神经网络与深度学习

作者：邱锡鹏
链接：https://nndl.github.io/

全书摘要

第 1 章是绪论，概要介绍人工智能、机器学习和深度学习，使读者全面了解相关知识．第 2、3 章介绍机器学习的基础知识．第 4～6 章分别讲述三种主要的神经网络模型：前馈神经网络、卷积神经网络和循环神经网络．第 7 章介绍神经网络的优化与正则化方法．第 8 章介绍神经网络中的注意力机制和外部记忆．第 9 章简要介绍一些无监督学习方法．第 10 章介绍一些模型独立的机器学习方法：集成学习、自训练和协同训练、多任务学习、迁移学习、终身学习、元学习等，这些都是目前深度学习的难点和热点问题．第 11 章介绍概率图模型的基本概念，为后面的章节进行铺垫．第 12 章介绍两种早期的深度学习模型：玻尔兹曼机和深度信念网络．第 13 章介绍最近两年发展十分迅速的深度生成模型：变分自编码器和生成对抗网络．第 14 章介绍深度强化学习的知识。第 15 章介绍应用十分广泛的序列生成模型

基础理论

- 机器学习
- 统计分析方法
- 神经网络
- 正则化
- 注意力机制
- 外部记忆机制
- 无监督学习
- 集成学习
- 概率图模型
- 深度信念网络
- 生成模型
- 强化学习
- 序列生成模型

技术框架

- 基本的深度学习相当于函数逼近问题，即函数或曲面的拟合，所不同的是，这里用作基函数的是非线性的神经网络函数
- Theano
- Caffe
- Tensorflow
- Pytorch

- 飞桨 PaddlePaddle
- MindSpore
- Chainer
- MXNet
- Keras
- DGL

研究方法

全面地介绍了神经网络、机器 学习和深度学习的基本概念、模型和方法，同时也涉及深度学习中许多最新进 展．书后还提供了相关数学分支的简要介绍

前沿进展

1. **卷积神经网络（Convolutional Neural Networks, CNNs）**:
    - 进展：已成为计算机视觉的主流模型，引入跨层的直连边，能够训练上百或上千层的网络，越来越多地使用 1×1 和 3×3 的小卷积核，出现不规则卷积操作如空洞卷积和可变形卷积。
    - 方向：全卷积网络（FCN），减少汇聚层和全连接层。
2. **注意力机制**:
    - 进展：在多个任务上表现出色，如语音识别、图像标题生成、阅读理解、文本分类和机器翻译。
    - 方向：自注意力机制，有效处理长距离依赖问题。
3. **神经网络与概率图模型结合**:
    - 进展：结合神经网络的表示能力和拟合能力来建模图模型中的推断、生成或势能函数问题，如变分自编码器、生成对抗网络。
    - 方向：图神经网络（GNN）和结构化注意力。
4. **生成对抗网络（Generative Adversarial Networks, GANs）**:
    - 进展：突破传统概率模型的最大似然估计限制，DCGAN 可以生成逼真图像，结合 GAN 和强化学习建立文本生成模型。
    - 方向：W-GAN 使用 Wasserstein 距离代替 JS 散度进行训练。
5. **深度强化学习**:
    - 进展：使用策略网络和值网络，演员-评论员算法，确定性策略梯度，深度确定性策略梯度。
    - 方向：异步优势的演员-评论员算法（A3C）。
6. **序列生成**:
    - 进展：基于循环神经网络的序列到序列模型用于机器翻译，引入注意力模型基于卷积神经网络的序列到序列模型。
    - 方向：全连接的自注意力模型，如 Transformer。

创新建议

第一部分 机器学习基础

第 1 章　绪论

[1] 深度学习问题是一个机器学习问题，指从有限样例中通过算法总结出一般性的规律，并可以应用到新的未知数据上

[2] 深度学习采用的模型一般比较复杂，指样本的原始输入到输出目标之间的数据流经过多个线性或非线性的组件（component）。这进一步产生贡献度分配问题（Credit Assignment Problem，CAP）

[3] 目前比较好解决贡献度分配问题的模型是人工神经网络。神经网络和深度学习并不等价．深度学习可以采用神经网络模型，也可以采用其他模型（比如深度信念网络是一种概率图模型）．但是，由于神经网络模型可以比较容易地解决贡献度分配问题，因此神经网络模型成为深度学习中主要采用的模型

在本章中，先介绍人工智能的基础知识，然后再介绍神经网络和深度学习的基本概念．

1.1　人工智能

- 人工智能（Artificial Intelligence，AI）就是让机器具有人类的智能
- 图灵测试：“一个人在不接触对方的情况下，通过一种特殊的方式和对方进行一系列的问答．如果在相当长时间内，他无法根据这些问题判断对方是人还是计算机，那么就可以认为这个计算机是智能的
- John McCarthy 提出了人工智能的定义：人工智能就是要让机器的行为看起来就像是人所表现出的智能行为一样
- 人工智能的主要领域大体上可以分为以下几个方面：　年图灵奖得主．
  - **感知**：模拟人的感知能力，对外部刺激信息（视觉和语音等）进行感知和加工．主要研究领域包括语音信息处理和计算机视觉等．
  - **学习**：模拟人的学习能力，主要研究如何从样例或从与环境的交互中进行学习．主要研究领域包括监督学习、无监督学习和强化学习等．
  - **认知**：模拟人的认知能力，主要研究领域包括知识表示、自然语言理解、推理、规划、决策等．

1.1.1　人工智能发展历史

（1）　推理期

- 1956 年的达特茅斯会议之后
- 基于人类经验、逻辑或者事实归纳出的规则
- 几何定理证明器、语言翻译器等
- 推理规则过于简单，对项目难度评估不足，AI 的研究陷入低谷

（2）　知识期

- 20 世纪 70 年代
- 知识对于 AI 系统的重要性
- 出现各类专家系统[亦称为基于知识的系统]（知识库+推理机）
- 领域专家级认识
- 模拟专家思维
- 达到专家级的水平
- prolog 语言作为开发工具
- 基于逻辑学理论而创建的逻辑编程语言，最初被用于 NLP、逻辑推理等领域

（3） 学习期

- 20 世纪 80 年代
- 知识+推理很难实现例如语言理解、图像理解等智能系统
- 机器学习

1.1.2 人工智能流派

（1） 符号主义

- 逻辑主义、心理学派或计算机学派
- 假设：1、信息可以用符号表示；2、符号可以通过显式的规则（如逻辑运算）来操作；
- AI 的推理期和知识期，符号主义为主
- 优点：可解释

（2） 连接主义

- 亦仿生学派或生理学派；
- 人类的认知过程是由大量简单神经元构成的神经网络中的信息处理过程，而不是符号运算。
- 特性：非线性、分布式、并行化、局部性及自适应性。
- 缺点：缺乏解释性

1.2 机器学习

1.2.1 浅层学习/传统机器学习

- 传统的机器学习主要关注如何学习一个预测模型
- 需要首先将数据表示为一组特征（Feature），特征的表示形式可以是连续的数值、离散的符号或其他形式
- 浅层学习的一个重要特点是不涉及特征学习，其特征主要靠人工经验或特征转换方法来抽取

1.2.2 机器学习模型步骤

（1） 数据预处理

（2） 特征提取

（3） 特征转换

- 降维：特征抽取、特征选择
- 升维
- 常用方法：主成分分析 PCA、线性判别分析 LDA

（4） 预测

图 1.2 传统机器学习的数据处理流程

1.3 表示学习

- **表示**：要将输入信息转换为有效的特征
- **表示学习**：法可以自动地学习出有效的特征，并提高最终机器学习模型的性能的算法
- **语义鸿沟**：指输入数据的底层特征和高层语义信息之间的不一致性和差异性。如果可以有一个好的表示在某种程度上能够反映出数据的高层语义特征，那么我们就能相对容易地构建后续的机器学习模型。解决语义鸿沟是表示学习的关键问题。
- **核心问题：**
  - 什么是好的表示
    i. 具有很强的表示能力，即同样大小的向量可以表示更多的信息
    ii. 好的表示应该使后续的学习任务变得简单，即需要包含更高层的语义信息
    iii. 具有一般性（期望学习到的表示比较容易地迁移到其他任务上）
  - 如何学习到好的表示
- **表示特征方法**
  - 局部表示/离散表示/符号表示
    i. 优点：良好的可解释性；稀疏的二值向量，用于线性模型时计算效率高
    ii. 缺点：向量维数很高、不可扩展；向量间相似度为 0
  - 分布式表示：低维的稠密向量
    i. 优点：表示能力强、维度低；向量间相似度可计算

1.4 深度学习

- 深度学习是将原始的数据特征通过多步的特征转换得到一种特征表示，并进一步输入到预测函数得到最终结果．和浅层学习不同，深度学习需要解决的关键问题是贡献度分配问题（即一个系统中不同的组件（component）或其参数对最终系统输出结果的贡献或影响）
- 深度学习可以看作一种强化学习（Reinforcement Learning，RL），每个内部组件并不能直接得到监督信息，需要通过整个模型的最终监督信息（奖励）得到，并且有一定的延时性
- 目前，深度学习采用的模型主要是神经网络模型，其主要原因是神经网络模型可以使用误差反向传播算法，从而可以比较好地解决贡献度分配问题

图 1.4　深度学习的数据处理流程

1.4.1　端到端学习

端到端学习（End-to-End Learning），也称端到端训练，是指在学习过程中不进行分模块或分阶段训练，直接优化任务的总体目标

1.5　神经网络

在机器学习领域，神经网络是指由很多人工神经元构成的网络结构模型，这些人工神经元之间的连接强度是可学习的参数

1.5.1　人脑神经网络

- 细胞体、细胞突起、树突、轴突、突触
- 赫布规则/赫布型学习：如果两个神经元总是相关联地受到刺激，它们之间的突触强度增加

1.5.2　人工神经网络

人工神经网络是为模拟人脑神经网络而设计的一种计算模型，它从结构、实现机理和功能上模拟人脑神经网络

- 首个可学习的人工神经网络是赫布网络，采用一种基于赫布规则的无监督学习方法
- 感知器是最早的具有机器学习思想的神经网络，但其学习方法无法扩展到多层的神经网络上
- 1980 年左右，反向传播算法有效地解决了多层神经网络的学习问题

1.5.3　神经网络发展历史

一、模型提出

第一阶段为 1943 年～1969 年，是神经网络发展的第一个高 潮期．在此期间，科学家们提出了许多神经元模型和学习规则。

- MP 模型
- 图灵机
- 感知器

二、冰河期

人们发现神经网络的两个关键缺陷：一是感知器无法处理“异或”回路问题；二是当时的计算机无法支持处理大型神经网络所需要的计算能力。但也出现了很多重要的模型算法

- 反向传播算法
- 新知机

三、反向传播算法引起的复兴

反向传播算法重新激发了人们对神经网络的兴趣

- Hopfield 联想记忆神经网络
- 玻尔兹曼机
- 分布式并行处理

四、流行度降低

1995 年至 2006 年，在此期间，支持向量机和其他更简单的方法（例如线性分类器）在机器学习领域的流行度逐渐超过了神经网络.

