10 选 40%

8 计t问 60%

2 6

概念.是什么

概念.计算方式

概念.效果.好处

第二讲 机器学习基础

选择

考概念

《深度学习》

南开大学 人工智能学院

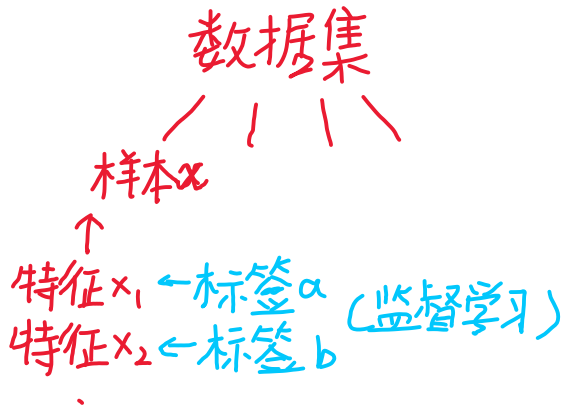
2 机器学习

- 机器学习是不显式编程地赋予计算机能力的研究领域
- 核心：具有自学习能力
  - 利用经验提升系统对某种任务的处理能力
  - 经验以数据形式存储
- 本质：通过数据产生模型
  - 学习到的模型在面对经验中没有出现的情况时，也能提供相应的判断或预测
- 机器学习问题分类
  - 监督学习
  - 无监督学习
  - 强化学习

2 机器学习

- 机器学习使用流程
  - 数据准备和预处理
  - 模型构建
  - 模型训练
  - 测试与泛化

2.1 数据准备和预处理



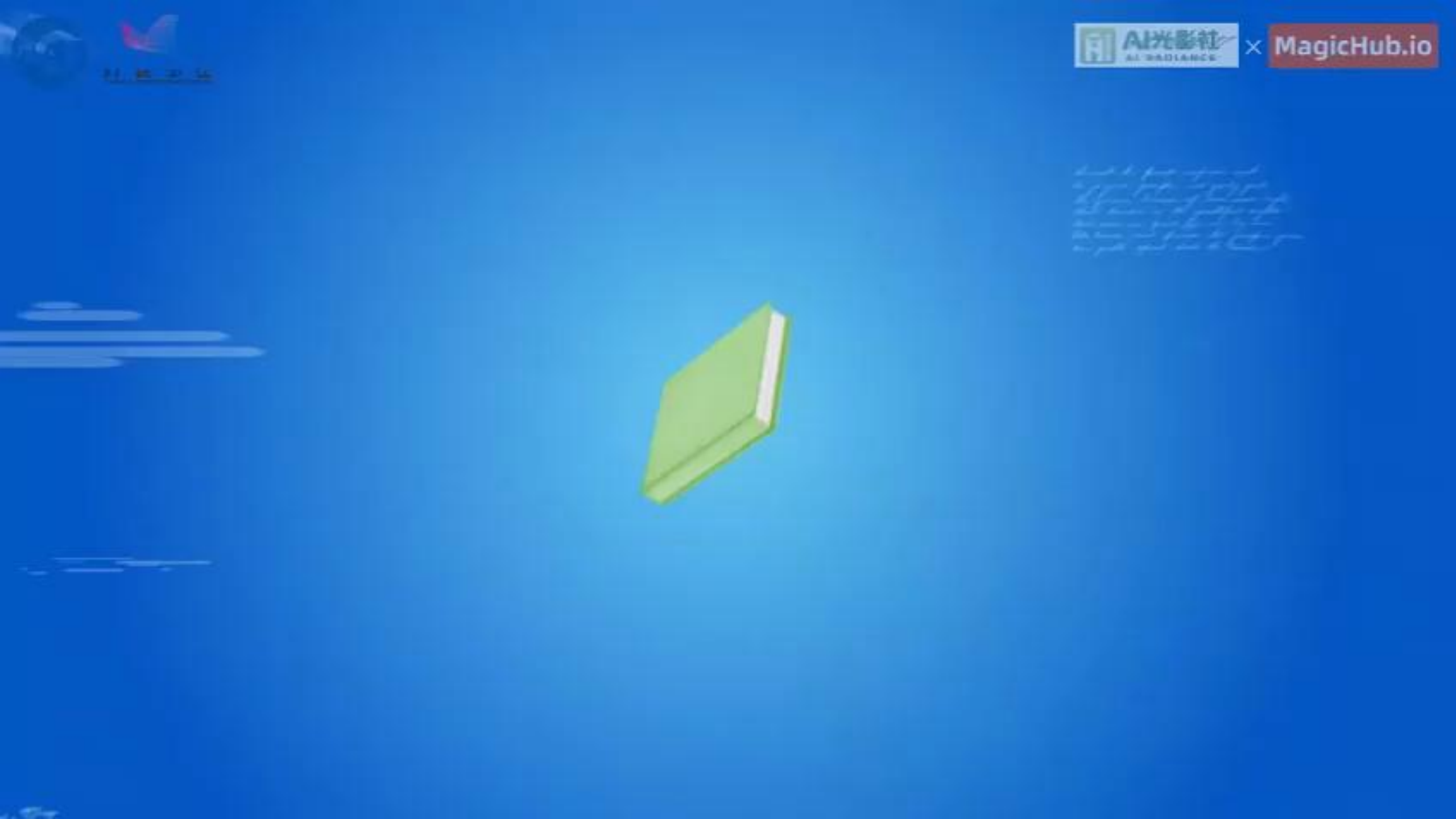
- 经验来源于数据
  - **数据集**由一个个**样本**组成
  - 每个样本由一组称为**特征**的属性组成，通常将一个样本表示成一个向量$x \in \mathbb{R}^n$，其中向量的每一个元素x_i是一个**特征**
    - 例如，一张图片的特征通常是指这张图片的像素值；医疗患者的特征往往是一组标准的生命特征（如年龄、生命体征和检查结果）
  - 在监督学习问题中，要预测的是一个特殊的属性，被称为**标签** 人为标注
  - 在监督学习中，样本为（特征，标签）

