是最成功的训练多层前馈神经网络的学习算法。BP算法基于梯度下降策略，以目标的负梯度方向对参数进行调整。

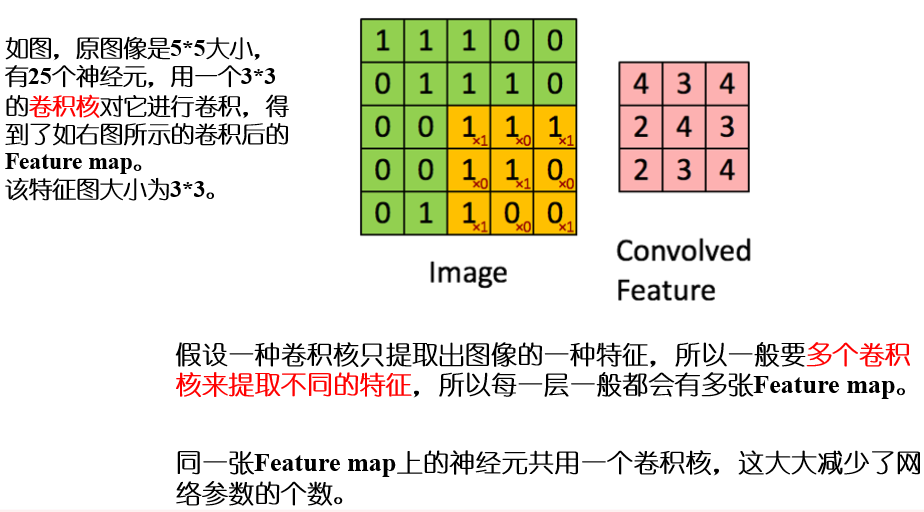




（掌握）卷积神经网络（CNN）学习算法

卷积神经网络是人工神经网络的一种，已成为当前语音分析和图像识别领域的研究热点。它的权值共享网络结构使之更类似于生物神经网络，降低了网络模型的复杂度，减少了权值的数量。该优点在网络的输入是多维图像时表现的更为明显，使图像可以直接作为网络的输入，避免了传统识别算法中复杂的特征提取和数据重建过程。卷积网络是为识别二维形状而特殊设计的一个多层感知器，是针对图像处理特殊设计的网络结构，这种网络结构对平移、比例缩放、倾斜或者共他形式的变形具有高度不变性。

Convolution过程：



Pooling过程：