五、深度学习崛起

从 2006 年开始至今，在这一时期研究者逐渐掌握了训练深层神经网络的方法，使得神经网络重新崛起

- 深度信念网络
- 神经网络在语音识别和图像分类上获得成功
- 计算能力提高

1.6 知识体系

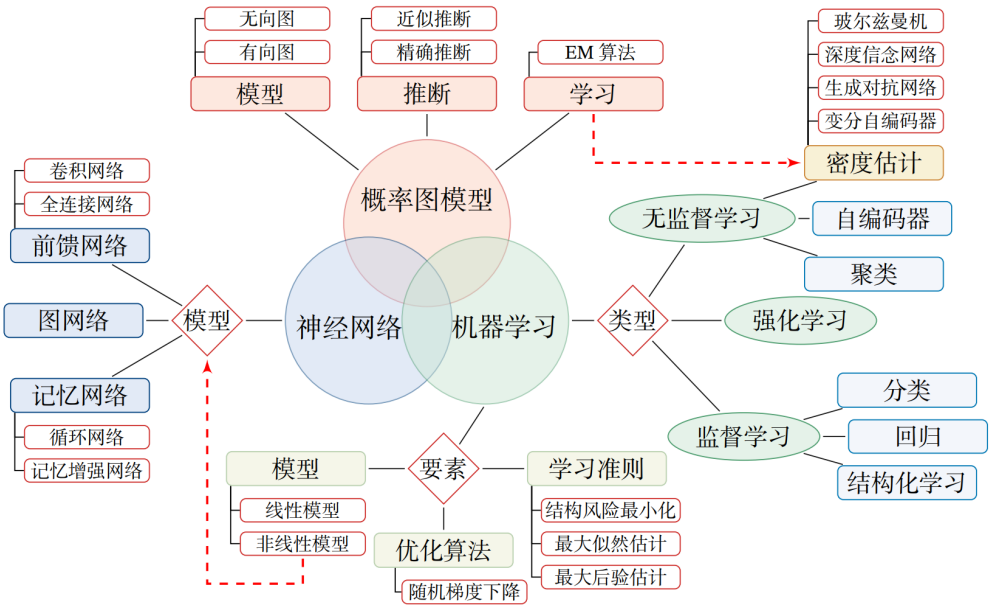


图 1.6 本书的知识体系

机器学习

机器学习可以分为监督学习、无监督学习和强化学习。第 2 章对机器 学习进行概述，使读者能够了解机器学习的基本概念以及三要素：模型、学习准则和优化算法，并以线性回归为例来讲述不同学习算法之间的关联。第 3 章主要介绍一些基本的线性模型. 这两章都以监督学习为主进行介绍. 第 9 章介绍了一些无监督学习方法，包括无监督特征学习和概率密度估计。第 10 章中介绍了一些和模型无关的机器学习方法. 第 14 章介绍了深度强化学习的知识.

神经网络

神经网络作为一类非线性的机器学习模型，可以更好地实现输入和输出之间的映射。第 4 章到第 6 章分别讲述三种主要的神经网络模型：前馈神经网络、卷积神经网络和循环神经网络. 第 6 章也简单介绍了一种更一般性的网络：图 网络. 第 7 章介绍神经网络的优化与正则化方法. 第 8 章介绍神经网络中的注意力机制和外部记忆.

概率图模型

概率图模型为机器学习提供了一个更加便捷的描述框架. 第 11 章介绍了概率图模型的基本概念，包括模型表示、学习和推断. 目前深度学习和概率图模型的融合已经十分流行. 第 12 章介绍了两种概率图模型：玻尔兹曼机和深度信念网络. 第 13 章和第 15 章分别介绍两种概率生成模型：深度生成模型和序列生成模型.

1.7 常用框架

- Theano
- Caffe
- Tensorflow
- Pytorch
- 飞桨 PaddlePaddle
- MindSpore
- Chainer
- MXNet
- Keras
- DGL

第 2 章 机器学习概述

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。

2.1 基本概念

- 样本（示例）：特征＋标签

- 数据集：训练集＋测试集
- 学习（训练）：给定一组训练集$\mathcal{D}$，通过某种学习算法$\mathcal{A}$，我们可以从函数集合$\mathcal{F}$中学习到一个最优的函数，然后可以用这个函数预测标签的值或标签的条件概率。

2.2 机器学习三个基本要素

机器学习是从有限的观测数据中学习（或“猜测”）出具有一般性的规律，并可以将总结出来的规律推广应用到未观测样本上。机器学习方法可以粗略地分为三个基本要素：模型、学习准则、优化算法。

2.2.1 模型

输入空间$\mathcal{X}$和输出空间$\mathcal{Y}$构成了一个样本空间，对于样本空间的样本$(x, y) \in \mathcal{X} \times \mathcal{Y}$，我们希望找到一个未知的真实映射函数$y = f(x)$或真实条件概率分布$p_r(y|x)$来描述，机器学习的目标就是找到这样一个模型来逼近真实映射函数$y = f(x)$或真实条件概率分布$p_r(y|x)$。

我们可以根据经验来设置一个函数集合$\mathcal{F}$，

$$\mathcal{F} = \{f(x;\theta)|\theta \in \mathbb{R}^D\}。$$

我们把$\mathcal{F}$称作**假设空间（Hypothesis Space）**，θ为模型的参数。

常见的假设空间可以分为线性和非线性两种，对应的模型也分别称为线性模型和非线性模型。

2.2.1.1 线性模型

$$f(x;\theta) = w^T x + b$$

2.2.1.2 非线性模型

广义的非线性模型可以写为多个非线性基函数$\phi(x)$的线性组合

$$f(x;\theta) = w^T \phi(x) + b，$$

其中$\phi(x) = [\phi_1(x), \dots, \phi_K(x)]^T$为$K$个非线性基函数组成的向量。

如果$\phi(x)$本身为可学习的基函数，比如

$$\phi_k(x) = h(w^T \phi'(x) + b_k)，\ \forall 1 \le k \le K，$$

其中$h(\cdot)$为非线性函数，$\phi'(x)$为另一组基函数，则$f(x;\theta)$就等价于神经网络模型。

2.2.2 学习准则

模型的好坏可以通过**期望风险（Expected Risk）** 来衡量，

$$\mathcal{R}(\theta) = \mathbb{E}_{(x,y)\sim p_r(x,y)}[\mathcal{L}(y, f(x;\theta)]。$$

2.2.2.1 损失函数

- 0-1 损失函数
- 平方损失函数
- 交叉熵损失函数
- Hinge 损失函数

2.2.2.2 风险最小化准则

一个好的模型应该有一个比较小的期望风险，但由于我们无法知道真实的数据分布，期望风险是无法计算的。这里我们给定一个训练集$\mathcal{D} = \{\left(x^{(n)}, y^{(n)}\right)\}_{n=1}^{N}$，我们可以计算**经验风险（Empirical Risk）**，即在训练集上的平均损失

$$\mathcal{R}_{\mathcal{D}}^{emp}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}\left(y^{(n)}, f\left(x^{(n)};\theta\right)\right)。$$

我们需要找到一组参数使得经验风险最小，这就是**经验风险最小化（Empirical Risk Minimization，ERM）** 准则。

由大数定理，当训练集大小$|\mathcal{D}|$趋向于无穷大，经验风险就趋向于期望风险。但在现实中，我们只有有限个样本，并且训练样本往往是真实数据的一个很小的子集或者包含一定的噪声数据，不能很好反应全部数据的真实分布，因此追求经验风险最小化原则容易导致模型在训练集上错误率很低，但在未知数据上错误率很高，这就是所谓的**过拟合（Overfitting）**。

> 定义-过拟合：给定一个假设空间$\mathcal{F}$，一个假设f属于$\mathcal{F}$，如果存在其他的假设f'也属于$\mathcal{F}$，使得在训练集上f的损失比f'的损失小，但在整个样本空间上f'的损失比f的损失小，那么就说假设f过度拟合训练数据。

过拟合问题的原因：

- 训练数据少
- 噪声
- 模型能力强

为解决过拟合问题，一般在经验风险最小化的基础上引入参数的**正则化（Regulazation）**来限制模型能力，使其不要过度地最小化经验风险。这种准则就是**结构风险最小化（Structure Risk Minimization，SRM）** 准则

和过拟合相反的一个概念是**欠拟合（Underfitting）**，即模型不能很好地拟合训练数据，在训练集上错误率较高。欠拟合一般是由于模型的拟合能力不足造成的。

机器学习的学习准则不仅仅是拟合训练集上的数据，同时也要使得泛化错误最低，使得模型能够对未知的样本进行预测。因此，机器学习可以被看作一个从有限、高维、有噪声的数据上得到更一般性的泛化规律。

2.2.3 优化算法

在确定了训练集$\mathcal{D}$、假设空间$\mathcal{F}$以及学习准则后，如何找到一个最优的模型$f(x;\theta^*)$就成了一个最优化问题。机器学习的训练过程就是最优化问题的求解过程。

在机器学习中，优化又可以分为参数优化和超参数优化．模型$f(x;\theta)$中的θ称为模型的参数，可以通过优化算法进行学习。除了可学习的参数θ之外，还有一类参数是用来定义模型结构或优化策略的，这类参数叫作**超参数（Hyper-Parameter）**。常见的超参数包括：聚类算法中的类别个数、梯度下降法中的步长、正则化项的系数、神经网络的层数、支持向量机中的核函数等。超参数的选取一般都是组合优化问题，很难通过优化算法来自动学习。因此，超参数优化是机器学习的一个经验性很强的技术，通常是按照人的经验设定，或者通过搜索的方法对一组超参数组合进行不断试错调整。