2.1 数据准备和预处理

收集 → 清洗 → 特征工程 → 划分 { 训练集 60-80% / 验证集 10-20% / 测试集 10-20% }

特征工程：标准化（统一量纲）、特征选择（去除无关）

- 数据准备与预处理主要包括:
  - 数据收集
    - 需要注意数据的代表性和多样性，来保证模型在未知数据上的表现，即泛化能力
  - 数据清洗
    - 处理缺失值、异常值、噪声、重复数据
  - 特征工程
    - 标准化：为了消除特征之间量纲不同导致的取值范围差异的影响，将数据按比例缩放，使其落入一个特定的区域
    - 特征选择：去掉某些对输出结果没有影响的特征，减少数据的维度
  - 数据集划分
    - 训练集：用于训练模型，占总数据集的60%-80%
    - 验证集：用于模型选择（比如决定哪个模型更好）和模型超参数调整，占总数据集的10%-20%
    - 测试集：在整个训练过程中未曾见过的数据，用于评估模型的效果，占总数据集的10%-20%



2.2 模型构建

- 在数据准备完成后，需要选择合适的模型
  - 模型的具体形式取决于我们选择的学习算法及超参数配置
  - 线型回归、决策树、SVM、神经网络……

$$-\frac{1}{n}\sum_{i=1}^{n}\left(y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)\right)$$

- 选择损失函数
  - 性能度量：来衡量模型的预测值与对应样本的真实标签之间的差别大小
  - 常用损失函数举例
    - **均方误差：**通过计算模型预测值$\hat{y}$与真实值y之间的平方差来衡量误差，也称为平方L_2损失

$$L_2 \text{ loss} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

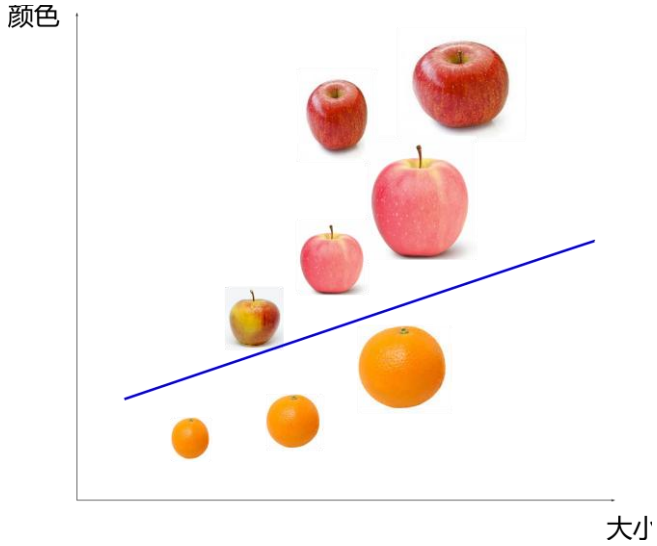
    - **交叉熵误差：**常用于分类问题，用于衡量模型预测的概率分布（$\hat{y}$）与真实标签的概率分布（y）之间的差异。对于二分类问题，交叉熵损失函数表示为：

$$\text{Binary Cross Entropy loss} = -\frac{1}{n}\sum_{i=1}^{n}\left(y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)\right)$$

其中 $y_i = 1$ 或 $y_i = 0$, $1 \geq \hat{y}_i \geq 0$

2.2.1 参数

- 什么是参数？
  - 参数可以被看作旋钮，旋钮的转动可以调整模型的行为
  - 举例：神经网络的权重、偏置
  - 在一个数据集上，可以通过最小化损失函数来学习模型参数的最佳值，找到最合适的模型



模型：

$$Y = A * \text{颜色} + B * \text{大小} + C$$

参数：

(A, B, C)

