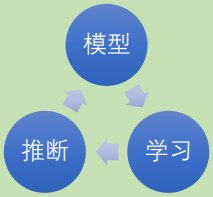
机器学习的理论框架

1. 概述

机器学习理论由紧密联系而又自成体系的三个模块所构成，分别是：模型、学习和推断。其中，**模型**为具体的问题域提供建模工具；**学习**是理论核心，为设定学习目标和学习效果提供理论保证；**推断**关注模型的使用性能和准确性。



本文的目的是对机器学习的总体理论框架做总结，提供原理性地解释。本文不会对具体的模型、算法做讲解（这些内容在相关的教材上都可以找到），只会对各模型、算法间的联系性做说明，以便帮助读者建立一个结构化的知识体系。目标读者需要对机器学习的基本原理、各种模型和算法已经有所了解和掌握。

2. 模型=假设=规律

2.1 模型的本质

形式上，模型就是一个带有参数的函数表达式，我们希望模型能够反映某种关系或规律。早在机器学习出现之前，人们就已经通过对周围世界的观察，归纳出了很多模型。例如，物理学中的万有引力定律$f = \frac{GMm}{r^2}$和质能方程式$\mathrm{E} = mc^2$。这些模型非常稳定（至少在我们所在的宇宙是这样子的）、适用性很强。然而，还有许许多多的规律（一般都局限在各个具体的研究领域），虽然适用性小一些，但却实实在在的影响了我们的生活，值得我们去发掘。发现这些规律的过程依赖于我们的观察。随着传感器技术的发展，人为的观察过程逐渐被机器所取代，观测的结果以电子化数据的方式被存储下来。基于统计学和计算机科学，人们希望从这些数据中发现某些统计规律，并将发现的规律应用于生产实践。由此，便产生了机器学习这一学科。

回到刚才说的模型，一个看起来显而易见但却是本质性的问题是：为什么需要模型？如果我们能够收集到包含足够信息的数据，是否可以直接发现规律，而跳过建模的过程呢？毕竟，机器学习领域已经提出了很多“model-free”的思路。

答案是：不能。

要理解这个问题的本质，首先要理解：什么是模型？模型是一组假设，这组假设一定是基于某个具体的问题所提出的。也就是说，当我们去完成一个机器学习的任务时，一定是为了解决某一个问题，而这个问题本身会对数据的类型、数据的收集过程、数据的预处理等作出限制，这些限制不能够直接通过观察数据本身得到（可以认为是一种 meta-information），所以就需要通过人为的假设来反映这些元信息。至于一些所谓的 model-free 的算法，如：聚类、强化学习的 model-free 方法，只是由于所处理的问题域本身对数据的限制较少，所以在某些方面不需要很强的假设，但仍然是有一些“隐假设”和“弱假设”的。所以，总体上来说，我们希望模型提供一个总体框架，然后用数据对框架进行微调，最终得到一个比较准确的可用的模型。用机器学习的术语来说，就是：先提供一个假设空间，然后在假设空间中搜索最合适的假设。

2.2 为不确定性建模

大多数情况下，我们希望模型具有一定的推断和预测能力，例如，给定一个用户信息与信用评级之间的关系模型：

$$y = f_\theta(X), \qquad X = (x_1, x_2 \dots x_n)$$

其中：y表示信用评级，取值为 1~k，代表 k 个信用等级，X表示由用户各项信息所组成的向量，θ是模型的未定参数。

现实中，我们不但希望知道给定用户的信用评级，而且希望知道给出这个评级时模型的确定性程度（也可称为：可信任度）。概率论为不确定性的建模提供了极好的工具，如下式所示：

$$p(y = f_\theta(X)) \Leftrightarrow p_\theta(\mathrm{y|X})$$

其中，p 表示概率分布函数，$p_\theta(\mathrm{y|X})$表示条件概率。

关于不确定性的完整建模，涉及到很多其他的理论和概念，包括信息论， 随机过程，混沌理论，贝叶斯定理，模型的固有局限性等，在此不做展开。

2.3 模型与数据结构

实际应用中的模型都比较复杂，往往包含成千上万的变量，即 X 的维数 n 很大，所需要确定的参数也同样是高维的，并且各个变量之间也有关联关系，因此我们需要利用先验知识为多个变量间的关系建模。高维空间相对于低位空间的处理来说比较困难，所以一个重要的直觉是：对高维空间分解成多个子空间，在各个子空间中处理得到解后，再合成高维空间中的解。条件概率的乘法公式可以实现这样的分解：