2.2.3.1 梯度下降法

为了充分利用凸优化中一些高效、成熟的优化方法，比如共轭梯度、拟牛顿法等，很多机器学习方法都倾向于选择合适的模型和损失函数，以构造一个凸函数作为优化目标．但也有很多模型（比如神经网络）的优化目标是非凸的，只能退而求其次找到局部最优解．在机器学习中，最简单、常用的优化算法就是梯度下降法

2.2.3.2 提前停止

针对梯度下降的优化算法，除了加正则化项之外，还可以通过提前停止来防 止过拟合． 在梯度下降训练的过程中，由于过拟合的原因，在训练样本上收敛的参数， 并不一定在测试集上最优．因此，除了训练集和测试集之外，有时也会使用一个验证集（Validation Set）来进行模型选择，测试模型在验证集上是否最优．在每次迭代时，把新得到的模型 $f(\boldsymbol{x};\theta)$ 在验证集上进行测试，并计算错误率． 如果在验证集上的错误率不再下降，就停止迭代．这种策略叫提前停止（Early Stop）．如果没有验证集，可以在训练集上划分出一个小比例的子集作为验证集． 图 2.4 给出了提前停止的示例．

2.2.3.3 随机梯度下降法

在公式 (2.27)的梯度下降法中，目标函数是整个训练集上的风险函数，这种方式称为批

量梯度下降法（Batch Gradient Descent，BGD）．批量梯度下降法在 每次迭代时需要计算每个样本上损失函数的梯度并求和．当训练集中的样本数 量N 很大时，空间复杂度比较高，每次迭代的计算开销也很大．在机器学习中，我们假设每个样本都是独立同分布地从真实数据分布中随 机抽取出来的，真正的优化目标是期望风险最小．批量梯度下降法相当于是从真实数据分布中采集 N 个样本，并由它们计算出来的经验风险的梯度来近似期 望风险的梯度．为了减少每次迭代的计算复杂度，我们也可以在每次迭代时只采集一个样本，计算这个样本损失函数的梯度并更新参数，即随机梯度下降法（Stochastic Gradient Descent，SGD）．

2.2.3.4 小批量梯度下降法

随机梯度下降法的一个缺点是无法充分利用计算机的并行计算能力．小批量梯度下降法（Mini-Batch Gradient Descent）是批量梯度下降和随机梯度下降的折中．每次迭代时，我们随机选取一小部分训练样本来计算梯度并更新参数，这样既可以兼顾随机梯度下降法的优点，也可以提高训练效率．

2.3 线性回归

线性回归（Linear Regression）是机器学习和统计学中最基础和最广泛应 用的模型，是一种对自变量和因变量之间关系进行建模的回归分析．自变量数量 为 1 时称为简单回归，自变量数量大于 1 时称为多元回归．

2.3.1 参数学习

给定一组包含 N 个训练样本的训练集 $\mathcal{D}=\{(\boldsymbol{x}(n),y(n))\}N\ n$=1，我们希望能够 学习一个最优的线性回归的模型参数$\boldsymbol{w}$． 我们介绍四种不同的参数估计方法：经验风险最小化、结构风险最小化、最大似然估计、最大后验估计．

2.3.1.1 经验风险最小化

由于线性回归的标签y和模型输出都为连续的实数值，因此平方损失函数非常合适衡量真实标签和预测标签之间的差异．这种利用梯度下降法来求解的方法也称为最小均方（Least Mean Squares，LMS）算法．

2.3.1.2 结构风险最小化

最小二乘法的基本要求是各个特征之间要互相独立，如果特征之间有较大的多重共线性（Multicollinearity），会使得数值上无法准确计算．数据集 $\boldsymbol{X}$ 上一些小的扰动就会导致大的改变，进而使得最小二乘法的计算变得很不稳定．为了解决这个问题， Hoerl 提出了岭回归（Ridge Regression），岭回归的解可以看作结构风险最小化准则下的最小二乘法估计

2.3.1.3 最大似然估计

最大似然估计（Maximum Likelihood Estimation，MLE）是指找到一组 参数 $\boldsymbol{w}$ 使得似然函数 $p(\boldsymbol{y}|\boldsymbol{X};\boldsymbol{w},\sigma)$ 最大，等价于对数似然函数 $\log p(\boldsymbol{y}|\boldsymbol{X};\boldsymbol{w},\sigma)$ 最大.

2.3.1.4 最大后验估计

估计参数 $\boldsymbol{w}$ 的后验概率分布的方法称为贝叶斯估计（Bayesian Esti mation），是一种统计推断问题．采用贝叶斯估计的线性回归也称为贝叶斯线 性回归（Bayesian Linear Regression）.

贝叶斯估计是一种参数的区间估计，即参数在一个区间上的分布．如果我们希望得到一个最优的参数值（即点估计），可以使用最大后验估计．最大后 验估计（Maximum A Posteriori Estimation，MAP）是指最优参数为后验分布 $p(\boldsymbol{w}|\boldsymbol{X},\boldsymbol{y};\nu,\sigma)$中概率密度最高的参数：

2.4 偏差-方差分解

为了避免过拟合，我们经常会在模型的拟合能力和复杂度之间进行权衡．拟合能力强的模型一般复杂度会比较高，容易导致过拟合．相反，如果限制模型的复杂度，降低其拟合能力，又可能会导致欠拟合.

如何在模型的拟合能力和复杂度之间取得一个较好的平衡，对一个机器学习算法来讲十分重要．偏差-方差分解（Bias-Variance Decomposition）提供了一个很好的分析和指导工具.

2.5 机器学习算法类型

2.5.1 根据函数类型

线性和非线性

2.5.2 根据学习准则

统计方法和非统计方法

2.5.3 按照训练样本提供的信息以及反馈方式的不同

2.5.3.1 监督学习

如果机器学习的目标是建模样本的特征 $\boldsymbol{x}$ 和标签 y 之间的关系：$y = f(\boldsymbol{x};\theta)$或$p(y|\boldsymbol{x};\theta)$，并且训练集中每个样本都有标签，那么这类机器学习称为监督学习（Supervised Learning）．根据标签类型的不同，监督学习又可以分为回归问题、分类问题和结构化学习问

题.

- 回归
  - 问题中的标签 y 是连续值（实数或连续整数），$f(\boldsymbol{x};\theta)$的输出也是连续值.
- 分类
  - 标签 y 是离散的类别（符号）. 在分 类问题中，学习到的模型也称为分类器（Classifier）. 分类问题根据其类别数量 又可分为二分类（Binary Classification）和多分类（Multi-class Classification） 问题.
- 结构化学习
  - 是一种特殊的分类问题. 在结构化学习中，标签$\boldsymbol{y}$通常是结构化的对象，比如序列、树或图等. 由于结构化学习的输出空间比较大，因此我们一般定义一个联合特征空间，将$\boldsymbol{x}, \boldsymbol{y}$映射为该空 间中的联合特征向量$\boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})$

2.5.3.2 无监督学习

无监督学习（Unsupervised Learning，UL）是指从不包含目标标签的训练样本中自动学习到一些有价值的信息. 典型的无监督学习问题有聚类、密度估计、特征学习、降维等.

2.5.3.3 强化学习

强化学习（Reinforcement Learning，RL）是一类通过交互来学习的机器学习算法. 在强化学习中，智能体根据环境的状态做出一个动作，并得到即时或延时的奖励. 智能体在和环境的交互中不断学习并调整策略，以取得最大化的期望总回报. 强化学习和监督学习的不同在于，强化学习不需要显式地以“输入/输出对”的方式给出训练样本， 是一种在线的学习机制.

表 2.1 三种机器学习类型的比较

	监督学习	无监督学习	强化学习
训练样本	训练集 $\{(\boldsymbol{x}^{(n)}, y^{(n)})\}_{n=1}^{N}$	训练集 $\{\boldsymbol{x}^{n}\}_{n=1}^{N}$	智能体和环境交互的轨迹 τ 和累积奖励 G_τ
优化目标	$y = f(\boldsymbol{x})$ 或 $p(y\|\boldsymbol{x})$	$p(\boldsymbol{x})$ 或带隐变量 $\boldsymbol{z}$ 的 $p(\boldsymbol{x}\|\boldsymbol{z})$	期望总回报 $\mathbb{E}_\tau[G_\tau]$
学习准则	期望风险最小化 最大似然估计	最大似然估计 最小重构错误	策略评估 策略改进

数据集一般都需要由人工进行标 注，成本很高. 因此，也出现了很多弱监督学习（Weakly Supervised Learning） 和半监督学习（Semi-Supervised Learning，SSL）的方法，希望从大规模的无标 注数据中充分挖掘有用的信息，降低对标注样本数量的要求

2.6 数据特征表示

- 图像特征：如果图像是一张大小为$M \times N$ 的图像， 其特征向量可以简单地表示为$M \times N$ 维的向量，每一维的值为图像中对应像素 的灰度值. 为了提高模型准确率，也会经常加入一个额外的特征，比如直方图、宽 高比、笔画数、纹理特征、边缘特征等
- 文本特征：为了将样本x 从文本形式转为向量形式，一种简单 的方式是使用词袋（Bag-of-Words，BoW）模型

直接使用原始特征的缺点：

a. 特征比较单一，需要进行 （非线性的）组合才能发挥其作用
b. 特征之间冗余度比较高
c. 并不是所有的 特征都对预测有用
d. 很多特征通常是易变的
e. 特征中往往存在一些噪声

2.6.1 传统特征学习

2.6.2 深度特征学习

2.6.2.1 特征选择

特征选择（Feature Selection）是选取原始特征集合的一个有效子集，使得 基于这个特征子集训练出来的模型准确率最高

- 子集搜索
  a. 过滤式方法
  b. 包裹式方法
- ℓ1 正则化

2.6.2.2 特征抽取

特征抽取（Feature Extraction）是构造一个新的特征空间，并将原始特征 投影在新的空间中得到新的表示

- 监督方法
  a. 线性判别分析 LDA
- 无监督方法
  a. 主成分分析 PCA
  b. 自编码器 AE

2.7 评价指标

对于分类问题，常见的评价标准有准确率、精确率、召回率和 F 值等

2.7.1 准确率

最常用的评价指标为准确率（Accuracy）：

$$\mathcal{A} = \frac{1}{N}\sum_{n=1}^{N} I\left(y^{(n)} = \hat{y}^{(n)}\right),$$

2.7.2 错误率

和准确率相对应的就是错误率（Error Rate）：

$$\mathcal{E} = 1 - \mathcal{A} = \frac{1}{N}\sum_{n=1}^{N} I\left(y^{(n)} \neq \hat{y}^{(n)}\right)$$

2.7.3 精确率和召回率

如果希望对每个类都进 行性能估计，就需要计算精确率（Precision）和召回率（Recall）对于类别c来说，模型在测试集上的结果可以分为以下四种情况：

（1） 真正例（True Positive，TP）：一个样本的真实类别为c并且模型正确地预测为类别c.