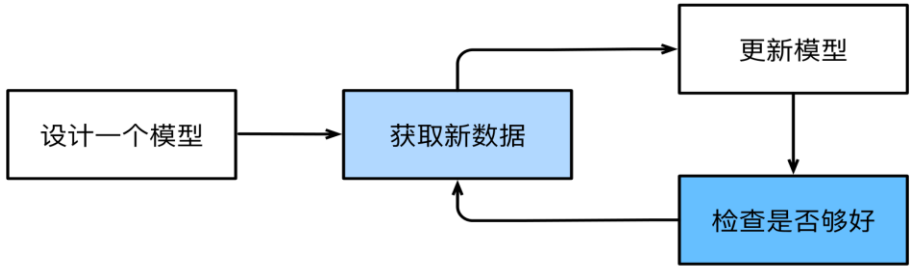
2.2.2 超参数

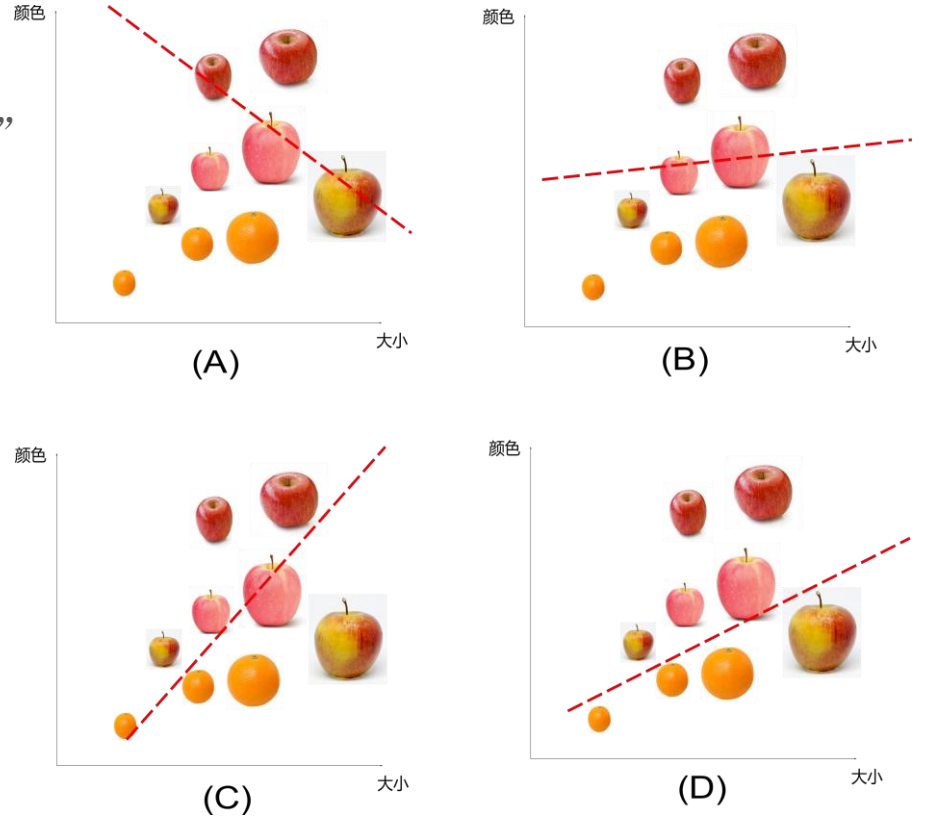
- 超参数是用于管理机器学习模型的外部配置变量
  - 举例：网络层数、宽度、学习率等
  - 超参数会决定模型架构和模型复杂性，比如神经网络中的层数和每一层的节点数
  - 超参数需要在训练前进行手动配置
    - 超参数的值不是通过学习算法学习出来的
    - 深度学习领域常说的“调参”，就是指超参数调优
  - 超参数调优可以手动进行，也可以使用自动算法完成
    - 验证集：用于挑选超参数的数据子集
    - 原则上，测试样本不能以任何形式参与到模型的选择中，包括设定超参数

