

第 7 章 专家系统与机器

学习

----- 机器学习



第7章 专家系统与机器学习

- 7.5.1 机器学习的基本概念
- 7.5.2 符号学习
- 7.5.3 深度学习

第7章 专家系统与机器学习

✓ 7.5.1 机器学习的基本概念

■ 7.5.2 符号学习

■ 7.5.3 深度学习

7.5.1 机器学习的基本概念

- 7.5.1.1 学习
- 7.5.1.2 机器学习
- 7.5.1.3 机器学习系统
- 7.5.1.4 机器学习的发展
- 7.5.1.5 机器学习的分类
- 7.5.1.6 机器学习的比较

7.5.1.1 学习

- (1) 学习是系统改进其性能的过程：西蒙，1980。
- (2) 学习是获取知识的过程。
- (3) 学习是技能的获取。
- (4) 学习是事物规律的发现过程。

- 学习：一个有特定目的的知识获取过程。
- 学习的内在行为：获取知识、积累经验、发现规律。
- 学习的外部表现：改进性能、适应环境、实现系统的自我完善。

7.5.1.2 机器学习

- **机器学习**: 计算机能模拟人的学习行为, 自动地通过学习获取知识和技能, 不断改善性能, 实现自我完善。

主要研究内容:

- (1) 学习机理
- (2) 学习方法
- (3) 学习系统

7.5.1.3 机器学习系统

1. 机器学习系统的定义

- **学习系统**：能够在一定程度上实现机器学习的系统。
- 萨利斯 (**Saris**) 的定义（**1973** 年）：能够从某个过程或环境的未知特征中学到有关信息，并且能把学到的信息用于未来的估计、分类、决策或控制，以便改进系统的性能。
- 施密斯等的定义（**1977** 年）：在与环境相互作用时，能利用过去与环境作用时得到的信息，并提高其性能。

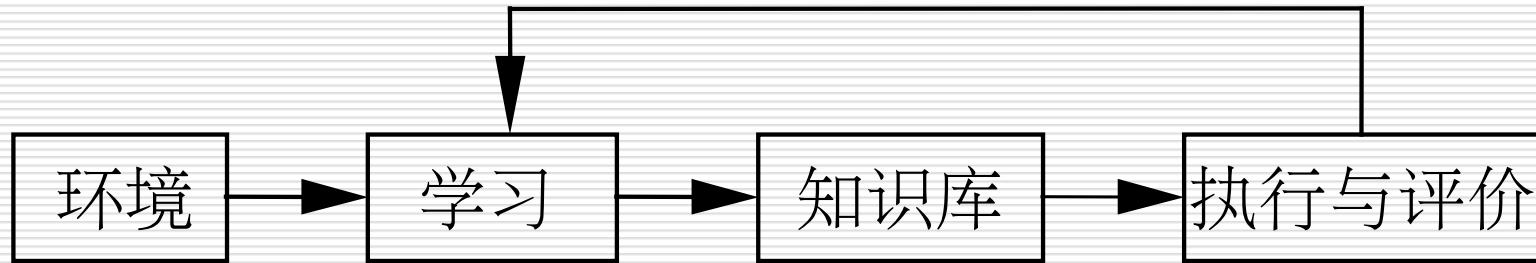
7.5.1.3 机器学习系统

2. 机器学习系统的条件和能力

- (1) 具有适当的学习环境
- (2) 具有一定的学习能力
- (3) 能应用学到的知识求解问题
- (4) 能提高系统的性能

7.5.1.3 机器学习系统

3. 机器学习系统的基本模型



学习系统的基本结构

7.5.1.4 机器学习的发展

1. 神经元模型的研究（20世纪50年代中期——）

- 主要研究工作：应用决策理论的方法研制可适应环境的通用学习系统（**general purpose learning system**）。
- 1957年，罗森勃拉特（**F. Rosenblatt**）提出感知器模型。
 -
- 1969年，明斯基和佩珀特（**Papert**）发表了论著《**Perceptron**》，对神经元模型的研究作出了悲观的论断。

7.5.1.4 机器学习的发展

2. 符号学习的研究（**20**世纪**70**年代中期-----）

- 莫斯托夫（**D. J. Mostow**）的指导式学习。
- 温斯顿（**Winston**）和卡鲍尼尔（**J. G. Carbonell**）的类比学习。
- 米切尔（**T. M. Mitchell**）等人的解释学习。

7.5.1.4 机器学习的发展

3. 连接学习的研究（20世纪80年代——）

- 连接学习是一种以非线性大规模并行处理为主流的神经网络的研究，特别是深度学习研究目前仍在继续进行之中。
- 符号学习的研究同时取得很大进展，它与连接学习各有所长，具有较大的互补性。
- 连接学习适用于连续发音的语音识别及连续模式的识别；而符号学习在离散模式识别及专家系统的规则获取。现在把两者结合起来研究。
- 1980**年召开了第一届机器学习国际研讨会。
- 1986**年创刊了第一本机器学习杂志《Machine Learning》。

7.5.1.5 机器学习的分类

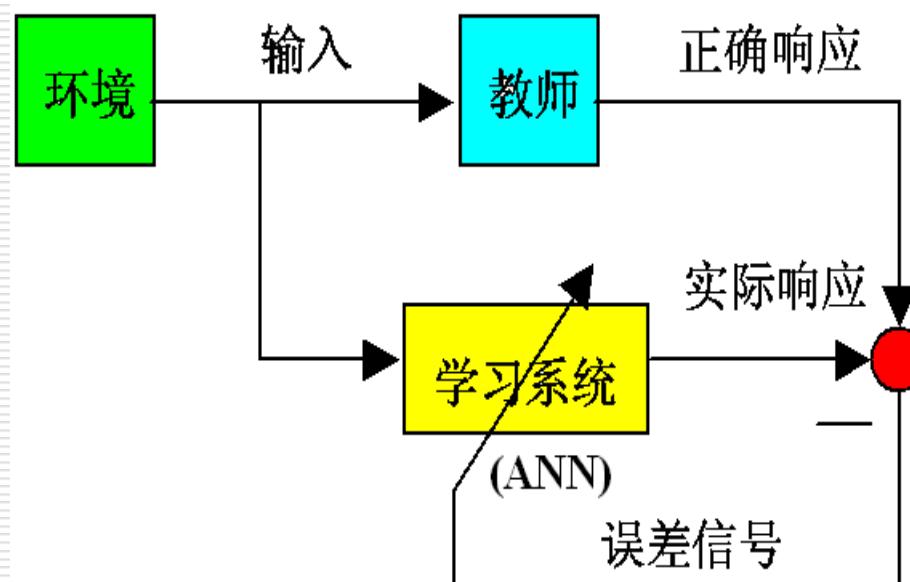
1. 按学习方法分类 (温斯顿, 1977) :

- 机械式学习
- 指导式学习
- 示例学习
- 类比学习
- 解释学习

7.5.1.5 机器学习的分类

2. 按学习能力分类:

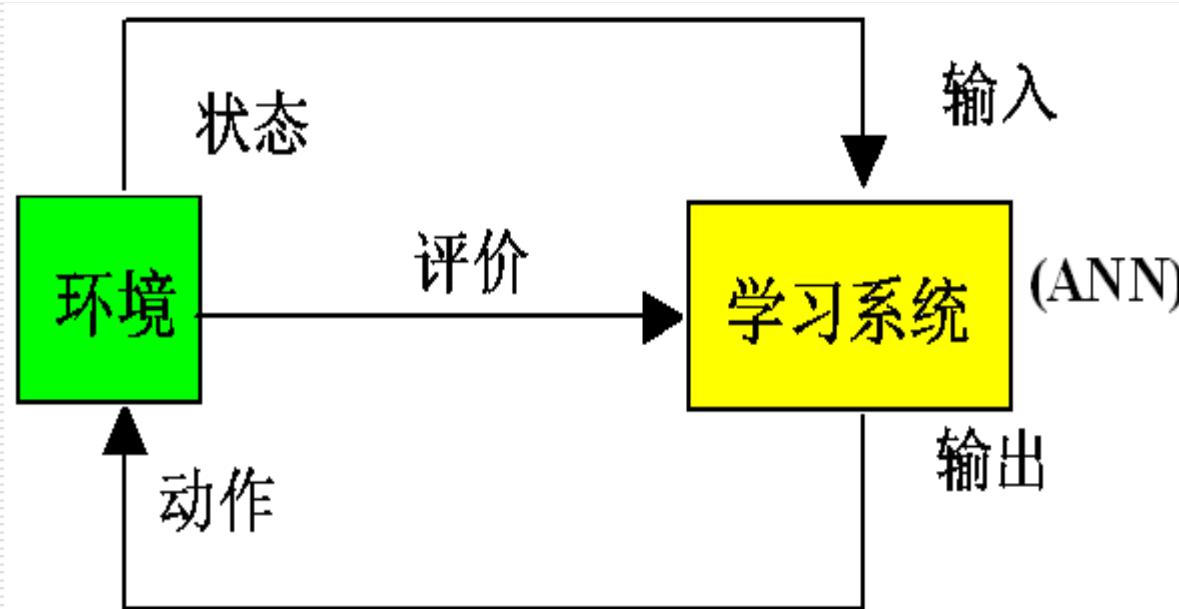
- 监督学习（有教师学习）



7.5.1.5 机器学习的分类

2. 按学习能力分类：

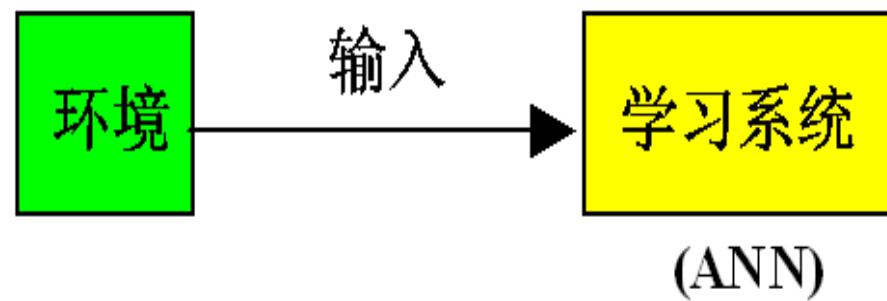
- 强化学习（Reinforcement Learning），又称为再励学习或增强学习



7.5.1.5 机器学习的分类

2. 按学习能力分类：

- 非监督学习（无教师学习）



7.5.1.5 机器学习的分类

3. 按推理方式分类:

基于演绎的学习（解释学习）。

基于归纳的学习（示例学习、发现学习等）。

4. 按综合属性分类:

归纳学习、分析学习、连接学习以及遗传算法等

◦

第7章 机器学习

- 7.5.1.1 学习
- 7.5.1.2 机器学习
- 7.5.1.3 机器学习系统
- 7.5.1.4 机器学习的发展
- 7.5.1.5 机器学习的分类
- 7.5.1.6 机器学习的比较

7.5.1.6 机器学习方法的比较

■ 以推理能力排列

- 机械式学习，指导式学习，解释学习，类比学习，示例学习，观察与发现学习。

■ 对领域理论的要求

- 示例学习、观察与发现学习：领域理论要求较少。
- 解释学习：要求提供完善的领域知识。

■ 适用领域

- 连接学习：模拟人类较低级的神经活动。
- 符号学习：模拟人类的高级思维活动。

7.5.1.6 机器学习方法的比较

■ 知识获取角度：

- 示例学习、观察与发现学习：通过学习可以产生新概念描述，可用于专家系统的知识获取。
- 解释学习的学习目标主要是改善系统的效率，而不扩充概念描述的范围。
- 指导式学习通过与指导者（如领域专家）的交互学习新知识，同时又可帮助指导追踪推理过程，发现其中的错误，找出产生错误的原因，然后由指导者进行修正。

机器学习的展望

- (1) 人类学习机制的研究。
- (2) 发展和完善现有的学习方法，并开展新的学习方法的研究。
- (3) 建立实用的学习系统，特别是多种学习方法协同工作的集成化系统的研究。
- (4) 机器学习的结构模型、计算理论、算法和混合学习的有关理论及应用的研究。

