模式识别与机器学习

Lecture Notes

笔记三

监督学习核心模型

目录

1 线性回归算法流程、平方和损失、最小二乘法与梯度下降

1.1 本节学习目标

掌握线性回归模型的基本形式与完整算法流程；理解平方和损失函数的理论依据、数学性质与工程价值；能够独立推导最小二乘解析解的完整过程；清晰掌握最小二乘法的核心局限性，理解梯度下降法的引入逻辑与核心原理；建立传统机器学习优化方法与大模型之间的关联认知，为后续复杂模型学习奠定基础。

1.2 线性回归的基本思想与模型表示

线性回归是有监督学习中最基础、应用最广泛的回归算法，核心用于解决连续型数值预测问题，典型场景包括房价预测、销量预估、金融收益率拟合、工业传感器数据建模、医学指标预测等。其核心思想是：假设输入特征与输出目标之间存在近似线性的映射关系，通过构建特征的线性组合来逼近真实输出，本质是在高维特征空间中寻找一个最优超平面，使得所有样本点到该超平面的整体误差最小化。

线性回归的核心优势在于模型结构简单、训练效率高、可解释性极强，是现代机器学习的“入门基石”。即便在大模型时代，线性模型依然是结构化数据预测、风控评分卡、运营分析等工业场景的首选基线方案，同时也是大模型微调、奖励模型训练、对齐机制等核心技术的理论基础。

线性回归的预测函数也称为假设函数,对于包含 n 个特征的单个样本 $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]^T$，模型的原始形式为：

$$h(\boldsymbol{x}) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n$$

其中 w_0 为偏置项（截距），用于拟合输出的基准值；$w_1, w_2, \ldots, w_n$ 为各特征对应的权重，用于衡量特征对输出的影响程度。为了简化矩阵运算与参数求解，通常在输入特征中补充一个常数项 $x_0 = 1$，将偏置项自然融入权重向量，此时模型可改写为紧凑的向量形式：

$$h(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x}$$

其中 $\boldsymbol{w} = [w_0, w_1, w_2, \ldots, w_n]^T$ 为 $(n+1)$ 维权重向量，$\boldsymbol{x} = [1, x_1, x_2, \ldots, x_n]^T$ 为补充偏置项后的特征向量。对于包含 m 个样本的训练集，特征矩阵 $\boldsymbol{X} \in \mathbb{R}^{m\times(n+1)}$、真实标签向量 $\boldsymbol{y} \in \mathbb{R}^{m\times 1}$，模型的矩阵形式为：

$$\boldsymbol{h} = \boldsymbol{X}\boldsymbol{w}$$

模型的核心学习目标，就是从训练数据中求解出一组最优权重 $\boldsymbol{w}^*$，使得模型预测值 $\boldsymbol{h}$ 与真实值 $\boldsymbol{y}$ 之间的整体误差尽可能小。

1.3 线性回归的完整算法流程

线性回归遵循标准的机器学习全流程，从原始数据到模型落地形成闭环，每个环节都直接影响模型的最终性能，具体流程拆解如下：

1.3.1 数据准备与预处理

数据是模型的基础，预处理的质量直接决定模型的上限。该环节的核心任务包括：

- 数据收集：获取与任务目标高度相关的标注数据集，确保数据分布与真实业务场景一致；
- 数据清洗：处理缺失值（删除、填充、插值）、剔除异常值（3σ 原则、箱线图法）、去重，消除数据噪声对模型的干扰；
- 特征工程：筛选与输出目标强相关的有效特征，剔除冗余特征；对特征进行标准化（Z-score 标准化）或归一化（Min-Max 归一化），消除不同特征量纲差异带来的权重偏差；
- 数据集划分：按照合理比例（通常为训练集 70%、验证集 15%、测试集 15%）将数据集划分为相互独立的三部分，训练集用于模型参数学习，验证集用于超参数调优，测试集用于最终泛化能力评估，严格避免数据泄漏。

1.3.2 模型假设与优化目标定义

确定线性回归的模型结构，明确假设函数 $h(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x}$，并将模型的核心目标数学化：寻找最优权重 $\boldsymbol{w}^*$，使得模型在训练集上的预测值与真实值的整体误差最小化，为后续损失函数的构建奠定基础。

1.3.3 损失函数与代价函数构建

损失函数用于量化单个样本的预测误差，代价函数用于量化全体训练样本的平均误差，是模型优化的核心目标。线性回归中采用平方损失函数与均方误差（MSE）代价函数，将拟合问题转化为可求解的最优化问题，具体定义与选择依据将在后续章节详细展开。

1.3.4 模型参数求解

针对构建好的代价函数，采用合适的优化算法求解最优权重 $\boldsymbol{w}^*$，线性回归主要有两种求解方案：

- 解析解法：最小二乘法，通过矩阵运算直接推导参数的闭式解，无需迭代；
- 迭代优化法：梯度下降法，通过沿负梯度方向迭代更新参数，逐步逼近最优解。

两种方法各有适用场景，将在后续章节详细对比分析。

1.3.5 模型评估与优化

使用训练好的模型在测试集上进行预测，计算泛化误差，常用评估指标包括均方误差（MSE）、均方根误差（RMSE）、平均绝对误差（MAE）、决定系数 R^2 等。通过对比训练误差与测试误差，判断模型是否存在欠拟合（偏差过大）或过拟合（方差过大）：若欠拟合，可增加特征数量、提升模型复杂度；若过拟合，可通过增加训练数据、加入 L1/L2 正则化、降维等方式约束模型复杂度，提升泛化能力。

1.3.6 模型部署与应用

将训练完成、评估达标的模型部署到实际业务系统中，对新的未知样本进行实时或批量预测，完成从数据到业务价值的转化，同时持续监控模型性能，定期用新数据 retrain 模型，保证模型的长期有效性。

1.4 线性回归中平方和损失函数的选择依据

损失函数是模型优化的核心，其选择直接决定模型的学习效果。线性回归中最常用的是平方损失函数与均方误差代价函数，具体定义如下：

对于第 i 个样本 $(\boldsymbol{x}^{(i)}, y^{(i)})$，平方损失函数（单样本误差）为：

$$L\left(h(\boldsymbol{x}^{(i)}), y^{(i)}\right) = \frac{1}{2}\left(h(\boldsymbol{x}^{(i)}) - y^{(i)}\right)^2$$

全体训练样本的均方误差代价函数（全局优化目标）为：

$$J(\boldsymbol{w}) = \frac{1}{2m}\sum_{i=1}^{m}\left(h(\boldsymbol{x}^{(i)}) - y^{(i)}\right)^2 = \frac{1}{2m}\sum_{i=1}^{m}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} - y^{(i)}\right)^2$$

其矩阵形式为：

$$J(\boldsymbol{w}) = \frac{1}{2m}\left\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\right\|_2^2$$

其中 $\|\cdot\|_2$ 为向量的 2-范数（欧氏范数），公式中的系数 $\frac{1}{2}$ 仅为了在求导时抵消平方项产生的系数 2，简化求导后的公式形式，不会改变最优参数 $\boldsymbol{w}^*$ 的求解结果，不同教材中系数设为 1 或 $\frac{1}{2}$ 均合理。

