Chapter 1 机器学习概述

1.1 什么是机器学习？

a) Samule：让计算机拥有自主学习的能力，而无须对其进行事无巨细的编程
b) Tom M. Mitchell：计算机程序如果通过某种方法，利用经验 E，提高在任务 T 上的性能(以 P 为评价标准)，则可认为该程序从经验 E 中进行了学习。
c) Nils J. Nilsson：机器在结构、程序、数据等方面发生了基于外部信息的某种改变，而这种改变可以提高该机器在未来工作中的预期性能。

总结：

- 上述这些定义本质上是一致的，即认为机器学习是通过接收外界信息(包括观察样例、外来监督、交互反馈等)，获得一系列知识、规则、方法和技能的过程

1.2 机器学习的基本框架

a) 知识 :人类已经获得的可形式化的某种理性表达,表达可以是确定的,也可以是概率的；可以是全局的，也可以是局部的（很多时候，这些知识被称为先验知识）
b) 经验：指机器在运行环境中得到的反馈，反馈不具有条理性，有有用的，也有没用的。

总结：

- 先验知识和后天经验相结合的信息处理方式是现代机器学习的基本特征之一；
- 在人类知识(Human Knowledge)和实际经验(Empirical Evidence)结合在一起的计算模式中，我们依赖知识设计合理的学习结构，利用实际经验对学习结构进行调整，实现既定学习目标最优化。

c) 学习目标分类：

应用角度：感知任务、归纳任务、生成任务
技术角度：预测任务（包括：回归、分类）、描述任务（聚类、概率估计）

目标函数：均方误差 MSE（回归任务）、交叉熵 cross entropy（分类任务）、Fisher 准则、稀疏性、信息量、最小因素错误准则。

d) 学习结构（一般称为 “模型”）：函数、网络（神经网络、概率图）、规则集、有限状态自动机、语法结构。

总结：定义学习结构、本身就是对先验知识进行形式化的过程。

e) 训练数据 :数据是经验的累积，利用数据对系统进行学习可以更新先验知识、提高系统可用性。数据的质量、数量和对实际场景的覆盖程度都会直接影响学习的结果。
f) 学习方法 :学习方法是学习过程的具体实现，即算法。一般将算法依是否需要**人为标注**分为**有监督学习**(Supervised Learning)、**无监督学习**(Unsupervised Learning)、**半监督学习**(Semi-Supervised Learning)和**增强学习**(Reinforcement learning) 。依**优化方法**分类，可分为**直接求解**(如 PCA 模型中求解数据协方差矩阵的特征向量)、**数值优化**(如神经网络中的梯度下降算法)和**遗传进化**(如协同学习中的鸟群算法)等。

1.3 机器学习流派

a) 符号学派：所有智能行为都可以被简化成在一个逻辑系统中的符号操作过程。
   - 优点：该方法在受限领域中表现出明显优势，因为在受限领域内，知识总结可以非常细致完备，而且很少有新知识加入，因此可以构造一个高精度的推理系统。事实上符号方法取得最大成功的也是在这些领域，如定理证明、路径优化、领域专家系统等。
   - 缺点：这些学习通常受到很严格的限制，无法摆脱既有知识框架的约束。当领域知识变得宽泛复杂以后，符号方法越来越难以实现；符号方法的一个明显缺陷是对不确定性的描述能力不足。
b) 贝叶斯学派：他们认为所有事件都是不确定的，因此要用随机变量来描述;同时，不同事件之间的关系也是不确定的，也应该用概率形式来描述。
   - 优点：和符号方法相比，贝叶斯方法引入的概念是革命性的:它引入了随机变量，对事件的随机性有了基本描述手段;它用条件概率来描述事件之间的关系，对规则上的不确定性具有天然描述能力;它将复杂事件之间的关系统一到概率框架中，将演绎过程归结为边际概率(Marginal Distribution)计算，将推理过程归结为后验概率(Posterior Distribution)计算，简洁而自洽
   - 缺点：推理过程中计算会比较复杂；简单的概率结构会降低模型对实际问题的描述能力；在复杂问题上，两个变量之间是否存在关系、存在何种关系，通常只有领域专家才能确定，给应用带来了某种局限性。
c) 连接学派：连接学派也称为神经网络学派，其基本思想是基于大量同质结点的连接网络来模拟智能行为。
   - 与贝叶斯学派的区别与共同点：
     - 共同点：都依赖一个结点网络
     - 区别：1、贝叶斯学派每个结点有清晰定义，是不同质的；而连接学派中的结点是同质的，不代表具体事件。2、贝叶斯中结点都是随机变量，具有概率意义；而连接学派中结点更像计算结点，较少具有概率意义。
   - 总结：

结点的特征	同质	不同质
随机	兼具贝叶斯和神经网络的概率模型(如：玻尔兹曼机)	贝叶斯方法
不随机	神经网络方法	符号学派方法

d) 进化仿生学派：在进化论者看来，真正有价值的学习在于自然选择;与其辛苦地做后验概率计算或反向递度传播，不如让电脑模拟出各种模型结构和参数，检测基于这些不同结构和参数得到的模型性能，优秀的留下，不好的淘汰。(没有学习结构，这是一种学习方法)

   - 遗传算法
   - 群体学习（social learning）
   - 协同学习（collaborative learning）

流派总结：1、技术发展是往前进的，该成为历史的注定会成为历史；2、技术发展具有一定曲折性，任何思想方法随着时代条件的变化会有不同的表现形式；3、未来各个学派之间的界限也许会越来越模糊。

1.4 机器学习的成功应用的例子（让人惊讶的学习）

- 机器人群体学习系统（google 机器手臂）
- 图片和文字理解
- 金融市场量化分析
- Alpha Go

1.5 机器学习基础

a) 训练、验证与测试：

- **训练**:给定一个包含 N 个样本的训练集$\{(x_1, t_1), \ldots, (x_N, t_N)\}$, 调整映射函数$f_w$的参数 w, 使得预测结果$f_w(x_n)$和标注$t_n$尽可能接近。
- **测试**:将映射函数f_w应用在一个独立的侧试集$\{(\tilde{x}_1, \tilde{t}_1), \cdots, (\tilde{x}_{N'}, \tilde{t}_{N'})\}$，验证所学到的映射函数能否在该数据集上做出准确预测，即$f_w(\tilde{x}_n)$是否与$\tilde{t}_n$接近。

  过拟合：当一个模型过于适应训练集时，往往会导致在新数据上性能下降，这一现象称为 “过拟合”（Over-Fitting ）。
  欠拟合 :在训练初期, 模型在训练和测试数据上的表达能力都比较差, 这时训练处于 “欠拟合” (Under-Fitting)状态;

- **验证**集(Validation Set)：是为了防止过拟合与欠拟合现象的发生。设置验证集，先在验证集上进行模型选择，最后在测试集上进行测试。

b) 模型的表达能力与泛化能力：一般来说，越简单的模型对数据的表达能力越弱，但泛化能力越强；反之，越复杂精细的模型对数据的表达能力越强，但越容易产生数据依赖，产生过拟合现象。

c) **Occam 剃刀原则**：考虑到模型表达能力和泛化能力的权衡，一般在保证较好表达能力的前提下尽量选择最简单的模型。一方而，简单模型具有较强的泛化能力，对数据变动比较鲁棒;另一方而简单模型计算更简单，训练起来也更容易。

d) **典型机器学习方法**：

- 监督学习与非监督学习：依据训练数据是否有人为标注
- 线性模型与非线性模型：所谓线性模型，一般是指模型参数与学习目标之间具有线性关系。不具有线性关系的模型称为非线性模型。
- 参数模型与非参数模型：参数模型是基于某一模型结构，由一组固定参数定义的模型；（有较强的模型假设，需要训练数据较少）非参数模型没有明确的参数集，参数量通常和训练数据相关。（非参数模型对测试数据的表达依赖训练数据本身，因此需要较多的训练数据以覆盖数据空间。）
- 生成模型与区分性模型：在分类任务中，常遇到生成模型和区分性模型。生成模型对每类数据分布进行建模$P(x|C_k)$，再依贝叶斯公式得到分类模型；区分性模型不考虑数据分布，仅关注分类而，直接对分类而建模
- 概率模型与神经模型：概率模型来源于贝叶斯学派，对所有因子建立显式的概率分布及相关性，具有较强的先验假设，需要的训练数据较少（所有推理任务归结为后验概率计算，结果具有明确概率意义）；神经模型来源于连接学派，不对每个因子建立明确的概率形式，而是通过多层非线性映射的组合来模拟复杂的映射关系，需要较多的训练数据来确定这一映射中的多个参数。（函数拟合）