第7章 机器学习

7.5.1 机器学习的基本概念

- 7.5.2 符号学习
- 7.5.3 知识发现与数据挖掘
- 7.5.3 深度学习

7.5.2 符号学习



7.5.2.1 机械式学习



7.5.2.2 指导式学习



7.5.2.3 归纳学习



7.5.2.4 示例学习



7.5.2.5 观察与发现学习



7.5.2.6 类比学习



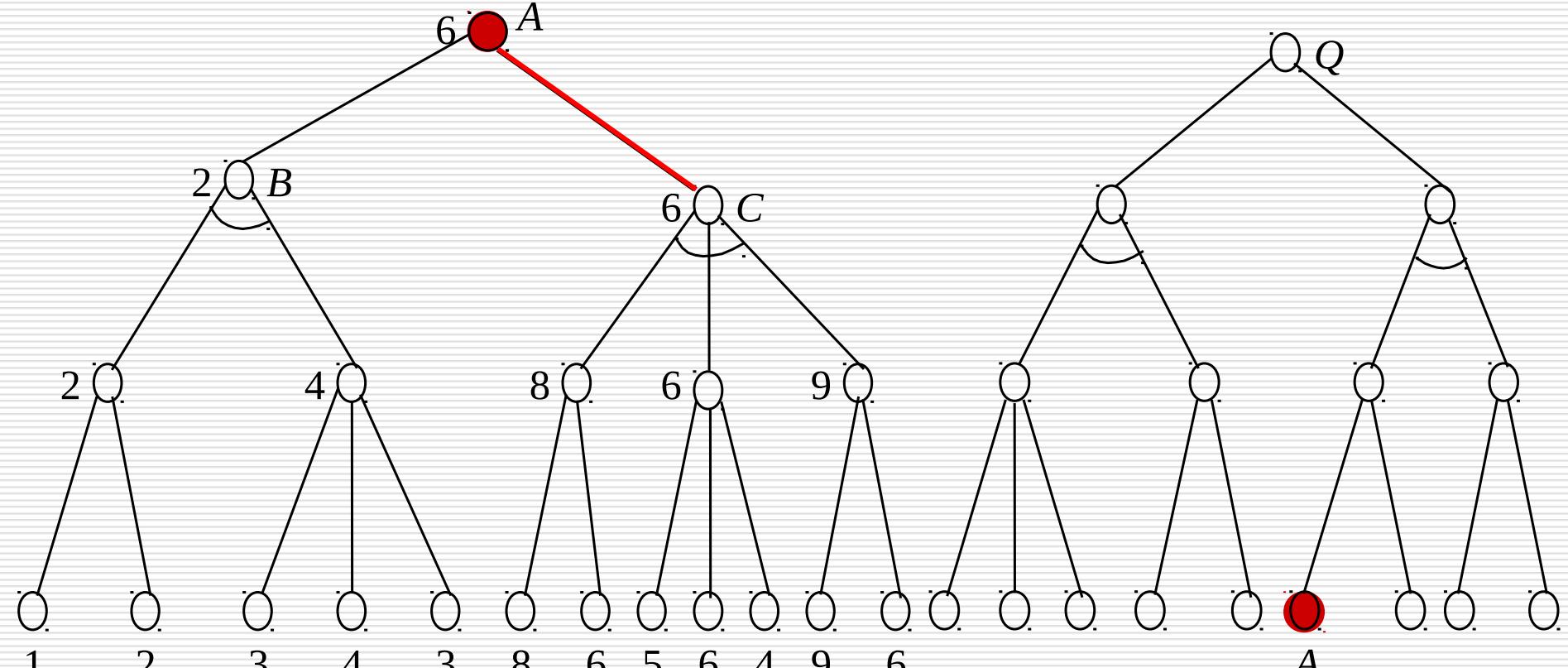
7.5.2.7 解释学习

7.5.2.1 机械式学习

- 机械式学习（**rote learning**）又称记忆学习，或死记式学习：通过直接记忆或者存储外部环境所提供的信息达到学习的目的，并在以后通过对知识库的检索得到相应的知识直接用来求解问题。
- 机械式学习实质是用存储空间来换取处理时间。

7.5.2.1 机械式学习

■ 塞缪尔的跳棋程序 **CHECKERS**



博弈搜索树

以 A 为结点的博弈树

7.5.2.2 指导式学习

- **指导式学习：**由外部环境向系统提供一般性的指示或建议，系统把它们具体地转化为细节知识并送入知识库中。在学习过程中要反复对形成的知识进行评价，使其不断完善。
- **指导式学习的学习过程：**征询指导者的指示或建议、把征询意见转换为可执行的内部形式、加入知识库、评价。

7.5.2.2 指导式学习

1. 征询指导者的指示或建议

- 简单征询：指导者给出一般性的意见，系统将其具体化。
- 复杂征询：系统不仅要求指导者给出一般性的建议，而且还要具体地鉴别知识库中可能存在的问题，并给出修改意见。
- 被动征询：系统只是被动地等待指导者提供意见。
- 主动征询：系统不只是被动地接受指示，而且还能主动地提出询问，把指导者的注意力集中在特定的问题上。

7.5.2.2 指导式学习

2. 把征询意见转换为可执行的内部形式

■ 学习系统应具有把用约定形式表示的征询意见转化为计算机内部可执行形式的能力，并且能在转化过程中进行语法检查及适当的语义分析。

3. 加入知识库

■ 在加入过程中要对知识进行一致性检查，以防止出现矛盾、冗余、环路等问题。

4. 评价

■ 评价方法：对新知识进行经验测试，即执行一些标准例子，然后检查执行情况是否与已知情况一致。

7.5.2.3 归纳学习

- 归纳学习是用归纳推理进行学习的一类学习方法。
- 归纳推理：应用归纳方法所进行的推理，即从足够多的事例中归纳出一般性的知识。
- 它是一种从个别到一般、从部分到整体的推理。
- 归纳推理的重要特征：归纳出的结论不能绝对保证它的正确性，只能以某种程度相信它为真。
- 由“麻雀会飞”、“鸽子会飞”、“燕子会飞”……
归纳结论：“有翅膀的动物会飞”、“长羽毛的动物会飞”
反例：鸵鸟不会飞

7.5.2.3 归纳学习

1. 枚举归纳

■ 从个别事例归纳出一般性知识的方法：

- 设 a_1, a_2, \dots, a_n : 某类事物 A 中的具体事物。
- 已知 a_1, a_2, \dots, a_n 都有属性 P ，并且没有发现反例。
- 当 n 足够大时，可得出：“ A 中所有事物都有属性 P ”。

7.5.2.3 归纳学习

1. 枚举归纳

■ 例如，设有如下已知事例：

张三是足球运动员，他的体格健壮。

李四是足球运动员，他的体格健壮。

.....

.....

刘六是足球运动员，他的体格健壮。

■ 事例足够多时，可归纳出一般性知识：

凡是足球运动员，他的体格一定健壮 **(0.9)**

7.5.2.3 归纳学习

2. 联想归纳

■ 已知两个事物 a 与 b 有 n 个属性相似或相同，即：

a 具有属性 P_1 ， b 也具有属性 P_1 。

a 具有属性 P_1 ， b 也具有属性 P_2 。

.....

.....

a 具有属性 P_1 ， b 也具有属性 P_n 。

且 a 具有属性 P_{n+1} ，则当 n 足够大时，可归纳出

b 也具有属性 P_{n+1} 。

7.5.2.3 归纳学习

3. 类比归纳

- 设: $A = \{a_1, a_2, \dots\}$ $B = \{b_1, b_2, \dots\}$
且 $P(a_i) \rightarrow Q(b_i)$ $i = 1, 2, \dots$
- 则当 A 与 B 中有新元素出现时 (设 A 中的 a_i 及 B 中的 b_j) , 若已知 a_i 有属性, 就可得出 b_j 有属性, 即
$$P(a') \rightarrow Q(b')$$

7.5.2.3 归纳学习

4. 逆推理归纳

■ 一般模式：

- (1) 若 H 为真时，则 $H \rightarrow E$ 必为真或以置信度 cf_1 成立。
- (2) 观察到 E 成立或以置信度 cf_2 成立。
- (3) 则 H 以某种置信度 (cf) 成立。

■ 用公式表示：
$$\frac{H \rightarrow E \quad cf_1}{E \quad cf_2} \quad cf$$

7.5.2.3 归纳学习

4. 逆推理归纳 (续)

■ cf 的计算方法：

$$cf_1' = P(H/E) = \frac{P(E/H) \times P(H)}{P(E)} = cf_1 \times \frac{P(H)}{P(E)}$$

$$cf = cf_1' \times cf_2$$

7.5.2.3 归纳学习

5. 消除归纳

■ 消除归纳：通过不断否定原先的假设来得出结论。

■ 已知： $A_1 \vee A_2 \vee \dots \vee A_i \vee \dots \vee A_n$

$$\neg A_1$$

⋮

$$\neg A_{i-1}$$

$$\neg A_{i+1}$$

⋮

$$\neg A_n$$

■ 结论：

$$A_i$$

7.5.2.3 归纳学习

演绎推理与归纳推理的比较

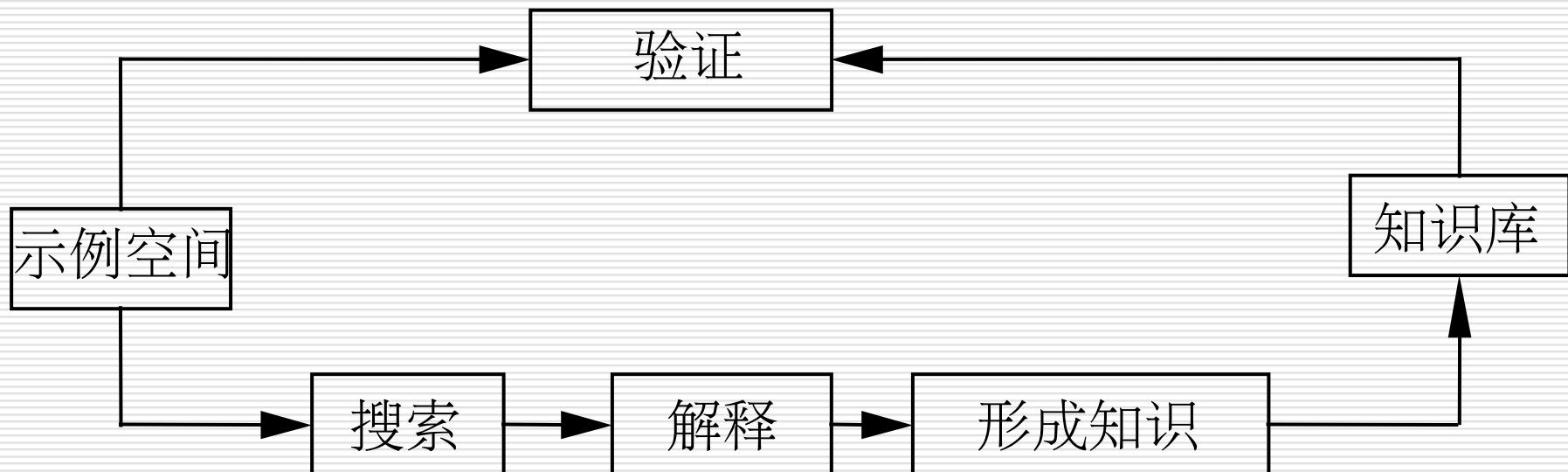
演绎推理	归纳推理
从一般到个别	从个别到一般
必然性推理	或然性推理
<ul style="list-style-type: none">■ 结论不会超出前提所断定的范围■ 不能获取新知识	<ul style="list-style-type: none">■ 结论适用于更大的范围■ 可获取新知识

7.5.2.4 示例学习

- **示例学习：**通过从外部环境中取得若干与某概念有关的例子，经归纳得出一般性概念的一种学习方法。
- **示例学习中，**外部环境（教师）提供一组例子（正例和反例），然后从这些特殊知识中归纳出适用于更大范围的一般性知识，它将覆盖所有的正例并排除所有反例。

7.5.2.4 示例学习

1. 示例学习的学习模型



7.5.2.4 示例学习

2. 形成知识的方法

(1) 变量代换常量

- 例如，假设有两个关于扑克牌“同花”概念的示例。

示例 1 花色 (c_1 , 梅花) \wedge 花色 (c_2 , 梅花) \wedge 花色 (c_3 , 梅花) \wedge 花色 (c_4 , 梅花) \rightarrow
同花 (c_1, c_2, c_3, c_4)

示例 2 花色 (c_1 , 红桃) \wedge 花色 (c_2 , 红桃) \wedge 花色 (c_3 , 红桃) \wedge 花色 (c_4 , 红桃) \rightarrow
同花 (c_1, c_2, c_3, c_4)

- 可得到一条一般性的知识：

规则 1：花色 (c_1, x) \wedge 花色 (c_2, x) \wedge 花色 (c_3, x) \wedge 花色 (c_4, x) \rightarrow
同花 (c_1, c_2, c_3, c_4)

7.5.2.4 示例学习

2. 形成知识的方法

(2) 舍弃条件

- 例如示例：

花色 (c_1 , 红桃) \wedge 点数 (c_1 , 2) \wedge
花色 (c_2 , 红桃) \wedge 点数 (c_2 , 4) \wedge
花色 (c_3 , 红桃) \wedge 点数 (c_3 , 6) \wedge
花色 (c_4 , 红桃) \wedge 点数 (c_4 , 8) \rightarrow
同花 (c_1, c_2, c_3, c_4)

- 可得到一条一般性的知识：

规则 1：花色 (c_1, x) \wedge 花色 (c_2, x) \wedge 花色 (c_3, x) \wedge 花色 (c_4, x) \rightarrow
同花 (c_1, c_2, c_3, c_4)

7.5.2.4 示例学习

2. 形成知识的方法

(3) 增加操作

- 前件析取法
- 例如关于“脸牌”示例：

示例1： 点数 (c_1, J) → 脸 (c_1)

示例2： 点数 (c_1, Q) → 脸 (c_1)

示例3： 点数 (c_1, K) → 脸 (c_1)

得到知识：

规则2： 点数 $(c_1, J) \wedge$ 点数 $(c_2, Q) \wedge$ 点数 $(c_3, K) \rightarrow$ 脸 (c_1)

7.5.2.4 示例学习

2. 形成知识的方法

(3) 增加操作

- 内部析取法
- 例如示例：

示例1： 点数 $(c_1) \in \{J\} \rightarrow$ 脸 (c_1)

示例2： 点数 $(c_1) \in \{Q\} \rightarrow$ 脸 (c_1)

示例3： 点数 $(c_1) \in \{K\} \rightarrow$ 脸 (c_1)

得到知识：

点数 $(c_1) \in \{J, Q, K\} \rightarrow$ 脸 (c_1)