线性回归选择平方和损失具有充分的理论依据、数学性质与工程价值，具体如下：

1.4.1 统计意义：高斯噪声下的极大似然估计

这是平方和损失最核心的理论支撑。我们假设样本的真实值与预测值之间的残差 $\varepsilon^{(i)} = y^{(i)} - \boldsymbol{w}^T\boldsymbol{x}^{(i)}$ 服从均值为 0、方差为 σ^2 的高斯分布（正态分布），即：

$$p(\varepsilon^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(\varepsilon^{(i)})^2}{2\sigma^2}\right)$$

代入残差定义，可得 $y^{(i)}$ 在给定 $\boldsymbol{x}^{(i)}$ 与 $\boldsymbol{w}$ 下的条件概率分布：

$$p(y^{(i)}|\boldsymbol{x}^{(i)};\boldsymbol{w}) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y^{(i)} - \boldsymbol{w}^T\boldsymbol{x}^{(i)})^2}{2\sigma^2}\right)$$

根据极大似然估计思想，我们需要找到参数 $\boldsymbol{w}$，使得所有样本观测值出现的联合概率最大，即最大化似然函数：

$$L(\boldsymbol{w}) = \prod_{i=1}^{m} p(y^{(i)}|\boldsymbol{x}^{(i)};\boldsymbol{w})$$

为简化计算，对似然函数取自然对数（对数函数单调递增，不改变最优解位置），将乘法转化为加法：

$$\begin{aligned}\ln L(\boldsymbol{w}) &= \ln\prod_{i=1}^{m}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y^{(i)} - \boldsymbol{w}^T\boldsymbol{x}^{(i)})^2}{2\sigma^2}\right)\\ &= \sum_{i=1}^{m}\left[\ln\frac{1}{\sqrt{2\pi}\sigma} - \frac{(y^{(i)} - \boldsymbol{w}^T\boldsymbol{x}^{(i)})^2}{2\sigma^2}\right]\\ &= m\ln\frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2}\sum_{i=1}^{m}(y^{(i)} - \boldsymbol{w}^T\boldsymbol{x}^{(i)})^2\end{aligned}$$

要最大化对数似然函数，前半部分为与 $\boldsymbol{w}$ 无关的常数，因此等价于最小化后半部分的求和项 $\sum_{i=1}^{m}(y^{(i)} - \boldsymbol{w}^T\boldsymbol{x}^{(i)})^2$，这正是线性回归的平方和损失函数。这一推导严格证明：**在高斯噪声假设下，最小化平方和损失等价于参数的极大似然估计**，为平方和损失的使用提供了坚实的概率理论支撑。

1.4.2 几何意义：最小化欧氏距离

从几何角度看，线性回归的目标是找到一个超平面，使得所有样本点到该超平面的欧氏距离之和最小。平方和损失的本质，就是最小化样本真实值与预测值之间的欧氏距

离平方和，物理意义直观明确，完全契合线性回归的几何目标，便于理解与解释。

1.4.3 数学性质：凸函数与可导性

平方和损失函数 $J(\boldsymbol{w})$ 是关于 $\boldsymbol{w}$ 的严格凸函数，凸函数的局部最小值就是全局最小值，保证优化过程能够收敛到全局最优解，不会陷入局部最优；同时，平方和损失是处处连续可导的光滑函数，梯度计算简洁，既支持解析求解（最小二乘法），也完美适配梯度下降等迭代优化算法，大幅降低了模型求解的难度。

1.4.4 工程特性：误差惩罚与鲁棒性

平方和损失对大误差的惩罚力度远大于小误差，符合绝大多数回归任务的业务需求：我们更希望模型避免出现严重的预测错误，对小误差有一定容忍度，能够有效约束模型极端偏差，提升模型稳定性与鲁棒性。同时，平方损失计算简单、实现便捷，是工业界最常用的损失函数之一，也广泛应用于大模型微调、奖励函数设计等场景。

1.5 最小二乘法原理与参数解析解推导

最小二乘法（Least Squares Method, LSM）是线性回归的经典解析求解方法，核心思想是直接最小化残差平方和，通过矩阵求导推导出参数的闭式解，无需迭代即可一次性得到全局最优解。

基于前文定义的平方和代价函数矩阵形式：

$$J(\boldsymbol{w}) = \frac{1}{2m} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 = \frac{1}{2m}(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})^T(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})$$

我们的目标是找到 $\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} J(\boldsymbol{w})$。对于凸可导函数，最小值出现在梯度为 0 的位置，因此对 $J(\boldsymbol{w})$ 关于 $\boldsymbol{w}$ 求偏导，并令偏导等于 0。

首先展开代价函数：

$$\begin{aligned} J(\boldsymbol{w}) &= \frac{1}{2m}\left[(\boldsymbol{X}\boldsymbol{w})^T\boldsymbol{X}\boldsymbol{w} - (\boldsymbol{X}\boldsymbol{w})^T\boldsymbol{y} - \boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{w}) + \boldsymbol{y}^T\boldsymbol{y}\right] \\ &= \frac{1}{2m}\left[\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - 2\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{y}\right] \end{aligned}$$

注：$(\boldsymbol{X}\boldsymbol{w})^T\boldsymbol{y}$ 与 $\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{w})$ 互为转置，且均为标量，因此两者相等，合并后系数为 2。

接下来对 $\boldsymbol{w}$ 求偏导，用到矩阵求导的核心法则：

- 对于二次型 $\frac{\partial \boldsymbol{w}^T\boldsymbol{A}\boldsymbol{w}}{\partial \boldsymbol{w}}$，若 $\boldsymbol{A}$ 为对称矩阵，则结果为 $2\boldsymbol{A}\boldsymbol{w}$；
- 对于线性项 $\frac{\partial \boldsymbol{w}^T\boldsymbol{a}}{\partial \boldsymbol{w}}$，结果为 $\boldsymbol{a}$（$\boldsymbol{a}$ 为与 $\boldsymbol{w}$ 无关的向量）。

由于 $\boldsymbol{X}^T\boldsymbol{X}$ 是对称矩阵，因此对 $J(\boldsymbol{w})$ 求偏导的结果为：

$$\begin{aligned} \frac{\partial J(\boldsymbol{w})}{\partial \boldsymbol{w}} &= \frac{1}{2m}\left[2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - 2\boldsymbol{X}^T\boldsymbol{y} + 0\right] \\ &= \frac{1}{m}\left(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}^T\boldsymbol{y}\right) \end{aligned}$$

令偏导数等于 0，求解 $\boldsymbol{w}$：

$$\begin{aligned} \frac{1}{m}\left(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}^T\boldsymbol{y}\right) &= 0 \\ \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} &= \boldsymbol{X}^T\boldsymbol{y} \end{aligned}$$

上式称为线性回归的**正规方程**。当矩阵 $\boldsymbol{X}^T\boldsymbol{X}$ 为可逆满秩矩阵时，两边同时左乘其逆矩阵，即可得到最小二乘法的解析解：

$$\boldsymbol{w}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

最小二乘法的核心特点：无需设置学习率等超参数，一次矩阵运算即可得到全局最优解，实现简单、结果精确；但仅适用于线性回归模型，无法直接扩展到逻辑回归、神经网络等其他模型，且在特征维度 n 较大时，矩阵求逆的计算成本极高（时间复杂度为 $O(n^3)$），实用性受限。

