CHW:

一、概论

1. 人工智能是由计算机科学、控制论、信息论、神经生理学、心理学、语言学等构成。

2. 智能科学研究**智能的基本理论**和**实现技术**，是由**脑科学、认知科学、人工智能**等学科构成的交叉学科。

3. 认知(cognition)是和**情感、动机、意志**等相对的理智或认识过程。**认知科学**是研究人类感知和思维信息处理过程的科学，包括从感觉的输入到复杂问题求解，从人类个体到人类社会的智能活动，以及人类智能和机器智能的性质。**思维**是客观现实的反映过程，是具有意识的人脑对于客观现实的**本质属性、内部规律性**的**自觉的、间接的和概括的**反映。**智能**是个体认识客观事物和运用知识解决问题的能力。

4. 人类思维的形态：感知思维、形象思维、抽象思维、灵感思维。

5. 神经网络基本特点：① 以**分布式**方式存储信息。② 以**并行方式**处理信息。③ 具有**自组织、自学习**能力。

符号智能：以知识为基础，通过推理进行问题求解。也即所谓的传统人工智能。

计算智能：以数据为基础，通过训练建立联系，进行问题求解。人工神经网络、遗传算法、模糊系统、进化程序设计、人工生命等都可以包括在计算智能

6. 符号智能与计算智能**区别**：符号智能就是传统人工智能，以知识为基础，通过推理求解问题；计算智能以数据为基础，通过训练建立联系，进行问题求解。人工神经网络，遗传算法、模糊等都是计算智能。

7. 非单调推理：一个正确的公理加到理论中，反而使得所得结论变无效。如**封闭世界假设CWA**，限定逻辑；**定性推理**：把物理系统分成子系统，对每个子系统之间的作用建立联系，通过局部因果性的行为合成获得实际物理系统的功能；**不确定性推理**：随机性、模糊性、不确定性。如 DS 证据、模糊集、粗糙集、贝叶斯。

8. 知识、知识表示及**运用知识的推理算法**是人工智能的核心， 而**机器学习**则是关键问题。机器学习的研究**四个阶段**：**①无知识的学习**：主要研究神经元模型和基于决策论方法的自适应和自组织系统。**②符号概念获取**：给定某一类别的若干正例和反例，从中获得该类别的一般定义。**③实例学习**：从实例学习结构描述。**④有知识的学习**：把大量知识引入学习系统做为背景知识

9. 机器学习的风范：**①归纳学习**：研究一般性概念的描述和概念聚类；**②分析学习**：在领域知识指导下进行实例学习， 包括基于解释的学习、 知识块学习等。**③发现学习**：根据实验数据或模型重新发现新的定律的方法。**④遗传学习**：模拟生物繁衍的变异和自然选择，把概念的各种变体当作物种的个体， 根据客观功能测试概念的诱发变化和重组合并， 决定哪种情况应在基因组合中予以保留。**⑤连接学习：**是神经网络通过典型实例的训练，识别输入模式的不同类别。

10. 分布式人工智能：研究在逻辑上或物理上分散的智能动作者如何协调其智能行为，即协调它们的知识、技能和规划，求解单目标或多目标问题，为设计和建立大型复杂的智能系统或计算机支持协同工作提供有效途径。

11. 人工思维将以**开放式自主系统**为基础，充分发挥各种处理范型的特长，实现**集体智能**，才能达到**柔性信息处理**，解决**真实世界**的问题。

12. 知识系统包括：**①专家系统**：专家系统是一种模拟人类专家解决领域问题的计算机程序系统。这类计算机程序包括两部分：知识库。它表示和存储由任务所指定领域知识的一组数据结构集合，包含有关领域的事实和专家水平的启发式知识。推理机，它是构造推理路径的一组推理方法集合，以便导致问题求解、假设的形成、目标的满足等。由于推理采用的机理、概念不同，推理机形成多种范型的格局。**②知识库系统**：把知识以一定的结构存入计算机，

进行知识的管理和问题求解，实现知识的共享。**③决策支持系统**是计算机科学（包括人工智能）、行为科学和系统科学相结合的产物，是以支持半结构化和非结构化决策过程为特征的一类计算机辅助决策系统，用于支持高级管理人员进行战略规划和宏观决策。其组成：数据库管理子系统、模型库管理子系统、方法库管理子系统、知识库管理子系统、会话子系统。