7.5.2.4 示例学习

2. 形成知识的方法

(4) 合取变析取

- 例如：“男同学与女同学可以组成一个班”。
- 归纳：“男同学或女同学可以组成一个班”。

(5) 归结归纳

- 例如：

$$P \wedge E_1 \rightarrow H$$

$$\neg P \wedge E_1 \rightarrow H$$

- 得到： $E_1 \vee E_2 \rightarrow H$

7.5.2.4 示例学习

2. 形成知识的方法

(6) 曲线拟合

- 设在示例空间提供了一批如下形式的示例: (x, y, z)

示例 1 : $(1, 0, 10)$

示例 2 : $(2, 1, 18)$

示例 3 : $(-1, -2, -6)$

- 应用曲线拟合法得到: $z=2x+6 y+8$

7.5.2.5 观察与发现学习

- 观察学习：用于对事例进行概念聚类，形成概念描述。
- 发现学习：用于发现规律，产生定律或规则。

7.5.2.5 观察与发现学习

1. 概念聚类

- **概念聚类：** 1980 年，米卡尔斯基（ R. S. Michalski ）。
- **概念聚类的基本思想：** 把事例按一定的方式和准则进行分组，如划分为不同的类，不同的层次等，使不同的组代表不同的概念，并且对每一个组进行特征概括，得到一个概念的语义符号描述。

7.5.2.5 观察与发现学习

1. 概念聚类

- 例如对如下事例：

喜鹊、麻雀、布谷鸟、乌鸦、鸡、鸭、鹅，…

- 分为两类：

鸟 = { 喜鹊， 麻雀， 布谷鸟， 乌鸦， … }

家禽 = { 鸡、鸭、鹅， … }

- 得知：

“鸟有羽毛、有翅膀、会飞、会叫、野生”。

“家禽有羽毛、有翅膀、会飞、会叫、家养”。

7.5.2.5 观察与发现学习

2. 发现学习

- **发现学习**：从系统的初始知识、观察事例或经验数据中归纳出规律或规则。——无教师指导的归纳学习
- **经验发现**：从经验数据中发现规律和定律。
- **知识发现**：指从已观察的事例中发现新的知识。

7.5.2.6 类比学习

■ 1. 类比推理

■ **类比推理**: 由新情况与记忆中的已知情况在某些方面相似，从而推出它们在其他相关方面也相似。

■ **源域 S** : 已经认识的域，包括过去曾经解决过且与当前问题类似的问题以及相关知识；

■ **目标域 T** : 当前尚未完全认识的域，遇到的新问题。

■ **类比推理的目的**: 从 S 中选出与当前问题最近似的问题及其求解方法来求解当前的问题，或者建立起目标域中已有命题间的联系，形成新知识。

7.5.2.6 类比学习

- 设 S_1 与 T_1 分别表示 S 与 T 中的某一情况，且 S_1 与 T_1 相似

假设 S_2 与 S_1 相关

则由类比推理可推出 T 中的 T_2 ，且 T_2 与 S_2 相似

推理过程：

1. 回忆与联想
2. 选择
3. 建立对应关系
4. 转换

7.5.2.6 类比学习

2. 属性类比学习

■ 属性类比学习：根据两个相似事物的属性实现类比学习的。

■ 1979 年，温斯顿研究开发了一个属性类比学习系统。

■ 源域和目标域都是用框架表示的，分别称为源框架和目标框架。框架的槽用于表示事物的属性。

■ 学习过程：把源框架中的某些槽值传递到目标框架的相应槽中去。

7.5.2.6 类比学习

(1) 从源框架中选择若干槽作为候选槽

- 候选槽：其槽值有可能要传递给目标框架的那些槽。
- 选择的方法：
 - (1) 选择具有极端槽值的槽作为候选槽。
 - (2) 选择已经被确认为“重要槽”的槽作为候选槽。
 - (3) 选择与源框架相似的框架中不具有的槽作为候选槽。
 - (4) 选择相似框架中不具有这种槽值的槽作为候选槽。
 - (5) 把源框架中的所有槽都作为候选槽。

7.5.2.6 类比学习

(2) 根据目标框架对候选槽进行筛选

■ 筛选规则：

- (1) 选择在目标框架中还未填值的槽。
- (2) 选择在目标框架中为典型事例的槽。
- (3) 选择与目标框架有紧密关系的槽，或者与目标框架的槽类似的槽。

7.5.2.6 类比学习

■ 3. 转换类比学习

- 在状态空间表示法的知识表示中，“状态”：描述问题在不同时刻的状况；“算符”：描述改变状态的操作。
- 当问题由初始状态变换到目标状态时，所用算符的序列就构成了问题的一个解。
- 如何使问题由初始状态变换到目标状态呢？
- “手段－目标分析”法（MEA）：纽厄尔等人在通用问题求解程序 GPS 中提出的一种问题求解模型。

7.5.2.6 类比学习

■ “手段—目标分析”法（MEA）求解问题的基本过程

:

- (1) 把问题的当前状态与目标状态进行比较，找出差异。
- (2) 根据差异找出一个可减小差异的算符。
- (3) 如果该算符可作用于当前状态，则用该算符把当前状态改变为另一个更接近于目标状态的状态；如果不能，则保留当前状态，并生成一个子问题，再对此子问题应用 MEA。
- (4) 当子问题被求解后，恢复保留的状态，继续处理原问题。

7.5.2.6 类比学习

■ 转换类比学习：由外部环境获得与类比有关的信息，学习系统找出与新问题相似的旧问题的有关知识，把这些知识进行转换使之适用于新问题，从而获得新的知识。

■ 回忆过程：找出新、旧问题间的差别，包括：

- (1) 初始状态的差别。
- (2) 目标状态的差别。
- (3) 路径约束的差别。
- (4) 求解方法可应用度的差别。

■ 转换过程：把旧问题的求解方法经适当变换使之成为求解新问题的方法。

7.5.2.7 解释学习

- **解释学习：**演绎学习方法。
- 它是通过运用相关的领域知识，对当前提供的单个问题求解实例进行分析，从而构造解释并产生相应知识的。
- **解释学习系统：**米切尔等人研制的 LEX 和 LEAP 系统、明顿（S. Minton）等人研制的 PRODIGY 系统等。

1. 解释学习的概念

- 解释学习：通过运用相关的领域知识及一个训练实例来对某一目标概念进行学习，并最终生成这个目标概念的一般性描述。
- 解释学习的一般性描述（米切尔，1986）：

给定：领域知识。

目标概念。

训练实例。

操作性准则。

找出：满足的关于的充分条件。

1. 解释学习的概念

■ 解释学习与示例学习的主要区别：

(1) 示例学习：系统要求输入一组实例。

解释学习：输入一个实例。

(2) 示例学习：归纳学习，不要求提供领域知识。

解释学习：演绎学习，要求提供完善的领域知识。

(3) 示例学习：概念的获取，即知识增加的一面。

解释学习：技能提高的一面。

2. 解释学习的学习过程

(1) 构造解释

- 构造解释的任务：证明提供给系统的训练实例为什么是满足目标概念的一个实例。
- 证明过程：通过运用领域知识进行演绎实现的，证明的结果是得到一个解释结构。

2. 解释学习的学习过程

(1) 构造解释

■ 例如，学习目标: $Safe - To - Stack(Obj_1, Obj_2)$

训练实例: $On(Obj_1, Obj_2)$

$Isa(Obj_1, book - AI)$

$Isa(Obj_2, table - book)$

$Volume(Obj_1, 1)$

$Density(Obj_2, 0.1)$

领域知识:

$\neg Fragile(y) \rightarrow Safe - To - Stack(x, y)$

$Lighter(x, y) \rightarrow Safe - To - Stack(x, y)$

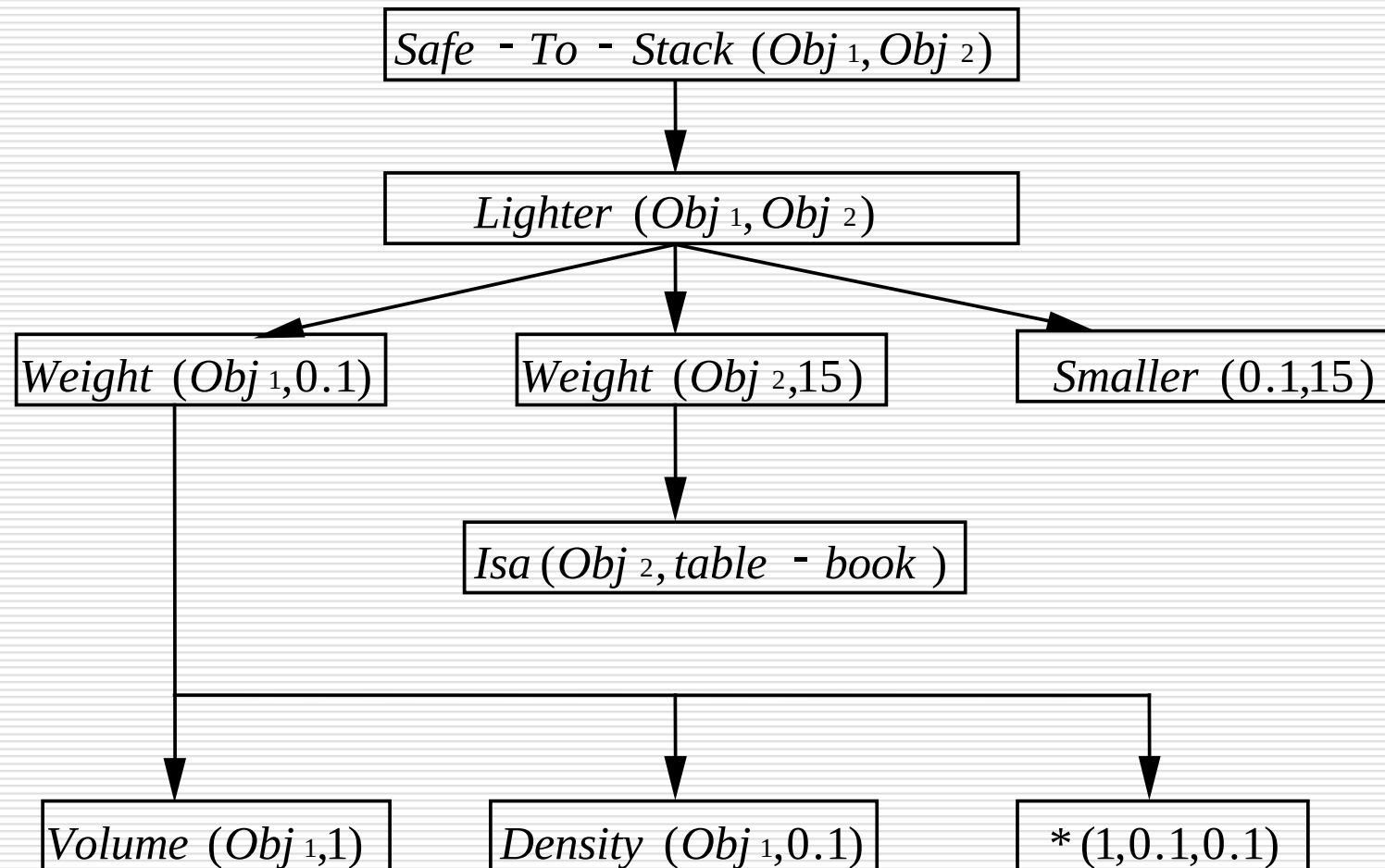
$Volume(p, v) \wedge Density(p, d) \wedge \star(v, d, w) \rightarrow Weight(p, w)$

$Isa(p, table - book) \rightarrow Weight(p, 15)$

$Weight(p_1, w_1) \wedge Weight(p_2, w_2) \wedge Smaller(w_1, w_2) \rightarrow Lighter(p_1, p_2)$