$$\mathrm{p}(x_1, x_2) = \mathrm{p}(x_1) \cdot \mathrm{p}(x_2|x_1)$$

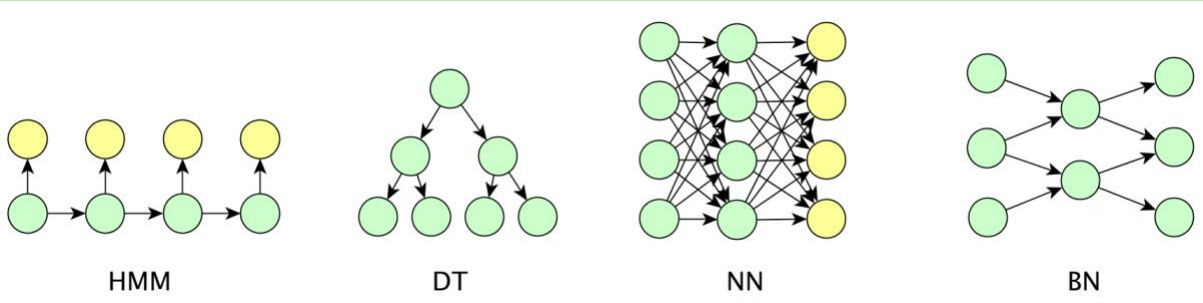
如果两个变量是相互独立的，那么可以进一步简化成：

$$\mathrm{p}(x_1, x_2) = \mathrm{p}(x_1) \cdot \mathrm{p}(x_2)$$

前面提到，机器学习是统计学和计算机科学的联姻，而计算机科学提供的最强大的两个工具是：数据结构和算法。在众多数据结构中，图是最通用的一种数据结构，如果我们能够把上述模型的数学公式表示成图的形式，不但可以提供直观的可视化展示，而且可以借助计算机科学中现有的算法来完成各类计算过程。

在上述的概率模型中，如果将变量用图的节点来表示，变量间的依赖关系用有向边来表示，那么就能够实现模型与图的对应。概率图模型就是这样一种思路。通过将模型与各种数据结构（图、树、森林、链表等）的结合，我们可以得到以下经典模型：贝叶斯网（BN），马尔科夫随机场（MRF），条件随机场（CRF），隐马尔科夫模型（HMM），决策树（DT），随机森林，神经网络（NN）等。而其他模型，如支持向量机（SVM），线性回归模型，逻辑斯底分类模型都是上述模型的进一步简化。

以下给出一些模型的简单图示：



2.4 小结

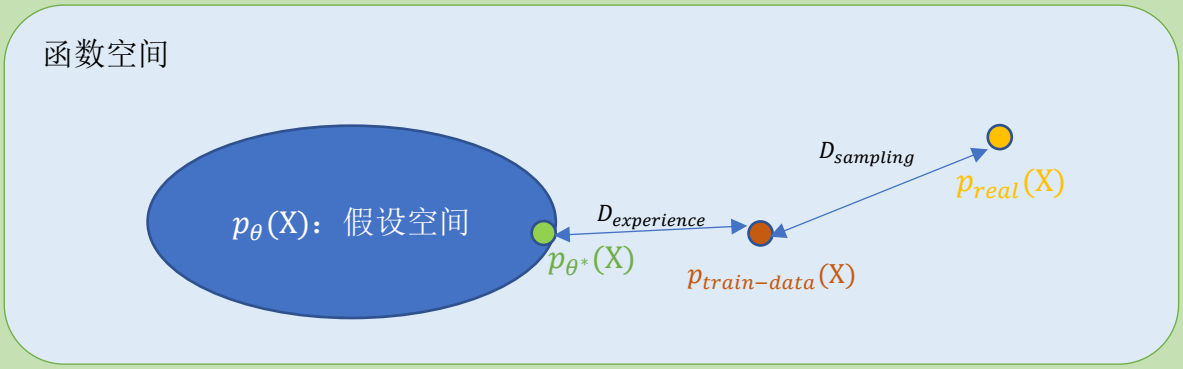
本节我们介绍了机器学习中模型的一个通用理论框架。接下来，我们需要理解如何通过数据来对模型做微调，即：确定参数θ。这个过程也称为“学习”或“训练”。

3 学习=训练=优化=拟合

3.1 “学习”的本质

“学习”是一个很能让人遐想的名称，它常常和智能联系在一起。认知科学试图直接从自然智能的原理出发，理解学习的本质，但这不是机器学习领域所采用的研究方法。机器的可学习性理论构建在统计学中的大数定律之上，具有严格的数学基础。

正如本节标题所表明的，在机器学习中，学习的本质是函数拟合，即：从可能的假设空间中，根据数据所提供的信息，找到一个和真实的函数最接近或相等的假设。以概率分布函数为例，$p_\theta(X)$是带有参数的概率分布，代表了模型所对应的假设空间，$p_{train-data}(X)$是由采样得到的数据的概率分布，$p_{real}(X)$是真实的概率分布。概率分布之间的相似度的度量是散度（divergence），其通用的定义为 f-divergence：$D_f(P \parallel Q) \equiv \int f(\frac{dP}{dQ})\,dQ$。（其中，比较常用的是 KL 散度，之后在讲最大似然估计的时候还会提到。）



如上图所示，我们的目标是求取一个θ^*，满足：

$$\theta^* = \underset{\theta}{\operatorname{argmin}}(D_{experience} + D_{sampling})。$$

$D_{sampling}$代表$p_{train-data}(X)$与$p_{real}(X)$之间的差异，根据大数定律，当所获取的样本数据满足独立同分布（简记为：i.i.d.），并且样本量足够大时，$D_{sampling} \approx 0$。这一定律一方面给我们提供了可学习性的理论保证，而另一方面也说明了数据收集过程的重要性：不管是直接获取的还是间接获取的数据，我们都要验证其是否满足 i.i.d.。