补充：
回归问题（regression）：即指我们的目标是预测连续值输出
分类问题（classification）：预测离散值的输出 （0，1，2，……）

Chapter 2 线性模型

2.1 线性预测模型

a) 多项式拟合

- 目标：找到参数 W，使得数据集 D 中的样本 x 的预测结果尽可能接近真实值 t
- 方法：采用平方误差为误差函数，误差函数是参数 W 的函数，对该误差函数进行优化，具体方法是对每个参数 W 都求偏导，并令其等于零。解得 W 的值。
- 问题：为什么采用平方误差做误差函数，而不是其它函数？

b) **线性回归（linear regression）**：用来预测连续的输出值

- 引入：通过引入概率模型来帮助我们描述不确定性，并由此得到基于概率的最优解。
- 假设：y（预测值）与 t（真实值）之所以存在差别，是因为观察值 t 由于种种原因是随机的，不确定的。不论产生这种随机性的原因是什么，我们假设这一随机性符合一个以 0 为均值，以β^{-1}为方差的高斯分布。引入一个随机变量ε来表示这一随机性。
- 最大似然（Maximum Likelihood，ML）准则：在给定输入变量 x 的请款下，我们得到了目标观察变量值 t 的生成概率，我们希望此概率值越大越好（即：似然函数值越大越好），这样模型对该数据集的描述能力越强，找到一组参数使得似然函数值最大，就是最大似然准则，相应的优化方法就是最大似然估计。
- **最大似然估计等价于线性拟合**，线性拟合中的平方误差事实上假设了目标观察值中的噪声符合高斯分布。

c) **Fisher 准则与线性分类**：

- 分类问题：给定输入向量 x 的特征，我们希望分类器能预测 x 所属的类别 t。具体有三种解决方法：
  - 区分函数法：Fisher 线性分类函数（没有考虑概率意义）
  - 生成性概率模型法：对每个类c_k都建立一个统计模型$p(\phi|C_k;w)$，分类时考察测试样本在每个模型上的概率，再基于贝叶斯公式得到属于某一类的后验概率$P(c_k|\phi)$。这种方法依赖模型假设与实际数据的契合程度，假设月合理，分类性能越好。
  - 区分性概率模型法：对后验概率（MAP）建模
- Fisher 准则：如果基于训练数据能学习一个优化的参数 W，使得不同类的训练样本在映射空间里的区分性最大。则能得到这样一个分类函数：

$$J(\mathbf{w}) = \frac{\mathbf{m}_2 - \mathbf{m}_1}{\mathbf{S}_1^2 + \mathbf{S}_2^2},$$

其中m_1，m_2，表示 c1，c2 的样本点在映射空间里的均值，S_1，S_2是映射空间里的方差。该公式表明类间距离越大，类内的分散程度越小，则依 Fisher 准则

这两个类的区分性越强。

- 基于 Fisher 准则的线性模型也称为**线性判别分析**(Linear Discriminate Analysis，LDA)
- 在分类问题上，Fisher 准则得到的映射函数和基于线性拟合得到的映射函数是等价的
- 线性拟合 (Fisher 准则) 存在的问题 :当出现奇异数据的时候，分类性能下降，因为基于最小平方误差准则会使这些奇异点的影响过大。

d) Logistic 回归

- 引入 :由于线性回归方法对奇异点敏感，所以提出对它的解决办法，即 :Logistic 回归。
- 方法 :对于二分类问题，假设真实值 t 符合伯努利分布，定义 logistic 函数为 :

$$\sigma(a) = \frac{1}{1+e^{-a}}.$$

该函数将实数域映射到开区间（0，1），起到非线性压缩作用。

- 过程：首先对输入 x 经过一个非线性映射ø（·）生成特征，再经由一个线性映射W^Tø投影到一个标量空间，再经过$\sigma(\cdot)$压缩到(0,1)之间，最后把该压缩值作为伯努利分布的参数生成目标 t。这一模型称为 Logistic 回归模型。比较 Logistic 模型和线性回归模型，可见二者具有相似性，差别只是一个非线性映射函数$\sigma(\cdot)$和伯努利分布假设。
- 使用梯度下降法来实现对交叉熵误差函数的最小化。

e) Softmax 回归

- 同上的方法可应用到多分类问题上，称为 Softmax 回归。基于多分类问题的性质，Softmax 回归将目标 t 由二分类问题中的伯努利分布扩展到多项分布(Multinomial Distribution)，相应的表示方法也由 0-1 表示扩展为 One-hot 表示。

2.2 线性概率模型

线性预测模型假设输入变量和输出变量是可见的，基于输入变量对目标变量进行预测。而线性概率模型中输入变量是不可见的随机变量。当输入变量变成隐变量后，我们能观察的数据只有目标变量 t，因此学习方式由监督学习变成无监督学习，推理过程也由前向预测变成了反向推理。

a) **主成分分析**(Principal Component Analysis，PCA)

- PCA 希望通过找到若干相互正交的方向，使得观察数据在这些方向上的映射最大可能地代表原数据的分布性质。每个方向称为一个主成分(Principle Component，PC)。依对原数据的代表能力排序，称为第一主成分，第二主成分等。一般来说，我们希望找到最具有代表性的几个主成分来代表数据，主成分的个数一般远小于数据维度，因此 PCA 可用在数据降维中。
- 两种准则评价主成分对数据的代表性：
  - 数据在主成分代表的映射空间里方差最大
  - 由映射恢复成原始数据的损失最小
- 所有主成分都是协方差的特征向量，且加入每一个主成分后所增加的方差等于该主成分对应的特征值。因此，若要求前 K 个主成分，只需选择特征值最大的 K 个特征向量即可。

b) **概率主成分分析**(Probabilistic Principal Component Analysis，PPCA)

- 引入：PCA 是经典的无监督学习方法，广泛应用于降维、正规化、流形学习等任务中。然而，PCA 的优化函数(映射空间方差最大或恢复误差最小)和主成分之间的正交限制很大程度上是种人为定义，这使得 PCA 的适用性缺少明确解释。而用线性概率模型来解释 PCA，这一方法称为概率主成分分析(Probabilistic PCA， PPCA)
- PCA 事实上假设了一个线性高斯模型，基于这一模型，观察数据由一个简单的各向同性的正态分布经过一个线性变换得到，因此观察变量 t 也应该符合高斯分布。当这一条件不满足时，PCA 的适用性将会降低。例如，当数据 t 明显不符合高斯分布时，PCA 的结果可能会产生较大偏差。

c) **概率线性判别分析** (Probabilistic Linear Detection Analysis，PLDA)