2. 解释学习的学习过程

(1) 构造解释



$\text{Safe} - \text{To} - \text{Stack} (\text{Obj}_1, \text{Obj}_2)$ 的解释结构

2. 解释学习的学习过程

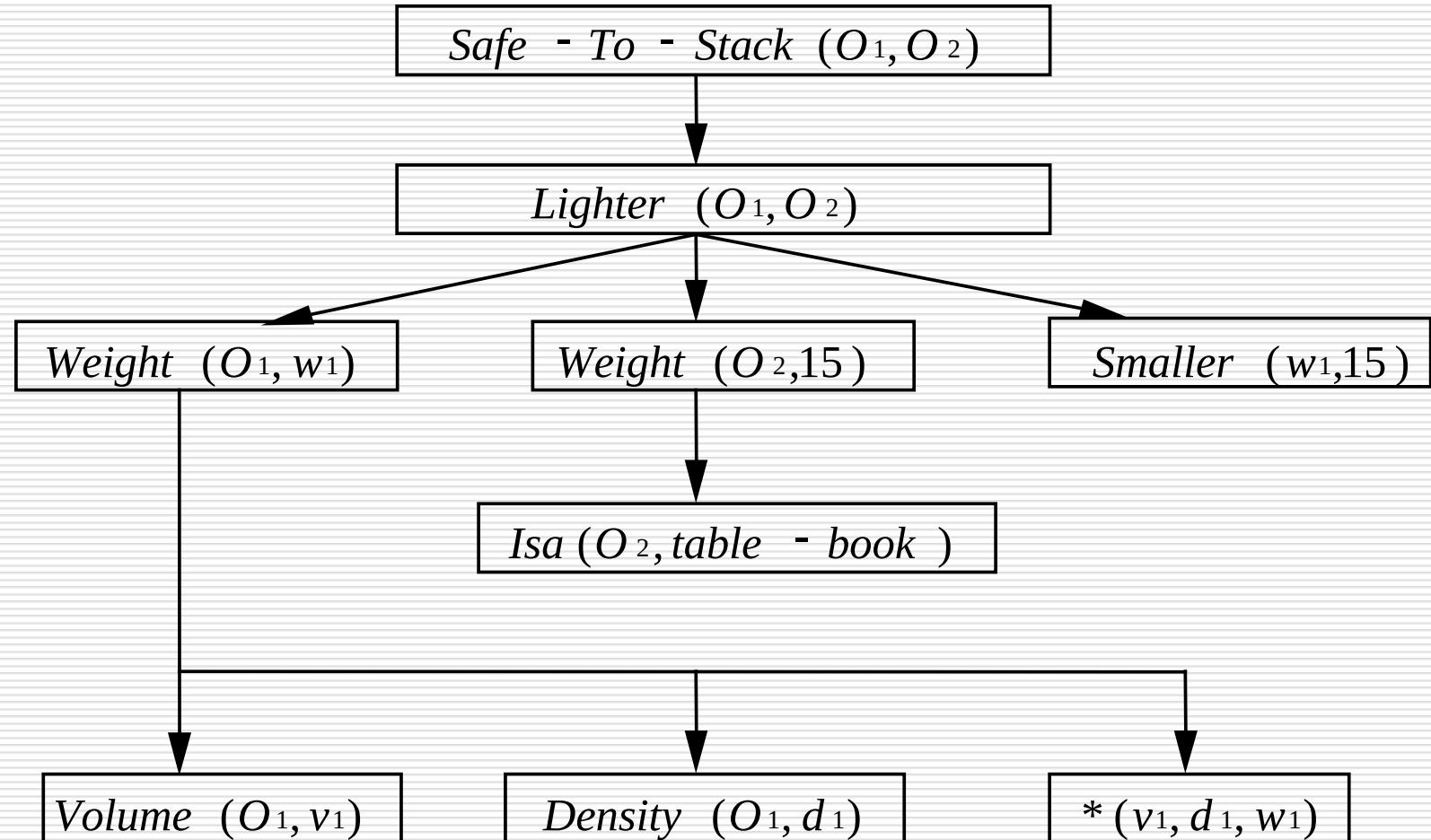
(2) 获取一般性的知

识 构造解释的任务：对上一步得到的解释结构进行一般化处理，从而得到关于目标概念的一般性知识。

处理的方法：把常量变换为变量，并把某些不重要的信息去掉，只保留那些对以后求解问题所必须的关键性信息。

2. 解释学习的学习过程

(2) 获取一般性的知识



Safe -To -Stack (O₁ , O₂) 一般化解释结构

3. 领域知识的完善性

■ 两种极端情况：

(1) 构造不出解释

- 原因：系统中缺少某些相关的领域知识，或者领域知识中包含了矛盾等错误。

(2) 构造出了多种解释

- 原因：领域知识不健全，已有的知识不足以把不同的解释区分开来。

第7章 机器学习

- 7.5.1 机器学习的基本概念
- 7.5.2 符号学习
- 7.5.3 深度学习

7.5.3 深度学习

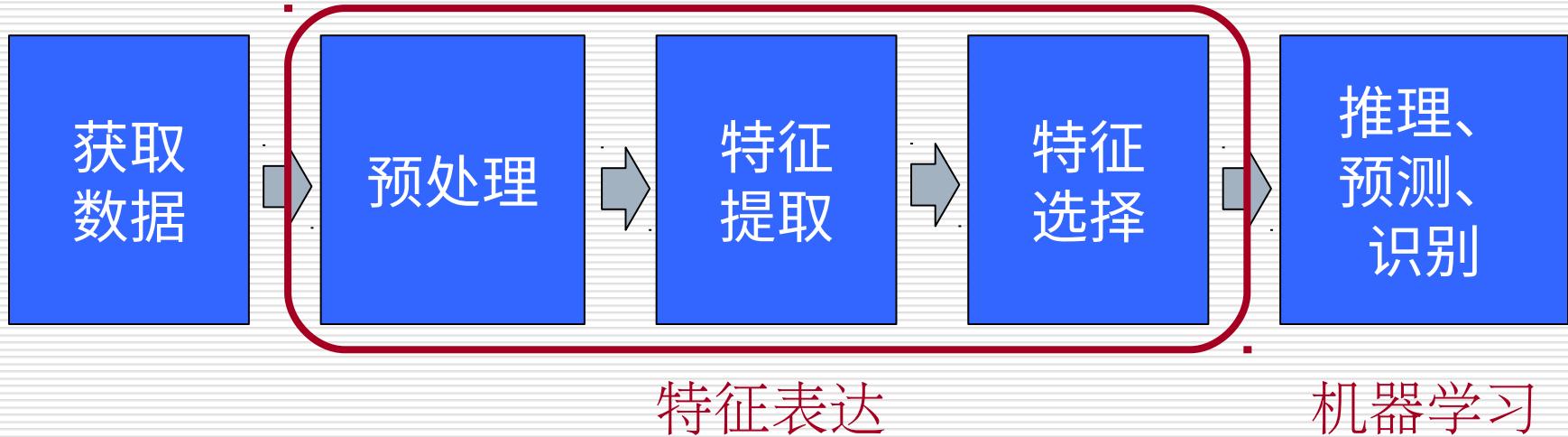
- 7.5.3.1 深度学习的提出
- 7.5.3.2 人脑视觉机理
- 7.5.3.3 特征
- 7.5.3.4 深度学习的基本思想
- 7.5.3.5 深度学习的训练过程
- 7.5.3.6 自动编码器
- 7.5.3.7 自动编码器的变体
- 7.5.3.8 受限玻尔兹曼机

7.5.3.1 深度学习的提出

- 2012年6月，《纽约时报》报道：**Google Brain**项目由斯坦福大学**Andrew Ng**（吴恩达）和**Jeff Dean**用**16000个CPU Core**的并行计算平台训练深度神经网络（**DNN**）。
- 2012年11月，微软在中国天津的一次活动上公开演示了一个全自动的同声传译系统，讲演者用英文演讲，后台的计算机自动完成语音识别、英中机器翻译和中文语音合成。
- 这些应用都基于**2006**年以来迅速发展的深度学习。

7.5.3.1 深度学习的提出

■ 模式识别：



- 良好的特征表达，是识别成功的关键；
- 识别系统的计算和测试工作耗时主要集中在特征表达部分；
- 传统的方法是手工设计良好的特征提取器，这需要大量的工程技术和专业领域知识。

能否自动地学习特征呢？



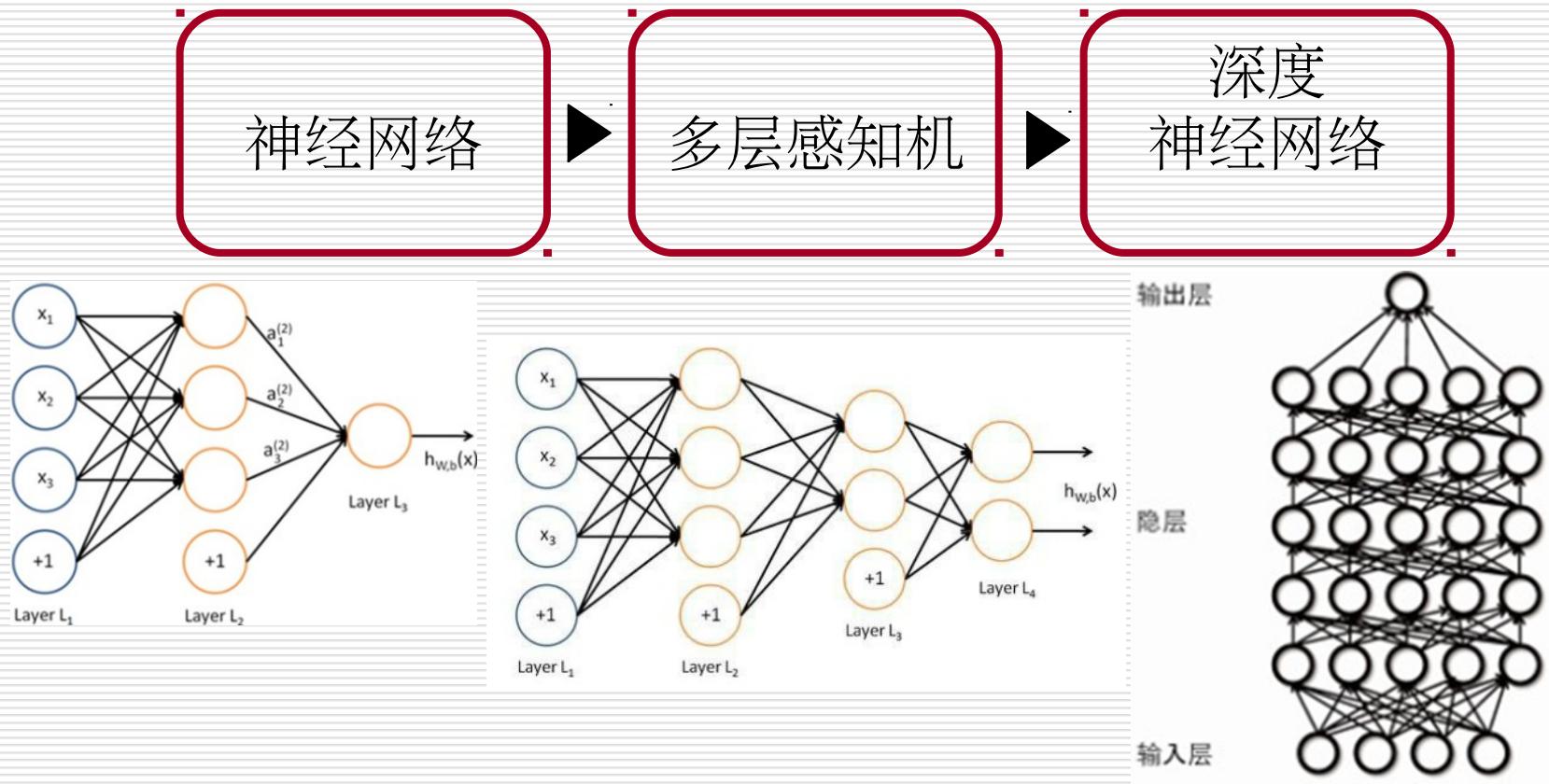
深度学习！

7.5.3.1 深度学习的提出

- 深度学习 (**Deep Learning**, **DL**)：是一种特征学习方法，把原始数据通过一些简单的但非线性的模型转变成为更高层次的、更加抽象的表达。通过足够多的转换组合，非常复杂的函数也可以被学习。
- 深度学习的核心是，使用一种通用的学习过程从数据中学习各层次的特征，而不是手工设计特征提取器。

7.5.3.1 深度学习的提出

■ 深度学习的概念源于人工神经网络的研究。含多隐层的多层感知器就是一种深度学习结构。

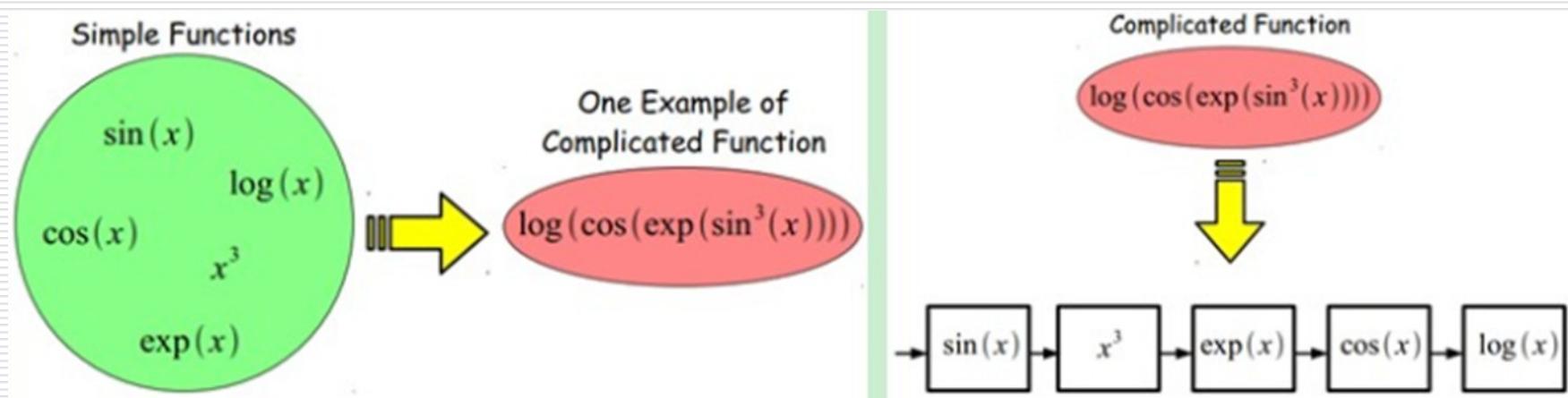


7.5.3.1 深度学习的提出

深度学习

结构：简单模块的多层栈，以及许多计算非线性输入输出的映射。栈中的每个模块将其输入进行转换，以增加表达的可选择性与不变性，大部分模块的目标是特征学习。

优点：可通过学习一种深层非线性神经网络结构，实现复杂函数逼近，表征输入数据分布式表示。



7.5.3.1 深度学习的提出

- 神经网络学习的历史
- 从机器学习的模型结构层次来分，机器学习经历了两次研究浪潮：
 - 第一次浪潮——浅层学习（**Shallow Learning**）
 - 第二次浪潮——深度学习

7.5.3.1 深度学习的提出

■ 浅层学习

- ✓ 人工神经网络（**BP** 算法）
 - 虽被称作多层感知机，但实际是一种只含有一层隐层节点的浅层模型。
- ✓ **SVM**（**Support Vector Machines**）、**Boosting**、最大熵方法（如 **Logistic Regression**）等。
 - 带有一层隐层节点（如 **SVM**、**Boosting**），或没有隐层节点（如 **Logistic Regression**）的浅层模型。

局限性：有限样本和有限计算单元情况下对复杂函数的表示能力有限，针对复杂分类问题其泛化能力受到一定制约。

7.5.3.1 深度学习的提出

■ 深度学习

- ✓ 2006 年，加拿大多伦多大学教授 **Geoffrey Hinton** 和他的学生在 **Science** 上发表的文章 [1] 掀起了深度学习的浪潮。
- ✓ 这篇文章提出了深度学习的两个重要观点：
 - ① 多隐层的人工神经网络具有优异的特征学习能力，学习得到的特征对数据有更本质的刻画，从而有利于可视化或分类；
 - ② 深度神经网络在训练上的难度，可以通过“逐层初始化”（**layer-wise pre-training**）来有效克服。逐层初始化可以通过无监督学习来实现。

[1] G.E. Hinton, et al. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504-507.