由于真实世界的分布$p_{real}(X)$是无法知道的，当我们确认了数据的有效性之后，我们的目标就变成了最小化$D_{experience}$，即：

$$\theta^* = \underset{\theta}{\operatorname{argmin}}(D_{experience})。$$

其中，$D_{experience}$代表$p_{\theta^*}(X)$与$p_{train-data}(X)$之间的差异。只要我们给出$D_{experience}$的数学表达式，那么上述问题就可以转化为一个传统的最优化问题。（最优化理论是一个独立的学科分支，这里不做展开，读者可以在任何一本相关的教科书中获取全面的知识。）一般来说，运用梯度下降法，就可以求解这个最优化问题。在某些特殊的问题上，也可能会用到 EM，二次规划和其他一些启发式的优化算法。

3.2 差异性的度量准则

在数学中，测度论为属性的度量提供了基础理论保证，在此之上得以建立对差异性的形式化定义 。根据处理对象的不同，差异性会具体化成不同的概念。
例如：

1. 在欧式空间，我们使用**欧式距离**来度量空间中的两个点的位置的差异。
2. 在线性空间，我们用**范数，余弦距离**来度量两个向量之间的差异。
3. 在概率分布函数空间中，我们用**散度**来度量两个分布之间的差异。
4. 在信息论中，我们用**交叉熵**来度量两个信息源所包含信息量的差异。
5. 在决策理论中，我们用**损失函数，风险** 来度量两个不同的决策的有效性之间的差异。
6. 在统计学中，我们用**似然性**来度量模型与数据之间的切合程度，本质上也是一种差异性的度量。

理解了上述各种差异性的度量方法，我们就不难理解为什么目标函数$D_{experience}$在不同的假设条件下会有那么多不同的名称和组成部分：f-散度，损失函数， 风险，似然，交叉熵。

3.3 损失函数与 f-散度的对偶性
损失函数和 f-散度在决定学习目标时，实际上说的是同一件事情。也就是说，当我们从损失的角度，定义了优化目标（如：最小化均方误差），其实本质上也是在最小化某个 f-散度（如：K-L 散度）。同样的，如果我们定义的学习目标是最小化某个 f-散度，其实质也是在最小化某个损失函数。具体从哪个角度来定义学习目标，要看情况而定。一般来说，损失函数比较直观，可以优先考虑。f-散度比较抽象，如果对数据的分布有可靠的了解的话，可以直接用散度来定义学习目标。

例如，如果我们定义损失函数为负对数似然，那么经过数学推导，我们可以证明:最小化负对数似然的损失与最小化 K-L 散度的等价性。证明如下：
给定：

$$loss(\theta) = \sum_{i=1}^{N} -\log p_\theta(x_i)$$

$p_{train-data}(\mathrm{X})$简记为$p_t(\mathrm{X})$，$loss(\theta)$称为经验风险。$-\log p_\theta(x_i)$是损失函数。

$$\begin{aligned}
\theta^* &= \underset{\theta}{\operatorname{argmin}}\, loss(\theta) = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{N} -\log p_\theta(x_i) \\
&= \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{N} -p_t(x_i)\log p_\theta(x_i) = \underset{\theta}{\operatorname{argmin}} - E_{x\sim p_t}(\log p_\theta(x)) \\
&= \underset{\theta}{\operatorname{argmin}} \int -p_t(x)\log p_\theta(x)dx \\
&= \underset{\theta}{\operatorname{argmin}} \left(\int -p_t(x)\log p_\theta(x)dx + \int p_\theta(x)\log p_\theta(x)dx \right) \\
&= \underset{\theta}{\operatorname{argmin}} \left(\int -p_t(x)\log p_\theta(x)dx + \int p_t(x)\log p_t(x)dx \right) \\
&= \underset{\theta}{\operatorname{argmin}} \int p_t(x)\,(\log \frac{p_t(x)}{p_\theta(x)})dx = \underset{\theta}{\operatorname{argmin}} E_{x\sim p_t}\left(\log \frac{p_t(x)}{p_\theta(x)}\right) \\
&= \underset{\theta}{\operatorname{argmin}}\, D_{kl}(p_t \parallel p_\theta)
\end{aligned}$$

同时，根据交叉熵的定义，可知：

$$\mathrm{H}(p_t, p_\theta) = E_{x\sim p_t} -\log p_\theta(x) = E_{x\sim p_t}\left(\log\frac{p_t(x)}{p_\theta(x)} - logp_t(x)\right)$$

$$= E_{x\sim p_t}\left(\log\frac{p_t(x)}{p_\theta(x)}\right) + E_{x\sim p_t}\left(-logp_t(x)\right) = D_{kl}(p_t \parallel p_\theta) + H(p_t)$$

因为当前数据的概率分布p_t是给定的，所以$H(p_t)$与θ无关，从而可得：

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\, loss(\theta) = \underset{\theta}{\operatorname{argmin}}\, D_{kl}(p_t \parallel p_\theta) = \underset{\theta}{\operatorname{argmin}}\, \mathrm{H}(p_t, p_\theta)$$

同时，我们也可以看到：

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\, loss(\theta) = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{N} -\log p_\theta(x_i) = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} \log p_\theta(x_i)$$

$$= \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{N} \log p_\theta(x_i) = Maximum\ liklihood$$