（2） 假负例（False Negative，FN）：一个样本的真实类别为c，模型错误地预测为其他类

（3） 假正例（False Positive，FP）：一个样本的真实类别为其他类，模型错误地预测为类别c

（4） 真负例（True Negative，TN）：一个样本的真实类别为其他类，模型也预测为其他类．这类样本数量记为TNc

这四种情况的关系可以用如表 2.3 所示的混淆矩阵（Confusion Matrix）来 表示．

表 2.3 类别 c 的预测结果的混淆矩阵

		预测类别	
		$\hat{y} = c$	$\hat{y} \neq c$
真实类别	$y = c$	TP_c	FN_c
	$y \neq c$	FP_c	TN_c

2.7.3.1 精确率

也叫精度或查准率，类别 c 的查准率是所有预测为类 别c的样本中预测正确的比例：

$$\mathcal{P}_c = \frac{TP_c}{TP_c + FP_c}.$$

2.7.3.2 召回率

也叫查全率，类别c的查全率是所有真实标签为类别c的样本中预测正确的比例：

$$\mathcal{R}c = \frac{TP_c}{TP_c + FN_c}$$

2.7.3.3 F 值

是一个综合指标，为精确率和召回率的调和平均：

$$\mathcal{F}c = \frac{(1+\beta^2)\times\mathcal{P}_c\times\mathcal{R}_c}{\beta^2\times\mathcal{P}_c+\mathcal{R}_c}$$

2.7.3.4 宏平均和微平均

宏平均是每一类的性能指标的算术平均值
微平均是每一个样本的性能指标的算术平均值

2.8 理论和定理

在机器学习中，有一些非常有名的理论或定理

2.8.1 PAC 学习理论

计算学习理论（Computational Learning The□ory）是机器学习的理论基础，其中最基础的理论就是可能近似正确（Probably Approximately Correct，PAC）学习理论．PAC 学习理论也可以帮助分析一个机器学习方法在什么条件下可以学习到 一个近似正确的分类器。

PAC 可分为两部分

- 近似正确：一个假设$f \in \mathcal{F}$ 是“近似正确” 的，是指其在泛化错误$\mathcal{G}_{\mathcal{D}}(f)$小于一个界限
- 可能：一个学习算法$\mathcal{A}$ 有“可能”以 $1-\delta$ 的概率学习到这样一个“近似正确”的假设

2.8.2 没有免费午餐定理

没有免费午餐定理（No Free Lunch Theorem，NFL）是由 Wolpert 和 Mac□erday 在最优化理论中提出的．没有免费午餐定理证明：

对于基于迭代的最优化算法，不存在某种算法对所有问题（有限的搜索空间内）都有效．如果一个算法 对某些问题有效，那么它一定在另外一些问题上比纯随机搜索算法更差．也就是说，不能脱离具体问题来谈论算法的优劣，任何算法都有局限性．必须要“具体问 题具体分析”．

2.8.3 奥卡姆剃刀原理

奥卡姆剃刀（Occam’s Razor）原理是由 14 世纪逻辑学家 William of Occam 提出的一个解决问题的法则：“如无必要，勿增实体”．奥卡姆剃刀的思想和机器 学习中的正则化思想十分类似：简单的模型泛化能力更好．如果有两个性能相近的模型，我们应该选择更简单的模型．因此，在机器学习的学习准则上，我们经常会引入参数正则化来限制模型能力，避免过拟合．

奥卡姆剃刀的一种形式化是最小描述长度（Minimum Description Length， MDL）原则，即对一个数据集$\mathcal{D}$，最好的模型$f \in \mathcal{F}$会使得数据集的压缩效果最好，即编码长度最小。

2.8.4 丑小鸭定理

丑小鸭定理（Ugly Duckling Theorem）是 1969 年由渡边慧提出的：

“丑小鸭与白天鹅之间的区别和两只白天鹅之间的区别一样大”. 因为世界上不存在相似性的客观标准，一切相似性的标准都是主观的．

2.8.5 归纳偏置

在机器学习中，很多学习算法经常会对学习的问题做一些假设，这些假设就 称为归纳偏置（Inductive Bias）。在朴素贝叶斯分类器中，我们会假设每个特征的条件概率是互相独立的．归纳偏置在贝叶斯学习中也经常称为先验（Prior）．

第 3 章 线性模型

线性模型（Linear Model）是机器学习中应用最广泛的模型，指通过样本特 征的线性组合来进行预测的模型

3.1 线性判别函数和决策边界

一个线性分类模型（Linear Classification Model）或线 性分类器（Linear Classifier），是由一个（或多个）线性的判别函数 $f(\boldsymbol{x}; \boldsymbol{w}) = \boldsymbol{w}\mathrm{T}\boldsymbol{x} + b$ 和非线性的决策函数 $g(\cdot)$ 组成．我们首先考虑二分类的情况，然后再扩展到多分类的情况．

3.1.1 二分类

在二分类问题中，我们只需要一个线性判别函数 $f(\boldsymbol{x};\boldsymbol{w})=\boldsymbol{w}\mathrm{T}\boldsymbol{x}+b$．特征空间 $\mathbb{R}D$ 中所有满足 $f(\boldsymbol{x};\boldsymbol{w})=0$ 的点组成一个分割超平面（Hyperplane），称 为决策边界（Decision Boundary）或决策平面（Decision Surface）．决策边界将特征空间一分为二，划分成两个区域，每个区域对应一个类别．

3.1.2 多分类

多分类（Multi-class Classification）问题是指分类的类别数 C 大于 2．多分 类一般需要多个线性判别函数，但设计这些判别函数有很多种方式． 假设一个多分类问题的类别为 $\{1,2,\cdots,C\}$，常用的方式有以下三种：

（1）“一对其余”方式：把多分类问题转换为 C 个“一对其余”的二分类问 题．这种方式共需要 C 个判别函数，其中第 c 个判别函数 fc 是将类别 c 的样本和 不属于类别c的样本分开．

（2）“一对一”方式：把多分类问题转换为 $C(C-1)/2$ 个“一对一”的二分 类问题．这种方式共需要$C(C-1)/2$ 个判别函数，其中第(i,j)个判别函数是把类 别i 和类别j 的样本分开．

（3）“argmax”方式：这是一种改进的“一对其余”方式，共需要 C 个判 别函数

3.2 Logistic 回归

Logistic 回归（Logistic Regression，LR）是一种常用的处理二分类问题的线性模型。为了解决连续的线性函数不适合进行分类的问题，我们引入非线性函数g ： $\mathbb{R}D\rightarrow(0,1)$来预测类别标签的后验概率，在 Logistic 回归中，我们使用 Logistic 函数来作为激活函数g

Logistic 回归采用交叉熵作为损失函数，并使用梯度下降法来对参数进行 优化．

3.3 Softmax 回归

Softmax 回归（Softmax Regression），也称为多项（Multinomial）或多类 （Multi-Class）的 Logistic 回归，是 Logistic 回归在多分类问题上的推广。

Softmax 回归使用交叉熵损失函数来学 习最优的参数矩阵$\boldsymbol{W}$．

3.4 感知器

感知器（Perceptron）是最简单的人工神经网络，只有一个神经元．感知器是对生物神经元的简单数学模拟，有与生物神经元相对应的部件，如权重（突触）、偏置（阈值）及激活函数（细胞体），输出为+1 或−1．

感知器的学习算法是一种错误驱动的在线学习算法．先初始化一个权重向量（通常是全

零向量），然后每次分错一个样本时，就用这个样本来更新权重.

3.5 支持向量机

支持向量机（Support Vector Machine，SVM）是一个经典的二分类算法， 其找到的分割超平面具有更好的鲁棒性，因此广泛使用在很多任务上，并表现出了很强优势.

给定一个二分类器数据集，如果两类样本是线性可分的，即存在一个超平面将两类样本分开。

支持向量机的主优化问题为凸优化问题，满足强对偶性，即主优化问题可以通过最大化对偶函数 $\max\lambda\geq 0\ \Gamma(\lambda)$ 来求解。

支持向量机还有一个重要的优点是可以使用核函数（Kernel Function）隐 式地将样本从原始特征空间映射到更高维的空间，并解决原始特征空间中的线 性不可分问题

第二部分 基础模型

第 4 章 前馈神经网络

人工神经网络（Artificial Neural Network，ANN）是指一系列受生物学和 神经科学启发的数学模型．这些模型主要是通过对人脑的神经元网络进行抽象， 构建人工神经元，并按照一定拓扑结构来建立人工神经元之间的连接，来模拟生物神经网络．在人工智能领域，人工神经网络也常常简称为神经网络（Neural Network，NN）或神经模型（Neural Model）.

在本章中，我们主要关注采用误差反向传播来进行学习的神经网络，即作为 一种机器学习模型的神经网络。神经网络一般可以看作一个非线性模型，其基本组成单元为具有非线性激活函数的神经元，通过大量神经元之间的连接，使得神经网络成为一种高度非线性的模型

4.1 神经元

人工神经元（Artificial Neuron），简称神经元（Neuron），是构成神经网络的基本单元，其主要是模拟生物神经元的结构和特性，接收一组输入信号并产生输出.

假设一个神经元接收D 个输入$x1$, $x2$, ⋯ , xD，令向量$\boldsymbol{x}$ = [$x1$; $x2$; ⋯ ; xD]来 表示这组输入，并用净输入（Net Input）$z \in \mathbb{R}$ 表示一个神经元所获得的输入信号$\boldsymbol{x}$的加权和。

净输入z在经过一个非线性函数$f(\cdot)$后，得到神经元的活性值（Activation） α，非线性函数$f(\cdot)$称为激活函数（Activation Function）.