测试集

2.3 模型训练

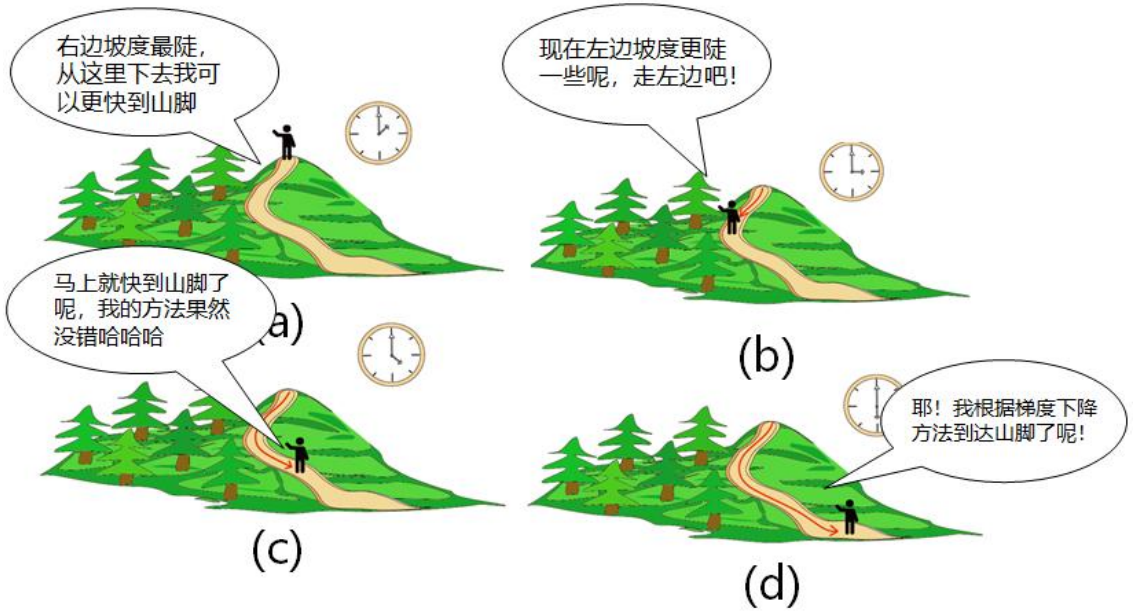
- 从数据中进行学习，获得模型最优参数的过程称为训练
  - 经过训练，我们可以发现合适的参数集，从而使模型执行所需的行为
- 训练过程通常包含如下步骤：
  - 从一个参数随机初始化的模型开始，这个模型基本没有“智能”
  - 获取一些数据样本（例如，图片以及对应的标签）
  - 调整参数，使模型在这些样本中表现得更好
  - 重复第2步和第3步，直到模型在任务中的表现令人满意





2.3.1 优化算法

- 优化算法：搜索最佳模型参数，以最小化损失函数
  - 优化算法通过不断调整模型的参数，使得在训练数据上的损失函数值最小
  - 深度学习中，大多流行的优化算法通常基于一种基本策略——梯度下降
    - 使用梯度下降的过程就像是在一个山谷中寻找最低点，每次都沿着最陡峭的方向（即梯度）走一步，直到到达梯度几乎为0的最低点
    - 深度学习中常用的优化算法：SGD、Adam



2.4 测试与泛化

- 在训练数据上表现良好的模型，不一定在测试数据上有同样的性能
- “一个模型在测试数据集上的性能”可以想象成“一个学生在期末考试中的分数”
- “一个模型在训练数据集上的性能”可以想象成“一个学生在模拟考试中的分数”
  - 可以用模拟考成绩作为期末考试的参考，但可能有较大偏差
  - 模拟考试考得好，期末考试不一定考得好
- 当一个模型在训练集上表现良好，但不能推广到测试集时，这个模型被称为**过拟合**
  - 机器学习的目标：从训练样本中学习适用于所有潜在样本的普遍规律
  - 过拟合：模型过于关注训练数据，把训练数据本身的一些特点当做了所有潜在样本都会具有的一般性质
  - 极端例子：只要输入出现在了训练集中，就把训练集中的标签作为对应输出；如果输入没有出现在训练集中，就输出一个随机值
    - 该模型在训练数据集上的损失为0，但没有意义
    - 仅仅是记住了训练集，没有挖掘训练集中的规律和模式

2.4.1 泛化

- 机器学习的主要挑战：算法必须能够在未观测的新输入上表现良好，而不只是在训练集上表现良好。在未观测到的输入上表现良好的能力称为**泛化**能力
- **训练误差**：模型在训练数据集上计算得到的误差
  - 训练集上的均方误差：

$$\mathrm{MSE}_{train} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i^{train} - \hat{\mathbf{y}}_i^{train})^2$$

- **泛化误差/测试误差**：模型应用在从原始样本的分布中抽取的无限多数据样本时，模型误差的期望
  - 无限多样本上的均方误差：

$$\mathrm{MSE}_{test} = \mathbb{E}_{y^{test}}\left[(y^{test} - \hat{y}^{test})^2\right] \approx \frac{1}{m}\sum_{i=1}^{m}(\mathbf{y}_i^{test} - \hat{\mathbf{y}}_i^{test})^2$$