由此得出如下洞察：

（1） 最大似然估计是经验风险在损失函数取负对数似然情况下的一个特例；

（2） 以负对数似然作为损失的经验风险等价于经验分布和模型分布之间的交叉熵。

更进一步，如果给定模型分布为高斯分布，，可以证明：负对数似然的损失函数与均方误差是等价的（从略）。

3.4 结构风险与后验概率的等价性

上一节中，我们假设如果数据量“足够大”时，只需要通过最小化经验损失就能够得到较好的模型。然而在实际中，由于各种技术条件的限制，数据量往往不能够达到“足够大”。尤其是当模型比较复杂时，其对应的假设空间相对于数据量来说太大，这样就会给优化算法寻找良好假设的效率和准确性造成影响。要么学习的速度很慢，要么即使找到了满足数据的假设，但离正确的假设仍然差距很大（称为：过拟合）。所以，我们需要更多的限制条件来对参数作出限制，指导优化算法更快更好的找到正确的模型。我们称这一过程为：正则化。由此，我们可以定义结构风险，作为学习的目标：

$$Loss(\theta) = \sum_{i=1}^{N} -\log p_\theta(x_i) + \lambda J(p_\theta)$$

$J(p_\theta)$表示模型的复杂度，称为正则项，对其意义的解释也颇具哲学意味：如果我们使经验风险非常小时，说明模型对数据的拟合很好，这时模型一般都比较复杂。但是我们并不希望拟合太好，因为模型可能学到了一些只有当前的采样数据所具有的特征，并不具备通用性。所以过多的复杂性对最后模型的性能来说，也是一种损失，需要被考虑到总的风险（结构风险）中。至于复杂度与损失之间的具体关系以及应该要施加多大的惩罚，需要根据具体的问题来设计λ和J。而对具体问题的理解，代表了我们的先验知识。对于概率模型来说，我们也希望能够加入这些先验知识。贝叶斯定理为注入先验知识提供了理论框架。

根据贝叶斯学习理论，学习过程就是用获得的数据来推测模型的参数，其形式化定义如下：

$$p(\theta|\mathrm{X}) = \frac{p(X|\theta)p(\theta)}{p(X)} = \alpha\, p(X|\theta)p(\theta)$$

$p(\theta|\mathrm{X})$称为后验概率，$p(X|\theta)$是似然函数，$p(\theta)$是先验分布。

当X给定时，$p(X)$与θ无关，所以用一正数α表示。直观上可知，给定数据下的最大后验概率所对应的θ为要找的参数 。即：

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\, p(\theta|\mathrm{X})$$

$$= \underset{\theta}{\operatorname{argmax}}\, \alpha\, p(X|\theta)p(\theta) = \underset{\theta}{\operatorname{argmax}}\; p(X|\theta)p(\theta) = \underset{\theta}{\operatorname{argmax}}\, (\log p(X|\theta)$$

$$+ \log p(\theta)) = \underset{\theta}{\operatorname{argmax}}\, (\sum_{i=1}^{N} - \log p_\theta(x_i) + \log p_\theta)$$

可以得到，此时的损失函数为：

$$\mathrm{Loss}_{Bayes}(\theta) = \sum_{i=1}^{N} - \log p_\theta(x_i) + \log p_\theta$$

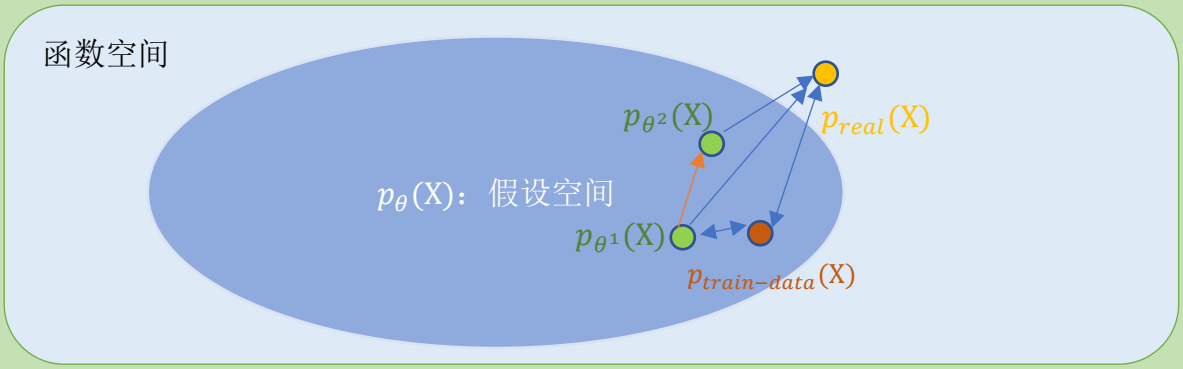
一方面，$p(\theta)$是θ的先验分布，另一方面，从结构风险的角度，我们得到：

$$\lambda J(p_\theta) = \log p_\theta$$

即证明了：最大后验概率与最小化结构风险之间的等价性。

假设p_θ~N(0,1)，可得：$\lambda J(p_\theta) = -log\sqrt{2\pi} - \frac{1}{2}\theta^2 \triangleq \| \theta \|_2$

可以看出，如果参数θ的先验分布为高斯分布，这时对应的是L^2正则化。



如上图所示，通过引入正则项（或先验分布），我们可以在不完全拟合采样数据的情况下，得出更优的结果，即：p_{θ^2}(X)相较于p_{θ^1}(X)更接近于真实分布。当然，前提条件是，我们注入的先验是合理的。

4. 推断=预测

当模型训练完成，即参数确定后，就可以用模型做预测或推断的任务。正如第 1 节所示，模型的本质是函数关系，那么推断任务的本质就是函数调用：给定一个输入变量，经过计算后，就会得到输出。