- 引入：PLDA 解决了非参数模型中的两个严重问题，一是如果数据中包含的类很多，则不仅计算开销增加，对数据的利用也不够充分(例如对那些样本很小的类，基于该模型得到均值u_k的会产生偏差);二是无法处理在测试时遇到的新类，因为这些类在训练数据中没有出现过，因此可能不适合当前模型。
- 使用期望最大化 (Expectation Maximization，EM)算法来对 PLDA 的模型参数进行估计，每一轮迭代分为期望计算（E）和期望优化（M）两步。EM 算法是解决包含隐变量的概率模型问题的基本方法，可以证明该算法总会收敛到局部最优。
- 比较：PLDA 与 PPCA 都基于最大似然准则，都是线性高斯模型；PCA 用于描述任务，LDA 用于分类任务，PCA 用于非监督学习，LDA 用于监督学习。都是概率模型，如果数据不符合高斯分布，性能都会显著下降。

2.3 贝叶斯方法

- 引入：上面方法的问题是没法引入有价值的先验知识，因为模型中的参数都是确定的。而贝叶斯方法将模型参数看作随机变量，将对参数取值范围的先验知识转化成参数的先验概率。引入随机参数不仅可以利用人们对模型的先验知识，更重要的是对模型本身的改变:对模型的优化不再是寻找一个最优参数(如最大似然估计)，而是对参数的后验概率进行估计。因此，即使引入的先验是一个无信息先验(如大范围的平均分布)，贝叶斯方法依然有重要价值。
- 最大后验估计依然是点估计，即选择某一确定的参数进行预测和推理。事实上，后验概率$\mathrm{p(w|D)}$提供了一种更有价值的预测和推理方式，即在预测或推理时考虑所有可能的参数 W，这样得到的结果会更加可靠。这一方法称为贝叶斯方法。以预测任务为例，贝叶斯方法可写成下式:

$$t = \int t\,p(t|\mathbf{w},\mathbf{x})\,p(\mathbf{w}|D)\mathrm{d}\mathbf{w}.$$

2.3 小结：

线性模型	输入变量是否可见	是否是监督学习
预测模型	可见	是
描述模型	不可见	否

线性模型	方法
线性预测模型	linear regression、logistic regression、softmax regression
线性概率模型	PCA、PPCA、LDA、PLDA

Chapter 3 神经网络

- 非线性变换学习方法包括：参数学习和非参数学习
  - 参数学习 :预先定义一个变换的函数形式，再对该形式中的参数进行学习。如：人工神经网络模型（ANN）
  - 非参数学习：模型没有一个确定的参数集，参数的多少与训练数据相关，如：支持向量机（SVM）模型

1. 神经网络概述

a) 对人工神经网络的研究可分为两个方向：

i. 如何描述人类大脑的实际运作方式：激励方式、抑制机理、传导模型；对这样的研究结果，可设计相似的人工结构对其进行模仿。（对人的神经网络进行模仿）

ii. 关注神经网络的表达能力，关注通过神经网络可实现的功能。（机器学习中对神经网络的研究主要采用这种方式）

b) 人工神经网络的定义：

i. Wikipedia：在机器学习和认知科学中，人工神经网络(ANN)是一个受生物神经网络(动物的，特别是大脑的中枢神经系统)启发而提出的统计学习模型家族。该网络可用来估计或近似那些未知的、能够根据大量输入而产生反馈的生物神经网络的一些功能。

ii. Simon Hayki 给出的工程化定义:神经网络是由简单处理单元构成的大规模并行分布式处理器，天然地具有存储经验知识并对其进行运用的能力。神经网络在两方面对人脑进行模拟：

1. 知识通过学习从环境中获得；
2. 知识被存储在神经元之间的连接权重中。

iii. 神经网络的主要特性：

1. 同质性:神经网络中的处理单元(神经元)是简单的、同质的，不同单元不论从信息接收、信息处理、激发模式等方面都具有高度一致性;
2. 连接性:神经网络中的神经元之间是互联的，通过组成网络来存储知识和模拟推理过程;
3. 可学习性:神经网络是可学习的，通过改变神经元之间连接的权重，实现网络学习，适应外来数据。

iv. 人工神经网络的功能：包括记忆、归纳（抽象）、演绎（预测）

v. 神经网络结构分为：**基于映射的、基于记忆的、基于过程的、模拟人类大脑的神经图灵机。**

2. 基于映射的神经模型

输入为一个特征向量，输出为基于该输入的预测目标。目标可以是回归任务中的一个预

测值，也可以是分类任务中的后验概率。

a) 模型优化方法：
   i. 梯度下降法（Gradient Descent，GD）
   ii. SGD
   iii. BP（Back Propagation）

b) 多层感知器（multiple layer perceptron，MLP）：
   i. 输出函数：为线性函数或 Logistics 函数，因为这些函数是连续可导的，因此可基于梯度下降算法进行优化
   ii. 模型结构：前向网络，（激活函数在回归任务中一般取线性；分类任务中一般取 Logistics 或 softmax 函数）
   iii. 训练方法：多层结构不能直接得到解析解，所以通常采用数值解法，即：梯度下降法（GD）、SGD、BP。
   iv. 训练技巧：
      1. 输入/输出正规化和特征变换(Feature Normalization and Transfer)
      2. 选择合适的激发函数(Appropriate Activation Function)
      3. 连接权重初始化(Weight Initialization)
      4. 二阶信息(Second Order Information)
      5. 使用动量(Using Momentum)
      6. 课程学习(Curriculum Learning)
      7. 迁移学习(Transfer Learning)
      8. 正则化(Regularization)

c) **径向基函数网络（Radio Basis Function，RBF）**：MLP 是基于函数嵌套设计非线性变换，每一层变换函数是简单的线性映射附加一个非线性激发函数。而径向基函数网络则是在映射空间设计一系列标识点（Anchor Points）v_j，基于这些标识点，可以将每一个采样点 x 用该点到 v 之间的距离表示出来，实现了由原始空间到变换空间的非线性变换$\emptyset(x)$。每个标识点v_j代表变换空间中的一个基（Basis），基于v_j的距离函数称为一个径向基函数，即 RBF。（其核心思想是：利用映射空间里一些具有代表性的点作为基来实现原始数据的非线性变换，一次这些表示点的选择至关重要）
   i. RBF 网络与训练方法：
   ii. RBF 网络的意义：从插值分析到 RBF、从核回归理解 RBF、从分类任务理解 RBF;
   iii. 多层感知器与径向基函数网络的比较：
      1. 相同点：当隐藏结点足够多时，均可描述任意连续函数
      2. 不同点：采用不同的非线性扩展方法得到特征映射；MLP 模型张的每个隐藏结点的“等激发线”是一个平面$C = w^T x$，而 RBF 模型中的每个隐藏结点的“等激发线”是一个球面$C = \|x - v_j\|$。这说明 MLP 中的 x 的变动对隐藏层的影响是全局的，而 RBF 中 x 的变动对隐藏层的影响是局部的。因此 MLP 是一个参数高度共享的网络，而 RBF 则不同，对任何输入 x，其信息只通过少数和它临近的 vj 所对应的隐藏结点向后传递，对应的，梯度回传也仅与这些隐藏结点相关，因此 RBF 的参数共享性较弱。所以 RBF 需要更多隐藏结点，且泛化能力较弱，在没有被 RBF 有效覆盖的数据空间通常预测性能较差；从**训练角度**看，RBF 网络可用无监督学习来训练 RBF 函数，极大降低了训练难度。即便是监督学习，因为 RBF 的局部特性，参数共享较小，参数更新时的互相影响(Co-adaptation 较小，训练起来也比参数高度共享的 MLP 要容易。

d) 神经网络模型与先验知识
   i. 引入：标准 MLP 模型是一种全连接结构，具有强大的学习能力，但这一模型缺少先验知识，是一种纯数据驱动方法。这导致网络训练通常需要大量数据，且容易产生过训练或欠训练。但很多实际应用问题中，我们对问题本身是有一定认识的。
   ii. 结构化模型与卷积神经网络：（将先验知识与神经网络相结合的方式）
      1. 典型结构化特性包括：空间结构、时序结构、频域结构
      2. 卷积神经网络（CNN）：
      3. 混合密度网络
      4. 贝叶斯方法：