  - 无法计算，需要遍历无穷多样本
  - 在随机选取的、未曾在训练集中出现的测试集样本上估计

2.4.2 独立同分布

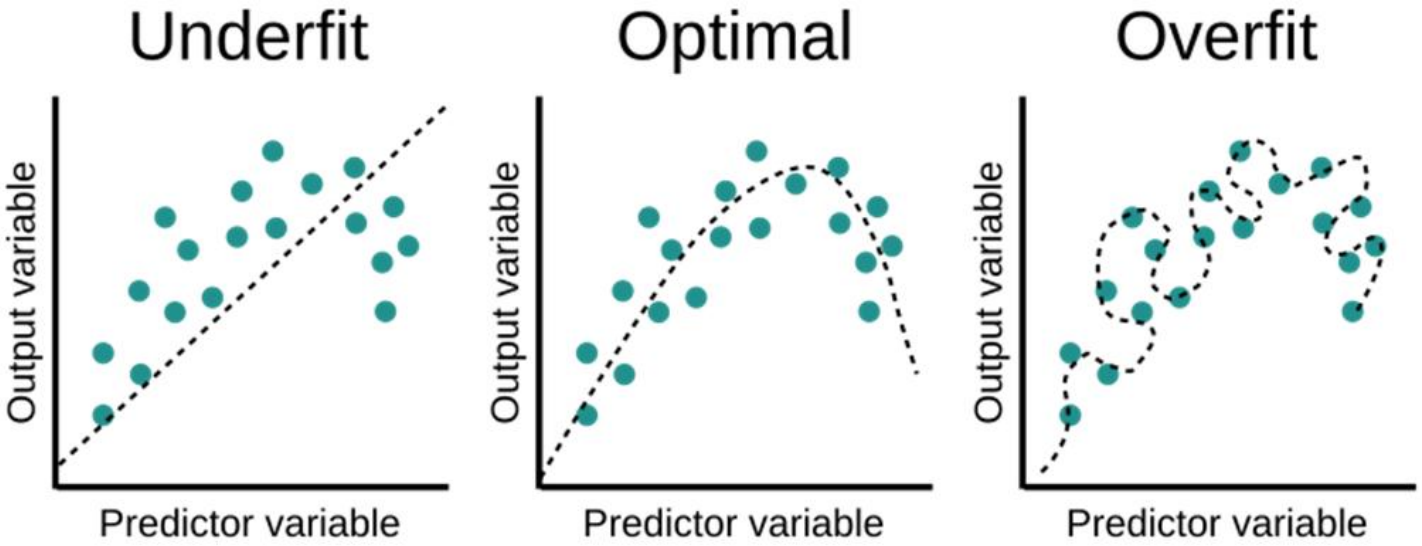
- 困难：当我们训练模型时，我们能访问的训练样本只是数据中的小部分样本
  - 当只能观测到有限的训练集时，如何才能保证测试集的性能呢？
- **独立同分布假设**
  1. 数据集中的每个样本都是彼此相互独立的
  2. 训练集和测试集是同分布的，即采样自相同的分布
     - 相同的分布用来生成每一个训练样本和每一个测试样本

数据集中的样本相互独立
训练集与测试集同分布

2.4.3 过拟合、欠拟合

- 决定机器学习模型效果的因素：
  - 能否降低训练误差，得到最佳参数
  - 能否缩小训练误差和测试误差的差距
- 机器学习的两个主要挑战：**欠拟合**和**过拟合**
  - 欠拟合是指模型不能在训练集上获得足够低的训练误差，对训练样本的一般性质都没有学到
  - 过拟合是指训练误差和和测试误差之间的差距太大
- 示例：学生利用往年的考试题目来辅助学习
  - 欠拟合：学生没有好好学习，在往年的考试题目上成绩糟糕
  - 过拟合：学生可能试图通过死记硬背考题答案来学习，他甚至可以完全记住往年考试的答案，但是在即将到来的考试中成绩不理想
- 机器学习的目标：发现某些模式，这些模式捕捉到了训练集潜在的规律，而不是记住了训练集
  - 避免欠拟合和过拟合

2.4.3 过拟合、欠拟合



线性函数　　二次函数　　高阶的多项式

- （左）用线性函数拟合数据导致欠拟合，它无法捕捉数据中的曲率信息
- （中）用二次函数拟合数据并在未观察到的点上泛化得很好，不会导致明显的欠拟合或者过拟合
- （右）一个高阶的多项式拟合数据导致过拟合，模型能够精确地穿过所有的训练点，但无法提取有效的结构信息