1.6 最小二乘法的局限性与梯度下降法的引出

最小二乘法虽有闭式解的优势，但在实际工程应用中存在显著的局限性，这些局限性直接推动了梯度下降法的引入与发展。

1.6.1 最小二乘法的核心局限性

- 高维特征下计算代价极高最小二乘法的核心是计算矩阵 $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$，矩阵求逆的时间复杂度为 $O(n^3)$，其中 n 为特征维度。当特征维度 n 较小时（如 $n < 1000$），计算效率较高；但当特征维度达到数万、数十万甚至更高时（如深度学习中的百万级特征），矩阵求逆的计算量呈立方级增长，耗时极长，甚至无法在有限时间内完成，完全不适用高维场景。

- 矩阵不可逆时无法直接求解当特征矩阵存在多重共线性（特征之间高度线性相关），或样本数量 m 小于特征维度 n 时，矩阵 $\boldsymbol{X}^T\boldsymbol{X}$ 会成为奇异矩阵（不满秩），不存在逆矩阵，此时最小二乘法的闭式解无法直接计算，必须通过加入正则化（岭回归）、降维等方式修正，增加了实现复杂度。

- 泛用性极差，仅适用于线性模型最小二乘法的闭式解仅在线性回归 + 平方和损失的特定组合下存在，对于逻辑回归、支持向量机、神经网络等绝大多数非线性模型，无法推导出闭式解，完全不适用，无法作为通用优化算法使用。

- 无法处理大规模流式数据最小二乘法需要一次性加载全部训练数据进行矩阵运算，当数据量极大无法一次性加载到内存，或数据以流式不断更新时，最小二乘法完全无法使用，无法适配大数据与实时学习场景。

1.6.2 梯度下降法的核心原理与引入

针对最小二乘法的上述局限性，梯度下降法作为通用数值优化算法被引入，完美解决了其核心痛点。梯度下降法的核心思想是：**沿着代价函数下降最快的方向（负梯度方向），迭代更新模型参数，直到代价函数收敛到最小值**。

其核心参数更新公式为：

$$\boldsymbol{w} := \boldsymbol{w} - \alpha \cdot \frac{\partial J(\boldsymbol{w})}{\partial \boldsymbol{w}}$$

其中 α 为学习率（步长），用于控制每次迭代的参数更新幅度，是梯度下降法的核心超参数。将线性回归的代价函数梯度代入，可得具体的参数更新公式：

$$\boldsymbol{w} := \boldsymbol{w} - \alpha \cdot \frac{1}{m} \left(\boldsymbol{X}^T \boldsymbol{X} \boldsymbol{w} - \boldsymbol{X}^T \boldsymbol{y} \right)$$

梯度下降法的核心优势：

- 无需矩阵求逆，计算复杂度仅为 $O(n)$，高维特征下依然高效；
- 适用于几乎所有可导的机器学习模型，泛用性极强；
- 支持批量梯度下降（BGD）、小批量梯度下降（MBGD）、随机梯度下降（SGD）等多种形式，可灵活处理大规模数据与流式数据；
- 可通过调整学习率、动量等优化策略，提升收敛速度与稳定性。

1.6.3 最小二乘法与梯度下降法的核心对比

表 1: 最小二乘法与梯度下降法对比

最小二乘法	梯度下降法
解析解法，一次计算得出全局最优解	迭代解法，多次迭代逼近最优解
无需设置学习率等超参数	需要选择合适的学习率，调参成本较高
需计算矩阵逆，n 大时代价极高，复杂度 $O(n^3)$	无需矩阵求逆，高维特征下效率高，复杂度 $O(n)$
仅适用于线性回归模型，泛用性差	适用于几乎所有可导模型，泛用性极强
一次性加载全部数据，无法处理大规模/流式数据	支持批量、小批量、随机更新，适配大数据场景

最小二乘法适合特征维度低、数据量小的线性回归场景；梯度下降法适合高维、大数据、复杂模型场景，二者共同构成了机器学习优化的核心基础，也是大模型训练、微调与对齐的核心算法支撑。

1.7 小结

线性回归通过特征的线性组合实现连续值预测，以平方和损失为核心优化目标，可通过最小二乘法直接求解解析解，也可通过梯度下降法迭代优化。平方和损失的选择具有严格的统计、几何、数学与工程依据，最小二乘法是线性回归的经典解析方法，而梯度下降法则解决了最小二乘法的高维、大数据等局限性，成为通用优化算法。理解线性回归的完整流程、损失函数原理、两种求解方法的推导与对比，是掌握传统机器学习、深度学习与大模型优化的重要基础。

2 线性模型详解与房价预测案例

2.1 线性模型基本概念

2.1.1 线性模型的严格定义

在机器学习语境下，**线性模型**指对参数呈线性的模型。其一般形式为：

$$f(\mathbf{x};\mathbf{w},b)=w_1x_1+w_2x_2+\cdots+w_px_p+b=\mathbf{w}^\top\mathbf{x}+b$$

其中：

- 输入向量 $\mathbf{x}=(x_1,x_2,\ldots,x_p)^\top\in\mathbb{R}^p$ 是一个样本的特征列向量。
- 参数向量 $\mathbf{w}=(w_1,w_2,\ldots,w_p)^\top\in\mathbb{R}^p$ 是权重向量。
- 偏置项 $b\in\mathbb{R}$ 是标量。

注意：我们称其为“线性模型”是因为它对**参数** $\mathbf{w},b$ 是线性的，而非对**特征** x_j 线性。例如，特征可以经过非线性变换（如 $x_1^2,\log x_2$）再代入，但只要这些变换后的新特征仍以加权和形式出现，模型依然是线性的。因此，线性模型的**表达能力**可以通过特征工程大幅扩展。

2.1.2 增广表示法（简化符号的关键技巧）

为了方便矩阵运算，我们常将偏置项 b 吸收进权重向量，并在特征向量中补入一个常数 1。即令：

$$\tilde{\mathbf{x}} = \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_p \end{pmatrix} \in \mathbb{R}^{p+1}, \quad \tilde{\mathbf{w}} = \begin{pmatrix} b \\ w_1 \\ \vdots \\ w_p \end{pmatrix} \in \mathbb{R}^{p+1}$$

则模型化为单一内积形式：

$$f(\mathbf{x}; \tilde{\mathbf{w}}) = \tilde{\mathbf{w}}^\top \tilde{\mathbf{x}}$$

在后续推导中，我们默认使用这种增广表示，**直接记 $\mathbf{x}$ 为增广特征，$\mathbf{w}$ 为增广参数**，以保持符号整洁。

2.1.3 几何意义

- **一维特征**：模型 $y = wx + b$ 是一条直线的方程，w 为斜率，b 为截距。
- **高维特征**：模型 $\hat{y} = \mathbf{w}^\top \mathbf{x}$ 定义了一个超平面（hyperplane），$\mathbf{w}$ 是该超平面的法向量方向，$\mathbf{w}^\top \mathbf{x}$ 正比于样本点到超平面的**带符号距离**。