人工智能	符号主义	连接主义	行为主义
认识层次	离散	连续	连续
表示层次	符号	连接	行动
求解层次	自顶向下	由底向上	由底向上

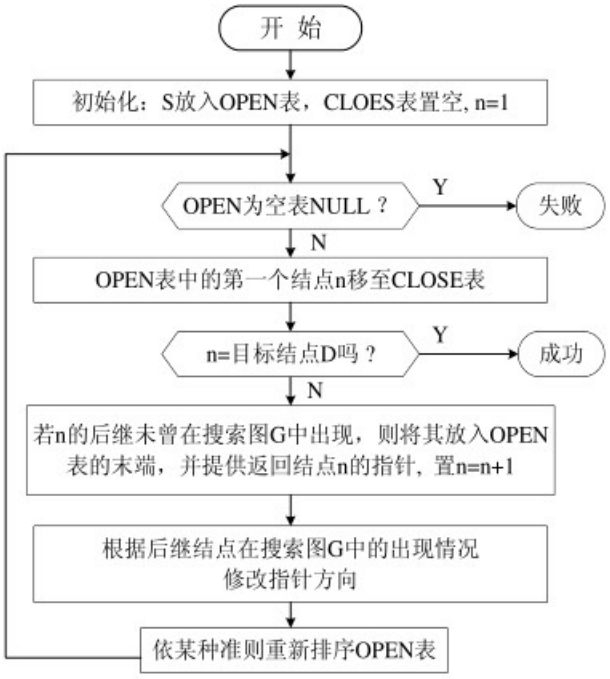
处理层次	串行	并行	并行
操作层次	推理	映射	交互
体系层次	局部	分布	分布
基础层次	逻辑	模拟	直觉判断

二、问题求解

1. 问题表达及其变换 1 同构同态变换 2 问题分解法 3 状态空间求解法 4 问题演绎法 5 博弈问题求解法

问题的状态空间：〈S，F，G〉 S 是初始状态集；F 是操作的集合；而 G 为目标状态集。

☆状态空间搜索算法流程：



2. 问题的状态空间可用有向图来表达，又常称其为状态树

状态空间图在计算机中有两种存储方式：一种是图的显式存储，另一种是图的隐式存储。**图的显式存储:**（1）概念：所谓图的显式存储，即把问题的全部状态空间图直接都存于计算机中的方式。诸如一般计算机文件、程序文件和库文件的存储等，均为图的显式存储方式。（2）适用条件：通常图的显式存储方式需要占据计算机的大量存储空间和处理时间，故这种存储方式仅适用于状态空间十分有限以及较简单的问题求解。（3）特点：其优点是直观、明了，但缺点是占据存储空间大。**图的隐式存储:**（1）概念：仅在计算机中存储关于要求解问题的相关各种知识，只在必要时再由相关的信息和知识逐步生成状态空间图的方式称为图的隐式存储。（2）适用条件：适用于一般问题求解，尤其适宜于状态空间庞大的情况。（3）特点：占据空间小，但不够直观，与图的显式存储刚好特点互补。

3. 深度搜索（盲搜索、非启发、深度越大越晚优先级越高、不完备、非算法）；**广度/宽度搜索**（盲搜索、非启发、深度越小越早优先级越高，完备）

4. 广度优先算法搜索原则：1 深度越**小**、越**早**生成结点的优先级越高。2 当最低层不止一个

结点时，它选择最先生成的结点进行搜索。 广度优先算法步骤：(1) 初始结点 S 加入到队列 OPEN 的尾部；(2) 若 OPEN 为空，则搜索失败，问题无解；(3) 取出 OPEN 队头的结点 n，并放入 CLOSE 队列中；(4) 若 n 是目标结点 D，则搜索成功，问题有解；(5) 若 n 是叶结点，则转(2)；(6) 扩展 n 结点(即找出它的所有直接后继)，并把它的诸子结点依次加入 OPEN 队尾，修改这些子结点的返回指针，使其指向结点 n。转(2)。

5. 深度优先搜索搜索原则：1 深度越**大**、越**晚**产生结点的优先级越高。2 深度优先搜索是不完备的，属于非算法的搜索过程。深度优先算法步骤：(1) 初始结点 S 放入堆栈 OPEN 中；(2) 若 OPEN 为空，则搜索失败，问题无解； (3) 弹出 OPEN 表中最顶端结点放到 CLOSE 表中，并给出顺序编号 n； (4) 若 n 为目标结点 D，则搜索成功，问题有解； (5) 若 n 无子结点，转(2)；(6) 扩展 n 结点，将其所有子结点配上返回 n 的指针，并按次序压入 OPEN 堆栈，转(2)。

6. 有界深度优先搜索特点：引入搜索深度**限制值 d**，使深度优先搜索过程具有完备性。有界深度优先算法步骤：(1)初始结点 S 放入堆栈 OPEN 中；(2)若 OPEN 为空，则搜索失败，问题无解；(3)弹出 OPEN 中栈顶结点 n，放入 CLOSE 表中，并给出顺序编号 n；(4)若 n 为目标结点 D，则搜索成功，问题有解；(5)若 n 的深度 d(n)=d，则转(2) ；(6)若 n 无子结点，即不可扩展，转(2) ；(7)扩展结点 n，将其所有子结点配上返回 n 的指针，并压入 OPEN 堆栈，转(2)。

7. 代价推进搜索特点：节点间有向边的代价不同。(1)广度优先搜索法：每次从 OPEN 表中取出具有**最小代价**的节点进行扩展。(2)有界深度优先搜索法：用代价限制 g 代替深度限制 d，用代价 g(n)代替节点深度 d(n)。