4.1.1 常用激活函数

4.1.1.1 Sigmoid 型函数

Sigmoid 型函数是指一类 S 型曲线函数，为两端饱和函数．常用的 Sigmoid 型函数有 Logistic 函数和 Tanh 函数． 对于函数 $f(x)$，若 $x \to -\infty$ 时，其导数 f

4.1.1.2 Hard-Logistic 函数和 Hard-Tanh 函数

Logistic 函数和 Tanh 函数都是 Sigmoid 型函数，具有饱和性，但是计算开销 较大．因为这两个函数都是在中间（0 附近）近似线性，两端饱和．因此，这两个函 数可以通过分段函数来近似．

4.1.1.3 ReLU 函数

ReLU（Rectified Linear Unit，修正线性单元）[Nair et al., 2010]，也叫 Rectifier 函数[Glorot et al., 2011]，是目前深度神经网络中经常使用的激活函数．ReLU 实际上是一个斜坡（ramp）函数

- 优点：

采用 ReLU 的神经元只需要进行加、乘和比较的操作，计算上更加高效． ReLU 函数也被认为具有生物学合理性（Biological Plausibility），比如单侧抑 制、宽兴奋边界（即兴奋程度可以非常高）

- 缺点：

ReLU 函数的输出是非零中心化的，给后一层的神经网络引入偏置偏移， 会影响梯度下降的效率． ReLU 神 经 元 指 采 用 ReLU 作为激活函数的神经元． 此外，ReLU 神经元在训练时比较容易“死亡”．如果参数在一次不恰当的更新后，第一个隐藏层中的某个 ReLU 神经元在所有的训练数据上都不能被激活，那么这个神经元自身参数的梯度永远都会是 0，在以后的训练过程中永远不能被激活．这种现象称为死亡 ReLU 问题。

4.1.1.4 Swish 函数

4.1.1.5 GELU 函数

4.1.1.6 Maxout 单元

4.2 网络结构

4.2.1 前馈网络

- 前馈网络中各个神经元按接收信息的先后分为不同的组．每一组可以看作一个神经层
- 整个网络中的信息是朝一个方向传播，没有反向的信息传播
- 前馈网络包括全连接前馈网络和卷积神经网络等
- 前馈网络可以看作一个函数，通过简单非线性函数的多次复合，实现输入空间到输出空间的复杂映射．网络结构简单，易于实现．

4.2.2 记忆网络

记忆网络也称反馈网络

- 网络中的神经元不但可以接收其他神经元的信息，也可以接收自己的历史信息．和前馈网络相比，记忆网络中的神经元具有记忆功能，在不同的时刻具有不同的状态．记
- 忆神经网络中的信息传播可以是单向或双向传递
- 记忆网络包括循环神经网络、Hopfield 网络、玻尔兹曼机、受限玻尔兹曼机等．
- 记忆网络可以看作一个程序，具有更强的计算和记忆能力．
- 为了增强记忆网络的记忆容量，可以引入外部记忆单元和读写机制，用来保存一些网络的中间状态，称为记忆增强神经网络（Memory Augmented Neural Network，MANN），比如神经图灵机和记忆网络

4.2.3 图网络

- 图网络是定义在图结构数据上的神经网络．图中每个节点都由一个或一组神经元构成
- 节点之间的连接可以是有向的，也可以是无向的．每个节点可以收到来自相邻节点或自身的信息．
- 图网络是前馈网络和记忆网络的泛化，包含很多不同的实现方式，比如 图卷 积网络（Graph Convolutional Network，GCN）、图注意力网络（Graph Attention Network，GAT）、消息传递神经网络（Message Passing Neural Network，MPNN）等．

4.3 前馈神经网络

前馈神经网络（Feedforward Neural Network，FNN）是最早发明的简单 人工神经网络．前馈神经网络也经常称为多层感知器（Multi-Layer Perceptron， MLP）．但多层感知器的叫法并不是十分合理，因为前馈神经网络其实是由多层 的 Logistic 回归模型（连续的非线性函数）组成，而不是由多层的感知器（不连 续的非线性函数）组成 [Bishop, 2007]． 在前馈神经网络中，各神经元分别属于不同的层．每一层的神经元可以接收 前一层神经元的信号，

并产生信号输出到下一层．第 0 层称为输入层，最后一层称 为输出层，其他中间层称为隐藏层．整个网络中无反馈，信号从输入层向输出层 单向传播，可用一个有向无环图表示．

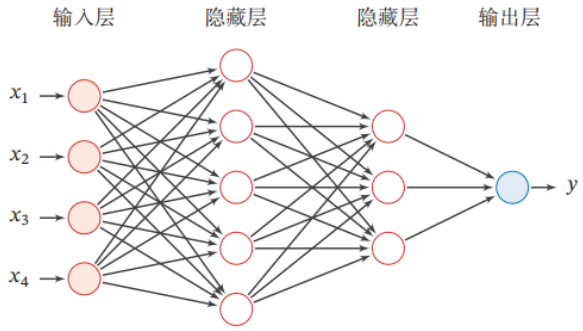


图 4.7　多层前馈神经网络

4.4　反向传播算法

第l 层的误差项可以通过第$l+1$ 层的误差项计算得 到，这就是误差的反向传播（BackPropagation，BP）．反向传播算法的含义是： 第 l 层的一个神经元的误差项（或敏感性）是所有与该神经元相连的第 $l+1$ 层 的神经元的误差项的权重和．然后，再乘上该神经元激活函数的梯度

4.5　自动梯度计算

神经网络的参数主要通过梯度下降来进行优化．当确定了风险函数以及网络结构后，我们就可以手动用链式法则来计算风险函数对每个参数的梯度，并用代码进行实现．

4.6　优化问题

神经网络的参数学习比线性模型要更加困难，主要原因有两点：1）非凸优 化问题和 2）梯度消失问题（误差经过每一层传递都会不断衰减．当网络层数很深时，梯度就会不停衰减，甚至消失，使得整个网络很难训练）

第 5 章　卷积神经网络

卷积神经网络（Convolutional Neural Network，CNN 或 ConvNet）是一种 具有局部连接、权重共享等特性的深层前馈神经网络．

卷积神经网络最早主要是用来处理图像信息．在用全连接前馈网络来处理 图像时，会存在以下两个问题：

（1）　参数太多

（2）　局部不变性特征

卷积神经网络是受生物学上感受野机制的启发而提出的．感受野（Receptive Field）机

制主要是指听觉、视觉等神经系统中一些神经元的特性，即神经元 只接受其所支配的刺激区域内的信号

目前的卷积神经网络一般是由卷积层、汇聚层和全连接层交叉堆叠而成的 前馈神经网络．卷积神经网络有三个结构上的特性：局部连接、权重共享以及汇聚．这些特性使得卷积神经网络具有一定程度上的平移、缩放和旋转不变性．和 前馈神经网络相比，卷积神经网络的参数更少．

5.1 卷积

卷积（Convolution），也叫褶积，是分析数学中一种重要的运算．在信号处 理或图像处理中，经常使用一维或二维卷积．

5.2 卷积神经网络

卷积神经网络一般由卷积层、汇聚层和全连接层构成．

5.3 参数学习

在卷积网络中，参数为卷积核中权重以及偏置．和全连接前馈网络类似，卷 积网络也可以通过误差反向传播算法来进行参数学习． 在全连接前馈神经网络中，梯度主要通过每一层的误差项 δ 进行反向传播， 并进一步计算每层参数的梯度

5.4 典型卷积神经网络

- LeNet-5
- AlexNet
- Inception 网络
- 残差网络

5.5 其他卷积方式

- 转置卷积、反卷积
- 微步卷积
- 空洞卷积

第 6 章 循环神经网络

在前馈神经网络中，信息的传递是单向的，这种限制虽然使得网络变得更容易学习，但

在一定程度上也减弱了神经网络模型的能力．在很多现实任务中， 网络的输出不仅和当前时刻的输入相关，也和其过去一段时间的输出相关．当处理这一类和时序数据相关 的问题时，就需要一种能力更强的模型．

循环神经网络（Recurrent Neural Network，RNN）是一类具有短期记忆能 力的神经网络．在循环神经网络中，神经元不但可以接受其他神经元的信息，也 可以接受自身的信息，形成具有环路的网络结构。

循环神经网络的参数学习可以通过随时间反向传播算法来学习．随时间反向传播算法即按照时间的逆序将错误信息一步步地往前传递．当输入序列比较长时，会存在梯度爆炸和消失问题，也称为长程依赖问题．为了解决这个问题，人们对循环神经网络进行了很多的改进，其中最有效的改进方式引入门控机制．

此外，循环神经网络可以很容易地扩展到两种更广义的记忆网络模型：递归神经网络和图网络．

6.1 给网络增加记忆能力

为了处理这些时序数据并利用其历史信息， 我们需要让网络具有短期记忆能力：

- 延时神经网络
- 有外部输入的非线性自回归模型
- 循环神经网络

6.2 简单循环网络

简单循环网络（Simple Recurrent Network，SRN）[Elman, 1990] 是一个非 常简单的循环神经网络，只有一个隐藏层的神经网络．在一个两层的 前馈神经网 络中，连接存在相邻的层与层之间，隐藏层的节点之间是无连接的．而简单循环 网络增加了从隐藏层到隐藏层的反馈连接．

6.3 应用到机器学习

循环神经网络可以应用到很多不同类型的机器学习任务．根据这些任务的 特点可以分为以下几种模式：

- 序列到类别模式
- 同步的序列到序列模式
- 异步的序列到序列模式．

6.4 参数学习

循环神经网络的参数可以通过梯度下降方法来进行学习

6.5 长程依赖问题

如果时刻 t 的输出 yt 依赖于时刻 k 的输入 $\boldsymbol{x}k$，当间隔 $t-k$ 比较大时，简单神经网络很难建模这种长距离的依赖关系，称为长程依赖问题（Long-Term Dependencies Problem）.

6.6 基于门控的循环神经网络

为了改善循环神经网络的长程依赖问题， 一种非常好的解决方案是在公 式(6.50)的基础上引入门控机制来控制信息的累积速度，包括有选择地加入新的信息，并有选择地遗忘之前累积的信息．这一类网络可以称为基于门控的循环神经网络（Gated RNN）

6.7 深层循环神经网络

如果将深度定义为网络中信息传递路径长度的话，循环神经网络可以看作 既“深”又“浅”的网络．一方面来说，如果我们把循环网络按时间展开，长时间间 隔的状态之间的路径很长，循环网络可以看作一个非常深的网络．从另一方面来 说，如果同一时刻网络输入到输出之间的路径 $\boldsymbol{x}t \rightarrow \boldsymbol{y}t$，这个网络是非常浅的．

6.8 扩展到图结构

如果将循环神经网络按时间展开，每个时刻的隐状态 $\boldsymbol{h}t$ 看作一个节点，那 么这些节点构成一个链式结构，每个节点 t 都收到其父节点的消息（Message）， 更新自己的状态，并传递给其子节点．而链式结构是一种特殊的图结构，我们 可以比较容易地将这种消息传递（Message Passing）的思想扩展到任意的图结 构上．

第 7 章 网络优化与正则化

应用神经网络模型到机器学 习时依然存在一些难点问题．主要分为两大类：

(1) 优化问题

深度神经网络的优化十分困难．首先，神经网络的损失函数 是一个非凸函数，找到全局最优解通常比较困难．其次，深度神经网络的参数通常非常多，训练数据也比较大，因此也无法使用计算代价很高的二阶优化方法，而一阶优化方法的训练效率通常比较低.此外，深度神经网络存在梯度消失或爆炸问题，导致基于梯度的优化方法经常失效.