2.2 线性回归基本概念

2.2.1 问题形式化

给定训练集 $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^n$，其中 $\mathbf{x}^{(i)} \in \mathbb{R}^{p+1}$（已增广），$y^{(i)} \in \mathbb{R}$。线性回归的目标是找到一个参数向量 $\mathbf{w}$，使得预测值 $\hat{y}^{(i)} = \mathbf{w}^\top \mathbf{x}^{(i)}$ 与真实值 $y^{(i)}$ 之间的误差平方和最小。

经验风险最小化（Empirical Risk Minimization）视角下的损失函数为：

$$J(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} \left(y^{(i)} - \mathbf{w}^\top \mathbf{x}^{(i)} \right)^2$$

省略常数因子不影响最优解，实际常直接使用残差平方和（Residual Sum of Squares, RSS）：

$$\mathrm{RSS}(\mathbf{w}) = \sum_{i=1}^{n} (y^{(i)} - \mathbf{w}^\top \mathbf{x}^{(i)})^2$$

2.2.2 矩阵向量化与正规方程的推导

将所有样本写成矩阵形式。设计矩阵（Design Matrix）$X \in \mathbb{R}^{n \times (p+1)}$ 每一行是一个增广特征向量的转置：

$$X = \begin{pmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \dots & x_p^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \dots & x_p^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(n)} & x_2^{(n)} & \dots & x_p^{(n)} \end{pmatrix}$$

真实值向量 $\mathbf{y} = (y^{(1)}, y^{(2)}, \dots, y^{(n)})^\top \in \mathbb{R}^n$。则残差向量 $\mathbf{e} = \mathbf{y} - X\mathbf{w}$，损失函数为：

$$\mathrm{RSS}(\mathbf{w}) = \|\mathbf{y} - X\mathbf{w}\|_2^2 = (\mathbf{y} - X\mathbf{w})^\top (\mathbf{y} - X\mathbf{w})$$

对 $\mathbf{w}$ 求梯度并令其为零向量：

$$\nabla_{\mathbf{w}} \mathrm{RSS} = -2X^\top (\mathbf{y} - X\mathbf{w}) = \mathbf{0}$$

得到正规方程（Normal Equation）：

$$X^\top X \mathbf{w} = X^\top \mathbf{y}$$

若 $X^\top X$ 可逆（满秩），则唯一解析解为：

$$\mathbf{w}^* = (X^\top X)^{-1} X^\top \mathbf{y}$$

这就是**最小二乘估计**（Ordinary Least Squares, OLS）。

2.2.3 梯度下降解法（大规模数据下的必然选择）

当特征维数 p 很大（例如数万维）或样本数 n 极大时，计算 $(X^\top X)^{-1}$ 的时间复杂度 $O(p^3)$ 不可接受，此时改用**迭代优化**。

批量梯度下降（Batch Gradient Descent）的更新公式（针对未增广的 $\mathbf{w}$ 和 b）：

$$w_j := w_j - \alpha \frac{\partial J}{\partial w_j} = w_j + \alpha \frac{2}{n} \sum_{i=1}^{n} (y^{(i)} - \hat{y}^{(i)}) x_j^{(i)}$$

$$b := b - \alpha \frac{\partial J}{\partial b} = b + \alpha \frac{2}{n} \sum_{i=1}^{n} (y^{(i)} - \hat{y}^{(i)})$$

向量化形式（更高效）：

$$\mathbf{w} := \mathbf{w} + \alpha \frac{2}{n} X^\top (\mathbf{y} - X\mathbf{w})$$

其中 α 是学习率（learning rate），需人工设定。

2.2.4 统计学视角下的线性回归假设与性质

高斯-马尔可夫定理（Gauss-Markov Theorem）指出：在以下假设成立时，最小二乘估计 $\mathbf{w}^*$ 是**最佳线性无偏估计**（BLUE）。

1. **线性性**：$y = \mathbf{w}^\top \mathbf{x} + \varepsilon$，其中 ε 为随机误差。
2. **零均值**：$\mathbb{E}[\varepsilon | X] = 0$。
3. **同方差性**：$\mathrm{Var}(\varepsilon | X) = \sigma^2$ 为常数，与 X 无关。
4. **无自相关**：$\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$，对 $i \neq j$。
5. 误差正态性（仅用于推断，不影响无偏性与一致性）：$\varepsilon \sim \mathcal{N}(0, \sigma^2)$。

2.3 符号约定

为了能够流畅阅读教材和代码，以下符号及其维度要刻在脑子里。

符号	含义	维度（未增广/增广后）
n	样本数量	标量
p	原始特征数量	标量
X	设计矩阵（每行一样本，每列一特征）	$n \times (p+1)$（增广后）
$\mathbf{y}$	真实值向量	$n \times 1$
$\mathbf{w}$	参数向量（增广后包含偏置）	$(p+1) \times 1$
$\hat{\mathbf{y}}$	模型预测向量 $\hat{\mathbf{y}} = X\mathbf{w}$	$n \times 1$
$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$	残差向量	$n \times 1$
$X^\top X$	Gram 矩阵（对称半正定）	$(p+1) \times (p+1)$
$(X^\top X)^{-1} X^\top$	伪逆矩阵（Moore-Penrose）的一种形式	$(p+1) \times n$

2.4 房价预测案例介绍

为直观呈现线性回归的建模过程与内在局限，我们构造一个包含五个样本的简化房价数据集。令特征向量 $\mathbf{x} = (x_1, x_2)^\top$，其中 x_1 表示房屋面积（单位：m^2），x_2 表示房龄（单位：年），目标变量 y 为成交价格（单位：万元）。数据详见表 1。

表 3: 微型房价数据集

样本编号 i	$x_1^{(i)}$（面积/m^2）	$x_2^{(i)}$（房龄/年）	$y^{(i)}$（价格/万元）
1	50	5	160
2	80	10	240
3	110	15	300
4	140	20	350
5	170	25	390

为了便于说明线性回归模型的解释性、拟合效果与局限性，下面选取一组具有明确

经济含义的示例参数作为分析对象：

$$\hat{\mathbf{w}} = \begin{pmatrix} b \\ w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} 80 \\ 2.0 \\ -2.0 \end{pmatrix}$$

对应的示例回归方程为：

$$\hat{y} = 80 + 2.0\,x_1 - 2.0\,x_2$$

该方程不再表述为该数据集的唯一解析解，而是作为后续分析模型解释性、残差表现和外推行为的一组示例系数。其经济含义为：

- 截距项 $b = 80$：当面积与房龄均为零时的基准价格（在实际中无对应物理意义，仅为回归平面的截距）；
- 面积系数 $w_1 = 2.0$：在房龄保持不变的条件下，面积每增加 1 m^2，预计价格上升 2.0 万元；
- 房龄系数 $w_2 = -2.0$：在面积保持不变的条件下，房龄每增加 1 年，预计价格下降 2.0 万元。

2.5 模型优势的实证展示

2.5.1 结构透明，参数具备直接经济解释

参数向量 $\hat{\mathbf{w}}$ 的每个分量均对应一个明确的经济含义，使得模型在房地产估值、信贷审批等需要向非技术受众解释的场景中具有不可替代的优势。