8. 广度、深度、有界深度优先搜索和代价推进搜索法等，它们的**局限性及其特点**为：①基本搜索策略普遍适用于树状问题求解，控制性知识简单，容易编程在计算机上实现。但是它一般必须知道问题的全部状态空间，搜索效果差，求解能力弱，故常被称为**弱方法**。②它们都是依据某种固定规则运行的搜索，均属于**非启发的强力搜索**，没有表现出智能搜索的活跃性与灵活性。③由于基本搜索策略疏忽了对给定问题的特有知识的分析,没有充分考虑所要求解问题的自身发展规律和特性，搜索完全是按事先确定好的固定排序来进行的，这是一种穷尽遍历的大海捞针式的策略，没有任何启发式信息引导，往往使得实际搜索效率很低，故又被称为**盲目搜索**。

9. 启发式搜索策略：利用与问题解有关的启发信息来作引导的搜索策略。特点是：由于充分考虑到问题求解所应用到的各种启发信息及知识，包括利用常识性推理和专家经验等信息与知识，启发式搜索能够动态地确定操作排序，优先调用较合适的操作规则，扩展、比较并选择最有希望的节点，使搜索尽可能以最快的速度，最短的距离，最小的代价，朝着最有利于达到目标节点的方向推进。即以智能思想调节搜索区，使尽量沿着最有希望找到解的路径方向上，纵向小范围地搜索前进。

采用启发式搜索策略：（1）使用了控制性知识中的启发信息，因而弥补了略去的部分状态空间所带来的有用信息丢失；（2） 启发式搜索往往是深度优先搜索法的改进。只需知道问题的部分状态空间，就可以求解智能问题。（3）与基本搜索相比，启发式搜索最大特点就是搜索效率要高得多。**搜索效率及其评价：P=L/T** L 是从根节点到达目标节点的深度；T 是在整个搜索过程中产生节点总数(不计根节点)，因此，这里 P 反映的是朝着目标搜索时的搜索宽度。

10. 启发式搜索方法分为局部择优搜索和全局择优搜索两大类。瞎子爬山局部择优搜索过程：每搜索到达一个节点，其随后选择的下一节点不是按规则预定的或盲目选定的，而是按经验进行智能处理，使用估计函数 f(x)来搜索当前最优的节点。

优点：方法简单，由于取消了 OPEN 表，要处理的数据量减少了，所以占用内存空间少、

速度快。缺点：瞎子爬山法主要只在单因素，单极值的情况下使用，而在多极值情况下会遇到许多困难，导致找不到最佳解，会遇到“多峰”、“盆地”或“平台”、“山脊”问题等。

11. 瞎子爬山法进行人机交互搜索（智能搜索）的主要思路和步骤如下：（1）分析搜索的性质 （2）是否有许多可供选择的路径和方案，各种选择是否有优劣之分； （3）有哪些可供利用的启发信息；（4）启发信息可否编制成操作简便、易于判别、便于实现的规则；（5）编程完成使用规则的程序，用之进行择优搜索。

12. 与/或树的有序搜索，即采用启发式搜索策略，求出能够使得根节点可解的最小搜索树的解树。一般来说，局部择优搜索与全局择优搜索仅适用于状态空间是**代价树**的搜索求解，而有序搜索既可适用于**代价树**的搜索求解，又能适用于**有向图**的搜索求解。

13. 博弈树：在博弈过程中，按照博弈规则和步法状态过程分析，客观评判博弈双方在各自分枝节点上所获得的分数估计值，并依照其中任何一方的角度而依次生成具有得分值表示的与/或搜索树。**博弈原理：**博弈的各方总是要挑选对自己最为有利而对对方最不利的那个行动方案。

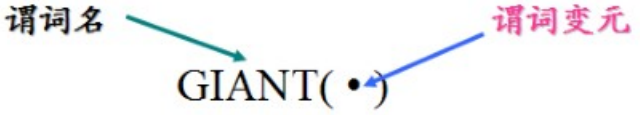
14. α 值—取子节点中的**最大倒推值**，作为**当前下界“≥ •”**，□，MAX 方控制的**或**节点。

β 值—取子节点中的**最小倒推值**，作为**当前上界“≤ •”**，○，MIN 方控制的**与**节点。

三、逻辑、推理与知识

1. 非单调推理指的是一个正确的公理加到理论中，反而会使预先所得到的一些结论变得无效了（限制理论、缺省理论、自认知理论、CWA）。**非单调推理过程**：建立假设，进行标准逻辑意义下的推理， 若发现不一致，进行回溯，以便消除不一致，再建立新的假设。大致分为两类：一类基于最小化语义，称为最小化非单调逻辑；另一类基于定点定义，称为定点非单调逻辑；最小化非单调逻辑可以分为基于最小化模型和基于最小化知识模型。前者主要有封闭世界假设、限制逻辑等，后者包括忽略逻辑等。

2.