7.5.3.1 深度学习的提出

■ 深度学习 V.S. 浅层学习

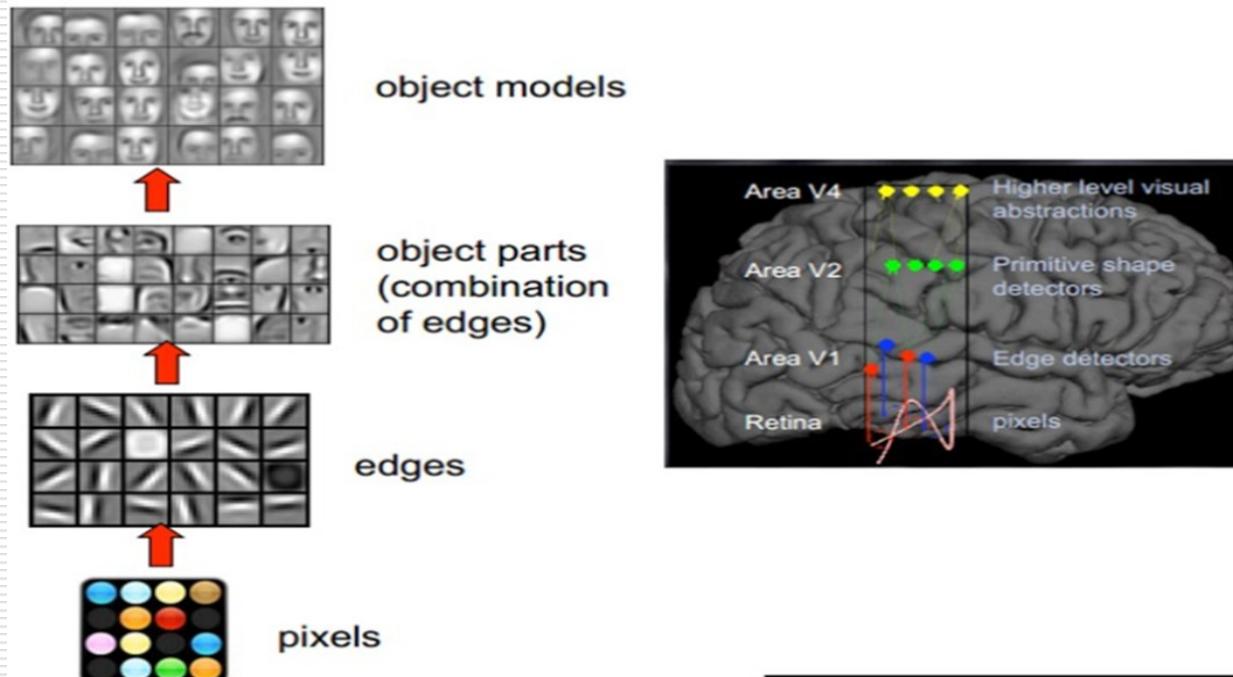
- ✓ **实质**: 通过构建多隐层的模型和海量训练数据（可为无标签数据）来学习更有用的特征，从而提升分类或预测的准确性。“深度模型”是手段，“特征学习”是目的。
- ✓ **与浅层学习的区别**:
 - n 强调了模型结构的深度: 通常有 **5 层、6 层**，甚至 **10** 多层的隐层节点；
 - n 突出了特征学习的重要性: 通过逐层特征变换，将样本在原空间的特征表示变换到一个新特征空间，从而使分类或预测更加容易。与人工规则构造特征的方法相比，利用大数据来学习特征，更能够刻画数据的丰富内在信息。

7.5.3.1 深度学习的提出

■ 3类深度神经网络:

- ✓ 前馈深度网络: 多层感知机、卷积神经网络等。
- ✓ 反馈深度网络: 反卷积神经网络、层次稀疏编码网络等。
- ✓ 双向深度网络: 深度玻尔兹曼机、深度信念网络、栈式自编码器等。

7.5.3.2 人脑视觉机理



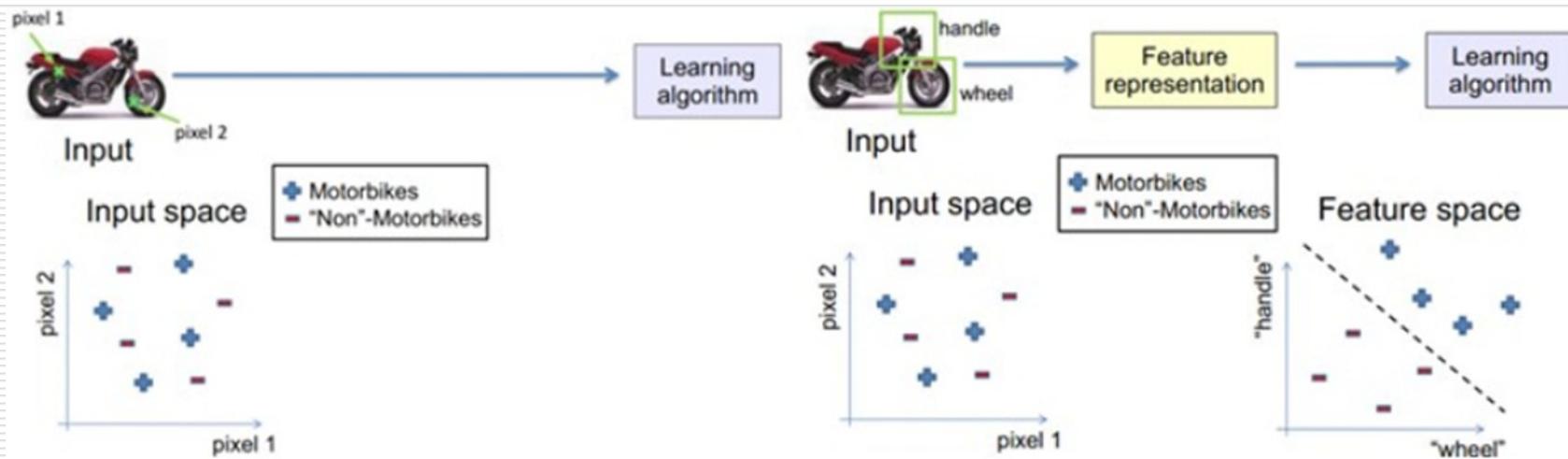
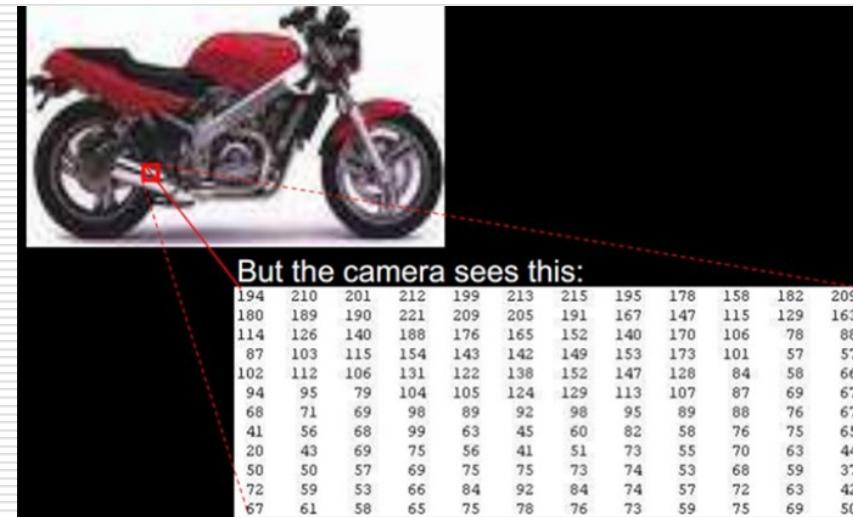
■ 诺贝尔医学奖获得者美国神经生物学家 **David Hubel** 和 **Torsten Wiesel** 发现：人的视觉系统的信息处理是分级的。

■ 高层的特征是低层特征的组合，从低层到高层的特征表示越来越抽象，越来越能表现语义或者意图

■ 抽象层面越高，存在的可能猜测就越少，就越利于分类

7.5.3.3 特征

- 如果数据被很好地表达成了特征，通常线性模型就能达到满意的精度。
- 什么样的特征表示粒度，才能更有利于分类呢？——具有结构性（或者语义）的高层特征。



7.5.3.3 特征

■ 浅层特征

稀疏编码（Sparse Coding）

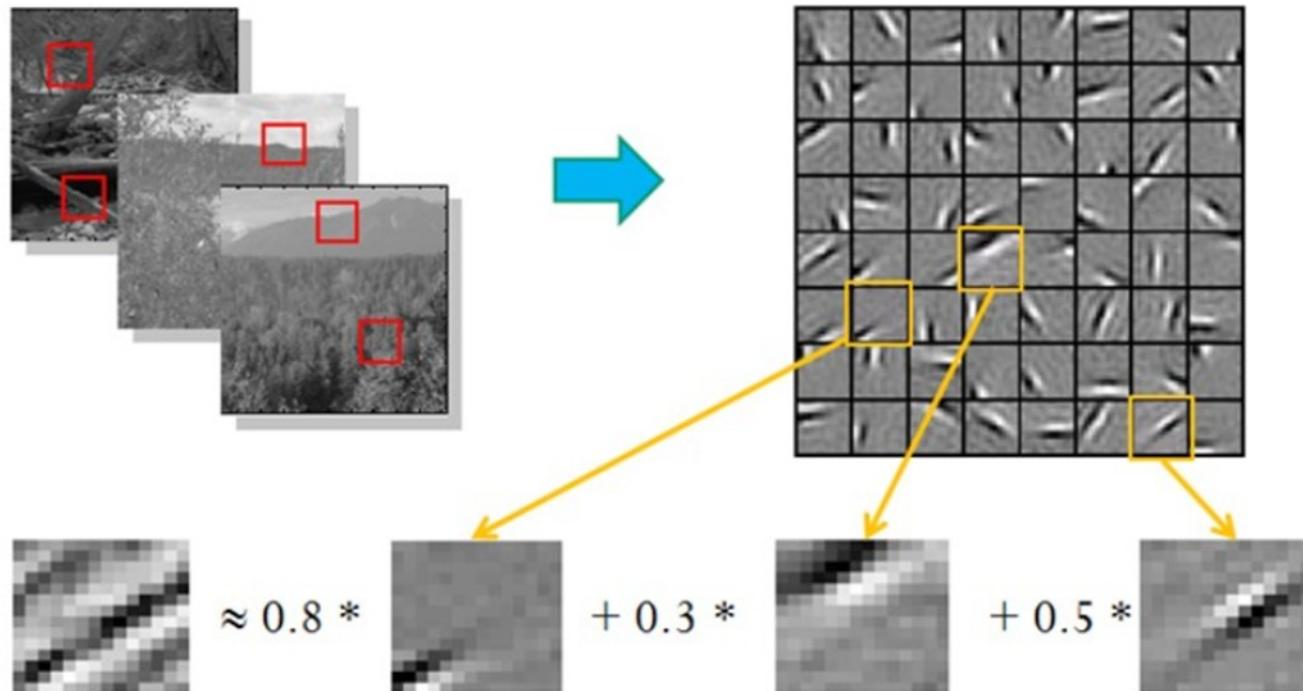
问题描述： $\text{Sum}_k (a[k] * S[k]) \rightarrow T$

迭代：

- 1) 选择一组 $S[k]$ ，然后调整权重系数 $a[k]$ ，使得 $\text{Sum}_k (a[k] * S[k])$ 最接近目标 T 。
- 2) 固定住 $a[k]$ ，在 n 个碎片中，选择其它更合适的碎片 $S'[k]$ ，替代原先的 $S[k]$ ，使得 $\text{Sum}_k (a[k] * S'[k])$ 最接近 T 。