3. 基于记忆的神经模型：
   a) 引入：在实际生活中，还有另一类问题，在这些问题中仅有数据 X 而没有明确的数据标记 t，我们希望设计一个模型可以描述 X 的分布，或者得到代表 X 的抽象特征。对这类任务，基于映射的神经网络并不适合。研究者提出各种基于记忆的神经网络来处理这一问题。当网络训练完成后，对于一个测试样本，可以基于网络中的记忆信息得到该样本的概率，采样出与之近似的训练样本，或得到该样本背后的隐藏结构。
   b) Konomen 网络：又称自组织映射（Self Organization Map，SOM）。其基本思想是将高维数据映射到一个低维空间，使得在高维空间中的分布结构在低维空间得以保持。
      i. Kohonen 网络是一种**局部映射**。高维空间中的某个数据点只与和它最相似的网络结点有关，而与大多数结点无关。而 PCA 是全局线性映射，只对高斯分布数据有效。
      ii. Kohonen 网络可发现数据分布的基础模式，并可用于简单的特征提取。
   c) Hopfield 网络：
      i. 引入：Kohonen 网络的表达能力与矢量量化(Vector Quantitization, VQ)相似，不同结点有独立的代表向量，结点间缺少参数共享，没有形成结构化协同表示，因此记忆能力很低。
         1. 该网络包含若干二值的神经元接待你（取+1 或-1），每一对结点间通过有向边相连，我们将所有结点组成一个向量，该向量的某种取值方式称为一个 “模式”。
         2. Hopfield 网络的学习过程即是对训练数据所代表的模式进行记忆的过程。
      ii. 模型学习：
         1. Hebbian 准则 :同时激发的单元互相连接。(同时激发的神经元连接增强)
      iii. Hopfield 网络的记忆功能：该网络具有抗噪功能；但其记忆能力有限，记忆效率低下。
   d) 玻尔兹曼机（Boltzmann Machine）：
      i. 引入 :Hopfield 网络有两个特点:一是其所有神经元结点都是可见的，二是结点的取值是确定的。这两点限制了该网络的表达能力。玻尔兹曼机（Boltzmann Machine)引入隐藏结点和结点取值的随机性来解决这一问题。这一模型事实上是一个无向图，亦称为马尔可夫随机场(Markov Random Field, MRF)。
      ii. 缺点：很难训练
      iii. 这一模型，连接了贝叶斯方法和神经模型方法两大学派；三是这一模型连接了物理学中某些简单的动态系统(如铁磁化过程)和机器学习中的概率模型，揭示

了概率方法与物理学的某些深刻联系。

e) 受限玻尔兹曼机（Restricted Boltzmann Machine，RBM）

i. 引入：通用玻尔兹曼机很难训练，但如果对其结构进行若干限制，则可得到有效的训练方法。一种限制结构是将可见结点和隐藏结点分为两组，只有不同组的两个结点可以相互连接。这个结构就称为限制性玻尔兹曼机。

ii. RBM 的应用：

1. RBM 可以认为是一种模式学习方法，其隐变量可以表达数据的显著模式。
2. 如果因变量比较少，可以看作是一种数据降维方法
3. 因为 RBM 可以学习数据中的主要特征，去掉干扰，因此可用作特征提取模型。
4. RBM 多用于非监督学习，但如果可见变量中包含某些目标变量，则 RBM 也可用于监督学习，如分类任务。

iii. RBM 训练

iv. 对比散度训练（算法流程#）

f) 自编码器（Auto Encoder，AE）是另一种基于记忆的神经模型。

i. AE 与其他模型的关系：

1. AE 与 PCA：PCA 训练目标是使得线性变换后得到特征在对输入进行重构时误差最小，因此 PCA 可以认为是一种特殊的 AE，其中编码器和解码器是线性的，且二者共享参数。显然，AE 的结构比 PCA 更灵活，允许非线性编码和解码，允许解码器和编码器有独立的参数，允许更灵活的目标函数。因此，AE 具有比 PCA 更强大的学习能力。
2. AE 与 RBM：由 RBM 的对比散度训练过程可知，RBM 的训练目标也是对原始输入数据的重构误差最小，这与 AE 的训练目标非常相似 ;AE 与 RBM 的区别在于 RBM 的编码与解码都是随机的，而 AE 的编码过程是确定的

ii. 其它约束： AE 的学习目标是对数据进行重构，因此需要一定的约束条件才可避免学习到平凡解(等值映射)。

4. 基于过程的模型

a) 引入：上文我们提到的映射模型和记忆模型可以认为是一种“静态模型”，即仅描述数据的分布特性。在实际应用中，我们还常遇到另一种问题，在这些问题中，样本的出现具有很强的序列性，且序列中的样本间存在很强的时序相关性，例如语音信号中不同时刻的采样点，自然语言理解任务中的文本序列，股票交易信号中不同时刻的交易记录，脑电波信号中不同时段的样本等。这类和时序相关的问题称为**序列问题**。解决序列问题的模型通常称为动态模型或过程模型。

b) 解决思路：解决序列问题的基本思路是使模型本身带有时序性，使之可以描述序列信号中的动态发展。传统方法包括各种动态概率模型，如离散状态空间的隐马尔科夫模型(HMM) 、连续状态空间的线性卡尔曼滤波器(Kalman Filter)，或更通用的动态贝叶斯模型(Dynamic Bayesian Network, DBN)方法等。

c) 代表方法：基于过程(序列)的神经模型利用神经网络来模拟这种动态性。在这种神经网络中，网络输出不仅依赖当前输入，还可以依赖前序所有输入和输出，因而可学习数据中的序列相关性。这种网络通常称为递归神经网络 (Recurrent Neural Network, RNN)。值得注意的是，序列问题通常是和时间序列相关的，因此 RNN 通常用在时序信号建模上，但 RNN 可以处理更广义上的序列，如逻辑序列。比如我们解一道数学题，完成一个化学实验，这些任务一般需要几个步骤，这些步骤之间固然有时序性，但更重要的是逻辑上的先后性。

d) Elman RNN:
e) 门网络
f) 序列对序列网络
g) 基于 Attention 模型的诗词生成

5. 神经图灵机
6. 总结：

神经模型	映射模型	记忆模型	动态模型	神经图灵机
功能	学习输入输出关系	学习内部模式	学习时序过程	学习复杂操作

Chapter 4 深度学习

1. 从浅层学习到深度学习

a) 深度学习的优势：
   i. 首先，深层网络有比浅层网络更强大的函数表达能力;
   ii. 第二，深层网络的层次学习方式可以从原始数据中抽象出典型特征，这和人类神经系统处理信息的方式一致;
   iii. 第三，深度学习提供了基于非监督方法进行特征学习的有效方式;
   iv. 第四，深度学习已经超越特征学习和任务决策的范畴，成为一种知识积累和过程学习的有效工具。