(2) 泛化问题

由于深度神经网络的复杂度比较高，并且拟合能力很强，很容易在训练集上产生过拟合.因此在训练深度神经网络时，同时也需要通过一定的正则化方法来改进网络的泛化能力.

本章从网络优化和网络正则化两个方面来介绍在神经网络的表示能力、复杂度、学习效

率和泛化能力之间找到平衡的方：

- 在**网络优化**方面，介绍一些常用的优化算法、参数初始化方法、数据预处理方法、逐层归一化方法和超参数优化方法.
- 在**网络正则化**方面，介绍一些提高网络泛化能力的方法， 包括 $\ell 1$ 和 $\ell 2$ 正则化、权重衰减、提前停止、丢弃法、数据增强和标签平滑.

7.1 网络优化

网络优化是指寻找一个神经网络模型来使得经验（或结构）风险最小化的过程，包括模型选择以及参数学习等

7.2 优化算法

梯度下降法可以分为：

- 批量梯度下降
- 随机梯度下降
- 小批量梯度下降

7.3 参数初始化

神经网络的参数学习是一个非凸优化问题. 当使用梯度下降法来进行优化网络参数时，参数初始值的选取十分关键，关系到网络的优化效率和泛化能力． 参数初始化的方式通常有以下三种：

- 预训练初始化
- 随机初始化
- 固定值初始化

7.4 数据预处理

归一化：归一化（Normalization）方法泛指把数据特征转换为相同尺度的方法，比如把数据特征映射到 $[0, 1]$ 或 $[-1, 1]$ 区间内，或者映射为服从均值为 0、方差为 1 的标准正态分布。

常用归一化方法：

- 最小最大值归一化
- 标准化
- 白化

7.5 逐层归一化

逐层归一化（Layer-wise Normalization）是将传统机器学习中的数据归一 化方法应用到深度神经网络中，对神经网络中隐藏层的输入进行归一化，从而使 得网络更容易训练.

7.6 超参数优化

常见的超参数有以下三类：
（1） 网络结构，包括神经元之间的连接关系、层数、每层的神经元数量、激活函数的类型等.
（2） 优化参数，包括优化方法、学习率、小批量的样本数量等.
（3） 正则化系数.

超参数优化（Hyperparameter Optimization）主要存在两方面的困难：
1） 超参数优化是一个组合优化问题，无法像一般参数那样通过梯度下降方法来优 化，也没有一种通用有效的优化方法；
2）评估一组超参数配置（Configuration） 的时间代价非常高，从而导致一些优化方法（比如演化算法（Evolution Algorithm））在超参数优化中难以应用.

超参数的配置，比较简单的方法有:

- 网格搜索
- 随机搜索
- 贝叶斯优化
- 动态资源分配
- 神经架构搜索

7.7 网络正则化

正则化（Regularization）是一类通过限制模型复杂度，从而避免过拟合，提高泛化能力的方法，比如引入约束、增加先验、提前停止等.

- 传统的机器学习中，提高泛化能力的方法主要是限制模型复杂度，比如采用 $\ell 1$ 和 $\ell 2$ 正则化等方式
- 在训练深度神经网络时，特别是在过度参数化 （Over-Parameterization）时，$\ell 1$ 和 $\ell 2$ 正则化的效果往往不如浅层机器学习模型中显著. 因此训练深度学习模型时，往往还会使用其他的正则化方法，比如
  - 数据增强
  - 提前停止

- 丢弃法
- 集成法

第 8 章　注意力机制与外部记忆

简化神经网络结构的方法：

- 局部连接
- 权重共享
- 汇聚操作

对人脑的生物神经网络研究注意到，人脑在有限的资源下，并不能同时处理大量过载的输入信息。大脑神经系统有两个重要机制可以解决信息过载问题：注意力和记忆机制

- 注意力机制，通过自上而下的信息选择机制来过滤掉大量的无关信息
- 外部记忆，优化神经网络的记忆结构来提高神经网络存储信息的容量

8.1　认知神经学中的注意力

人脑可以有意或无意地从这些大量输入信息中选择小部分的有用信 息来重点处理，并忽略其他信息．这种能力就叫作注意力（Attention）

注意力一般分为两种：
（1） 自上而下的有意识的注意力，称为聚焦式注意力（Focus Attention）．聚焦式注意力是指有预定目的、依赖任务的，主动有意识地聚焦于某一对象的注 意力．
（2） 自下而上的无意识的注意力，称为基于显著性的注意力（SaliencyBased Attention）．基于显著性的注意力是由外界刺激驱动的注意，不需要主动 干预，也和任务无关．

8.2　注意力机制

在计算能力有限的情况下，注意力机制（Attention Mechanism）作为一种 资源分配方案，将有限的计算资源用来处理更重要的信息，是解决信息超载问题 的主要手段．

8.3　自注意力模型

使用全连接网络建立输入序列之间的长距离依赖关系，全连接网络是一种非常直接的建模远距离依赖的模型，但是无法处理变长的输入序列．不同的输入长度，其连接权重的大小也是不同的．这时我们就可以利用注意力机制来“动态”地生成不同连接的权重，这就是自注意力模型（Self-Attention Model）

8.4 人脑中的记忆

表 8.1 不同领域中记忆模型的不严格类比

记忆周期	计算机	人脑	神经网络
短期	寄存器	短期记忆	状态（神经元活性）
中期	内存	工作记忆	外部记忆
长期	外存	长期记忆	可学习参数
存储方式	随机寻址	内容寻址	内容寻址为主

8.5 记忆增强神经网络

为了增强网络容量，我们可以引入辅助记忆单元，将一些和任务相关的信息 保存在辅助记忆中，在需要时再进行读取，这样可以有效地增加网络容量．这个引入的辅助记忆单元一般称为外部记忆（External Memory），以区别于循环神经网络的内部记忆（即隐状态）．

这种装备外部记忆的神经网络也称为记忆增强神经网络（Memory Augmented Neural Network，MANN），或简称为记忆网络 （Memory Network，MN）．

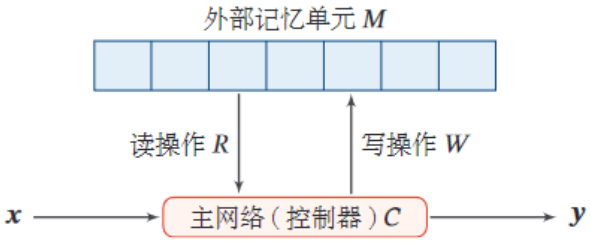


图 8.6 记忆网络的典型结构

8.6 基于神经动力学的联想记忆

将基于神经动力学（Neurodynamics）的联想记忆模型引入到神经网络以增加网络容量．联想记忆模型（Associative Memory Model）主要是通过神经网络的动态演化来进行联想，有两种应用场景：

1）输入的模式和输出的模式在同一空间，这种模型叫作自联想模型（AutoAssociative Model）．自联想模型可以通过前馈神经网络或者循环神经网络来实现，也常称为自编码器（Auto-Encoder，AE）．

2）输入的模式和输出的模式不在同一空间，这种模型叫作异联想模型（Hetero-Associative Model）．从广义上讲，大部分机器学习问题都可以被看作异联想，因此异联想模型可以作为分类器使用．

联想记忆模型可以利用神经动力学的原理来实现按内容寻址的信息存储和检索．一个经典的联想记忆模型为 Hopfield 网络．

第 9 章　无监督学习

- 无监督学习（Unsupervised Learning，UL）是指从无标签的数据中学习出 一些有用的模式
- 无监督学习算法一般直接从原始数据中学习，不借助于任何人工给出标签或者反馈等指导信息
- 无监督学习就是发现隐藏的数据中的有价值信息，包括有效的特征、类别、结构以及概率分布等

典型的无监督学习问题分类：

（1） 无监督特征学习（Unsupervised Feature Learning），是从无标签的训练数据中挖掘有效的特征或表示。无监督特征学习一般用来进行**降维、数据可视化或监督学习前期的数据预处理**

（2） 概率密度估计（Probabilistic Density Estimation），简称密度估计，是根据一组训练样本来估计样本空间的概率密度。密度估计可以分为参数密度估 计和非参数密度估计：

i. 参数密度估计是假设数据服从某个已知概率密度函数形式的分布（比如高斯分布），然后根据训练样本去估计概率密度函数的参数.

ii. 非 参数密度估计是不假设数据服从某个已知分布，只利用训练样本对密度进行估计，可以进行任意形状密度的估计．非参数密度估计的方法有**直方图、核密度估计**等.