在谓词逻辑演算中，最重要的有三大类：命题逻辑演算、一阶谓词逻辑演算和二阶谓词演算。符号**“→”**称为“条件”(Conditional)或者“蕴涵”(Implication)，它表示 **“如果……，则……”**的定义关系。

3. 其他逻辑系统：“非二值”逻辑、逻辑程序设计、非单调逻辑、封闭世界假设、情景演算、动态描述逻辑 DDL

封闭世界假设 CWA 的思想：如果无法证明 P，则就认为它是否定的。即：如果从知识库中无法证明 P 或者¬P，则就向基本信念集合 KB 中增加¬P。

4. 知识按**问题求解要求**分为：**叙述型知识、过程型知识、控制型知识。**知识按**其作用**分为：**描述性知识、判断性知识、过程性知识。**知识按**其描述对象**分为：**对象级知识、元知识。**

5. 产生式推理系统图：如果一个产生式的前提包含了几个事实，那么它的结论对应着这些事实的合取；如果同一个结论可以由多个产生式得到，则这个结论对应着这些产生式的析取。

6. 知识及其表示：不确定性知识表示、状态空间表示法、与或图表示法、知识的逻辑表示法、产生式表示法、语义网络表示法、框架表示法、Petri 网知识表示法。

7. 粗糙集理论在知识发现中的作用：在数据预处理过程中，**粗糙集理论可以用于对遗失数据的填补。**在数据准备过程中，利用粗糙集理论的数据约简特性，**对数据集进行降维操作。**在数据挖掘阶段，可将粗糙集理论**用于分类规则的发现。**

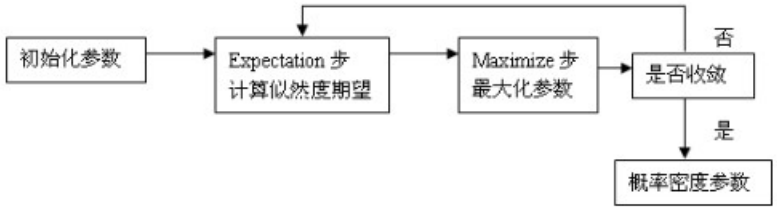
8. 支持向量机方法 SVM 是建立在**统计学习理论**的 **VC 维理论**和**结构风险最小原理**基础上的，根据有限的样本信息在模型的**复杂性（即对特定训练样本的学习精度）**和**学习能力（即**

无错误地识别任意样本的能力）之间寻求最佳折衷，以期获得最好的推广能力。它在解决**小样本**、**非线性**及**高维模式识别**中表现出许多特有的优势

9. $P(A_i \mid B) = P(A_i)P(B \mid A_i) \Big/ \sum_{j=1}^{n} P(A_j)P(B \mid A_j)$ **它是在观察到事件 *B* 已发生的条件下，寻找导致 *B* 发生的每个原因的概率。**

10. 贝叶斯网络是用来表示变量间连接概率的图形模式，它提供了一种自然的表示因果信息的方法，用来发现数据间的潜在关系。在这个网络中，用节点表示变量，有向边表示变量间的依赖关系。

11. EM 算法：E 步是计算待估计参数的似然期望；M 步在假设 E 步所得到的概率分布是正确的情况下，通过极大似然估计对参数进行更新以得到最大的参数 。



12. 粗糙集理论反映了人们用粗糙集方法处理不分明问题的常规性，即以不完全信息或知识去处理一些不分明现象的能力，或依据观察、度量到的某些不精确的结果而进行分类数据的能力。基本粗糙集理论认为知识就是人类和其他物种所固有的分类能力，粗糙集理论利用集合（下近似集和上近似集）处理含糊和不精确性问题。粗糙集主要优点包括：除数据集之外，无需任何先验知识（或信息）；对不确定性的描述与处理相对客观。

13. 模糊集理论利用模糊隶属度来表示自然界模糊现象，从研究集合与元素的关系入手研究不确定性。广泛应用于专家系统和智能控制中。模糊集是不可计算的，即没有给出数学公式描述这一含糊概念，故无法计算出它的具体的含糊元素数目，如模糊集中的隶属函数 μ 和模糊逻辑中的算子 λ 都是如此。

四、机器学习

1. 奥卡姆剃刀原则—优先选择与数据一致的最简单假设。

2. 归纳学习通常是从假设空间中选出一个假设对新实例进行分类预测；**集体学习**是从假设空间中选择一个作为整体的假设集合称为集体，将它们对新实例的分类预测进行合成，然后再输出结果。

ID3 是一种自顶向下增长树的贪婪算法，在每个节点选取能最好的分类样例的属性。继续这个过程直到这棵树能完美分类训练样例，或所有的属性都已被使用过。

熵与不确定性信息论中用熵表示事物的不确定性，同时也是信息含量的表示—熵值越大，表示不确定性越大，同时信息量越多；反之则不确定性越小，信息量越小**熵的单位=比特**