经过几次迭代后，最佳的 $S[k]$ 组合被遴选出来了，其中被选中的 $S[k]$ 基本上都是照片上不同物体的边缘线，这些线段形状相似，区别在于方向。

7.5.3.3 特征

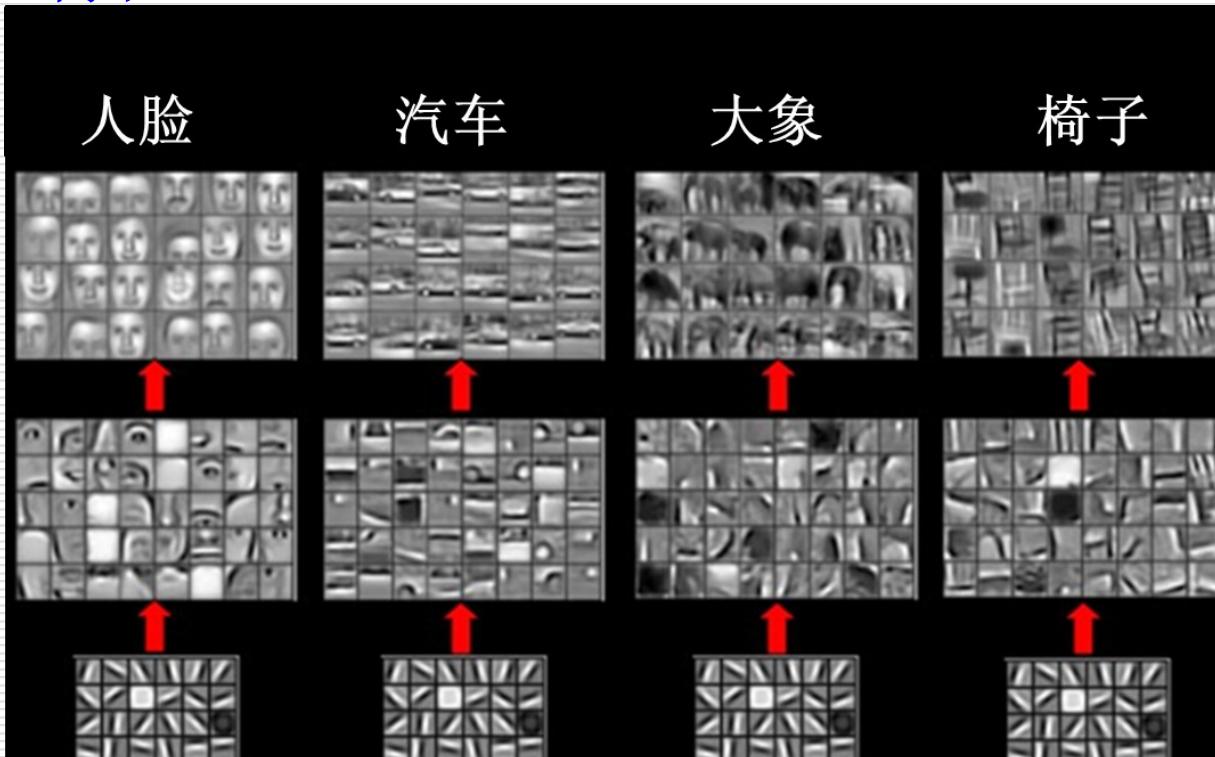


$[a_1, \dots, a_{64}] = [0, 0, \dots, 0, \mathbf{0.8}, 0, \dots, 0, \mathbf{0.3}, 0, \dots, 0, \mathbf{0.5}, 0]$
(feature representation)

- 高层特征（或图像），往往是由一些浅层特征（基本结构）组成的。

7.5.3.3 特征

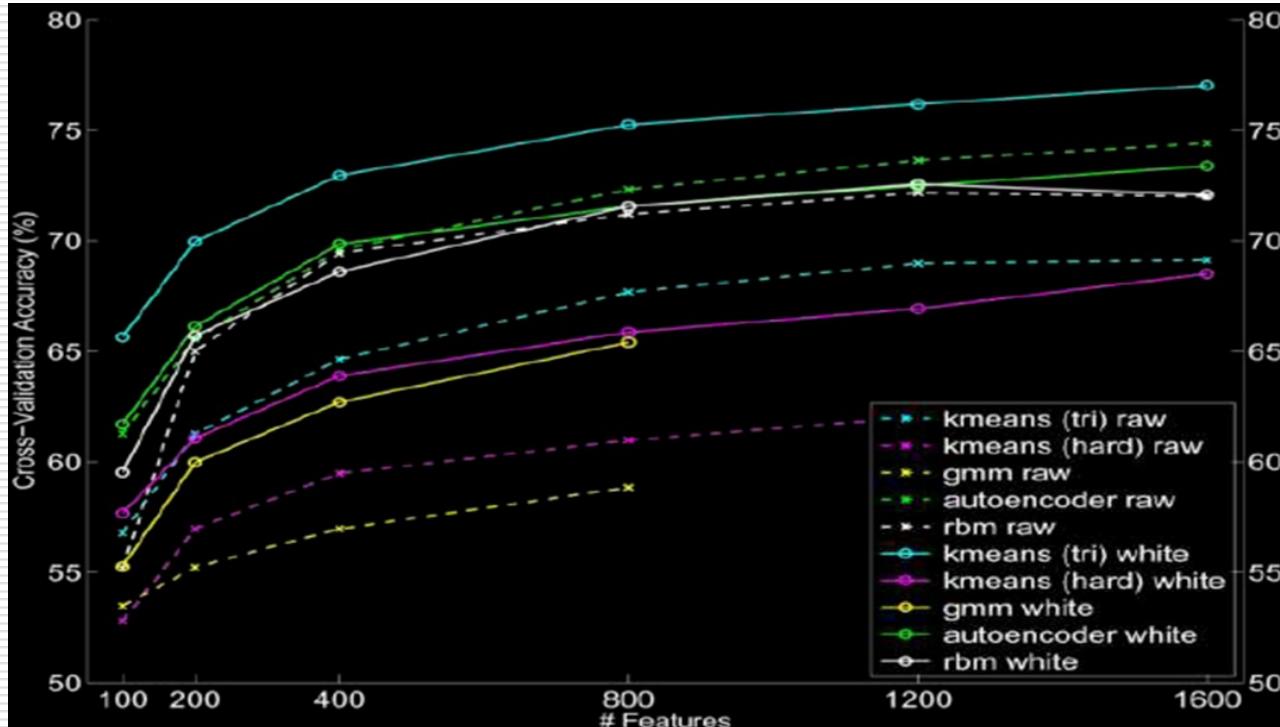
■ 结构性特征



- 高层的特征是低层特征的组合，从低层到高层的特征表示越来越抽象，越来越能表现语义或者意图。
- 抽象层面越高，存在的可能猜测就越少，就越利于分类

7.5.3.3 特征

■ 需要多少特征?



- ✓ 特征越多，给出信息就越多，识别准确性会得到提升；
- ✓ 但特征多，计算复杂度增加，探索的空间大，可以用来训练的数据在每个特征上就会稀疏。

7.5.3.4 深度学习的基本思想

- 假设系统 **S** 有 **n** 层 (**S₁, ..., S_n**) , 它的输入是 **I** , 输出是 **O** , 表示为: **I => S₁ => S₂ => ... => S_n => O** 。如果调整系统中参数, 使得它的输出 **O** 等于输入 **I** , 那么就可以自动地获得输入 **I** 的一系列层次特征, 即 **S₁ , ... , S_n** 。
- 通过这种方式, 就可以实现对输入信息进行分级表达了。
- 输出严格地等于输入的要求太严格, 可以要求输入与输出的差别尽可能地小。

7.5.3.5 深度学习的训练过程

■ 深度学习不采用 **BP** 算法的原因：

- ✓ **BP** 算法随机设定初始值，当初始值是远离最优区域时易收敛至局部最小；
- ✓ 对于一个深度网络（7层以上），误差校正信号传播到最前面的层已经变得太小，出现所谓的梯度扩散（**gradient diffusion**）；
- ✓ **BP** 算法需要有标签数据来训练，但大部分数据是无标签的；

7.5.3.5 深度学习的训练过程

■ 逐层初始化（**layer-wise pre-training**）

- ✓ 自下而上的无监督学习——特征学习
- ✓ 自上而下的监督学习——调优

■ 回顾一下监督学习和无监督学习：



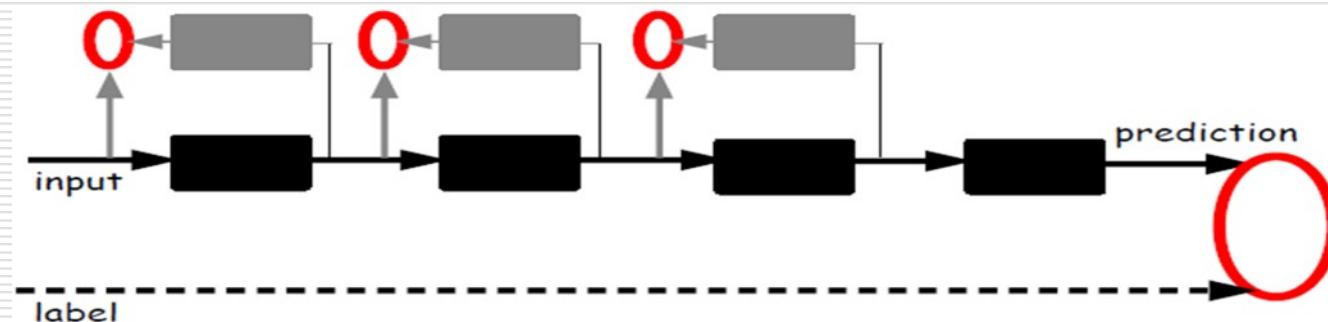
7.5.3.5 深度学习的训练过程

■ 自下而上的无监督学习：

- ✓ 采用无标签数据分层训练各层参数，这一步可以看作是特征学习的过程（与神经网络的训练方法最大的区别）
- ✓ 从底层开始，自下而上地训练，分别得到各层参数。

■ 自上而下的监督学习：

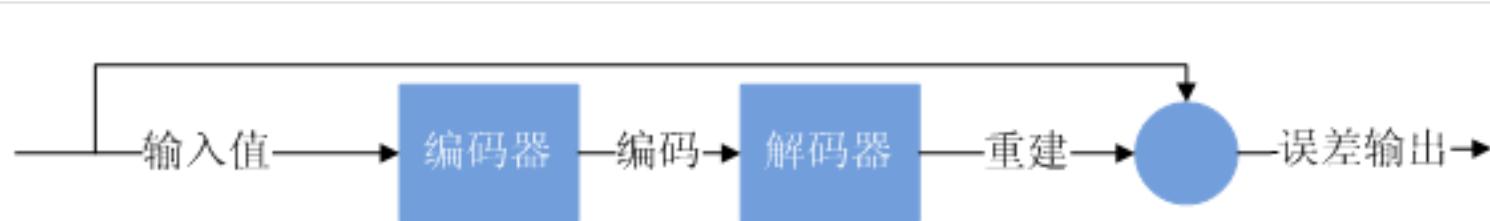
- ✓ 基于第一步得到的各层参数进一步调整整个多层模型的参数；
- ✓ 通过有标签数据来训练，误差信号自上而下传输，对网络进行微调。



7.5.3.5 深度学习的训练过程

■ 无监督学习的实现过程：

- ✓ 将样本输入数据输入到一个编码器就会得到一个编码，这个编码也就是样本输入的一种表示。为了验证这个编码是样本输入，将这个编码输入到一个解码器。如果解码器输出的信息和输入样本信息相似（理想情况下相同），说明这个编码是合适的。所以，可以通过调整编码器和解码器的参数，使得重构误差最小，就得到了样本输入的编码表示了。
- ✓ 因为是无标签数据，所以误差的来源就是通过重构信息与原输入相比较得到。



7.5.3.5 深度学习的训练过程

- 深度学习的第一步实质上是一个网络**参数初始化**的过程。
- 传统神经网络的随机初始化，使得当初始值远离最优解时易收敛到局部最优。而深度学习模型是通过无监督学习得到初始值的，因而这个**初始值比随机初始值更接近全局最优**，从而能够取得更好的效果。
 -

深度学习的局限性

What's wrong with Deep Learning ? —— Yann Lecun

■ 缺乏理论支持

对于深度学习架构，存在一系列的疑问：卷积神经网络为什么是一个好的架构（事实上其存在梯度扩散等缺点），深度学习的结构需要多少隐层，在一个大的卷积网络中到底需要多少有效的参数（很多权重相互之间似乎都存在冗余），等等。

■ 缺乏推理能力

尽管深度学习和简单推理已经应用于语音和手写字识别很长一段时间了，我们仍需要在大的向量上使用新的范式来代替基于规则的字符表达式操作。

深度学习的局限性

What's wrong with Deep Learning ? —— Yann Lecun

■ 缺乏短时记忆能力

例如在自然语言理解的许多任务（例如问答系统）中需要一种方法临时存储分隔的片段，正确解释视频中的事件并能回答有关问题需要记住的视频中发生的事件的抽象表示。包括深度学习系统，都不能很好地存储多个时间序列上的记忆。

■ 缺乏无监督学习的能力

虽然无监督学习可以帮助特定的深度网络进行“预训练”，但最终绝大部分能够应用于实践的深度学习方法都是使用有监督学习。还没有找到很合适的非监督学习算法，非监督学习在未来存在巨大的研究空间。尤其是对视频的理解。

深度学习在应用上的困难

- 深度神经网络模型复杂，训练数据多，计算量大。语音识别目前通常使用样本量达数十亿，以 CPU 单机需要数年才能完成一次训练，用流行的 GPU 卡也需要数周才能完成一次训练。
- 如何支持大模型。**ImageNet 2012 竞赛冠军的网络占用 3.99 GB 的显存**，已接近主流 GPU 的显存容量，试图增大模型则会超过 GPU 显存范围。因此，如何支持更大模型是一个大的挑战。
- 深度神经网络训练收敛难，需要反复多次实验。深度神经网络是非线性模型，其代价函数是非凸函数，容易收敛到局部最优解。深度神经网络的数学基础研究不足。虽可以通过 RBMs 等生成式建模方法初始化网络模型，但仍然不是彻底的解决方案。

