第一章

（ 深度学习 ）是机器学习的一个分支。

在深度学习中，（ 贡献度分配 ）是一个很关键的问题

一个较好地解决贡献度分配问题的模型是（人工神经网络 ）。

复杂神经网络的学习可以看成是一种深度的机器学习，即（ 深度学习 ）。

所谓“**深度**”是指（ 原始数据进行非线性特征变换的次数 ）

深度学习需要解决的关键问题：（ 贡献度分配问题 ），即一个系统中不同的组件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（ 一种端到端学习 ）

深度学习采用的模型主要是（ 神经网络模型 ）**：**

目前，大部分采用神经网络模型的深度学习可以看作一种（端到端学习）

从系统观点看，人工神经网络是由大量神经元通过极其丰富和完善的连接构成的（ 自适应非线性系统 ）

1980年代，（ ）才有效地解决了多层神经网络的学习问题，并称为最流行的神经网络学习算法

第二章

预测连续值：（ ）问题； 预测离散值：（ ）问题

**数据集**(data set)

一组样本构成的集合，数据集分为训练集D和测试集D′

**训练集**(training set)  
用来训练模型的样本集合，也叫训练样本(training sample)  
**测试集**(test set)

用来检验模型好坏的样本集合，也叫测试样本(test sample)

**独立同分布：**样本的选取需满足独立同分布，即（ 独立地 ）从（ 相同数据分布 ）中抽取！

机器学习的三个基本要素：（ 学习准则 ）**、**（ 优化算法 ）**、**（ 模型 ）

f(x;θ)是参数为θ的函数，也称为（ 模型 ），D为参数的数量

模型f(x;θ)的好坏可以通过（ 期望风险 ）R(θ)来衡量  
损失函数是一个（ 非负 ）的函数，用来（ 量化模型预测和真实数据之间的差异 ）

0-1损失函数，经常用（ 连续可微 ）的损失函数替代。

平方损失函数，常用于（ 回归任务 ）（预测（ 预测标签y为连续值 ）的任务）。

交叉熵损失函数一般用于（ 离散任务 ）。

**经验风险最小化(ERM)准则**：一个可行的的学习准则是找到一组参数θ^∗使得经验风险最小当训练集大小趋向于无穷大时，经验风险就趋向于期望风险

**过拟合**：经验风险最小化准则很容易导致（ 过拟合 ），**模型在训练集上错误率很低，但是在未知数据上错误率很高**

过拟合问题造成原因：（ 训练数据过少 ）和（ 噪声 ）以及（ 模型能力过强 ）等

**欠拟合**(underfitting)：

模型不能很好地拟合训练数据，在训练集上的错误率较高；

欠拟合一般是由于模型能力不足造成的

（ 泛化 ）用于衡量模型对未见过数据样本的预测能力。

不管训练集上表现如何，只有在新的未知的样本集上有较好的表现，模型才是真的好，这就是模型的泛化能力。

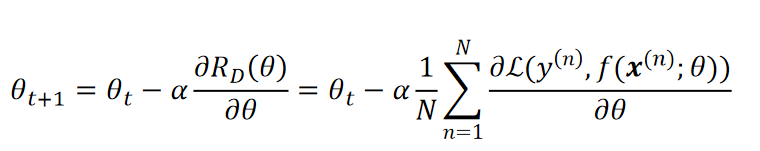
在确定了训练集D、假设空间F以及学习准则后，如何找到最优的模型f(x;θ^∗)就成了一个（ 最优化 ）问题