熵和决策树 1 开始，决策树的树根对应于最大的不确定状态，表示在分类之前对被分类的对象一无所知 2 随着每个属性的不断判断，向树的叶子方向前进，即相当于选择了其中的一棵子树，其不确定状态就减小了 3 到达叶子节点，分类完成，此时不确定性为零要提高决策树的分类效率，就相当于要求熵值下降的更快 / 这样，ID3 算法的实质就是构造一棵熵值下降平均最快的决策树

3. 过拟合：给定假设空间 H 和一个假设 h， 如果存在其他假设 h'，使得在训练样例上 h 的错误率比 h'小，但在整个实例分布上 h'错误率比 h 小，则称 h 过拟合训练样例。

过拟合问题的解决解决方法：决策树剪枝— 2 剪枝 / 防止用不相关的属性来划分决策树交叉验证—适用于任何学习算法 / 预留部分已知数据(将训练集中的一小部分留出来做测

试集用),利用其测试其余已知数据归纳出来的假设的性能 / k-交叉检验(每次预留 1/k 数据)。

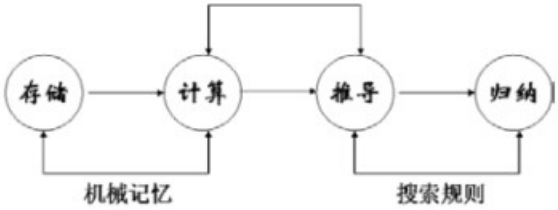
4. 集体学习（boosting）：归纳学习通常是从假设空间中选出一个假设对新实例进行分类预测。集体学习是从假设空间中选择一个作为整体的假设集合称为集体，将它们对新实例的分类预测进行合成，然后再输出结果。Boosting 算法多次调用这个弱学习算法，用不同的训练子集(每个样例赋予不同权值)每次从被调用算法获得一个弱假设 / 算法最终把上述弱假设组合起来，产生一个更加精确的单一假设，以获得更好的预测结果。

5. AdaBoost 总结 Adaboost 算法中不同的训练集是通过调整每个样本对应的权重来实现的。开始时，每个样本对应的权重是相同的，即 其中 n 为样本个数，在此样本分布下训练出一弱分类器 。对于分类错误的样本，加大其对应的权重；而对于分类正确的样本，降低其权重这样分错的样本就被突出出来，从而得到一个新的样本分布。在新的样本分布下，再次对分类器进行训练，得到弱分类器。依次类推，经过 T 次循环，得到 T 个弱分类器把这 T 个弱分类器按一定的权重叠加（boost）起来，得到最终想要的强分类器。

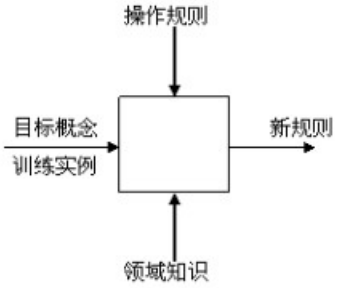
6. 强化学习是指从环境状态到行为映射的学习，以使系统行为从环境中获得的累积奖励值最大。**试错搜索**和**延期强化**这两个特性是强化学习中两个**最重要的特性。**强化学习技术的**基本原理**是：如果系统某个动作导致环境正的奖励，那么系统以后产生这个动作的趋势便会加强。反之系统产生这个动作的趋势便减弱。这和生理学中的条件反射原理是接近的。

7. 常见的几种学习方法：

机械学习：



基于解释的学习：



基于事例的学习 ；基于概念的学习；基于类比的学习：(1) 回忆与联想；(2) 选择；(3) 建立对应映射；(4) 转换

8. ID3 算法（1）创建根节点，先计算 Gain(X)，A 为类标号属性，对 X 划分 C={C1, C2}，

$H(X,C)=-p(C_1)\log p(C_1)-p(C_2)\log p(C_2)$,属性 A 对 X 的划分为…得 C_1^1,C_1^2,C_2^1,C_2^2，

$$E(A)=\frac{|[YELLOW]|}{|X|}H([YELLOW],\{C_1^1,C_1^2\})+\frac{|[PURPLE]|}{|X|}H([PURPLE],\{C_1^1,C_1^2\})$$

$$=-\frac{6}{1}(\frac{4}{60}\log\frac{4}{6}+\frac{2}{6}\log\frac{2}{6})-\frac{4}{1}(\frac{3}{40}\log\frac{3}{4}+\frac{1}{4}\log\frac{1}{4})$$