2.5.2 训练集内拟合效果可接受

将训练样本代入回归方程，得到预测值与残差见表 2。

表 4: 训练集预测结果与残差分析

i	$y^{(i)}$	$\hat{y}^{(i)} = 80 + 2x_1 - 2x_2$	残差 $e^{(i)} = y^{(i)} - \hat{y}^{(i)}$
1	160	170	-10
2	240	220	+20
3	300	270	+30
4	350	320	+30
5	390	370	+20

模型在训练集上的均方根误差（RMSE）为：

$$\text{RMSE} = \sqrt{\frac{(-10)^2 + 20^2 + 30^2 + 30^2 + 20^2}{5}} \approx 23.45 \text{ 万元}$$

相对于房价均值（288 万元），误差率约 8.1%，表明线性模型作为初步估值工具具备合理的参考价值。

2.6 模型局限性的数据与图像论证

同一组数据与回归方程同样清晰地暴露了线性回归的若干结构性缺陷。以下逐一分析。

2.6.1 非线性模式欠拟合（边际效应递减）

观察实际样本点 (x_1, y) 的分布：面积从 50 m^2 增至 80 m^2（+30 m^2），价格上升 80 万元；而从 140 m^2 增至 170 m^2（同样 +30 m^2），价格仅上升 40 万元。现实市场中，面积对价格的边际贡献通常呈**递减趋势**。若将数据点绘于二维平面（横轴 x_1，纵轴 y），可发现数据点大致沿一条**上凸的曲线**分布，而模型 $\hat{y} = 80 + 2.0x_1 - 2.0x_2$ 在给定 x_2 的条件下（例如固定房龄为 15 年），简化为 $\hat{y} = 50 + 2.0x_1$，这是一条直线。直线无法弯曲以贴合曲线的弧度，导致中间样本（80–140 m^2）的系统性低估（残差均为正值），两端样本相对高估。此即线性模型对非线性关系欠拟合的典型表现，需引入多项式特征（如 x_1^2）或非线性变换（如 $\log x_1$）加以修正。

2.6.2 对离群值的极度敏感性

假设将样本 5 的价格由 390 万元人为修改为 590 万元（模拟极端高价豪宅），则新的数据集 $\mathcal{D}'$ 将迫使最小二乘估计发生剧烈变化。重新求解正规方程（推导从略），可得新参数为：

$$\hat{\mathbf{w}}' \approx \begin{pmatrix} 120 \\ 3.5 \\ -1.0 \end{pmatrix}$$

面积系数由 2.0 飙升至 3.5，房龄系数由 -2.0 骤变为 -1.0，截距亦显著漂移。单一异常样本通过平方损失函数放大了自身权重，严重扭曲了对大多数普通住宅的价格评估逻辑。在图像上，回归平面将明显向离群点倾斜，牺牲对主体数据的拟合优度。在实际建模中，若不对异常交易（如法拍房、赠与过户）加以清洗或采用稳健损失函数（如 Huber Loss），模型的泛化能力将大打折扣。

2.6.3 多重共线性致使参数估计不稳定

从表 1 可以看出，面积 x_1 与房龄 x_2 随样本编号同步增大，二者具有明显的正相关趋势。虽然这组数据并不满足严格的线性关系，因此不能据此直接推出 $X^\top X$ 奇异，但它已经提示一个常见问题：当不同特征之间高度相关时，参数估计会变得不稳定。此时，模型虽然仍可能给出可接受的预测结果，但每个系数的单独经济解释会受到干扰。

更一般地说，若特征之间存在严格线性相关，则设计矩阵 X 的列向量线性相关，矩阵 $X^\top X$ 不可逆，正规方程不能直接求得唯一解；若只是高度相关而非完全相关，则矩阵虽然可逆，但条件数可能很大，参数会对样本微小扰动十分敏感。真实房产数据中，面积、卧室数、卫生间数、套内面积等指标之间往往存在较强相关性，这会导致权重的标准误膨胀、符号异常或数值波动明显，从而削弱模型解释的可靠性。

2.6.4 外推预测违背常识约束

将模型应用于远离训练数据分布的新样本：

- 情形 A：$x_1 = 300, x_2 = 40$（远超最大面积 170、最大房龄 25）

$$\hat{y} = 80 + 2.0 \times 300 - 2.0 \times 40 = 600 \text{ 万元}$$

预测值尚在合理范围，但已缺乏实际数据支撑。

- 情形 B：$x_1 = 200, x_2 = 150$（历史保护建筑，房龄极大）

$$\hat{y} = 80 + 2.0 \times 200 - 2.0 \times 150 = 80 + 400 - 300 = 180 \text{ 万元}$$

180 万元的估值可能严重低估历史建筑的文化溢价，且若房龄继续增加，预测价格将持续线性递减，甚至出现负值，这违背了房地产价值的基本常识。

- 情形 C：$x_1 = 500, x_2 = 0$（超大面积全新豪宅）

$$\hat{y} = 80 + 2.0 \times 500 - 0 = 1080 \text{ 万元}$$

由于面积系数沿用普通住宅的 2.0 万元/m^2，该预测很可能大幅低估顶级豪宅的真实市价（其单价往往远超普通住宅）。

线性模型的外推本质上是对训练数据习得的线性函数进行无限延拓，而数据分布边缘之外的真实函数形态往往是未知且非线性的。当新样本的特征组合显著偏离训练集的凸包时，外推结果极易失实。

2.7 总结与延伸

以上基于参数化微型数据集的案例研究，生动诠释了线性回归的“双刃剑”特性：它结构简明、解释性强、计算高效，是绝大多数回归任务的基准起点；但面对非线性模式、离群噪声、特征共线及分布外泛化时，其刚性假设又构成显著的性能瓶颈。

2.8 案例延伸与相关补充

前四部分已经给出了线性模型与线性回归的基本框架。为了使笔记从“会看公式”进一步过渡到“会理解模型、会判断模型是否适用”，这里补充六个常用但容易遗漏的基础内容。

2.8.1 线性模型与线性回归的关系

线性模型是一个更广的模型族，其共同特征是模型输出可以写成特征的加权和，因此模型对参数保持线性形式。在线性模型框架下，根据任务类型不同，可以得到不同的具体模型。在线性回归中，目标变量是连续实数，模型输出直接作为预测值；在线性分类中，线性函数通常还需要进一步经过符号函数或概率映射后再用于判别类别。

因此，线性回归不是独立于线性模型之外的新模型，而是线性模型在回归任务中的一个具体实例。把这层关系说明清楚后，可以更自然地理解为什么线性模型既能用于连续值预测，也能扩展到分类、排序和表示学习等问题。

2.8.2 预测函数、损失函数与优化方法的区别

在线性回归中，至少要区分三个层次：模型、损失和优化。

第一层是预测函数，即模型本身：

$$\hat{y} = \mathbf{w}^\top \mathbf{x} + b$$

它回答的是“输入进入模型后如何得到输出”。

第二层是损失函数，用于衡量预测值与真实值之间的偏差。例如平方损失可以写成：

$$L(y, \hat{y}) = (y - \hat{y})^2$$

在训练集中，把所有样本的损失加总后，就得到经验风险或目标函数。

第三层是优化方法，用于寻找使损失函数尽可能小的参数。正规方程属于解析求解

方法，梯度下降属于迭代求解方法。三者分别回答“模型长什么样”“误差怎么定义”“参数怎么求”，逻辑上应当分开理解。初学时若把这三件事混在一起，往往会误把某种损失函数当成模型本身，或误把优化算法当成模型结构。