2.4.4 模型复杂度

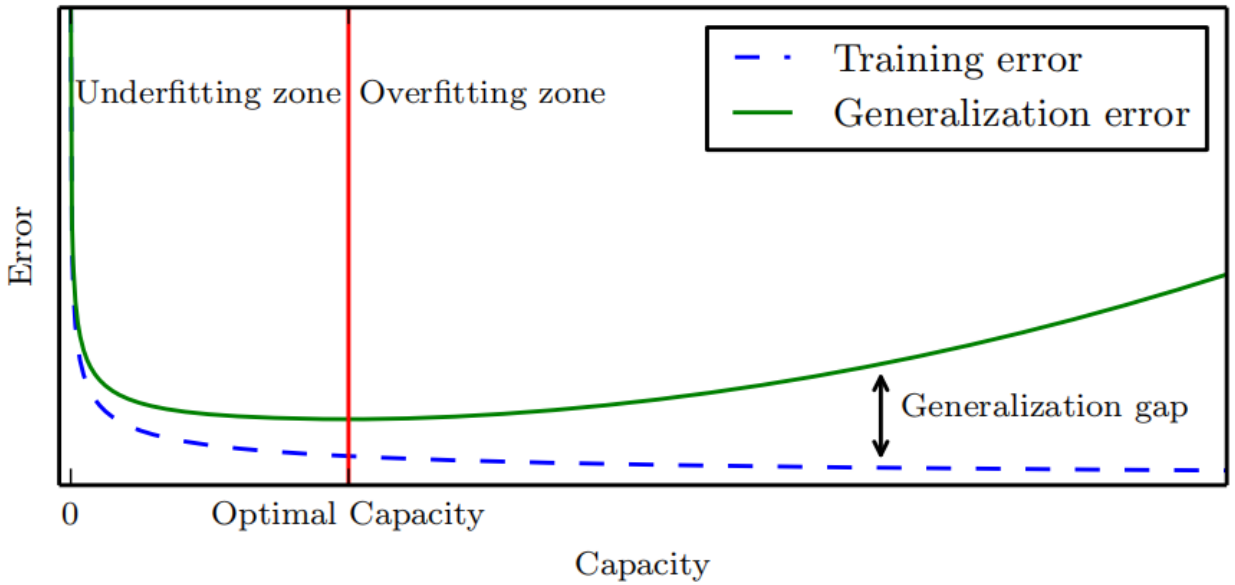
- 模型的**容量**（capacity，学习能力，拟合能力），是指其拟合各种函数的能力
  - 通过调整模型的容量，我们可以控制模型是否偏向于过拟合或者欠拟合
  - 线性模型的容量是线性函数空间
  - 深度神经网络的容量是非线性函数空间
  - 欠拟合：容量低的模型可能很难拟合训练集
    - 前例中的线性函数
  - 过拟合：容量高的模型可能记住了不适用于测试集的训练集性质
    - 前例中的高阶函数
- 当机器学习算法的容量适合所执行任务和所提供训练数据时，算法效果最佳
  - 容量低的模型不能解决复杂任务
  - 容量高的模型能解决复杂任务，但当其容量高于任务所需时，过拟合