（3） 聚类（Clustering），是将一组样本根据一定的准则划分到不同的组（也称为簇（Cluster））。

无监督学习的三个基本要素：

- ✓ 模型
- ✓ 学习准则：最大似然估计、最小重构错误等
  - ■ 在无监督特征学习中，经常使用的准则为最小化重构错误，同时也经常对特征进行一 些约束，比如独立性、非负性或稀释性等.
  - ■ 在密度估计中，经常采用最大似然估 计来进行学习.
- ✓ 优化算法

9.1　无监督特征学习

无监督特征学习是指从无标注的数据中自动学习有效的数据表示，从而能够帮助后续的机器学习模型更快速地达到更好的性能．无监督特征学习主要方法有：

- 主成分分析
- 稀疏编码
- 自编码器

9.2 概率密度估计

概率密度估计（Probabilistic Density Estimation），简称密度估计（Density Estimation），是基于一些观测样本来估计一个随机变量的概率密度函数．密度估计在数据建模、机器学习中使用广泛． 密度估计方法可以分为两类：

- ❖ 参数密度估计
- ❖ 非参数密度估计

第 10 章 模型独立的学习方式

很多场合中，机器学习的应用会受到局限，由于以下可能的原因：

- 要准备一定规模的训练数据，这 些训练数据需要和真实数据的分布一致，然后设定一个目标函数和优化方法，在 训练数据上学习一个模型
- 不同任务的模型往往都是从零开始来训练的， 一切知识都需要从训练数据中得到．这也导致了每个任务都需要准备大量的训 练数据

于是，出现了一些“模型独立的学习方式”，用于使模型能够快速适应新的任务。（模型独立，指学习方式不限于具体的模型）：

- ❖ 集成学习
- ❖ 协同学习
- ❖ 自训练
- ❖ 多任务学习
- ❖ 迁移学习
- ❖ 终身学习
- ❖ 小样本学习
- ❖ 元学习

10.1 集成学习

集成学习（Ensemble Learning）就是通过某种策略将多个模型集成起来， 通过群体决策来提高决策准确率．集成学习首要的问题是如何集成多个模型．比 较常用的集成策略有直接平均、加权平均等．

10.2 自训练和协同训练

监督学习往往需要大量的标注数据，而标注数据的成本比较高．因此，利用 大量的无标注数据来提高监督学习的效果有着十分重要的意义．这种利用少量 标注数据和大量无标注数据进行学习的方式称为半监督学习（Semi-Supervised Learning，SSL）．本节介绍两种半监督学习算法：自训练和协同训练．

10.3 多任务学习

多任务学习（Multi-task Learning）是指同时学习多个相关任务，让这些 任务在学习过程中共享知识， 利用多个任务之间的相关性来改进模型在每个 任务上的性能和泛化能力. 多任务学习可以看作一种归纳迁移学习（Inductive Transfer Learning），即通过利用包含在相关任务中的信息作为归纳偏置（Inductive Bias）来提高泛化能力

10.4 迁移学习

迁移学习是指两个不同领域的知识迁移过程，利用源领域（Source Domain） $\mathcal{DS}$ 中学到的知识来帮助目标领域（Target Domain）$\mathcal{DT}$ 上的学习任务. 源领域 的训练样本数量一般远大于目标领域.

10.5 终身学习

终身学习（Lifelong Learning），也叫持续学习（Continuous Learning）， 是指像人类一样具有持续不断的学习能力，根据历史任务中学到的经验和知识 来帮助学习不断出现的新任务，并且这些经验和知识是持续累积的，不会因为新的任务而忘记旧的知识

10.6 元学习

在面对不同的任务时，人脑 的学习机制并不相同. 即使面对一个新的任务，人们往往也可以很快找到其学习 方式. 这种可以动态调整学习方式的能力，称为元学习（Meta-Learning），也称 为学习的学习（Learning to Learn）

第三部分 进阶模型

第 11 章 概率图模型

概率图模型（Probabilistic Graphical Model，PGM），简称图模型（Graphical Model，GM），是指一种用图结构来描述多元随机变量之间条件独立关系的概率模型，从而给研究高维空间中的概率模型带来了很大的便捷性.

此外，很多机器学习模型都可以归结为概率模型，即建模输入和 输出之间的条件概率分布. 因此，图模型提供了一种新的角度来解释机器学习模 型，并且这种角度有很多优点，比如了解不同机器学习模型之间的联系，方便设 计新模型等

当概率模型中的变量数量比较多时，其条件依赖关系也比较复杂. 我们可以使用图结构的方式将概率模型可视化，以一种直观、简单的方式描述随机变量 之间的条件独立性，并可以将一个复杂的联合概率模型分解为一些简单条件概率模型的组合. 图中每个节点表示一个变量，每条连边表示变量之间的依赖关系.

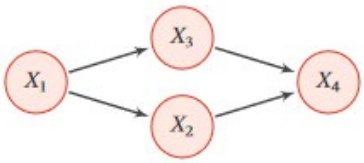


图 11.1　变量X_1, X_2, X_3, X_4之间条件独立性的图形化表示

图模型有三个基本问题：

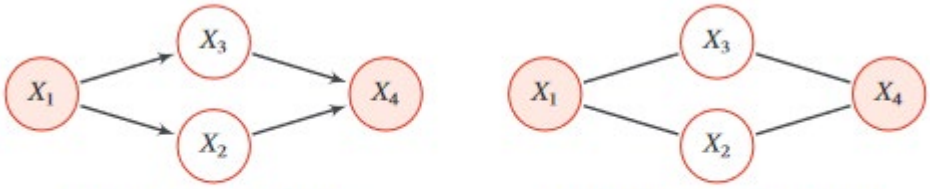
（1） **表示问题**：对于一个概率模型，如何通过图结构来描述变量之间的依赖关系.

（2） **学习问题**：图模型的学习包括图结构的学习和参数的学习．在本章中， 我们只关注在给定图结构时的参数学习，即参数估计问题.

（3） **推断问题**：在已知部分变量时，计算其他变量的条件概率分布.

11.1 模型表示

常见的概率图模型可以分为两类：有向图模型和无向图模型



(a) 有向图:贝叶斯网络　　(b) 无向图:马尔可夫随机场

图 11.2　有向图和无向图示例

11.2 学习

图模型的学习可以分为两部分：一是网络结构学习，即寻找最优的网络结 构；二是网络参数估计，即已知网络结构，估计每个条件概率分布的参数.

11.3 推断

在图模型中，推断（Inference）是指在观测到部分变量 $\boldsymbol{e} = \{e1, e2, \cdots, eM\}$ 时，计算其他变量的某个子集 $\boldsymbol{q} = \{q1, q2, \cdots, qN\}$ 的条件概率 $p(\boldsymbol{q}|\boldsymbol{e})$. 在图模型中，常用的推断算法可以分为**精确推断算法**和**近似推断算法**两类.

11.4 变分推断

变分推断（Variational Inference）是变分法在推断问题中的应用，是寻找一个简单分布 $q*(\boldsymbol{z})$ 来近似条件概率密度 $p(\boldsymbol{z}|\boldsymbol{x})$，也称为变分贝叶斯（Variational Bayesian）. 这样，推断问题转换为一个泛函优化问题

11.5 基于采样法的近似推断

在很多实际机器学习任务中，推断某个概率分布并不是最终目的，而是基于 这个概率分布进一步计算并作出决策．通常这些计算和期望相关．

不失一般性，假设要推断的概率分布为 $p(x)$，并基于 $p(x)$ 来计算函数 $f(x)$ 的期望，当 $p(x)$ 比较复杂或难以精确推断时，我们可以通过采样法来近似计算期望 $\mathbb{E}p[f(x)]$ 的解

第 12 章 深度信念网络

- ❖ 深度信念网络可以有效学习变量之间复杂依赖关系的概率图模型
- ❖ 两种相关基础模型：玻尔兹曼机和受限玻尔兹曼机
- ❖ 深度信念网络中包含很多层的隐变量，可以有效地学习数据的内部特征表示，也可以作为一种有效的非线性降维方法.这些学习到的内部特征表示包含了数据的更高的、有价值的信息，因此十分有助于后续的分类和回归等任务．

1. 玻尔兹曼机和深度信念网络都是生成模型，借助隐变量来描述复杂的数据分布．
2. 作为概率图模型，玻尔兹曼机和深度信念网络的共同问题是推断和学习问题．因为这两种模型都比较复杂，并且都包含隐变量，它们的推断和学习一般通过 MCMC 方法来进行近似估计．
3. 这两种模型和神经网络有很强的对应关系，在 一定程度上也称为随机神经网络（Stochastic Neural Network，SNN）．

12.1 玻尔兹曼机

玻尔兹曼机（Boltzmann Machine）是一个随机动力系统（Stochastic Dynamical System），每个变量的状态都以一定的概率受到其他变量的影响． 玻尔兹曼机可以用概率无向图模型来描述．一个具有 K 个节点（变量）的玻尔兹曼机满足以下三个性质：

（1） 每个随机变量是二值的，所有随机变量可以用一个二值的随机向量$\boldsymbol{X} \in \{0,1\}K$ 来表示，其中可观测变量表示为 $\boldsymbol{V}$ ，隐变量表示为 $\boldsymbol{H}$．

（2） 所有节点之间是全连接的．每个变量 Xi 都依赖于所有其他变量 $\boldsymbol{X}\mathrm{K}i$．

（3） 每两个变量之间的互相影响（$Xi \rightarrow Xj$ 和 $Xj \rightarrow Xi$）是对称的．

12.2 受限玻尔兹曼机

受限玻尔兹曼机（Restricted Boltzmann Machine，RBM）是一个二分图结构的无向图模型，．受限玻尔兹曼机中的变量也分为隐变量和可观 测变量．我们分别用可观测层和隐藏层来表示这两组变量．同一层中的节点之间 没有连接，而不同层一个层中的节点与另一层中的所有节点连接，这和两层的全 连接神经网络的结构相同．

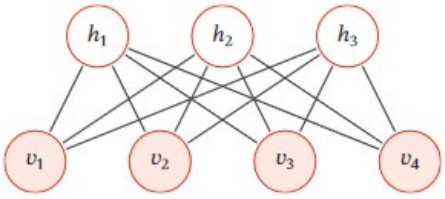


图 12.3　一个有 7 个变量的受限玻尔兹曼机

12.3 深度信念网络

深度信念网络（Deep Belief Network，DBN）是一种深层的概率有向图模 型，其图结构由多层的节点构成．和全连接的前馈神经网络结构相同．每层节点的内部没有连接，相邻两层的节点之 间为全连接．网络的最底层为可观测变量，其他层节点都为隐变量．最顶部的两 层间的连接是无向的，其他层之间的连接是有向的．