2. 深度神经网络训练
   a) 基础训练算法：
      i. 梯度下降法：GD、SGD、Mini-Batch SGD
      ii. 二阶方法：
         1. Newton 法： Newton 方法不能判断驻点是否是极小值点，一旦进入驻
         2. 点即认为已经完成优化，因而无法摆脱马鞍点的吸引，这是 Newton 方法的一个主要缺点；牛顿方法的另一个显著缺陷是需要计算 Hessian 矩阵的逆矩阵。
         3. Quasi-Newton 法
         4. Truncated Newton：其基本思路是将优化问题转换成解以 Hessian 矩阵为参数的线性方程组，而在解这一方程组时不必依赖 Hessian 矩阵 H，只需依赖 H 与偏移量 d 的乘积 Hd 即可。
         5. 自然梯度下降（NSGD）
      iii. DNN 训练的困难
         1. 局部极值问题
         2. 梯度奇异问题
         3. 激发函数非线性饱和区
         4. 梯度爆炸和梯度消失：
         5. 马鞍点 ：是指目标函数仅在某些方向局部最小，而在另一些方向是局部最大或非极值的点。这些马鞍点是驻点，对典型的优化方法(如 SGD, Newtown 等)具有“吸引力”，使得这些方法趋向收敛到这些点，但这些点又不是局部极小值点，因此会导致训练进入非优化状态。

6. 马鞍点现象的理论解释

iv. DNN 训练技巧

1. 参数初始化：一般希望输入特征的方差可以在前向传递时得以保持。如果对特征进行了 Mean-Variance 正规化，则输入特征的方差为 1。如果希望每一层 DNN 的输出(经过非线性变换)保持这一方差，则需对初始参数的大小有一定约束。
2. 学习率调整
   a) 一个简单的策略是在训练开始时设置较大的学习率，并在迭代过程中逐渐减小。
   b) 还可以根据学习效果对学习率进行调整。
   c) 二阶方法设置参数相关学习率
      i. 动量（Momentum）
      ii. 学习率自适应：包括 AdaGrad、RMSProp、AdaDelta
      iii. 学习率自适应+动量：ADAM 方法
3. Batch Norm :我们设想可以通过一个标准化过程来解决 “Covariance Shift” 现象，即对于每一个 Mini-Batch，将每个网络的每一层输入都标准化到一个均值为 0，方差为 1 的分布。通过这一标准化，每一层输出的分布情况保持稳定，从而提高后一层的学习效率。这一方法称为 Batch Norm（BN)
4. Dropout :Dropout 是 DNN 练中另一种简单有效的方法。所谓 Dropout，是指在训练过程中随机将某些隐藏结点的输出置零，使得其连接权重不会更新。Dopout 方法可使不同结点间的依赖性降低，学得相对独立的有效特征。

3. 神经网络正则化

i. 结构化网络与参数共享
   1. 卷积神经网络
   2. 递归神经网络
   3. 神经自回归分布估计（NADE）

ii. 范式约束与稀疏网络：在目标函数中引入约束项(正则项)是防止模型过拟合的通常做法。从贝叶斯角度看，引入这些约束项相当于引入先验知识，通常可减少 DNN 学习的盲目性。
   1. L2 约束：防止过拟合
   2. L1 约束 :与 I2 约束类似，I1 约束将模型参数约束在零值附近。不同的是，I1 约束倾向于拉开不同参数之间的距离，让某些参数首先接近零值。
   3. Sparse Coding 稀疏编码

iii. 加噪训练与数据扩增：增强数据的泛化能力，提高系统的鲁棒性

iv. 联合训练 :联合训练(Joint Training)是另一种隐性正则化方法。所谓联合训练，是指在训练中不仅关注目标任务，同时也关注相关任务，从而避免在目标任务上过度训练引起的过拟合。
   1. 一种联合训练方法是将有监督学习和无监督学习结合起来，学习目标任务(如分类任务)的同时，通过一个 AE 对原始输入进行重构，这一学习方式的优点在于可以利用大量无标注数据来学习有效特征。
   2. 另一种联合训练方法是特征共享学习。如果若干任务具有相似的前端处理过程，则这些前端处理模块可以共享。
   3. Deeply Supervised Net（DSN）

v. 知识迁移 :迁移学习(Transfer Learning)是指将在一个任务上学到的知识迁移到其它任务上。

4. 生成模型下的深度学习

a) 引入：当前深度学习多基于确定性网络，即网络中各个神经元的激发值是由输入和参数决定的，缺少随机性。这种确定性网络适合分类、回归等可描述为映射函数的任务，但不擅长描述复杂的概率关系，因此对生成问题能力有限。相对应的，概率模型，特别是贝叶斯模型，对变量之间的概率关系直接建模，因而是天然的生成模型。如何将神经网络和贝叶斯模型结合起来，使得神经网络具有更强的概率意义，不仅可以提高在生成任务上的能力，还可以提高模型本身的泛化能力。

i. 神经网络的简要概率表达

1. 神经网络输出的概率意义
2. 神经网络与概率模型相结合：Neuro-CRF、RNN-RBM
3. 记忆神经网络

ii. 后验拟合（Posterior Approximation）与 Variational AE：用于时序建模

iii. Variational RNN

5. 计算图与复杂神经网络

i. 由 Chain Rule 到计算图：采用动态规划算法实现这一转变。

ii. 基于计算图的参数优化

iii. 计算图的模块化：计算图本质上是层次化和模块化的。

iv. 计算图与深度神经网络

6. 计算平台与方法

当前深度学习研究中解决计算问题的方法包括三个主要方向:

- 一是引入专用计算设备，对深度学习中需要的主要运算操作进行优化，以提高基础计算能力;
- 二是采用并行计算，协调多个计算进程进行学习，提高数据的吞吐能力；
- 三是对模型进行剪裁、精减，以节约计算资源。这三个方向并不矛盾，在解决实际问题时经常同时使用。

i. GPU 与 TPU

ii. 并行计算

1. 模型并行和数据并行
2. 同步更新和异步更新
3. DNN 并行方法

iii. 模型压缩

1. Optimal Brain Surgeon(OBS)
2. 对矩阵进行结构化处理
3. 参数量化
4. 参数共享方法

b) 深度学习的应用

i. 语音信号处理

1. 语音识别
2. 说话人识别

ii. 自然语言处理：这些成就大程度上要归功于 Bengio 提出的 Word Embedding 概念。所谓 Embedding ,是指将离散符号映射到连续向量空间，用向量代表符号。Word Embedding 即是用连续向量代表词，所以也称为 “词向量”。词向

量化具有重要意义，因为经过向量化后，词与词之间的距离变得可以度量，这为深度学习在自然语言处理领域的应用铺平了道路。

1. SMT
2. NMT

iii. 计算机视觉

1. GAN 模型的应用

Chapter 5 核方法

- 引入：
  - 无论是线性回归模型还是线性分类模型，当数据的线性关系不显著时，线性模型会出现较大偏差，为解决这一问题，一种方法是对 x 做非线性映射，这时我们来看其非线性映射和预测目标是否有线性关系或者线性可分。但设计一个合理的映射并不容易，特别是当我们对任务本身的知识相对有限的时候
  - 另一种解决问题的办法就是使用神经网络去学习这种映射，以避免人为设计的困难，但神经网络的缺点是：
    - 首先，特征学习需要对原始数据有很明确的向量表达，但在很多实际应用中很难将每个对象表达成数值向量。
    - 第二，特征学习方法对特征空间的大小有严格限制，特征空间维度过高会导致学习困难，但一些复杂数据必须在较高维的特征空间上才能表现出线性。
    - 第三，特征学习，特别是基于复杂函数(如 DNN)的特征学习是一个非凸问题，训练存在很大困难，容易发生过拟合或欠拟合。
  - 基于以上这些原因，核方法提供了另一个选择。核方法是另一种映射函数的生成方法。与特征学习不同，核方法不对映射函数做显示的表示或学习，而是通过数据间的相关性函数对映射函数进行隐式定义。其中的相关性函数称为核函数。
  - 核方法的优势：
    - 这一方法只关注数据间的关系，而不是数据本身，因此特别适合数据样本难以用向量明确表达的任务。
    - 由和函数引导出来的特征空间可能具有非常高的维度，甚至是无限维，因此可满足对复杂数据分布的线性化要求
    - 最后特征空间中的模型是线性的，因此模型训练是一个凸优化问题，可保证得到全局最优解。