总之，深度学习是一个效果很好但门槛极高的方向，如何落地产生实际应用效果成为关注的焦点。

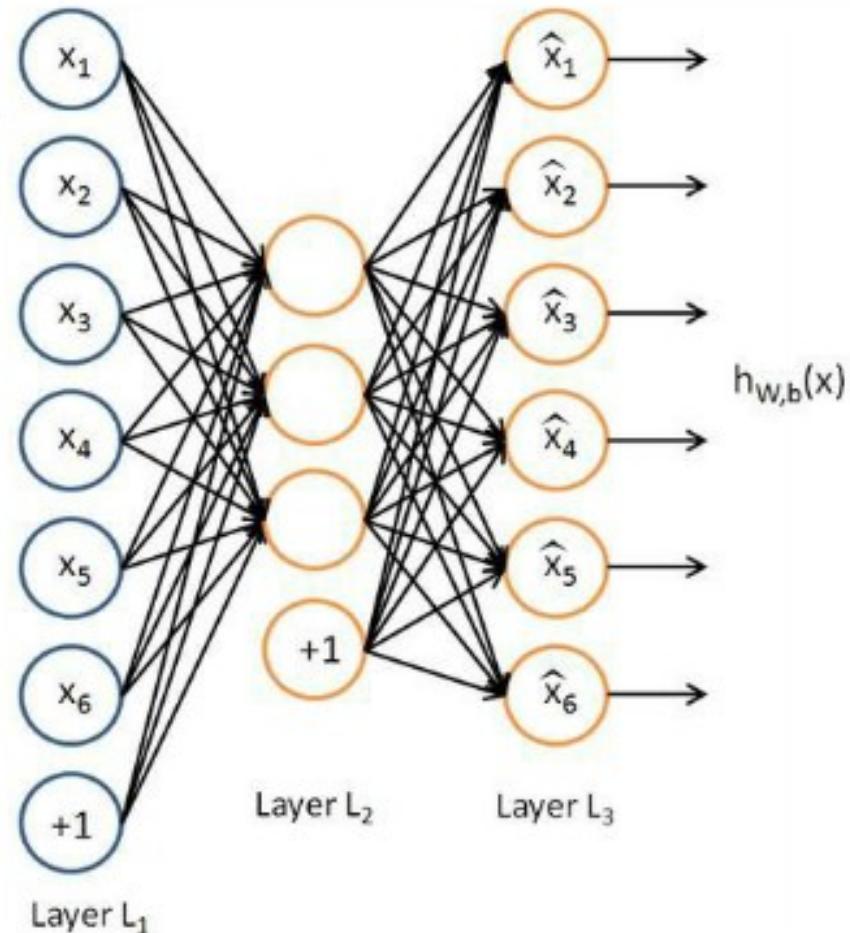
7.5.3.6 自动编码器

■ 自动编码器（Auto Encoder）是一种尽可能复现输入信号的神经网络。

■ 学习函数： $h_{w,b}(x) \approx x$

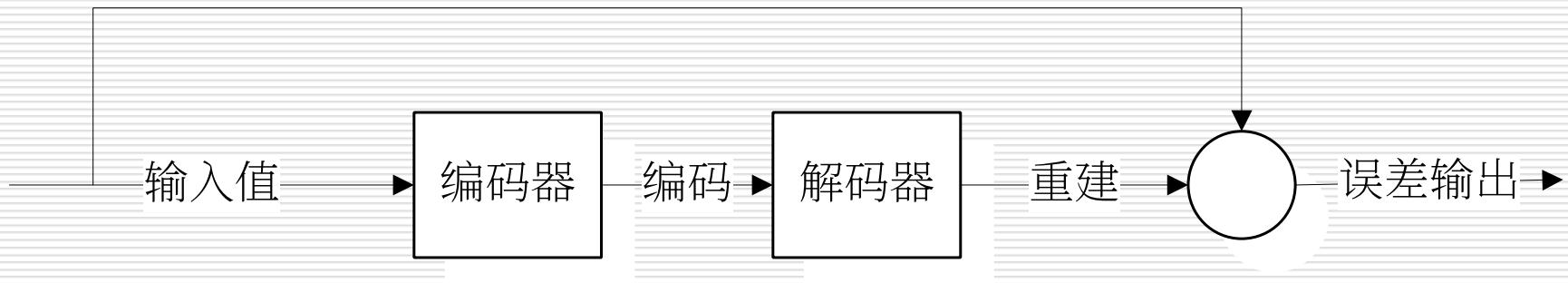
■ 自动编码器能够找到可以代表原信息的主要成分。

■ 自动编码器是一种非监督学习方法。

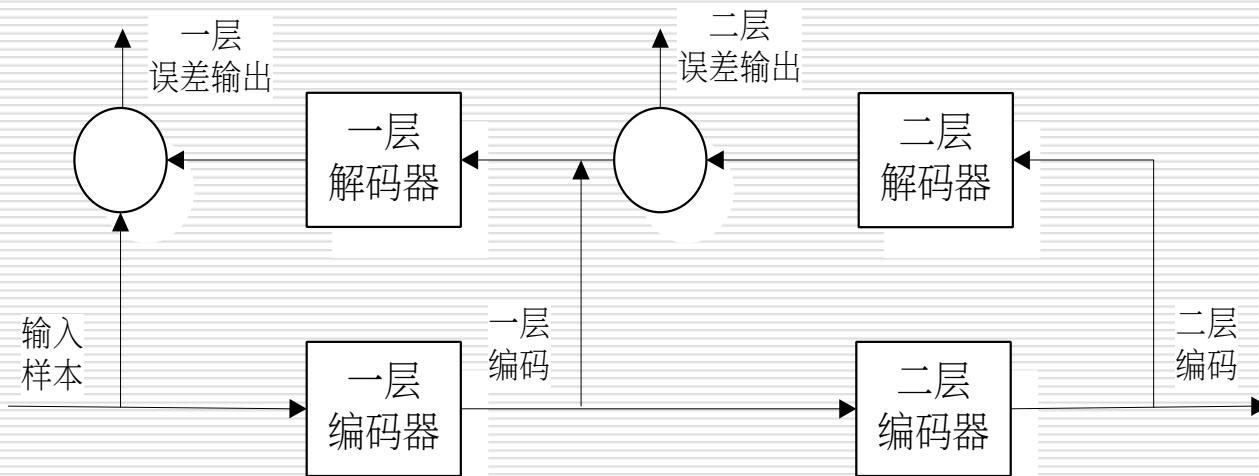


7.5.3.6 自动编码器

■ (1) 给定无标签数据，用非监督学习来学习特征；

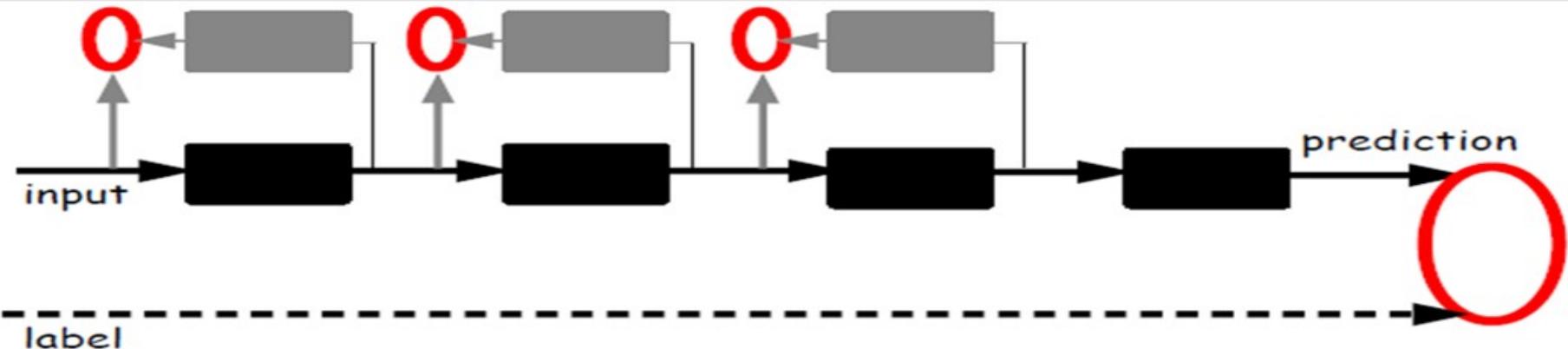


■ (2) 通过编码器产生特征，然后逐层训练下一层；



7.5.3.6 自动编码器

(3) 有监督微调



经过（1） - （3），自动编码器能够获得代表输入的特征，这个特征可以最大程度上代表原输入信号。为了实现分类，需要在最后一个编码器的后面添加一个分类器（例如 **BP** 神经网络、**SVM** 等），然后通过监督训练方法（梯度下降法）去训练这个分类器。

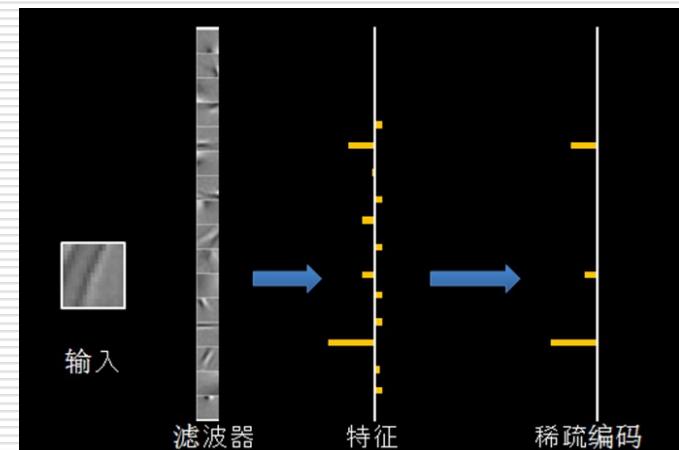
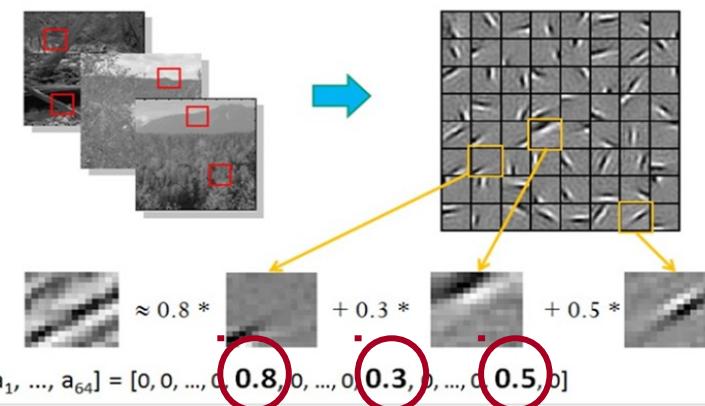
7.5.3.7 自动编码器的变体

1. 稀疏自动编码器

- ✓ 稀疏自动编码器（**Sparse Auto Encoder**）是在自动编码器的基础上加上稀疏性约束，即约束每一层中的大部分节点都要为 **0**，只有少数不为 **0**。



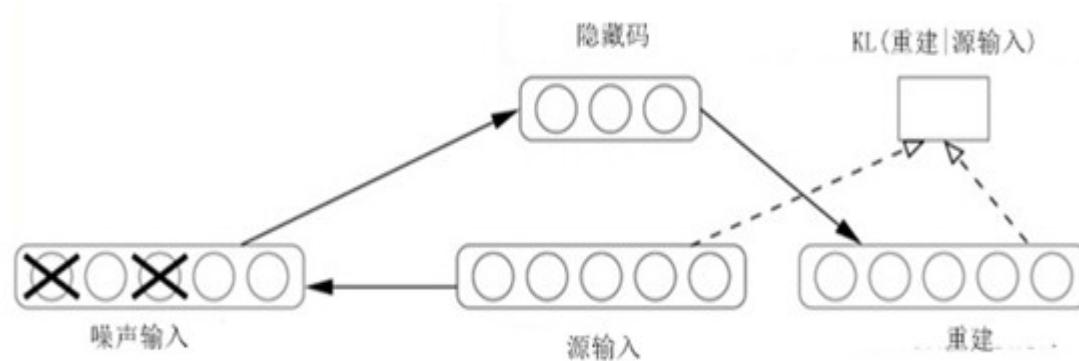
- ✓ 限制每一次得到的特征表达尽量稀疏，这样就可以简洁的表示原信息的主要成分。



7.5.3.7 自动编码器的变体

2. 降噪自动编码器

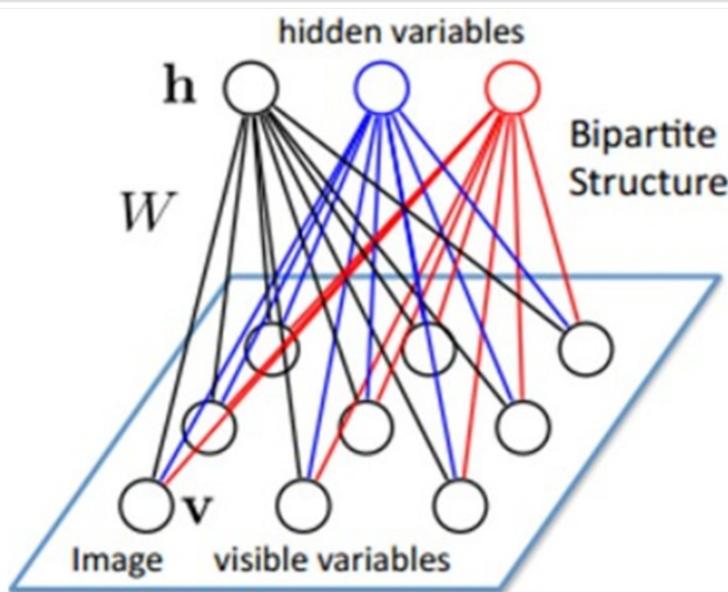
- ✓ 降噪自动编码器（**Denoising AutoEncoders, DAs**）是在自动编码器的基础上，**在训练数据中加入噪声**，让自动编码器学习去除这种噪声而获得实际输入。