2.8.3 模型评估与泛化能力

评价线性回归模型时，不能只看训练集上的残差大小，还要关注模型对未见样本的预测能力。训练误差描述模型对已有数据的拟合程度，测试误差更能反映模型的泛化能力。若训练误差很小而测试误差明显偏大，通常说明模型对训练集过于贴合，但对新样本缺乏稳定预测能力。

常用评价指标包括：

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

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y^{(i)} - \hat{y}^{(i)}|$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})^2}{\sum_{i=1}^{n}(y^{(i)} - \bar{y})^2}$$

其中，RMSE 对较大误差更敏感，MAE 更直观，R^2 用于衡量模型对样本波动的解释程度。在线性回归的实际应用中，通常需要将数据划分为训练集与测试集，分别用于参数学习与性能验证。

2.8.4 可逆性、伪逆与最小二乘解

正规方程

$$X^\top X\mathbf{w} = X^\top \mathbf{y}$$

建立在线性代数的基础上。当 $X^\top X$ 可逆时，最小二乘解唯一存在，可写为

$$\mathbf{w}^* = (X^\top X)^{-1} X^\top \mathbf{y}$$

当设计矩阵的列向量线性相关时，$X^\top X$ 不可逆，正规方程不能直接求逆，此时最小二乘解不再唯一。为了得到一组稳定的解，常引入 Moore–Penrose 伪逆：

$$\mathbf{w}^* = X^+ \mathbf{y}$$

伪逆的意义在于：即使矩阵不可逆，仍可以在最小二乘意义下求得一组范数最小的解。这样一来，“最小二乘问题能否求解”和“正规方程能否直接求逆”就被区分开了。前者通常仍可处理，后者则要求更强的可逆性条件。这个补充对于理解多重共线性、欠定系统以及高维回归都很重要。

2.8.5 正则化的基本思想

在线性回归中，如果特征较多、特征之间相关性较强，或者希望限制参数过大，就可以在原有损失函数上加入正则化项。最常见的两种形式是岭回归与 Lasso 回归。

岭回归在目标函数中加入参数平方和惩罚：

$$J(\mathbf{w}) = \|\mathbf{y} - X\mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_2^2$$

Lasso 回归在目标函数中加入参数绝对值和惩罚：

$$J(\mathbf{w}) = \|\mathbf{y} - X\mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_1$$

其中，λ 为正则化系数，用于控制拟合误差与参数复杂度之间的权衡。岭回归通常用于缓解多重共线性、提高参数稳定性；Lasso 回归除抑制参数波动外，还可能将部分参数压缩为 0，从而起到特征选择的作用。对于基础学习而言，掌握正则化的核心思想比死记公式更重要：它本质上是在“拟合数据”和“避免模型过于复杂”之间做平衡。

2.8.6 特征缩放与标准化

线性回归的数学形式虽然简单，但在数值计算时，特征尺度差异会直接影响求解效果。若一个特征的量级远大于其他特征，例如面积以平方米计、房价以万元计、距离以米计，则梯度在不同维度上的更新速度可能差异很大，导致收敛变慢，甚至出现震荡。

一种常见处理方式是对特征做标准化：

$$x_j' = \frac{x_j - \mu_j}{\sigma_j}$$

其中，μ_j 和 σ_j 分别表示第 j 个特征的均值与标准差。标准化后，各维特征处于相近尺度，更有利于迭代优化，也便于比较参数的相对影响大小。虽然正规方程从理论上不要求必须标准化，但在实际数值计算中，适当的尺度处理通常能够提升稳定性。

2.8.7 对房价案例的理解方式

房价预测案例把线性模型的优点和局限都表现得比较集中。示例方程

$$\hat{y} = 80 + 2.0\,x_1 - 2.0\,x_2$$

能够直接给出面积和房龄对价格的影响方向与影响幅度，说明模型结构清晰、参数明确、计算简单，在需要快速建模的场景中具有明显优势。对训练样本进行预测时，模型也能给出可接受的近似结果，说明它适合作为回归分析的基础模型。

但从同一个案例也能看到它的缺点。样本中的价格变化并不完全服从线性规律，导致模型在部分区间出现系统性偏差；一旦加入异常样本，平方损失会放大离群点的影响，使回归结果明显偏移；当特征之间相关性较强时，参数估计会变得不稳定，系数的数值和经济解释都可能发生波动；对于远离训练样本范围的新房源，线性外推还可能得到与实际市场不符的结果。由此可见，线性模型的优势在于简单、透明、易于计算，局限则在于表达能力有限，对数据分布和样本质量较为敏感。

3 分类问题

分类问题是监督学习中最核心的任务之一，目的是根据输入特征将样本划分到预先定义好的离散类别中。

3.1 二分类问题

二分类是最简单、最基础的分类形式，只有两个类别。

- 标签取值：$y \in \{0,1\}$ 或 $y \in \{-1,1\}$，通常 0 为负类，1 为正类。
- 二分类任务仅需构建 1 个分类器、完成 1 次分类决策，即可完成全量样本的类别划分。
- 典型应用：垃圾邮件识别、疾病诊断、交易风控、用户流失预测。

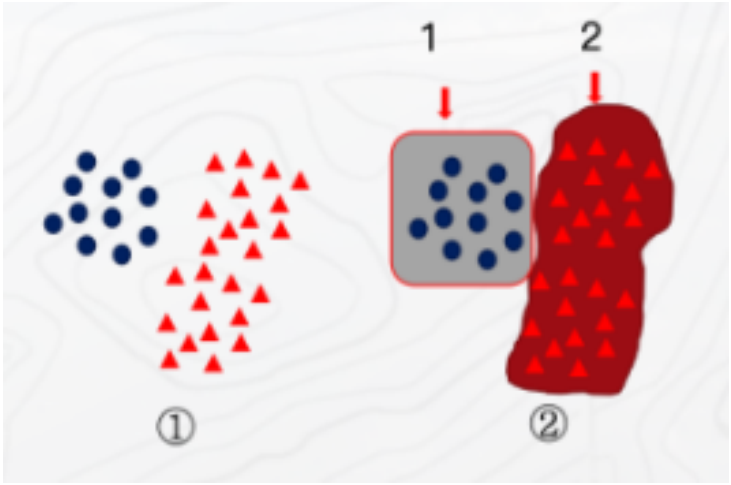


图 1: 二分类任务示意图

3.2 多分类问题

多分类包含三个及以上类别，类别标签一般用 $y \in \{0,1,2,\ldots,n-1\}$（$n \geq 3$）表示。工程上常用 **One-vs-All（一对多）**策略将其转化为若干二分类问题，步骤如下：