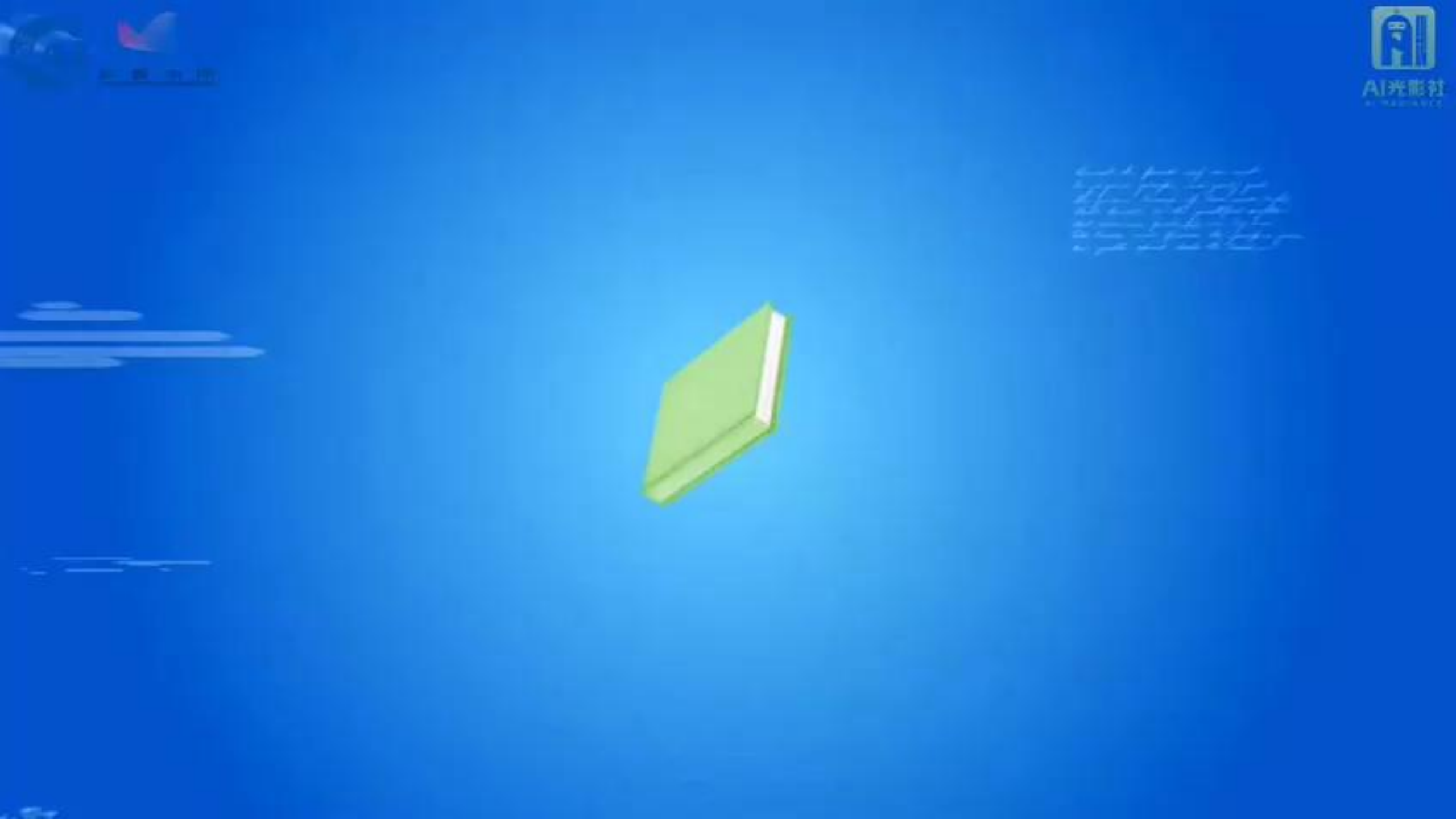
2.4.4 模型复杂度

- 影响模型泛化的因素举例
  - 参数的数量
    - 参数的数量（有时称为**自由度**）很大时，容量很大，模型往往更容易过拟合
      - 解决方法：减少参数，DNN v.s. CNN
  - 参数的取值范围
    - 当参数的取值范围较大时，容量更大，模型可能更容易过拟合
      - 解决方法：正则化（权重衰减），对参数大小进行限制
  - 训练样本的数量
    - 样本数越多，越不容易发生过拟合
    - 即使模型很简单，也很容易过拟合只包含一两个样本的数据集
    - 而过拟合一个有数百万个样本的数据集则需要一个极其灵活的模型
- 奥卡姆剃刀原则：在同样能够解释已知观测现象的模型中，挑选最简单的那一个
  - 二阶函数和高阶多项式都可以拟合训练数据，挑选最简单的那一个

2.4.4 模型复杂度

- 容量和误差之间的典型关系
  - 在图的左端，训练误差和泛化误差都非常高。这是**欠拟合机制**
  - 当增加容量时，训练误差减小
  - 当增加容量时，泛化误差先减小，后增大，呈U形曲线，此时进入到了**过拟合机制**





2.5 微积分基础

- 一个函数在某一点的**导数**描述了这个函数在这一点附近的变化率，也称为函数在该点的切线斜率
  - 假设函数 $f(x)$ 在某一区间内有定义，x_0 及 $x_0+\triangle x$ 在该区间内，则定义函数 $f(x)$ 在 x_0 点处的导数 $f'(x_0)$ 为

$$f'(x_0)=\lim_{\triangle x\to 0}\frac{f(x_0+\triangle x)-f(x_0)}{\triangle x}$$

- 泰勒展开：一种常见的函数近似方法，其本质是利用幂函数来近似任意函数
  - 假设 $f(x)$ 在 x_0 处有 n 阶导数，则泰勒公式为

$$f(x)=f(x_0)+\frac{f'(x_0)}{1!}(x-x_0)+\frac{f''(x_0)}{2!}(x-x_0)^2+\cdots+\frac{f^{(n)}(x_0)}{n!}(x-x_0)^n+o((x-x_0)^n)$$

  n 越大，近似效果越好

2.5 微积分基础

- 多元函数是自变量具有多个分量的函数，其一般形式为

$$y = f(x_1, x_2, \cdots, x_n)$$

- 多元函数y关于某一分量x_i的导数称为偏导数，记为$\frac{\partial y}{\partial x_i}$，具体定义如下

$$\frac{\partial y}{\partial x_i} = \lim_{\Delta x \to 0} \frac{f(x_1, \ldots, x_{i-1}, x_i + \Delta x, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_i, \ldots, x_n)}{\Delta x}$$

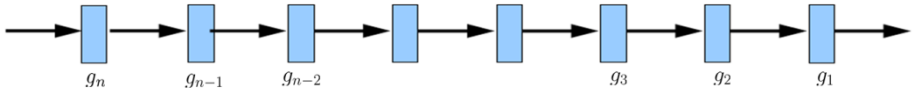
  - 即在求偏导的过程中，保持分量$x_1, \ldots, x_{i-1}, x_{i+1}, \cdots, x_n$恒定不变，仅求该函数关于分量$x_i$的导数即可

2.5 微积分基础

- 在深度学习中，函数通常是以嵌套的形式出现的，例如当$y=f(u)$，且$u=g(x)$时，y与x的关系表示为$y=f(g(x))$，我们将这类函数称为复合函数
- 对于复合函数求导，通常利用链式法则