- ✓ 这就迫使编码器去学习输入信号的更加鲁棒的表达，这也是它的泛化能力比一般编码器强的原因。

7.5.3.8 受限玻尔兹曼机

■ 定义：假设有一个二分图，同层节点之间没有连接，一层是可视层，即输入数据层 (**v**)，一层是隐藏层 (**h**)，如果假设所有的节点都是随机二值（**0, 1**）变量节点，同时假设全概率分布 $p(v, h)$ 满足 **Boltzmann** 分布，我们称这个模型是受限玻尔兹曼机 (**Restricted Boltzmann Machine, RBM**)。



7.5.3.8 受限玻尔兹曼机

1. 能量模型和概率分布

- ✓ 能量函数是描述整个系统状态的一种测度。系统越有序或者概率分布越集中，系统的能量越小。能量函数的最小值，对应于系统的最稳定状态。
- ✓ RBM 的**能量函数**（联合组态能量）：

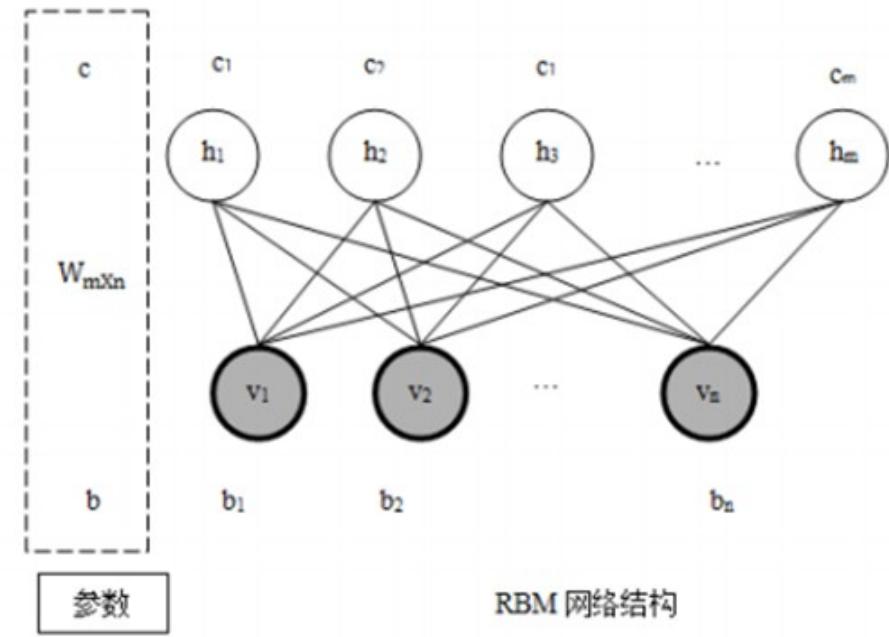
$$E(v, h) = - \sum_{i=1}^n \sum_{j=1}^m w_{ij} h_i v_j - \sum_{j=1}^m b_j v_j - \sum_{i=1}^n c_i h_i$$

- ✓ $\Theta=\{ W, b, c \}$ 是模型参数

7.5.3.8 受限玻尔兹曼机

2.RBM 的网络结构:

- ✓ v_i : 可视层中第 i 个神经元的状态
- ✓ h_j : 隐藏层中第 j 个神经元的状态
- ✓ b_i : 可视层中第 i 个神经元的偏置
- ✓ c_j : 隐藏层中第 j 个神经元的偏置
- ✓ W : 可视层与隐藏层之间的权重矩阵。其中, w_{ij} 表示可视层第 i 个神经元和隐藏层第 j 个神经元之间的连接权重。
- ✓ $\Theta = \{ W, b, c \}$: 表示 RBM 中的参数。



7.5.3.8 受限玻尔兹曼机

从能量模型到概率的转换

- ✓ 根据 Boltzmann 分布和 RBM 的能量函数，定义可视节点 v 和隐藏节点 h 的联合概率：

$$p(v, h) = \frac{e^{-E(v, h)}}{\sum_{v, h} e^{-E(v, h)}}$$

- ✓ 根据联合概率，可以得到边缘概率和条件概率：

$$p(v) = \frac{\sum_h e^{-E(v, h)}}{\sum_{v, h} e^{-E(v, h)}}, p(h) = \frac{\sum_v e^{-E(v, h)}}{\sum_{v, h} e^{-E(v, h)}}$$

$$p(v | h) = \frac{e^{-E(v, h)}}{\sum_v e^{-E(v, h)}}, p(h | v) = \frac{e^{-E(v, h)}}{\sum_h e^{-E(v, h)}}$$

7.5.3.8 受限玻尔兹曼机

从能量模型到概率的转换

- ✓ 当给定可视层上所有神经元状态时，隐藏层上某个神经元被激活（即取值为 1）的概率：

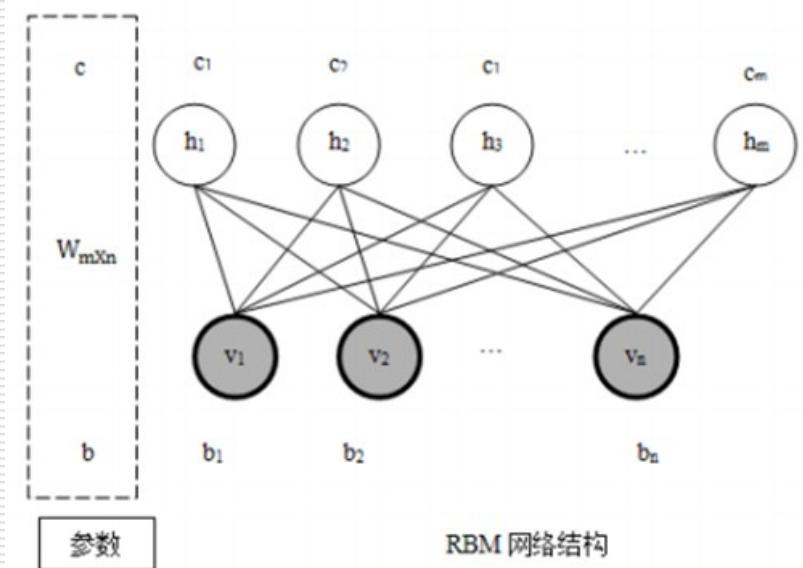
$$P(h_k = 1 \mid \mathbf{v}) = \text{sigmoid}(b_k + \sum_{i=1}^{n_v} w_{k,i} v_i),$$

- ✓ 当给定隐藏层上所有神经元状态时，可视层上某个神经元被激活（即取值为 1）的概率：

$$P(v_k = 1 \mid \mathbf{h}) = \text{sigmoid}(a_k + \sum_{j=1}^{n_h} w_{j,k} h_j).$$

7.5.3.8 受限玻尔兹曼机

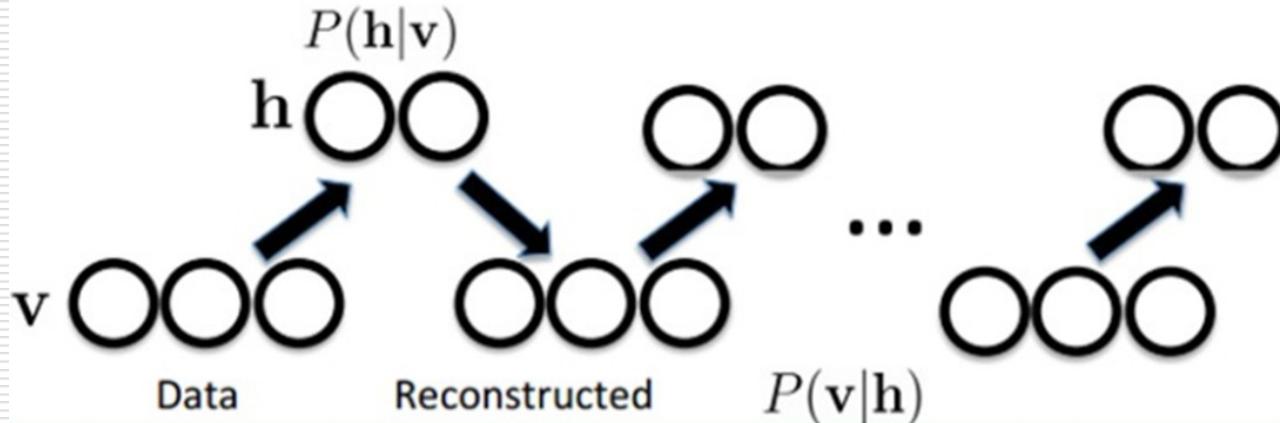
3. RBM 网络的特征学习过程



- ✓ RBM 网络的参数：可视层与隐藏层之间的权重矩阵 W ，可视节点的偏置 $b=(b_1, b_2 \cdots b_n)$ ，隐藏节点的偏置 $c = (c_1, c_2 \cdots c_m)$ 。
- ✓ 这几个参数决定了 RBM 网络将一个 n 维的样本编码成一个什么样的 m 维的样本。

7.5.3.8 受限玻尔兹曼机

3. RBM 网络的特征学习过程



当输入 v 的时候，通过 $p(h|v)$ 可以得到隐藏层 h ，而得到隐藏层 h 之后，通过 $p(v|h)$ 又能得到可视层 v_1 ，通过调整参数，如果能使得从隐藏层 h 得到的可视层 v_1 与原来的可视层 v 一样，那么得到的隐藏层就是可视层另外一种表达，从而隐藏层可以作为可视层的特征表示。

7.5.3.8 受限玻尔兹曼机

3. RBM 网络的特征学习过程:

- ✓ 一个训练样本 X 取值为 $x = (x_1, x_2 \dots x_n)$ ，根据 RBM 网络，可以得到这个样本的 m 维的编码后的样本 $y = (y_1, y_2 \dots y_m)$ ，这 m 维的编码也可以认为是抽取了 m 个特征的样本。
- ✓ 样本 y 的值（0 或 1）是按照下面的规则生成的：
(1) 根据 x 的值计算概率

$$p(h_j = 1 | v) = \sigma(\sum_{i=1}^n w_{ij} \times v_i + c_j)$$

(2) 产生一个 0 到 1 的随机数，如果它小于 $p(h_j = 1 | v)$ 的取值就是 1，否则就是 0。

- ✓ 解码过程，跟上面同理。

7.5.3.8 受限玻尔兹曼机

■ 4. RBM 的作用

- ✓ 对数据进行编码，然后由监督学习方法进行分类或回归。
- ✓ 得到权重矩阵和偏移量，供 BP 神经网络初始化训练。
- ✓ 作为一个生成模型使用：估计联合概率

$$P(v | h)$$

- ✓ 作为一个判别模型使用：计算条件概率

$$P(h | v)$$

7.5.3.8 受限玻尔兹曼机

- 5. 基于 **RBM** 的深度网络结构
- 如果把 **RBM** 的隐层数增加，可以得到深度波尔兹曼机（Deep Boltzmann Machine, DBM）。
- 如果在靠近可视层的部分使用贝叶斯信念网络即有向图模型，而在远离可视层的部分使用 **RBM**, 可以得到深信度网络（Deep Belief Nets）。

深度学习的应用

- 深度学习正在取得重大进展，解决了人工智能界的最大努力很多年仍没有进展的问题。
- 深度学习能发现高维数据中的复杂结构，因此它能够被应用于科学、商业和政府等领域，特别是自然语言理解。
- 深度学习算法需要采集到充分大的高维数据样本，才能缓解复杂模型的过度学习。
- 深度学习的可解释性不强，即很难对深度学习算法在具体问题上给出具体解释。

深度学习的应用

- 语音识别
- 图像识别和理解
- 自然语言处理（**NLP**）
- 视频内容分析
- 搜索广告 **CTR** 预估

搜索广告是搜索引擎的主要变现方式，而按点击付费（Cost Per Click，CPC）又是其中被最广泛应用的计费模式。在 CPC 模式下，预估的 CTR（pCTR）越准确，点击率就会越高，收益就越大。传统上，Google、百度等搜索引擎公司以 Logistic Regression（LR）作为预估模型。

2012 年以来，百度意识到模型结构对广告 CTR 预估的重要性：百度尝试将 DNN 作用于搜索广告，这套深度学习系统已于 2013 年 5 月开始服务于百度搜索广告系统，每天为数亿网民使用。

Reading List

Books

- Deep Learning, Yoshua Bengio, Ian Goodfellow, Aaron Courville, MIT Press, In preparation.

Review Papers

- Representation Learning: A Review and New Perspectives, Yoshua Bengio, Aaron Courville, Pascal Vincent, Arxiv, 2012.
- The monograph or review paper Learning Deep Architectures for AI (Foundations & Trends in Machine Learning, 2009).
- Deep Machine Learning – A New Frontier in Artificial Intelligence Research – a survey paper by Itamar Arel, Derek C. Rose, and Thomas P. Karnowski.
- Graves, A. (2012). Supervised sequence labelling with recurrent neural networks(Vol. 385). Springer.



THE END