1. 对 n 个类别，先定义其中一类为正类，其余 $n-1$ 个类别统一归为负类，完成一次二分类。

2. 去掉已划分的正类数据，对剩余样本重复上述过程，依次划分出剩余类别。

3. 共需完成 $n-1$ 次二分类决策。预测时，选择概率/得分最高的类别作为最终输出。

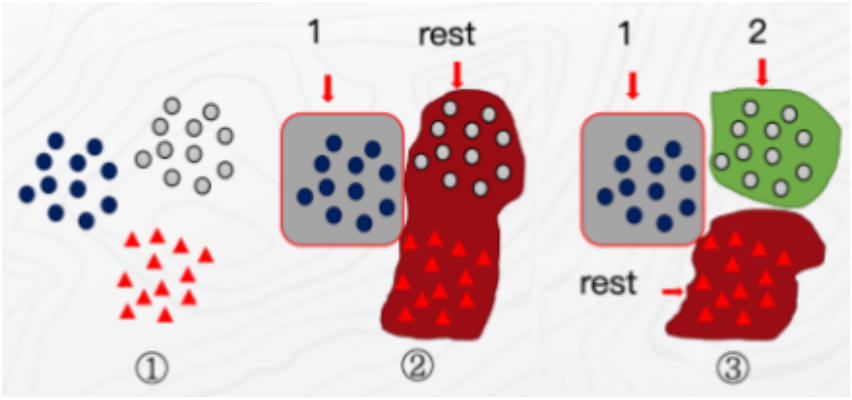


图 2: 多分类任务（One-vs-All 策略）示意图

4 Sigmoid 函数与逻辑回归

4.1 为什么需要 Sigmoid 函数

线性回归的核心公式为 $z = \boldsymbol{w}^T\boldsymbol{x} + b$，其输出 z 的取值范围为 $(-\infty, +\infty)$，无法直接解释为概率。为此，需要引入 Sigmoid 函数将输出压缩到 $(0, 1)$ 区间。

4.2 Sigmoid 函数定义

Sigmoid 函数（S 形单调递增函数）的数学表达式为：

$$\sigma(z) = \frac{1}{1 + e^{-z}}, \quad z = \boldsymbol{w}^T\boldsymbol{x} \tag{1}$$

关键性质：

- 定义域：$z \in \mathbb{R}$（全体实数）；值域：$\sigma(z) \in (0, 1)$，天然可表示概率。
- 中心点：$\sigma(0) = 0.5$；单调递增、光滑连续、处处可导。
- 导数满足：$\sigma'(z) = \sigma(z)(1 - \sigma(z))$，计算极方便。

- 当 $z \to +\infty$ 时 $\sigma(z) \to 1$；当 $z \to -\infty$ 时 $\sigma(z) \to 0$。

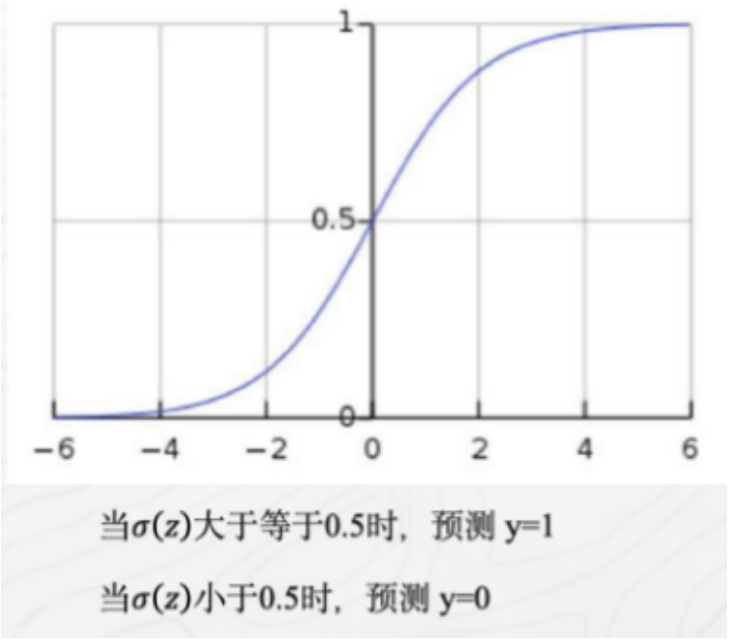


图 3: Sigmoid 函数曲线

4.3 逻辑回归模型

逻辑回归（Logistic Regression）是**线性模型 + Sigmoid 概率映射**，专门解决二分类问题。其假设函数（Hypothesis Function）表示样本属于正类的概率：

$$h_{\boldsymbol{w}}(\boldsymbol{x}) = \sigma(\boldsymbol{w}^T\boldsymbol{x}) = \frac{1}{1+e^{-\boldsymbol{w}^T\boldsymbol{x}}} \tag{2}$$

分类决策规则：

$$\hat{y} = \begin{cases} 1, & h_{\boldsymbol{w}}(\boldsymbol{x}) \geq 0.5 \;\Leftrightarrow\; \boldsymbol{w}^T\boldsymbol{x} \geq 0 \\ 0, & h_{\boldsymbol{w}}(\boldsymbol{x}) < 0.5 \;\Leftrightarrow\; \boldsymbol{w}^T\boldsymbol{x} < 0 \end{cases} \tag{3}$$

逻辑回归不是直接输出类别，而是先输出概率，再依据阈值进行判别，决策边界即为 $\boldsymbol{w}^T\boldsymbol{x} = 0$。

5 逻辑回归的损失函数推演

5.1 引言

逻辑回归虽然名称中含有” 回归”，但其最典型的应用场景是二分类问题。一个核心问题是：为什么其损失函数不是常见的平方误差，而是

$$L(\hat{y}, y) = -y \log \hat{y} - (1-y) \log(1-\hat{y})$$

本节将按照” 概率建模 $\rightarrow$ 似然函数 $\rightarrow$ 对数似然 $\rightarrow$ 损失函数 $\rightarrow$ 交叉熵” 的主线展开推导。

5.2 单样本概率的统一表达

在二分类任务中，标签 $y \in \{0, 1\}$。逻辑回归希望学习模型使其对输入 x 输出属于正类的概率：

$$P(y=1 \mid x; \boldsymbol{w}) = \hat{y} = h_{\boldsymbol{w}}(x), \quad P(y=0 \mid x; \boldsymbol{w}) = 1 - h_{\boldsymbol{w}}(x)$$

利用 $y \in \{0, 1\}$ 的性质，可将两种情况合并为一个统一表达式：

$$P(y \mid x; \boldsymbol{w}) = \big(h(x)\big)^{y} \big(1 - h(x)\big)^{1-y} \tag{4}$$

验证：当 $y = 1$ 时，得 $h(x)$；当 $y = 0$ 时，得 $1 - h(x)$，与原定义完全一致。

5.3 从单样本到似然函数

设训练集为 $\{(x^{(i)}, y^{(i)})\}_{i=1}^{m}$，假设样本之间相互独立，则整个训练集的联合概率（即**似然函数**）为：

$$L(\boldsymbol{w}) = \prod_{i=1}^{m} \big(h(x^{(i)})\big)^{y^{(i)}} \big(1 - h(x^{(i)})\big)^{1-y^{(i)}} \tag{5}$$