，所以 Gain(X) = H(X, C) – E(A)，同样方法得到其他的 Gain，最高信息增益，选为根节点。

（2）创建子节点，左边或右边的表作为新数据集，重复步骤（1），计算信息增益，构造内

节点，直到得到一颗完整的决策树。

五、人工神经网络

1. 根据**连接的拓扑结构不同**，神经网络可分为四大类：**分层前向网络（**输入层、中间层（又称隐层，可有一层或多层）和输出层，各层顺序连接；且信息严格地按照从输入层进，经过中间层，从输出层出的方向流动**）、反馈前向网络（**也是一种分层前向网络，但它的输出层到输入层具有**反馈连接**。反馈的结果形成封闭环路，具有反馈的单元也称为**隐单元**，其输出称为内部输出**）、互连前向网络（**也是一种分层前向网络，但它的**同层**神经元之间有相互连接。同一层内单元的相互连接使它们之间有彼此牵制作用**）、广泛互连网络（**指网络中任意两个神经元之间都可以或可能是可达的，*Hopfield 网络、波尔次慢机模型*）

2. 神经网络的工作过程主要由两阶段组成：第一个阶段是**学习期**，此时各计算单元状态不变，各连接权值通过学习样本可修改；第二个阶段是**工作期**，此时各连接权值固定，计算单元的状态变化，以求达到稳定状态。**网络学习**是利用一组称为样本的数据作为网络的输入(和输出)，网络按照一定的训练规则)自动调节神经元之间的连接强度或拓扑结构，当网络的实际输出满足期望的要求或者趋于稳定时，则认为学习成功。

神经网络的功能：1. 数学上的映射逼近 通过一组映射样本，网络以自组织方式寻找输入与输出之间的映射关系；2. 数据聚类 通过自组织方式对所选输入模式聚类；3. 联想记忆 实现模式完善、恢复，相关模式的相互回忆等。4. 优化计算和组合优化问题求解。5. 模式分类。6. 概率密度函数的估计

3. 神经网络模型按学习方式分类：**导师学习、强化学习和无导师学习；**按网络结构分类：**分层结构与互连结构；**按网络的状态分类：**连续时间变化状态、离散时间变化状态**；按网络的活动方式分类：**一种是由确定性输入经确定性作用函数，产生确定性的输出状态；另一种是由随机输入或随机性作用函数，产生遵从一定概率分布的随机输出状态。**

4. BP 网络就是多层前向网络，结构：输入层、一层或多层隐层、输出层，激发函数：S 形函数。**正向传播：**输入信息从输入层经隐层逐层处理后传至输出层，每层神经元的状态只影响下一层神经元的状态。**反向传播：**输出层得不到期望输出，把误差信号沿原连接路径返回，并修改各神经元的权值，使误差信号最小。**BP 学习算法：（1）选取比例参数；（2）进行下列过程直至性能满足要求为止，（a）对于每一训练（采样）输入：①按下式计算输出节点的值 *βz=dz-Oz* ②按下式计算全部其他节点** $\beta_j = \sum W_{j\to k} O_k (1-O_k)\beta_k$ **③按下式计算全部权值变化** $\Delta w_{i\to j} = rO_j(1-O_j)\beta_j$ **（b）对于所有训练（采样）输入，对权值变化求和，并修正各权值。**

5. Hopfield 网络是一种具有正反相输出的带反馈人工神经元。**Hopfield 网络系统**不仅能够实现联想记忆，而且能够执行线性和非线性规划等优化求解任务。

6. 进化计算 建立在进化理论基础上的计算，它是仿照生物生命发展过程而建立起来的计算理论。进化计算研究内容：包括**遗传算法**；进化策略；进化规划和进化编程共四方面的内容。

遗传算法原理：遗传算法基于达尔文进化论的观点，依照适者生存，优胜劣汰等自然进化法则，通过计算机来模拟生命进化的机制，进行智能优化计算和问题搜索求解。

GA 功能：在解决许多传统数学难题以及常规条件下明显失效的复杂问题时，遗传算法提供了一个行之有效的新途径。

7. 遗传算法目的：一方面是通过它的研究来进一步解释自然界的适应过程；另一方面是为了将自然生物系统的重要机理运用到人工系统的设计中。

遗传算法实现：本质上，所谓遗传算法，就是一个通过基因因子选择、重组、复制、评价计算，从而再循环繁殖、继承而不断地进化以接近于最佳种群的过程。换言之，这是一个自适应地逐渐找到最优解的组织实现过程。实现 GA 过程主要包括：编码；确定种群；遗传操作；优胜劣汰等运算过程。

8. 进化计算选择方法的改进：精英保护法、自适应变异概率选择法、移民算法、分布式遗传算法等。

比较项目	遗传算法 GA	进化策略 ES	进化规划 EP
个体表现形式	离散值	连续值	连续值
参数调整方法	无	标准偏差、协方差	方差
适应度评价方法	变换目标函数值	直接使用目标函数值	变换目标函数值
个体变异算子	辅助搜索方法	主要搜索方法	唯一搜索方法
个体重组算子	主要搜索方法	辅助搜索方法	不使用
选择复制算子	概率、保存	确切的、不保存	概率、不保存