1. 从线性回归到核方法：这一解法中，我们并不需要显式地求出模型参数 W，也不需要明确定义特征映射函数，只需知道训练数据之间的关系 K 和测试数据与训练数据的关系 k（x）。该函数称为核函数。相应的方法称为核方法。
2. 核函数的性质：
   a) 再生核希尔伯特空间与 Mercer 定理：
      i. 我们看到一个函数$\mathbf{k}(\mathbf{x},\mathbf{x}')$是合法核函数的充分必要条件是该函数是对称且半正定的;或者，对于任意 N 个{Xn；n=1，……，N}，由 k（·，·）导出的 Gram 矩阵是对称半正定矩阵。这一结论称为 Mercer 定理，发表于 1909 年。在构建核函数时，我们可以通过 Mercer 定理判断一个核函数是否合法，这一点在构造复杂核函数时非常有用。
   b) 核函数的基本性质：通过核函数的这些基本性质，我们可以从简单核函数生成复杂

核函数，这比直接构造复杂核函数要容易的多。

3. 常用核函数
   a) 引入： 由前而讨论可知，核函数的形式决定了映射函数的属性，不同的核函数将数据映射到不同的特征空间。在解决实际问题时，当然希望数据在特征空间的性质越简单越好(如线性可预测性，线性可区分性、高斯分布等)，因此对不同任务需要设计不同的核函数。
   b) 简单核函数：
      i. 线性核
      ii. 多项式核
      iii. 高斯核
      iv. 指数核（拉普拉斯核），与其类似的 Gamma 指数核
   c) 概率核：前而所述的核函数直接计算数据样本点之间的距离，不具有概率意义，对噪音比较敏感。序列样本)。概率核方法通过对数据建立概率模型，再基于该模型计算样本间的距离，从而可以解决上述问题。本质上，概率核是将生成模型和区分性模型结合起来的方法。
      i. 概率核设计思路：
         1. 基于模型距离的核：如 KL 核、Bhattacharyya 核
         2. 基于模型映射的核：Fisher 核
   d) 复杂对象上的核函数：核方法的一个重要价值是对复杂对象的建模。
      i. 集合上的核（距离替换法、标志集向量法）
      ii. 序列上的核：不同序列的长度不同，序列与序列之间可能有复杂的包含关系，这给设计带来困难，一种简单的处理方法是忽略序列中的元素顺序，这时一个序列退化为一个集合，序列上的核函数退化为集合上的核函数。文本处理中常用的词袋模型（Bag of word model），还有忽略全局顺序，只考虑局部顺序的 N-gram 词袋模型
      iii. 图上的核：卷积图核方法、基于最短路径的图核……

4. Kernal PCA

在之前提到过，PCA 是一个线性高斯模型，其基本假设是数据由一个符合正态分布的隐变量通过一个线性映射得到，因此可很好描述高斯分布的数据。然而，在很多实际应用中数据的高斯性并不能保证，这时用 PCA 建模通常会产生较大偏差。如图 5.2 所示，原始数据的样本点呈现明显的非高斯性，这时用传统 PCA 很难找到一个合适的主成分方向。为解决这一问题，我们可以设计一个合理的非线性映射，将原始数据映射到特征空间，使数据在该空间中的映射具有合理的高斯性，即可进行有效的 PCA 建模。

5. 高斯过程

高斯过程是随机过程的一种，一个随机过程可以认为是随机变量的扩展；随机变量是独立变量 x 的分布特性，随机过程是否一个变量集合 X 的分布特性；高斯过程可以认为是传统核方法的随即版本；同时也可以认为是贝叶斯线性回归方法的核函数版本。

6. 支持向量机（SVM）

不论是基于核函数的线性回归还是基于高斯过程的非参数模型，都需要计算 Gram 矩阵 K 及其逆矩阵。当训练集中的数据量较大时，这显然会带来非常高的计算复杂度和内存开销。同时，在预测过程中，测试样本要和训练集中的所有样本做核函数计算，同样带来较高的计算量。一种有效的解决方法是仅保留部分较重要的训练数据来进行预测，而将那些不重要的

数据丢弃。这些保留下来的训练样本称为支持向量，相应的模型称为支持向量机，即 SVM。

a) 线性可分的 SVM：KKT 条件
b) 线性不可分的 SVM
c) V-SVM

7. 相关向量机：基于贝叶斯框架的
8. 总结

本章从线性回归模型出发，推导出了该模型的对偶表达，从而引出了核函数的概念。我们介绍了核函数的性质和构造方法，并讨论了一些常用核函数形式。我们进一步讨论了主成分分析(PCA)的核版本，这一扩展使得 PCA 得以处理非高斯数据。进一步，我们讨论了高斯过程，并推导出基于该过程进行贝叶斯线性回归的表达方式。最后，我们讨论了支持向量机和相关向量机，这是两种典型的稀疏性核方法，该方法在预测时仅考虑最有代表性的训练数据(支持向量或相关向量)，因而可极大减小预测时的计算量。

Chapter 7 无监督学习

- 现实生活中常常会有这样的问题：缺乏足够的先验知识，因此难以人工标注类别或进行人工类别标注的成本太高。很自然地，我们希望计算机能代我们完成这些工作，或至少提供一些帮助。根据类别未知(没有被标记)的训练样本解决模式识别中的各种问题，称之为无监督学习

7. 无监督学习任务

聚类算法和流形学习是两种典型的无监督学习任务。

a) 聚类： 聚类是一种典型的无监督学习任务，其目的是将数据空间划分成若干子区域，使得每个子区域中的数据具有更强的内聚性，不同子区域之间具有明显的分离性。
b) 流形学习：指局部具有欧几里得空间性质的空间。流形（Manifold）是指局部具有欧几里得空间性质的空间。机器学习中的流形通常指数据空间中的一个子空间，数据在该子空间中具有较高的密度，在其余位置的密度相对较低。现实中几乎所有数据都具有典型的流形结构
c) 因子学习：聚类和流形学习本质上是同一种任务，即对数据的内在结构进行学习，这些结构本质上是由数据的生成机制决定的，因此和数据之间具有强烈的因果关系，我们称之为因子（Factor）。不论是聚类还是流形学习，其目的都是对产生数据的因子进行分析，分析产生观察数据的因子应具有的分布特性和结构特性。这一过程称为因子学习（Factor learning）。

8. 聚类方法
   a) 基于划分的聚类方法：k-mean、K-medois
   b) 基于连接的聚类方法：层次聚类、相关聚类、谱聚类、Affinity Propagation
   c) 基于密度的聚类方法：DBSCAN
   d) 基于模型的聚类方法：GMM
9. 流形学习：流形学习中讨论的 “流形” 主要指数据所集中分布的低维空间。流形学习（Manifold Learning）即是从原始高维数据中发现数据中低维结构的学习方法。具体而言，流形学习的目的是构造一个映射函数，将高维数据映射到某个低维空间中，使得原始数据的某些特性在低维空间中得到有效保持。这些特性包括数据分布形态、拓扑结构、

可区分性等。流形学习被大量应用在数据降维和可视化的任务当中。

a) 线性方法：主成分分析（PCA）、多维标度（MDS）
b) 非线形方法：等距映射（ISOMap）、自组织映射（SOM）、谱嵌入、局部线性嵌入算法（LLE）

10. 图模型与无监督学习
    a) 图模型下的聚类任务
    b) 图模型下的流形学习
    c) 图模型下的因子学习
11. 神经模型与无监督学习
    a) 特征学习任务中的因子学习：RBM、AE
    b) 生成任务中的因子学习：VAE、RNN
    c) 分类/回归任务中的因子学习
12. 总结心得：
    a) 无监督学习常常被用于数据预处理。一般而言，这意味着以某种平均-保留的方式压缩数据，比如 PCA 或 SVD；之后，这些数据可被用于深度神经网络或其它监督式学习算法。