一般来说，实践中的函数都比较复杂，会执行大量高维向量和矩阵之间的乘法操作。矩阵乘法的高效算法属于数值计算领域的问题，当前的主要方式是通过大规模的并行计算来高效实现，因此，GPU 在其中得到了广泛的应用。

能够执行矩阵乘法的前提条件是：求解的函数表达式有解析式，这个问题在对概率密度函数（以下简称为 PDF: Probability Density Function）的积分计算中变得很棘手。我

们知道，对于概率模型的推断往往都会涉及到对 PDF 的积分，但由于 PDF 非常复杂，对它做积分，一般都是没有闭式解的。因此，我们需要设计专门的算法来处理这个问题。当前，主要有以下几类方法和其变种：

（1） 变量消除

（2） 蒙特卡洛采样

（3） 变分法

这些算法按精确性来分，主要分为精确推断和近似推断。精确推断利用了计算过程中发现的冗余步骤，通过消除这些冗余步骤来提供算法效率。底层的通用算法基础是动态规划。近似推断中，主要有两类算法：蒙特卡洛采样和变分法。

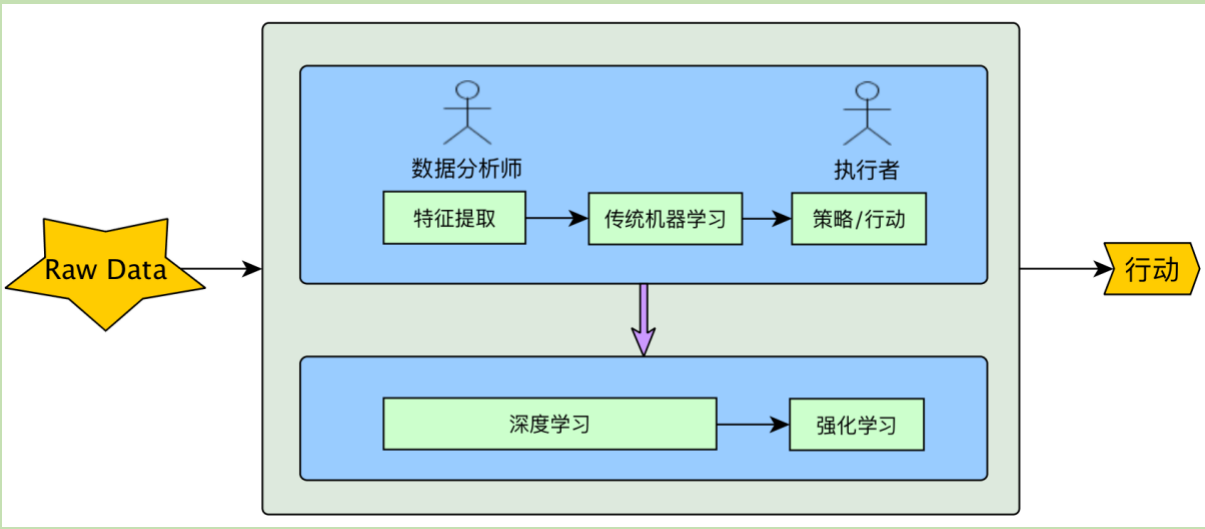
蒙特卡洛采样是一个神奇的算法，当我第一次听说这个算法的时候，简直是叹服。其理论基础也是大数定律，用样本均值来近似估计总体均值。关键技术是确保采样过程具有足够的随机性。随着计算机技术的发展，采样效率的提高，该算法在概率积分求解中起到非常大的作用。目前该算法的理论已经比较成熟，最常使用的是 Importance 采样，MCMC 和 Gibbs 采样。

变分法的数学基础是泛函，相对比较抽象。但直觉上还是比较容易理解的：既然现在的 PDF 很复杂，那么能否找一组简单的 PDF，用这组简单的 PDF 的线性组合来近似原来的 PDF，而这组简单的 PDF 对于要处理的积分是有解析式或者比较容易求解。而变分法就提供了找出这组简单 PDF 的工具。

5. 机器学习的分支领域

在 1~4 节中，我们已经完成了对机器学习的基础理论框架的全部论述。然而，随着机器学习的进一步发展，产生了两个有重要影响的子领域：深度学习和强化学习，虽然它们仍然是构建在统计学习理论之上的，但是在其内部已经形成了自己的理论架构，值得我们单独列出来进行阐述。

其中，深度学习试图解决手工特征提取的问题，而强化学习试图解决决策推理的问题。随着上述两个问题被逐步解决，就有希望消除人的参与，让机器自动适应环境并作出决策，最终使得真正的 AI 成为可能。



5.1 深度学习

5.1.1 感知计算问题

计算机技术完全构建在数学和逻辑之上，从它诞生的那刻起，就非常擅长于求解定义明确、输入输出可量化的形式化问题。而形式化问题（如：公式推导，矩阵运算）对大多数人来说是一件非常困难的事情，以至于人们曾一度乐观的认为，21 世纪之前就可以通过计算机实现真正的人工智能。直到研究人员在计算机视觉、语音识别、自然语言理解领域遭遇一次次的挫败，人们才渐渐意识到智能问题远比我们想象的复杂。