9. 群智能：群智能思想的产生主要源于**复杂适应系统理论**以及**人工生命**的研究。群智能理论：群内个体具有能执行简单的时间或空间上的评估和计算的能力；群内个体能对环境(包括群内其它个体)的关键性因素的变化做出响应；群内不同个体对环境中的某一变化所表现出的响应行为具有多样性；不是每次环境的变化都会导致整个群体的行为模式的改变。环境所发生的变化中，若出现群体值得付出代价的改变机遇，群体必须能够改变其行为模式。

10. 复杂适应系统（CAS）：主体的适应性，是指它能够与环境以及其它主体进行交流，在交流的过程中 “学习” 或 “积累经验”，并且根据学到的经验改变自身的结构和行为方式。**CAS 特点：**（1）主体是主动的、活的实体。（2）个体与环境(包括个体之间)之间的相互影响、相互作用是系统演变和进化的主要动力。（3）这种方法不是把宏观和微观截然分开，而是把它们有机地联系起来。（4）这种建模方法还引进了随机因素的作用，使它具有更强的描述和表达能力。**CAS 理论提供了模拟生物、 生态、 经济、 社会等复杂系统的巨大潜力。**

11. 人工生命包括两方面的内容：① 如何利用计算技术研究生物现象；② 如何利用生物技术研究计算问题。

12. 群智能 SI 定义：任何一种由昆虫群体或其它动物社会行为机制而激发设计出的算法或分布式解决问题的策略均属于群智能。

构建一个 SI 系统所应满足的五条基本原则:1.Proximity Principle: 群内个体具有能执行简单的时间或空间上的评估和计算的能力。2.Quality Principle: 群内个体能对环境(包括群内其它个体)的关键性因素的变化做出响应。3.Principle of Diverse Response: 群内不同个体对环境中的某一变化所表现出的响应行为具有多样性 4.Stability Principle: 不是每次环境的变化都会导致整个群体的行为模式的改变。5.Adaptability Principle: 环境所发生的变化中，若出现群体值得付出代价的改变机遇，群体必须能够改变其行为模式。

★13. 粒子群优化算法 PSO：在 PSO 系统初始化为一群随机粒子(随机解)，然后通过迭代找到最优解。在每一次迭代中，粒子通过跟踪两个“极值”来更新自己，同时也通过跟踪它们实现粒子间的信息交换。第一个就是粒子本身所找到的最优解，这个解叫作个体极值 pBest(“自身经验”)。另一个极值是整个群体目前找到的最优解，这个极值是群体极值 gBest(群体的“社会经验”)。

14. 人工鱼群算法 AFSA，它利用自上而下的寻优模式模仿自然界鱼群觅食行为，主要利用鱼的觅食、聚群和追尾行为，构造个体底层行为；通过鱼群中各个体的局部寻优，达到全局最优值在群体中凸现出来的目的。

15. 群智能算法与进化计算比较

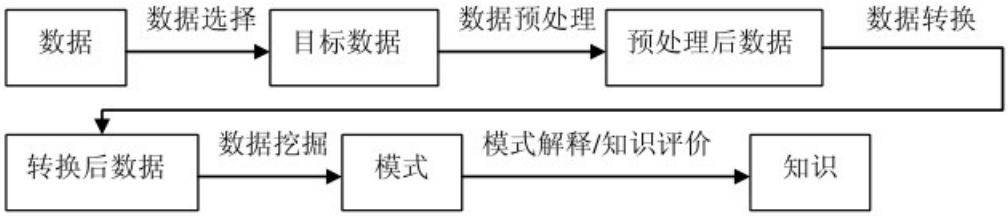
首先，EC 是模拟生物系统进化过程，其最基本单位是基因(Gene) ，它在生物体的每一代之间传播；已有的基于 SI 的优化算法都是源于对动物社会通过协作解决问题行为的模拟，它

主要强调对社会系统中个体之间相互协同作用的模拟，其最基本单位是敏因。其次，EC 中强调“适者生存”，不好的个体在竞争中被淘汰；SI 强调“协同合作”，不好的个体通过学习向好的方向转变,不好的个体被保留还可以增强群体的多样性。最后，EC 的迭代由选择、变异和交叉重组操作组成，而 SI 的迭代中的操作是“跟随”，ACO 中蚂蚁跟随信息素浓度爬行，PSO 中粒子跟随最优粒子飞行。

六、知识发现与数据挖掘

1. 数据库知识发现 KDD：指大量数据中获取有效的、新颖的、有潜在作用的和最终可理解的模式的非平凡过程。**数据集、新颖、潜在有用、可理解、模式、过程、非平凡、有效性**

2. 知识发现的处理过程：



3. 数据挖掘（Date Mining）是从大型数据库或数据仓库中提取人们感兴趣的知识，这些知识是隐含的、事先未知的、潜在的、有用的信息。**广泛观点的定义：**是从存放在数据库、数据仓库或其他信息库中的大量数据中挖掘有趣的知识过程。