**参数与超参数**：在机器学习中，优化又可以分为参数优化和超参数优化

**梯度下降法**

在机器学习中，最简单、常用的优化算法是梯度下降法。

首先初始化参数θ\_0，然后按下面的迭代公式来计算训练集D上风险函数的最小值



a为搜索步长，一般称为学习率(learning rate)。

上式的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(**BGD)。

针对梯度下降的优化算法，除了加正则化项之外，还可以（ 提前停止来防止过拟合 ）。

除了训练集和测试集之外，有时也会使用一个（ 验证集 ）来进行模型选择，测试模型在验证集上是否最优。

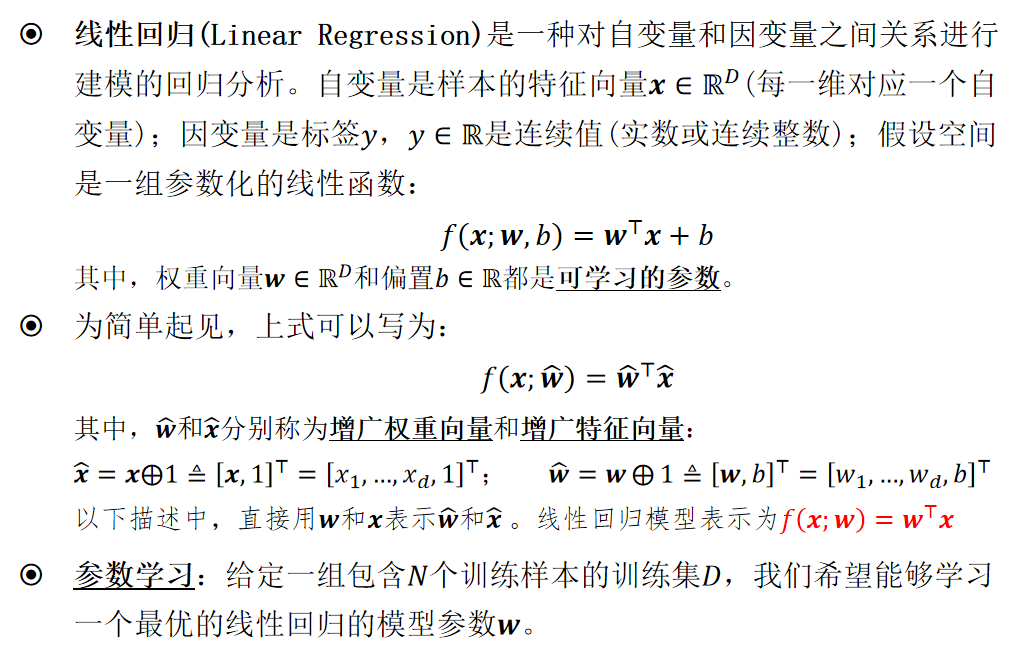
数据集划分一般按（**训练集:验证集:测试集=**（ 7：2：1 ））划分

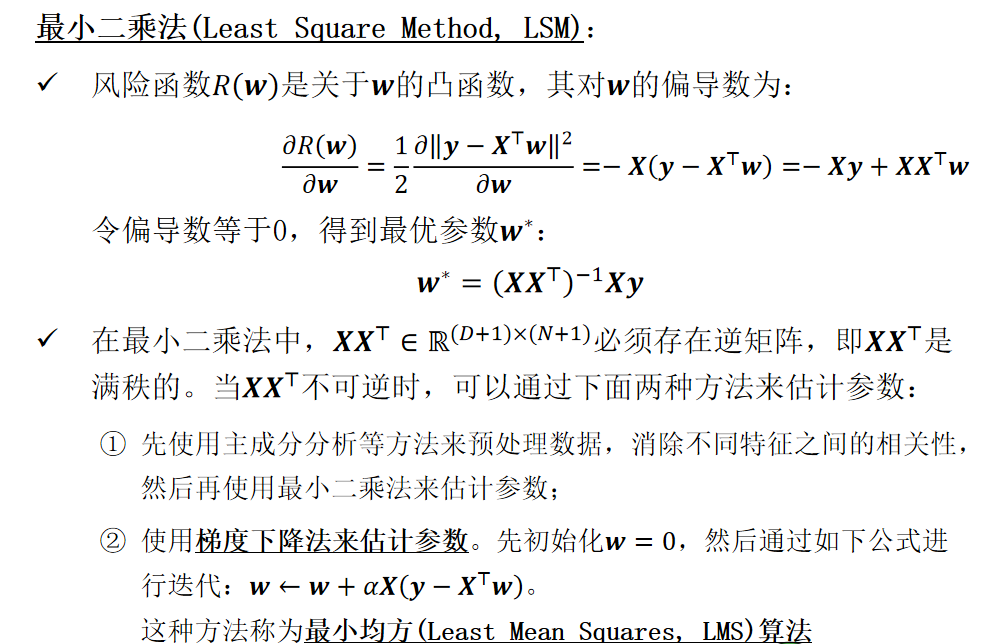
**提前停止**(early stop)策略：  
在每次迭代时，把新得到的模型f(x;θ)在验证集上进行测试，并计算错误率。如果在验证集上的错误率不再下降，就停止迭代。

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 ）

**小批量梯度下降法**是批量梯度下降和随机梯度下降的折中

在实际应用中，小批量随机梯度下降法有收敛快、计算开销小的优点，因此逐渐成为**大规模的机器学习中的**（ 主要优化算法 ）





（ 最大似然估计 ）是指找到一组参数w∗使得似然函数最大，等价于对数似然函数最大。

（ 最大后验估计 ）是指最优参数为后验分布p(w│X,y;v,σ)中概率密度最高的参数

为了避免过拟合，经常在模型的（ 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比较高，容易导致过拟合；