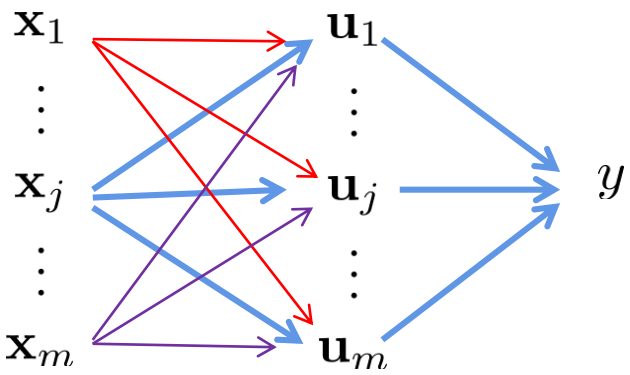
$$\frac{\partial y}{\partial x}=\frac{\partial y}{\partial u}\frac{\partial u}{\partial x}$$

$$f(x)=g_1(g_2(\ldots(g_n(x)\ldots))$$



- 链式法则对多元复合函数也成立
  - 假设有一可微多元函数$y=f(u_1,u_2,\cdots,u_m)$，对于任意可微函数u_i，都有变量$x_1,\cdots,x_n$，
  - 则同样可以利用链式法则求得多元复合函数y对任意自变量x_j的偏导数：

$$\frac{\partial y}{\partial x_j}=\frac{\partial f}{\partial u_1}\frac{\partial u_1}{\partial x_j}+\cdots+\frac{\partial f}{\partial u_m}\frac{\partial u_m}{\partial x_j}$$



2.6 线性代数基础

- 定义含有 n 个元素的向量 v 为

$$\boldsymbol{v} = \begin{bmatrix} \boldsymbol{v}_1 \\ \boldsymbol{v}_2 \\ \vdots \\ \boldsymbol{v}_n \end{bmatrix}$$

  - 对于多元函数 $f(x_1, \cdots, x_n)$，我们对其自变量的所有分量分别求偏导数，并把求得的偏导数写为（列）向量形式，称作该多元函数的梯度，记作 $\nabla f(x)$

$$\nabla f(x) = \left[\frac{\partial f}{\partial x_1}, \cdots, \frac{\partial f}{\partial x_n} \right]^T$$

- 定义 m 行 n 列的标量组成的矩阵 A 为

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{1,1} & \boldsymbol{A}_{1,2} & \cdots & \boldsymbol{A}_{1,n} \\ \boldsymbol{A}_{2,1} & \boldsymbol{A}_{2,2} & \cdots & \boldsymbol{A}_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{A}_{m,1} & \boldsymbol{A}_{m,2} & \cdots & \boldsymbol{A}_{m,n} \end{bmatrix}$$

2.6 线性代数基础

- 对于实对称矩阵$A \in \mathbb{R}^{n \times n}$，其特征值分解为

$$A = V \Lambda V^T$$

  - 其中 $V = (v_1, v_2, \cdots, v_n)$ 是由特征值向量构成的正交矩阵，满足 $VV^T = I$，$V^T V = I$
  - $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n)$ 是由特征值构成的对角矩阵

- 对于矩阵$A \in \mathbb{R}^{m \times n}$，其奇异值分解为

$$A = U \Sigma V^T$$

  - 其中 $U \in \mathbb{R}^{m \times m}$ 和 $V \in \mathbb{R}^{n \times n}$ 是由奇异值向量构成的正交矩阵
  - $\Sigma \in \mathbb{R}^{m \times n}$ 是由奇异值构成的对角矩阵

2.7 概率统计基础

- **期望**又称数学期望、均值，反映随机变量平均取值的大小
  - 对于离散型随机变量，$\mathbb{E}[X] = \sum_i p_i x_i$
  - 对于连续型随机变量，$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x) dx$
- **方差**又称变异数、变方，描述的是一个随机变量的离散程度，即该变量与其期望值的距离，是随机变量与其总体均值或样本均值的差的平方的期望值

$$V(X) = \mathbb{E}[(X - \mu)^2]$$

  - 方差的平方根称为该随机变量的**标准差**

2.7 概率统计基础

- 设A与B为样本空间中的两个事件，在事件B发生的条件下，事件A发生的条件概率，记为 $P(A|B)$ 具体定义如下：

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

- 假设A与B为随机事件，事件A在B已发生的条件下发生的概率，与B在A已发生的条件下发生的概率有确定的关系，该关系常被称为贝叶斯法则/公式：

$$P(A \mid B) = \frac{P(A)P(B \mid A)}{P(B)}$$

  - 式中 $P(A|B)$ 称为事件A的后验概率，实际上是已知B发生后A发生的条件概率
  - $P(A)$ 称为A的先验概率（或边缘概率），其不考虑任何有关B的因素
  - $P(B|A)$ 是A发生后B发生的条件概率，或称为“似然”或者“似然率”

总结

- 回顾了传统机器学习的基本概念，这将用于本课程其他章节中
  - 数据准备和预处理
  - 模型构建
  - 模型训练
  - 泛化、过拟合、欠拟合、模型复杂度
- 回顾了微积分、线性代数、概率统计中的基本概念