我们每天都在与他人进行互动，当我们走进一个会议的房间，立马就能够感受到氛围的变化，别人细微的一个表情或身体姿态的变化，我们立刻就能捕捉到。人类实在是太擅长这一技能了，以至于我们把它看做是想当然的简单的本能反应，而殊不知这种简单的本能反应是大自然给了我们数十亿年的时间，经过无数次的优胜劣汰才进化出来的！而计算机只有几十年的发展历程，对这一类问题（我们称之为：感知计算问题）是没有经验准备的。

5.1.2 感知计算问题的困难

混沌理论表明了现实环境内在的随机性和复杂性：无数变量在各个层面（原子，分子，细胞，生物，团体，经济体）通过各种关系（引力，化学键，生物电，血液循环，电子通信，货币）相互影响，构成一个嵌套的、多层次的动态系统。这样一个动态系统包含的信息量近乎无穷。而人类自身以及目前所发明的所有传感器，由于受到精度、存储、计算能力和成本的限制，只能收集到其中很小一部分的数据。更重要的是，所有和人类活动相关的现象，都受到人类的文化、观念、情绪的影响，而这一类变量是抽象的，压根就无法进行直接测量。退一步讲，即使我们能够获取所有我们想要的数据，如何分析这些数据呢？毕竟，计算机看到的都是 01 比特，如何才能将这些比特映射成合理的语义信息？

5.1.3 智能与表征学习

现实看起来如此艰难，但毕竟人类大脑还是克服了这个看似不可能的困难，成功的认识了这个世界。认知科学研究的核心是理解大脑对知识的表征形式，并认为智能的首要特征是：有一个良好的表征系统。这一洞见和我们对数据结构的认识是一致的：只有在良好定义的数据结构之上，才能开发出高效的算法。

通过对脑科学和人类视觉皮层的研究，我们得到了两个启发：

（1） 复杂的智能行为是由相对简单的神经元通过网络结构组合而产生的

（2） 不同的神经元负责感知不同的视觉元素，并通过层层传递的方式来构成更复杂、抽象的信息

深度神经网络就是由此启发而产生的，而其重点就是表征学习。其设计的核心理念有两条：

（1） 组合泛化

（2） 层次化

之后的所有架构设计都是围绕这两个理念，并结合具体的问题域而展开的。

5.1.4 架构设计

深度学习的核心组件是深度神经网络，“深”的意义在于提供了近乎无限的函数拟合能力。但是为了驾驭好这种能力，需要通过网络架构的设计，植入先验，引导神经网络拟合出我们想要的函数关系。

网络架构的设计属于模型设计的范畴，针对不同的问题领域，都需要有相适应的架构与之匹配。目前主流的架构主要包括以下 3 个系列：CNNs，RNNs，GANs。这些网络主要针对以下几类组件进行设计和选择：

（1） 神经元

神经元的设计就是激活函数的设计，目前主要的激活函数有：ReLU, ELU, LReLU, PReLU, CReLU。不同的激活函数都有各自的优缺点，但总的来说差别不大。

（2） 层内结构

主要包括：向量，树，图

（3） 网络宽度

和输入数据的维度（如：图像的大小，音频的长度，文字的含义复杂度）以及输出结果的维度（如：分类数）有关

（4） 网络层数

和数据所包含的概念、特征的抽象层次数成正比。

（5） 层之间的连接关系和顺序

跳跃连接，反馈连接，全连接，卷积连接（普通卷积，空洞卷积，3D 卷积等），池化连接，拼接与分拆（一对多，多对一）

（6） 多个网络的组合

在无监督表示学习中常被使用，如：自动编码器，GAN。

可以看出，对不同的架构设计选择进行排列组合后，就可以得到不同的网络。这些排列组合构成了一个巨大的超参数空间。如何在这个超参数空间中找出最优解，依赖于经验和领域知识。当然，也有人试图通过强化学习，来自动搜索出最优解，如：Google 的 NASNet.

5.1.5 学习范式

除了具体网络架构上的设计考量外，在学习范式上，有如下设计思路：

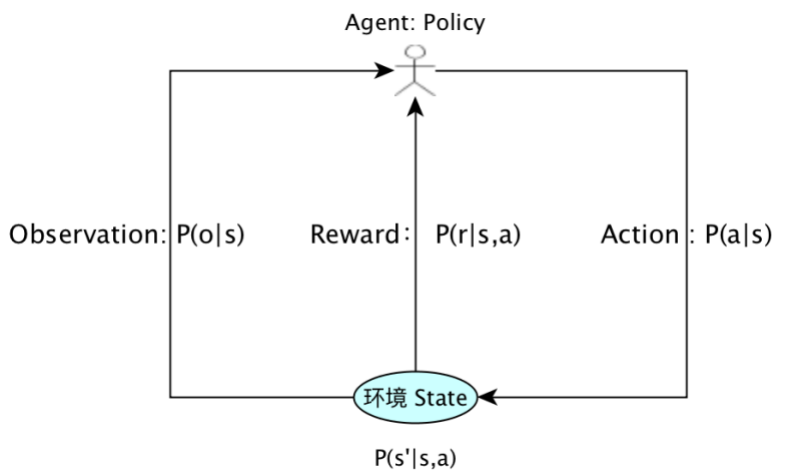
- 预训练
- 对抗训练
- 端到端学习
- 迁移学习
- 多任务学习
- 多模态学习
- 课程学习
- 自监督学习
- 主动学习