Chapter 8 非参数模型

- 参数模型：模型的参数化极大简化了学习过程:它相当于预先定义了一个知识表达形式，并通过对这一形式中的参数进行学习来确定具体模型，因此可以认为是一种将先验知识(模型形式)和经验学习相结合的方法。被参数完全定义的模型称为参数模型。参数模型的一个重要特点是:这一模型的知识表达形式是确定的，因此模型的规模也是确定的，不会随着训练样本的变化而发生改变。
- 非参数模型：参数模型的优势在于其对知识的抽象能力，但这一方法也存在一些问题。例如，当先验知识不足时，对模型形式的设计可能是不合理的;其次，当训练数据较丰富时，我们希望模型的规模可以相应增大，以描述更多细节，而参数模型不具有这种扩展能力;最后，训练数据在不同区域的分布可能是不均衡的，我们希望在训练数据较多的区域有更细节的模型，但参数模型通常是全局的，难以根据数据实际分布情况进行调节。总而言之，在一些学习任务中，我们希望模型的形式和复杂度由训练数据本身来确定，而不是预先设计的固定形式。这种由数据驱动的模型称为非参数模型(Non-Parametric Model)。

13. 简单非参数模型

    a) K 近邻，这一方法假设数据空间的分类是有连续性的，因而一个待考察点周围的点应具有相似的分类。
    b) 决策树(Decision Tree)。这一模型基于一定的准则，将数据自顶向下分裂成相对独立的子集，最终分成的子集多少和数据量及数据的分布情况直接相关。虽然分裂方式、分裂准则、分裂深度等都可能基于某些参数，但模型本身并不能写成这些参数的函数形式，模型规模也会随数据量的增长而变化，因此是一种非参数模型。
    c) 支持向量机（SVM）
    d) 非参数模型的特点：
        i. 对数据的分布不做过强的假设，让数据自己表达自身的分布情况和分类情况;

ii. 模型规模随训练数据的增长而增长;
iii. 很多模型保留全部或部分训练数据用于模型推理。这些特性显然是相关的，正因为不做过强的假设，因此需要保留部分训练数据来进行推理，从而导致模型规模的增长。

14. 高斯过程
    a) 高斯过程：如果参数 w 为高斯分布，则对应的随机函数在任意点集上的取值亦符合高斯分布，且这一分布的性质由协方差函数 k（xi，xj）描述。注意上述分布性质在任意点集上都成立，这事实上定义了一个随机过程，称为高斯过程。
    b) 高斯过程回归：
        i. 贝叶斯线性回归
        ii. 高斯过程回归建模
        iii. 高斯过程回归预测
    c) 高斯过程用于分类任务：高斯过程的本质是提供一簇函数的先验概率，基于这一先验概率，可实现对任何预测任务的非参数贝叶斯学习。前面介绍的高斯过程回归是一个典型的例子，同样的方法也可用于分类任务。
15. 迪利克雷过程

在聚类任务中，我们一般会定义一个贝叶斯生成模型，并确定模型中的聚类数 K，对该模型参数进行优化使其对训练数据的生成概率最大。这一方式显然是参数的。如果我们不是定义某一个聚类模型，而是定义一个在所有可能聚类方式上的先验概率，并基于训练数据得到在每个聚类方式上的后验，则可让数据自动选择出合理的模型复杂度，实现非参数聚类。

a) 高斯混合模型
b) 中国餐馆问题
c) 迪利克雷分布与性质
d) 迪利克雷过程
e) 迪利克雷过程的表示：狄利克雷过程的后验概率依然是一个狄利克雷过程，任何从 DP 中抽取出的采样都是一个离散分布，即使 H 是连续的。
f) 迪利克雷过程的构造
g) 推理方法
h) Hierarchical DP

Chapter 9 演化学习

- 推理学习：基于推理的优化方法效率较高，但受目标函数制约，普适性不强。首先，对多数问题，我们能得到的解空间信息只是局部的，因此只能得到局部最优解;另一方而，一些问题比较复杂，解空间信息很难计算。例如 SGD 需要当前解附近的梯度信息，但一些任务的目标函数是不连续的，这时梯度将无法计算。再如图模型的变分法中，如果模型比较复杂，则求目标函数对某些变量的期望很困难，使得迭代优化无法进行。
- 演化学习：演化学习(Evolutional Learning, EL)提供了另一种学习方法，这种学习方法不是通过推理逐渐逼近优化解，而是随机生成一些可能的解，再对这些解进行优化选择。这种生成-选择方式迭代进行，直到得到满意的解。这种 "Try-and-Error" 的学习方法可称为采样法。和推理法相比，采样法简单直观，事实上是生物界进化的基础方法。因此，EL 经常被一些人工智能研究者认为是实现智能机器的普适方

法。本章将主要讨论两种演化学习方法:遗传算法(GA)和遗传编程(GP)。同时，EL 方法中的两个主要成分:群体学习和随机优化也独立发展成两种常用的优化方法。

16. 基于采样的优化方法

a) 演化学习：(遗传算法、遗传编程)。
b) 群体学习与随机优化：
   i. 群体学习方法：蚁群算法、粒子群优化算法、人工蜂群算法
   ii. 随机优化方法：模拟退火算法、禁忌搜索算法、和声搜索算法

17. 遗传算法（GA）:这一算法模拟生物进化方式，首先随机生成一个种群，种群中每个个体代表一个目标问题的解。通过选择质量较高的个体对种群进行优化，再基于这些优质个体进行交叉繁衍和个体变异，生成新一代种群，这一新种群通常具有更高的质量。上述选择、繁衍过程迭代进行，经过若干代演化后即可得到优化的种群，其中的最优个体即对应目标问题的优化解。
    a) 算法框架：种群初始化、个体选择、种群繁衍。
    b) 编码方式：数据编码、过程编码
    c) 个体选择策略
    d) 繁衍策略：精英策略……
    e) 结束条件
18. 遗传编程

标准 GA 的操作对象是数据，如模型的参数或搜索任务中的路径。如果我们将操作对象换作一组连续操作，并基于类似的演化原则，即可实现对操作过程的学习。这种基于演化原则对操作过程进行学习的方法称为遗传编程

a) 算法基础
   i. 基因编码
   ii. 初始化
   iii. 个体选择：比武决胜法
   iv. 交叉与变异
b) GP 高级话题
   i. 线性编码，即二进制编码
   ii. 图编码
   iii. 概率 GP
c) 其他演化学习算法
   i. 进化编程
   ii. 进化策略
   iii. 基因表达编程
   iv. 差异进化
   v. 神经进化
   vi. 分类器学习系统
d) 群体学习方法
   i. 蚁群优化算法基本原理
      1. 蚂蚁在路径上释放信息素。
      2. 碰到还没走过的路口，就随机挑选一条路走。同时，释放与路径长度有关的信息素。
      3. 信息素浓度与路径长度成反比。后来的蚂蚁再次碰到该路口时，就选择信