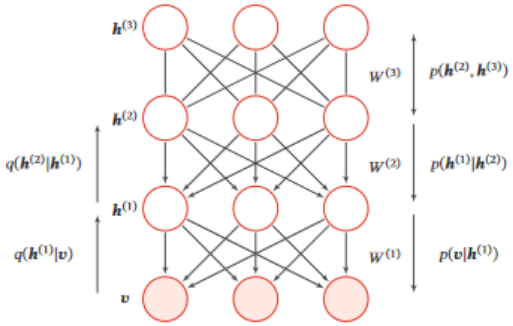


图 12.5　一个有 4 层结构的深度信念网络

第 13 章 深度生成模型

概率生成模型（Probabilistic Generative Model），简称生成模型，是概率统计和机器学习领域的一类重要模型，指一系列用于随机生成可观测数据的模型． 假设在一个连续或离散的高维空间 $\mathcal{X}$ 中，存在一个随机向量 $\boldsymbol{X}$ 服从一个未知的数据分布 $pr(\boldsymbol{x}), \boldsymbol{x} \in \mathcal{X}$．生成模型是根据一些可观测的样本 $\boldsymbol{x}(1), \boldsymbol{x}(2), \cdots, \boldsymbol{x}(N)$ 来 学习一个参数化的模型 $p\theta(\boldsymbol{x})$ 来近似未知分布 $pr(\boldsymbol{x})$，并可以用这个模型来生成 一些样本，使得“生成”的样本和“真实”的样本尽可能地相似．生成模型通常包含两个基本功能：概率密度估计和生成样本（即采样）．图 13.1 以手写体数字图像 为例给出了生成模型的两个功能示例，其中左图表示手写体数字图像的真实分 布 $pr(\boldsymbol{x})$ 以及从中采样的一些“真实”样本，右图表示估计出了分布 $p\theta(\boldsymbol{x})$ 以及从 中采样的“生成”样本．

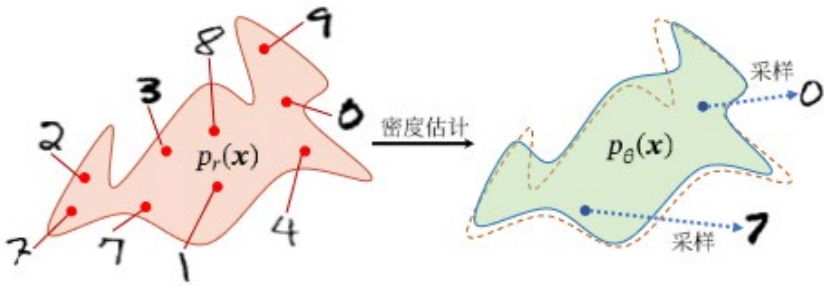


图 13.1　生成模型的两个功能

13.1 概率生成模型

生成模型一般具有两个基本功能：密度估计和生成样本.

13.2 变分自编码器

变分自编码器（Variational AutoEncoder，VAE）[Kingma et al., 2014] 是一种深度生成模型，其思想是利用神经网络来分别建模两个复杂的条件概率密度 函数.

（1） 用神经网络来估计变分分布 $q(\boldsymbol{z};\boldsymbol{\phi})$，称为推断网络. 理论上 $q(\boldsymbol{z};\boldsymbol{\phi})$ 可 以不依赖 $\boldsymbol{x}$. 但由于 $q(\boldsymbol{z};\boldsymbol{\phi})$ 的目标是近似后验分布 $p(\boldsymbol{z}|\boldsymbol{x};\theta)$，其和 $\boldsymbol{x}$ 相关，因此变分密度函数一般写为 $q(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\phi})$. 推断网络的输入为 $\boldsymbol{x}$，输出为变分分布 $q(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\phi})$.

（2） 用神经网络来估计概率分布 $p(\boldsymbol{x}|\boldsymbol{z};\theta)$，称为生成网络. 生成网络的输入为 $\boldsymbol{z}$，输出为概率分布 $p(\boldsymbol{x}|\boldsymbol{z};\theta)$.

13.3 生成对抗网络

生成对抗网络（Generative Adversarial Networks，GAN）是通过对抗训练的方式来使得生成网络产生的样本服从真实数据分布. 在生成对抗网络中，有两个网络进行对抗训练：

- 一个是**判别网络**，目标是尽 量准确地判断一个样本是来自于真实数据还是由生成网络产生；
- 另一个是**生成网络**，目标是尽量生成判别网络无法区分来源的样本. 这两个目标相反的网络不 断地进行交替训练. 当最后收敛时，如果判别网络再也无法判断出一个样本的来 源，那么也就等价于生成网络可以生成符合真实数据分布的样本.

第 14 章 深度强化学习

（1） 强化学习（Reinforcement Learning，RL），也叫增强学习，是指一类从（与环境）交互中不断学习的问题以及解决这类问题的方法.

（2） 强化学习问题可以描述为一个智能体从与环境的交互中不断学习以完成特定目标（比如取得最大奖励值）.

（3） 和深度学习类似，强化学习中的关键问题也是贡献度分配问题，每一个动作并不能直接得到监督信息，需要通过整个模型的最终监督信息（奖励）得到，并且有一定的延

时性.

14.1 强化学习问题

在强化学习中，有两个可以进行交互的对象：智能体和环境：

- **智能体**（Agent）可以感知外界环境的状态（State）和反馈的奖励 （Reward），并进行学习和决策. 智能体的决策功能是指根据外界环境的状态来做出不同的动作（Action），而学习功能是指根据外界环境的奖励来调整策略.
- **环境**（Environment）是智能体外部的所有事物，并受智能体动作的影响而改变其状态，并反馈给智能体相应的奖励.

强化学习的基本要素包括：

- 状态s 是对环境的描述，可以是离散的或连续的，其状态空间为 $\mathcal{S}$.
- 动作a 是对智能体行为的描述，可以是离散的或连续的，其动作空间为 $\mathcal{A}$.
- 策略$\pi(a|s)$ 是智能体根据环境状态 s 来决定下一步动作 a 的函数.
- 状态转移概率$p(s'|s,a)$ 是在智能体根据当前状态 s 做出一个动作 a 之 后，环境在下一个时刻转变为状态 s' 的概率.
- 即时奖励$r(s,a,s')$ 是一个标量函数，即智能体根据当前状态 s 做出动作 a 之后，环境会反馈给智能体一个奖励，这个奖励也经常和下一个时刻的状态 s' 有关.

策略

智能体的策略（Policy）就是智能体如何根据环境状态 s 来决定下一步的 动作 a，通常可以分为确定性策略（Deterministic Policy）和随机性策略（Stochastic Policy）两种

14.2 基于值函数的学习方法

- 值函数是对策略 π 的评估. 如果策略 π 有限（即状态数和动作数都有限）， 可以对所有的策略进行评估并选出最优策略$\pi*$.
- 基于值函数的策略学习方法中最关键的是如何计算策略 π 的值函数，一般有动态规划或蒙特卡罗两种计算方式.

14.3 基于策略函数的学习方法

- 强化学习的目标是学习到一个策略$\pi\theta(a|s)$ 来最大化期望回报．一种直接的方法是在策略空间直接搜索来得到最佳策略，称为策略搜索（Policy Search）.
- 策略搜索本质是一个优化问题，可以分为基于梯度的优化和无梯度优化.
- 策略搜索和基于值函数的方法相比，策略搜索可以不需要值函数，直接优化策略．参数化的策略能够处理连续状态和动作，可以直接学出随机性策略.

14.4 演员-评论员算法

演员-评论员算法（Actor-Critic Algorithm）是一种结合策略梯度和时序差分学习的强化学习方法.

- 其中演员（Actor）是指策略函数 $\pi\theta(a|s)$，即学习一个策略来得到尽量高的回报
- 评论员（Critic）是指值函数 $V\phi(s)$，对当前策略的值函数进行估计，即评估演员的好坏.

借助于值函数，演员-评论员算法可以进行单步更新参数，不需要等到回合结束才进行更新.

第 15 章 序列生成模型

在深度学习的应用中，有很多数据是以序列的形式存在，比如声音、语言、视频、DNA 序列或者其他的时序数据等. 以自然语言为例，一个句子可以看作符合一定自然语言规则的词（word）的序列.

将一个长度为 T 的文本序列看作一个随机事件，每个位置上的变量 Xt 的样本空间为一个给定的词表，则一个文本序列的概率大小可以用来评估它符合自然语言规则的程度.

序列概率模型有两个基本问题：

（1） 概率密度估计： 给定一组序列数据，估计这些数据背后的概率分布

（2） 样本生成：从已知的序列 分布中生成新的序列样本

序列数据一般可以通过概率图模型来建模序列中不同变量之间的依赖关系. 本章主要介绍在序列数据上经常使用的一种模型：自回归生成模型（AutoRegressive Generative Model）.

15.1 序列概率模型

在序列模型方式中，每一步都需要将前面的输出作为当前步的输入，是一种自回归（AutoRegressive）的方式. 因此这一类模型也称为自回归生成模型 （AutoRegressive Generative Model）. 主要介绍两种比较主流的自回归生成模型：N 元统计模型和深度序列模型.

15.2 N 元统计模型

由于数据稀疏问题，当 t 比较大时，依然很难估计条件概率 $p(xt|\boldsymbol{x}1：(t-1))$. 一个简化的方法是 N 元模型（N-Gram Model）

15.3 深度序列模型

深度序列模型（Deep Sequence Model）是指利用神经网络模型来估计条件概率 $p\theta(xt|\boldsymbol{x}1：$

$(t-1)$).

15.4 评价方法

- 困惑度（Perplexity）是信息论中的一个概念，可以用来衡量一个分布的不确定性.
- BLEU（BiLingual Evaluation Understudy）算法是一种衡量模型生成序列和参考序列之间的 N 元词组（N-Gram）重合度的算法
- ROUGE（Recall-Oriented Understudy for Gisting Evaluation）算法最早应用于文本摘要领域. 和 BLEU 算法类似，但 ROUGE 算法计算的是召回率（Recall）.

15.5 序列生成模型中的学习问题

使用最大似然估计来学习自回归序列生成模型时，会存在以下三个主要问题：

- 曝光偏差问题
- 训练目标不一致问题
- 计算效率问题

15.6 序列到序列模型

序列到序列（Sequence-to-Sequence，Seq2Seq）是一种条件的序列生成问题，给定一个序列 $\boldsymbol{x}1:S$，生成另一个序列 $\boldsymbol{y}1:T$. 输入序列的长度 S 和输出序列的长度 T 可以不同. 比如在机器翻译中，输入为源语言，输出为目标语言.

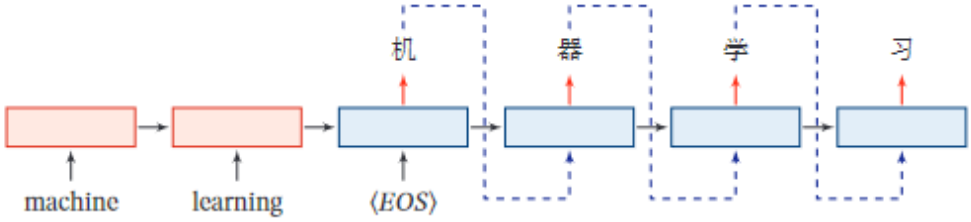


图 15.5 基于循环神经网络的序列到序列机器翻译