这些思路很多借鉴了人类的学习过程，主要的目的是：希望通过更少的数据，更少的人工标注，更高效的训练，学习到更好的表示，以便具有更好的泛化能力。这些范式不是深度学习所独有的，但当前主要的应用都是基于深度学习的。

5.2 强化学习

5.2.1 从观察者变为参与者

大部分的机器学习都是对外部世界的建模，其隐含假设是：观察者是独立于外部环境的，并且不会对当前要观察的环境产生影响。而现实中的智能体除了要理解外部环境的规律外，还需要与环境互动，在互动的过程中理解环境的动态性，并为制定决策提供经验和理论基础。强化学习就是为这类问题建模的工具。其基本建模框架可以用下图来表示：



从上图可以看出该模型的三个特点：

（1） 环境的动态性与交互性：P(s'|s,a)，环境的未来状态同时受到当前状态和 Agent 动作的影响。

（2） 不确定性：所有互动的结果都是不确定的，包括：对环境状态的观察结果 P(o|s)，采取动作后的奖励结果 P(r|s,a)和环境变化结果 P(s'|s,a)。

（3） 时序性：整个互动过程是循环推进的，形成一个不断重复的（State，Observation，Action，Reward）序列。

5.2.2 理论基础

除了统计理论外，强化学习的主要理论基础是决策理论。决策理论本身是一个庞大的学科分支，按照单人/多人，一次性/多次性，单因素/多因素这三个维度可以分成不同的子领域，举例如下：

（1） 简单决策

（2） 多人博弈

（3） 序列式决策

强化学习解决的就是序列式决策问题，既有单人的，也有多人博弈的情况。序列式决策的基本模型是 MDP（马尔科夫决策过程），如果模型的参数已知，则可以通过动态规划算法得出最优决策策略，如果模型参数未知，则 Agent 需要通过和环境互动，收集数据，从数据中学习到环境的一些性质，来帮助制定最优决策。其中，收集的数据是 State，Action，Reward 组成的序列，学习的方法可以大致分成三大类：

（1） Q-Learning, Sarsa：用收集到数据来求解某些统计量（如：不同状态下的期望回报），进而求解 Bellman 方程，以获得最优的值函数或动作-值函数，进而求出策略函数。

（2） Policy Gradient：先定义一个含参的策略函数，然后根据策略函数，设计一个目标函数（一般是某个总的价值量），然后利用梯度下降法，求出参数，进而得到策略函数。

（3） Model-Based: 先根据数据，求出 MDP 模型，然后根据模型，运用动态规划方法，直接设计出策略。

整个学习过程是迭代完成的，分成：Prediction 和 Control 两个阶段。在具体实操上，会有更多的设计权衡，主要归纳如下：

（1） On-Policy VS Off-Policy
（2） Monte Carlo VS TD
（3） Exploration VS Exploitation
（4） Experience Replay

5.2.3 与深度学习的结合

同其他传统的机器学习算法一样，强化学习在早期遇到的一个很重要的困难是：状态空间太大而没有一个很好的表示方法。例如：如果要通过强化学习让机器人学会摆放物品，机器人需要看到各种物品的图像。假设图像的大小是 400*400，那么我们所面临的状态空间的大小至少是：$256^{400*400}$（这还不包括其他机器人所需要感知到的状态，如当前的位置，电量等）。所以，我们需要通过深度学习学到一个关于环境状态的紧凑的表示，用这样一个表示作为策略函数的输入参数。除了节约了保存状态的内存和计算资源外，这样做的另一个好处是：为强化学习提供更好的环境适应能力（泛化），因为在现实环境中，Agent 几乎不可能遇到两次一模一样的情况，它需要能对“似曾相识”的情境做出有效的决策。

5.2.4 与监督学习的异同

强化学习和监督学习都需要收集很多数据来进行训练，但强化学习的数据是在 Agent 与环境的互动过程中，一边学习一边收集的。强化学习不依赖于数据标签，但依赖于奖励信号。可以将奖励信号看作是“延时的标签”，也正因为这样的标签是延时的，所以需要一个回溯迭代的过程来求得最优解，这是强化学习所特有的。

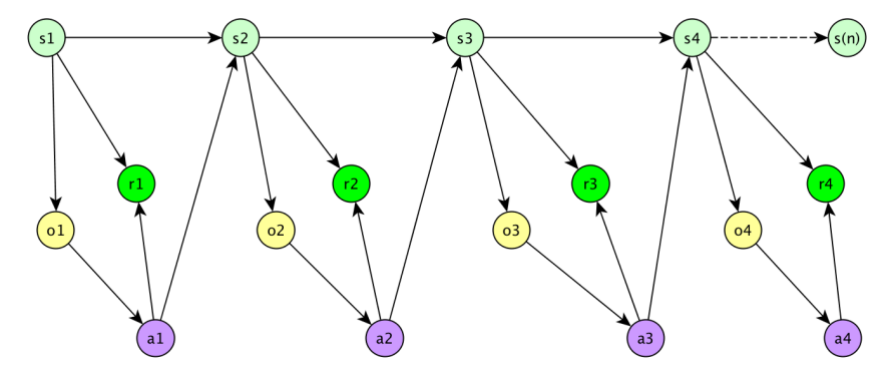
因为需要与环境互动来获得数据，强化学习往往比监督学习花费更多的时间来做训练， 所以有时候需要通过 off-policy 的学习方式来部分解决这个问题。