息素浓度较高路径。

4. 最优路径上的信息素浓度越来越大。
5. 最终蚁群找到最优寻食路径

ii. 人工蜂群算法：模拟蜂群的工作方式。一个蜂群系统包括三类职能的蜜蜂:采蜜蜂(Employed beef)、驻守蜂(Outlook bee)和巡逻蜂(Scoutbee)

iii. 粒子群算法：(Particle Swarm Optimization, PSO)是一种模拟羊群或鱼群的群体行为的优化方法。这些动物在行动的时候，单个个体的动作方向同时受到个体知识和全局认知的影响。

iv. 捕猎者搜索

v. 萤火虫算法

e) 随机优化方法

i. 模拟退火算法：设计一个温度调节优化的随机函数

ii. 杜鹃搜索

iii. 和声搜索

iv. 禁忌搜索：一种组合搜索算法

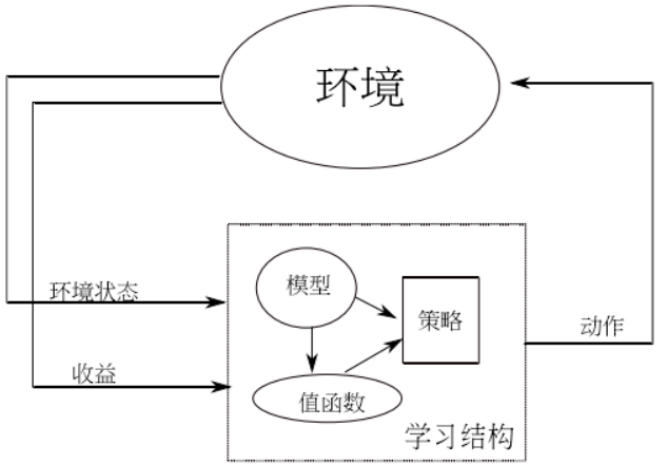
Chapter 10 强化学习

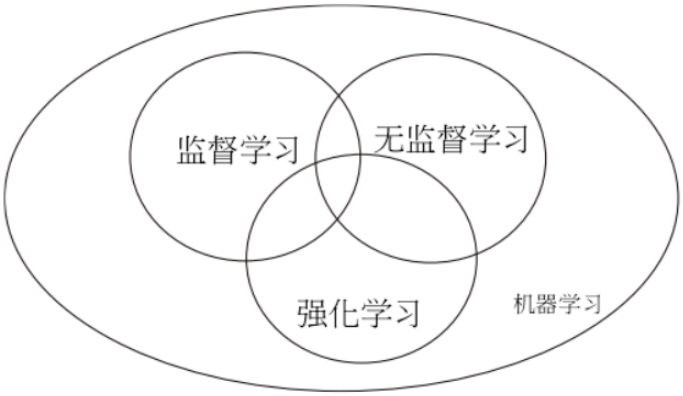
19. 强化学习

a) 什么是强化学习？强化学习(Reinforcement Learning)是一类学习方法的总称，这类学习方法通过和环境进行交互，利用环境给出的反馈信息进行学习。

- 特点：
  - 学习过程需要不断主动尝试;
  - 指导信息是一系列动作完成后的反馈，而非某个动作的具体对错;
  - 学习的目标是一系列动作完成后的总体收益最大化;
  - 学习过程中产生动作有可能对环境产生影响。

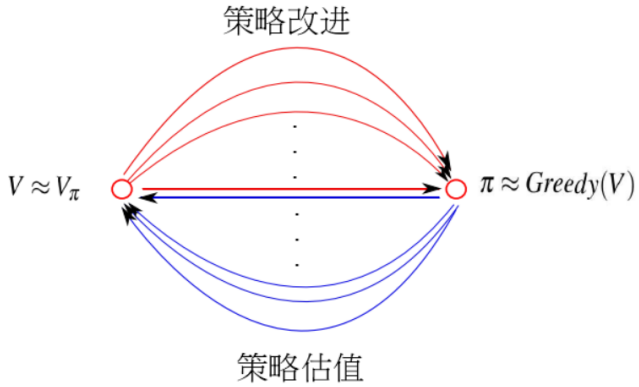
b) 与其他学习方法的区别





c) 强化学习的应用：机器人领域、金融、媒体、医疗。

20. 强化学习的基本要素：状态、动作和受益。



a) 强化学习算法分类：
   i. 规划任务
   ii. 学习任务：模型学习、值函数学习、策略学习

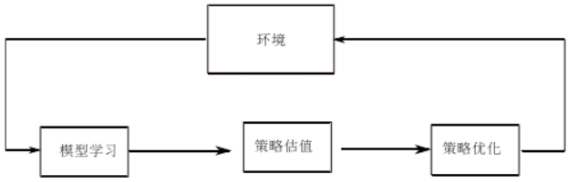
21. 值函数学习:基于模型的规划算法
   a) 马尔可夫决策过程（MDP）：策略估值，动态规划算法

22. 值函数学习:基于采样的蒙特卡洛方法
   a) 学习任务与采样方法：
   b) 蒙特卡洛策略估值（MC）：在多数实际任务中，生成所有可能的交互序列是不太现实的，但通过大量采样，总可以趋近于真实的值函数。这种基于全路径采样来估计值函数的方法称为蒙特卡洛 CMC)方法。
   c) 蒙特卡洛策略优化：off-policy

23. 值函数学习:基于采样的时序差分方法
   a) N-step TD 与 TD(λ)
   b) 三种策略估值算法的特点

	DP	MC	TD
规划任务	是	是	是
学习任务	是	是	是
多轮学习	是	是	是
连续学习	是	否	是
马尔可夫假设	是	否	是
采样	否	是	是
回溯深度	一步	全路径	一步或多步
回溯宽度	全状态	采样	采样
Bootstrapping	是	否	是

24. 模型学习
    a) 模型学习方法



    b) Dyna:混合学习方法
25. 函数近似与策略学习
26. 深度强化学习方法：深度强化学习包括两个部分，一是深度学习模型，利用 DNN(或其它深度网络)来近似值函数或策略函数，第二部分是强化学习方法，利用强化学习的目标函数对 DNN 的参数进行训练。
    a) AlphaGo
    b) Atari 游戏

Chapter 11 优化方法

1.1 函数优化

d) 优化问题分类
    a) 离散优化和连续优化
    b) 无约束优化和带约束优化
    c) 全局优化和局部优化
    d) 凸优化和非凸优化
e) 基础定理
    a) 泰勒定理：一阶必要条件和二阶充要条件

1.2 无约束优化问题

无约束优化问题不考虑约束条件，对 f(x)进行无限制优化。

g) 大部分迭代算法可分为两类：线性搜索和置信域优化
h) 线性搜索
    a) 梯度下降（GD）
    b) 牛顿法
    c) 拟牛顿法
    d) 共轭梯度法

i) 置信域优化
   a) Dogleg 算法：将一阶近似和二阶近似结合起来，得到的一种联合优化方法。
   b) 置信域调整
   c) 对线性搜索和置信域优化两种方法的比较：线性搜索中我们先找到一个搜索方向，再从该搜索方向上确定搜索步长；而置信域优化方法是先确定一个置信域，在该置信域内计算近似函数的最优解

1.3 带约束优化问题

是指在无约束问题的基础上加上一组限制条件

a) 拉格朗日乘子法
b) 对偶问题
c) 线性规划（LP）
   a) 单纯型法：是解决线性优化问题最常用的方法之一
   b) 内点法
d) 二阶规划
   a) 等式约束
   b) 不等约束
e) 一般非线性优化：对于这类问题，一般先将约束问题转化为无约束问题，再用五约束问题的求解方法求局部最优；另一种方法是将一般非线性带约束优化问题分解成一系列局部优化问题，每一个局部优化问题是一个相对简单的 LP 问题或 QP 问题。
   a) 惩罚法
   b) 增广拉格朗日方法：等式约束
   c) 增广拉格朗日方法：不等约束
   d) SQP：顺序二阶规划（Sequential Quadratic Programming）

问题：

1、 激活集算法推导过程？
2、 惩罚法实际应用中广泛吗？具体在什么时候？