如果限制模型的复杂度，降低其拟合能力，又可能导致欠拟合。

在（ 模型的拟合能力和复杂度之间取得较好的平衡 ）对ML算法十分重要

模型的拟合能力和复杂度之间进行权衡的分析和指导工具：**偏差-方差分解**

**偏差**(bias)：  
一个模型在不同训练集上的平均性能和最优模型的差异。

**方差**(variance)：  
一个模型在不同训练集上的差异，用来衡量一个模型是否容易过拟合。

**最小期望错误等价于**（ 最小化偏差和方差之和 ）

**机器学习模型的分析**：

当一个**模型在训练集上的错误率比较高**时，说明模型的拟合能力不够(**模型欠拟合**)，**偏差比较高**。  
 通过增加数据特征、提高模型复杂度、减小正则化系数等操作来改进。

当**模型在训练集上的错误率比较低，但验证集上的错误率比较高**时，说明**模型过拟合**，**方差比较高**。  
 方差一般会随着训练样本的增加而减少。  
 通过降低模型复杂度、加大正则化系数、引入先验等方法来缓解。  
 集成模型：通过多个高方差模型的平均来降低方差。

按照**函数f(x;θ)**的不同，可以分为（ 线性模型 ）和（ 非线性模型 ）；

按照**学习准则**的不同，可以分为（ 统计方法 ）和（ 非统计方法 ）；

按照**训练样本提供的信息以及反馈方式**的不同，可以分为（ 监督学习 ）、（ 无监督学习 ）和（ 强化学习 ）。

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**(regression)问题中的标签y是连续值,f(x;θ)的输出也是连续值。

**分类**问题中的标签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。分类问题根据其类别数量又可分为二分类和多分类问题

**结构化学习**(structured learning)问题是一种特殊的分类问题。在结构化学习中，标签y通常是结构化对象，如序列、树或图等

监督学习通常需要大量的有标签数据集，这些数据集一般都需要由人工进行标注，成本很高。

因此，出现了很多（ 弱监督学习 ）**和**（ 半监督学习 ）的方法

**无监督学习**(Unsupervised Learning，UL)

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

**强化学习(**Reinforcement Learning，RL)

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之处在于，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

**评价指标**

1. **准确率**
2. **错误率**
3. **精确率**(precision)和**召回率**(recall)

真正率(True Positive，TP)假负率(False Negative，FN)

假正率(False Positive，FP)真负率(True Negative，TN)

**精确率**(precision)也叫**精度**或**查准率**

**召回率**(recall)也叫**查全率**

**F值**(F measure)是综合指标，为精确率和召回率的调和平均

**宏平均**(Macro Average)和**微平均**

为了计算分类算法在所有类别上的总体精确率、召回率和F1值，经常使用两种平均方法，分别称为宏平均和微平均。

宏平均是每一类的性能指标的算术平均值

**交叉验证**

**线性模型**(Linear Model)：  
通过样本特征的线性组合来进行预测的模型。

**增广形式：**

线性回归可以直接用f(x;w)来预测输出目标y=f(x;w)。而线性分类不能直接用f(x;w)来预测输出目标y=f(x;w)