还有一个很重要的区别是：强化学习所收集到的数据不符合独立同分布的要求，所以，不能够直接用于回归。这里有两种处理方式：

（1） 在每个给定的 policy 训练完后，丢弃到之前的数据，然后重新采集。这样做比较直观，但同时也带来了 Sample Efficiency 的问题。
（2） 将所有采集到的数据建立一个 Buffer，然后采用随机抽样的方式来选出样本数据，这样抽取的样本是基本符合 i.i.d. 的，就可以用作回归学习了。

5.2.5 与动态贝叶斯网的关系

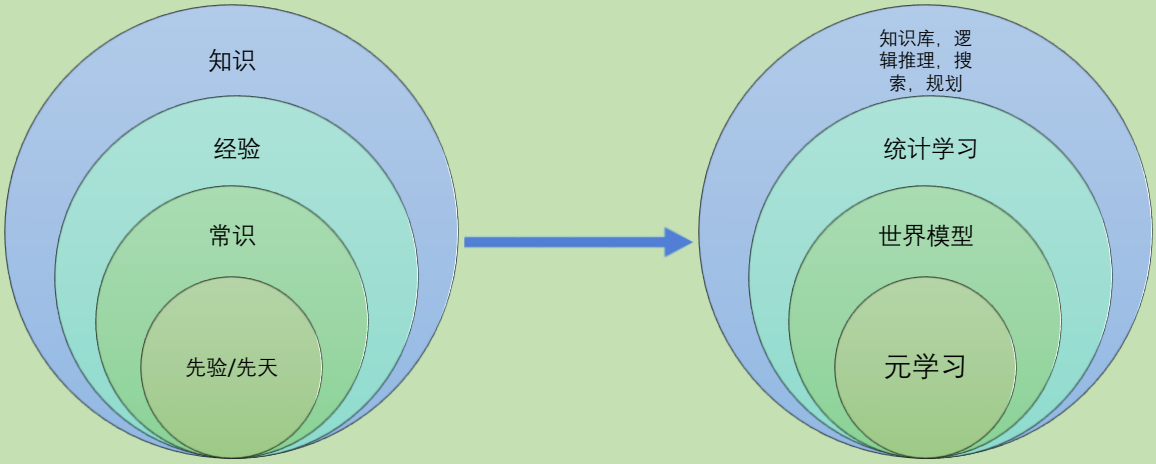
如果将 MDP 用概率图模型的方式表示出来，可以得到下图：



可以看出，MDP 本质上是就是一个动态贝叶斯网，或者进一步，也可以认为是一个决策网络。如果我们能够收集到足够的数据的话，也可以通过动态贝叶斯网的各种算法来求解出这个模型，然后用求出的模型做概率推断，来得到最优策略。但这就要求我们穷举出每一种 Policy，并针对每一种 Policy，收集足够多的数据。现实中，这样做的效率很低，成本是不可接受的。这里，问题的本质上是建模的有效性：MDP 针对序列式决策问题内建了很多假设条件，显著减小了不必要的假设空间的大小，所以相对于动态贝叶斯网，是一种更加高效、准确的模型。

6. 总结与展望

当我去理解机器学习乃至 AI 领域的发展历史和现状的时候，我一直会问自己这样的问题：我们现在在哪？我们是否走在正确的理论道路上？如果我们的方向是对的，那么就应该有一个整体性的框架来框定正确的方向，设定边界，为后续的研究提供理论保证。下图是我对当前机器学习和 AI 发展的看法：



如图所示，我认为过去 60 年人类对智能的理解就像是：剥洋葱，从外到内，不断取得突破，不断加深着我们对智能的理解。先是对知识本身，通过对各领域知识的分析、归纳，整体，然后将其编码进计算机系统，形成了各类知识库，逻辑推理程序，和各种高效的算法。当手工的知识编码遇到瓶颈时，统计学习方法登上舞台。然而，基于统计的学习算法（包括深度学习、强化学习）本质上还是函数拟合，它缺乏对这个世界的基本认识和常识。因此，我们需要一个关于世界基本运行规律的模型。而要能够建立这样一个世界模型，需要我们对智能的本质有更深刻的理解，我们需要理解人是如何学习的，人是如何**学会学习**的。人从生下来的时候，就开始观察这个世界，并不断学习，掌握知识。这种能够学习的能力本身也许是智能最本质的特征。我们把它称为：元学习。它是智能得以产生和发展的**内核**。

因此，要理解智能的本质，就需要理解人脑的先天机制。这是脑科学和认知科学需要研究的主题，也是智能科学赖于继续发展的前提，同时也是哲学所探讨的一个重要主题（“先验”一词最早出现在康德的《纯粹理性批判》，当时提出的主要目的是用于调和经验主义和理性主义的矛盾）。在心理学领域，精神分析学派的荣格提出了“集体无意识”，在我看来，也是对先验的另一种阐述。

这样一来，仿佛所有和人相关的学科都指向了同一个内核：人脑**先天**的运作原理。而要理解人脑的运作原理，需要从进化的角度，把智能看做是生命适应环境的产物。也许，一切都是自然而然产生的。**适应环境的过程就是学习的过程**，只要给足时间（50亿年？），任何智能体都会随着环境不断进化和升级。如果一定要给出一个关于学习的通用理论框架，我认为应该是这样的：

（1） 模型：存储+随机访问
（2） 学习目标：生存下去
（3） 学习算法：信息反馈+误差纠正