知识发现的方法：统计方法（传统方法、模糊集、支持向量机、粗糙集）、机器学习（规则归纳、决策树、范例推理、贝叶斯网络、科学发现、遗传算法）、神经计算和可视化方法（可视化（visualization）就是把数据、信息和知识转化为可视的表示形式的过程）。

4. 分类是找出描述并区分数据类或概念的分类函数或分类模型(也常常称作分类器)，该模型能把数据库中的数据项映射到给定类别中的某一个，以便能使用模型预测类标记未知的对象类。常用分类方法：信息论方法（ID3 方法； IBLE 方法）、集合论方法（粗集方法、概念格方法）、人工神经网络方法、遗传算法、统计分析方法。

5. 聚类：是把数据按照相似性归纳成若干类别，同一类中的数据彼此相似，不同类中的数据相异。聚类是一种无监督分类法， 没有预先指定的类。与分类的区别：分类依赖于预先定义的类和带类标号的训练实例，是一种观察式的学习；而聚类是找到这个簇的特征或者标号的过程。

6. K-means算法：从D中随机取k个元素，作为k个簇的各自的中心；分别计算剩下的元素到k个簇中心的相似度，将这些元素分别划归到相似度最高的簇；根据聚类结果，重新计算k个簇各自的中心；将D中全部元素按照新的中心重新聚类；重复第4步，直到聚类结果不再变化；将结果输出。

7. 序列模式：是指通过时间序列搜索出的重复发生概率较高的模式。时间序列模式根据数据随时间变化的趋势预测将来的值。主要算法：GSP、PrefixSpan 算法。

8. 粗糙集理论是一种研究不精确、不确定性知识的数学工具，这一方法在数据挖掘中具有重要的作用，通常处理含糊性和不确定的问题，发现不准确数据或噪音数据内在的结构关系，可用于特征的约简和相关分析中。**粗糙集方法优点：**不需要预先知道的额外信息，如统计中要求的先验概率和模糊集中要求的隶属度，算法简单，易于操作。

9. 关联规则分析就是发现关联规则，在交易数据、关系数据或其他信息载体中，查找存在于项目集合或对象集合之间的频繁模式、关联、相关性、或因果结构。

例：major(x, “CS”) ^ takes(x, “DB”) ® grade(x, “A”) [1%, 75%]

10. 数据挖掘任务分类：描述：了解数据中潜在的规律；预言：用历史预测未来

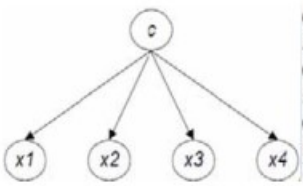
数据挖掘技术：概念/类描述（特征化和区分）；关联规则分析；分类（预言）；聚类；序列

模式；异常检测

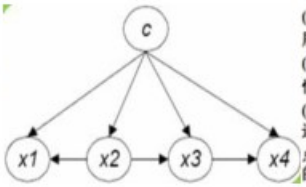
11. Apriori 算法基本思想：频繁项集的任何子集也一定是频繁的。**算法的核心：**用频繁的(k – 1)-项集生成候选的频繁 k-*项集*；用数据库扫描和模式匹配计算候选集的支持度。**算法瓶颈：**候选集生成；巨大的候选集；多次扫描数据库。

12. 贝叶斯公式 $P(c_j \mid x) = \dfrac{P(x \mid c_j)P(c_j)}{P(x)}$ 先验概率 P(cj)，联合概率 P(x|cj)，后验概率 P(cj|x)

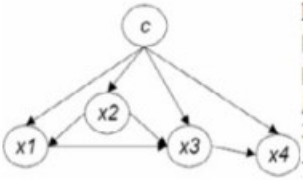
贝叶斯分类器（BNC）原理：通过某对象的先验概率，利用贝叶斯公式计算出其后验概率，即该对象属于某一类的概率，选择具有最大后验概率的类作为该对象所属的类。



(1)NB 中，分类节点是所有属性节点的父结点
(2) 所有的属性变量之间是相互独立的
(3)由于(1)(2)两点，NB 的结构比较容易构建，不需要结构学习算法



(1)TAN 是NB的一个扩展；
(2) 分类节点作为所有属性节点的父结点；
(3)除(2)之外，属性节点还允许有一个其它 属性节点作为其父节点，父节点的个数最多为2个；



BAN是NB和TAN的扩展，它与TAN的不同之处在于，属性节点之间的不再是树形结构，而一个图的结构。

GBN与前面的分类器不同之处在于，它把分类节点看成普通的节点，所有的节点构成一个*有向无环图*。GBN的结构学习常采用贝叶斯网络结构学习的算法，如K2算法。

13. Web 数据挖掘：Web 挖掘是对 Web 文档的内容、Web 上可利用资源的使用情况以及资源之间的关系进行分析，从中发现有效的、新颖的、潜在有用的、并且最终可理解的模式。

Web 数据挖掘流程：查找资源、信息选择及预处理、模式发现、模式分析。

Web 数据挖掘分类