在分类问题中，输出目标y是一些离散的标签，而f(x;w)的值域为实数；

需要引入一个（非线性决策函数）g(∙)来预测输出目标，其中f(x;w)也称为（ 判别函数 ）

对于二分类问题，g(∙)可以是（ 符号函数 ）

**Logistic回归、Softmax回归、感知器，**这些模型的区别主要在于使用了（ 不同的目标函数 ）

（ 二分类 ）问题的类别标签y只有两种取值，通常可以设为+1,−1或0,1

特征空间R^D中所有满足f(x;w)=0的点组成一个（ 特征向量 ）**，**称为（ ）

所谓**“线性分类模型”**就是指其**决策边界是**（ 超平面 ）**。**

在特征空间中，**决策平面与权重向量w正交。**

特征空间中**每个样本到决策平面的**（ 有向距离 ）

特征空间中**原点到决策平面的有向距离**

**Logistic回归(Logistic Regression，LR)**是一种常用的处理二分类问题的线性模型

非线性函数g(∙)通常称为（ 激活函数 ）

**Softmax回归：**也称为多项或多类的Logistic回归，是Logistic回归在多分类问题上的推广。

**感知器**由Frank Roseblatt于1957年提出，是一种广泛使用的线性分类器。

**感知器的收敛性**

对于两类问题，如果训练集是线性可分的，那么感知器算法可以在有限次迭代后收敛。如果训练集不是线性可分的，那么感知器算法则不能确保会收敛。

**投票感知器**

**平均感知器**

为了提高迭代速度，让更新只需要（ 在错误预测发生时 ）**。**

第三章

人工神经网络常常简称为（ 神经网络 ）或（ 神经模型 ）

神经网络最早是作为一种主要的连接主义模型使用。最流行的连接主义模型是（ 分布式并行处理 ）

**人工神经元，**简称**神经元**

**激活函数**在神经元中是非常重要的，为了（ 增强网络的表示能力和学习能力 ）

Tanh函数的输出是（ 零中心化的 ）

**非零中心化的输出**会使其后一层的神经元的输入发生（偏置偏移 ）

（ 前馈神经网络 ）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 ）

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**。

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

通用近似定理在**实数空间ℝ^D中的**（ 有界闭集 ）上依然成立

**三层神经网络可以近似任意连续函数**

所谓**“挤压”性质的函数**是指像Sigmoid函数的有界函数，但神经网络的通用近似性质被证明对于其他类型的激活函数(如ReLU)也是适用的

**多层前馈神经网络也可以看成是一种特征转换方法**，其输出ϕ(x)作为分类器的输入进行分类

BP神经网络算法中（信息的正向传播 ）实现分类功能；（ 误差的反向传播 ）调整网络权值。

BP算法的实现需要（ 输入信号前馈 ）和（ 误差信号反向传递 ）两个过程

自动梯度计算的方法可以分为以下3类：  
**数值微分、符号微分和自动微分**。

将复合函数f(x;w,b)分解为一系列的基本操作，并构成一个（ 计算图 ）

按照计算导数的顺序，自动微分可以分为（ 前向模式 ）和（ 反向模型 ）。

计算图按构建方式可以分为（ 静态计算图 ）和（ 动态计算图 ）

神经网络的参数学习比线性模型更加困难，主要原因有两点：  
**非凸优化问题和梯度消失问题**。

**非凸优化问题**

第一种损失函数为（ 平方误差损失 ）：ℒ(w\_1,w\_2)=(1−y^2)；  
第二种损失函数为（ 交叉熵损失 ）： ℒ(w\_1,w\_2)=logy。

**梯度消失问题**/**梯度弥散问题**：  
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。

第四章

适合计算机视觉的神经网络结构需具有（ 平移不变性 ）和（ 局部性 ）特点



上述公式表示一个卷积层。V被称为（ 卷积核 ）**或者**（ 滤波器 ）

两个（ 连续 ）函数(比如f,g: ℝ^d→ℝ)之间的卷积被定义为



当有（ 离散 ）对象时(即定义域为ℤ)，积分就变成求和，得到以下定义



由于输入图像是三维的，隐藏表示H也是一个三维张量。  
因此，可以把隐藏表示想象为一系列具有二维张量的（ 通道 ），这些通道有时也被称为（ 特征图/特征映射 ）

输出的卷积层有时被称为（ 特征映射/特征图 ），因为它可以被视为一个输入映射到下一层的空间维度的转换器

在CNN中，对于某一层的任意元素x，其（ 感受野 ）是指在前向传播期间可能影响x计算的所有元素(来自所有先前层)

当一个特征图中的任意元素需要检测更广区域的输入特征时，可以构建一个（更深的网络 ）

（ ）和（ ）也影响卷积的输出形状

在许多情况下，需要设置（ ）和（ ），使输入和输出具有相同的高度和宽度

（ ）：每次滑动元素的数量

池运算符是确定性的，通常计算池化窗口中所有元素的最大值或平均值。这些操作分别称为（ ）和（ ）

计算机视觉研究人员认为推动领域进步的是（ ），而不是学习算法。

计算机视觉研究人员相信，从对最终模型精度的影响来说，更大或更干净的（ ）、或是稍微改进的（ ），比任何学习算法带来的进步要大得多

（ NiN ）完全取消了全连接层。相反，NiN使用一个NiN块，其输出通道数等于标签类别的数量。最后放一个全局（ 平均池化层 ），生成一个对数几率

（ 批量归一化 ）是一种流行且有效的技术，可持续加速深层网络的收敛速度

对于深度神经网络，如果能将新添加的层训练成（ 恒等映射 ）

何恺明等人提出了（ 残差网络 ），并深刻影响了后来的深度神经网络的设计

**残差网络的核心思想**是：（ 每个附加层都应该更容易地包含原始函数作为其元素之一 ）

（ 残差块 ）拟合残差映射f(x)−x。

DenseNet主要由两部分构成：  
（ 稠密块 ） - 定义如何连接输入和输出；  
（ 过渡层 ） - 控制通道数量，使其不会太复杂