训练的自然目标是找到使训练数据出现概率最大的参数 $\boldsymbol{w}$，即**最大似然估计（MLE）**：

$$\arg\max_{\boldsymbol{w}} L(\boldsymbol{w})$$

5.4 对数似然与其优势

直接最大化连乘形式的似然函数不便计算，故对其取自然对数：

$$\ell(\boldsymbol{w}) = \log L(\boldsymbol{w}) = \sum_{i=1}^{m} \left[y^{(i)} \log h(x^{(i)}) + (1 - y^{(i)}) \log\big(1 - h(x^{(i)})\big) \right] \tag{6}$$

取对数的优势：

1. **化乘积为求和**，推导和优化更方便。
2. **数值更稳定**，避免多个小概率连乘后的数值下溢。
3. **单调性不变**，最大化 $L(\boldsymbol{w})$ 与最大化 $\ell(\boldsymbol{w})$ 等价。

5.5 从对数似然到损失函数

机器学习惯用" 最小化损失函数"，故对对数似然加负号并取平均，得到**代价函数（交叉熵损失）**：

$$J(\boldsymbol{w}) = -\frac{1}{m} \sum_{i=1}^{m} \left[y^{(i)} \log h_{\boldsymbol{w}}(x^{(i)}) + (1 - y^{(i)}) \log\big(1 - h_{\boldsymbol{w}}(x^{(i)})\big) \right] \tag{7}$$

单样本损失可写为：

$$L(\hat{y}, y) = -y \log \hat{y} - (1 - y) \log(1 - \hat{y})$$

- 当 $y = 1$：$L = -\log \hat{y}$，$\hat{y}$ 越接近 1，损失越小；$\hat{y}$ 越接近 0，损失迅速增大。
- 当 $y = 0$：$L = -\log(1 - \hat{y})$，$\hat{y}$ 越接近 0，损失越小；$\hat{y}$ 越接近 1，损失迅速增大。

该损失函数惩罚的是" **错误而且自信**" 的预测，比平方误差更适合分类问题。

5.6 交叉熵的信息论解释

在信息论中，交叉熵衡量真实分布 p 与预测分布 q 之间的差异：$H(p,q) = -\sum_i p_i \log q_i$。对于二分类，将真实分布 $p = (y, 1-y)$ 与预测分布 $q = (\hat{y}, 1-\hat{y})$ 代入，恰好得到：

$$H(p,q) = -y \log \hat{y} - (1-y) \log(1-\hat{y})$$

与逻辑回归单样本损失函数完全一致，因此逻辑回归的损失函数正是**二分类交叉熵**。

$$\boxed{\text{逻辑回归损失函数} = \text{平均负对数似然} = \text{二分类交叉熵损失}}$$

这三种说法来自不同视角（概率统计、优化、信息论），本质上描述的是同一个目标函数。

5.7 为什么不用平方误差

- **概率建模角度**：逻辑回归基于伯努利分布建模，由最大似然自然推出的损失就是负对数似然，而非平方误差。
- **优化行为角度**：交叉熵对高置信度错误预测给出更强的惩罚，能更快推动模型修正方向；且逻辑回归的交叉熵损失是**凸函数**，梯度下降可以找到全局最优解。

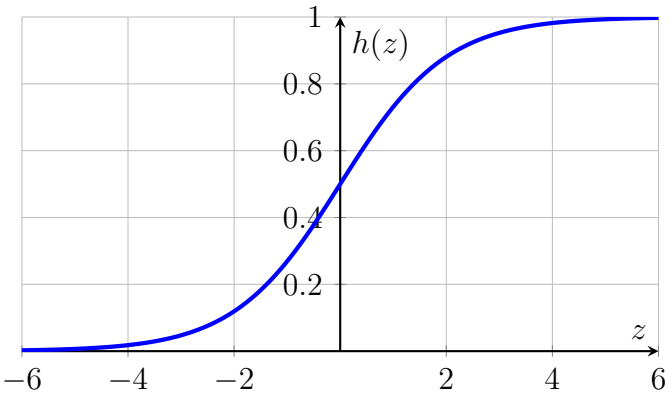
6 逻辑回归模型

逻辑回归用于解决二分类问题，其核心思想是通过线性组合后接一个 Sigmoid 函数，将结果映射为概率值：

$$h(x) = \frac{1}{1 + e^{-w^T x}}$$

其中，$w^T x$ 表示特征的线性组合，Sigmoid 函数将其压缩到 $(0,1)$ 区间。

6.1 Sigmoid 函数图像



Sigmoid 函数具有良好的性质：单调递增、连续可导，非常适合用于概率建模。

7 损失函数及其意义

逻辑回归采用对数似然损失函数（交叉熵损失）：

$$J(w) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log(h(x^{(i)})) + (1-y^{(i)})\log(1-h(x^{(i)}))\right]$$

解释：

- 当 $y=1$ 时，希望 $h(x)$ 越接近 1 越好；
- 当 $y=0$ 时，希望 $h(x)$ 越接近 0 越好；
- 该损失函数本质是在最大化数据出现的概率（最大似然估计）。

8 梯度下降法求解过程

逻辑回归没有解析解，因此需要使用梯度下降法求最优参数。

8.1 第一步：计算损失函数对参数的梯度

对 w_j 求偏导：

$$\frac{\partial J(w)}{\partial w_j} = \frac{1}{m}\sum_{i=1}^{m}\left(h(x^{(i)}) - y^{(i)}\right)x_j^{(i)}$$

理解：

- $h(x) - y$ 表示预测误差
- x_j 表示该特征对误差的贡献

8.2 第二步：更新参数

梯度下降更新规则：

$$w_j := w_j - \alpha\frac{\partial J(w)}{\partial w_j}$$

代入梯度：

$$w_j := w_j - \alpha \cdot \frac{1}{m}\sum_{i=1}^{m}\left(h(x^{(i)}) - y^{(i)}\right)x_j^{(i)}$$

8.3 第三步：迭代直到收敛

不断重复：

1. 计算预测值 $h(x)$
2. 计算损失函数
3. 计算梯度
4. 更新参数

直到：

- 损失函数变化很小

- 或达到最大迭代次数

9 分类决策边界

逻辑回归通过以下规则进行分类：

$$h(x) \geq 0.5 \Rightarrow y = 1, \quad h(x) < 0.5 \Rightarrow y = 0$$

对应决策边界：

$$w^T x = 0$$

10 逻辑回归的优点

1. **模型简单，易解释**每个 w_j 表示对应特征对结果的影响方向和强度，具有良好的可解释性。
2. **计算效率高**模型是线性的，训练和预测复杂度低，适合处理大规模数据。
3. **输出概率**输出结果为概率值，可以根据不同应用灵活调整分类阈值，而不是固定分类。
4. **不易过拟合**（在特征不多时）相比复杂模型，逻辑回归更稳定，泛化能力较好。
5. **可扩展性强**可以通过加入正则化（L_1/L_2）来防止过拟合。

11 逻辑回归的缺点

1. **线性模型，表达能力有限**只能学习线性决策边界，对于复杂非线性问题效果较差。
2. **依赖特征工程**需要人工构造特征（如多项式特征），否则模型性能受限。

3. **对异常值敏感**数据中的异常点可能对模型产生较大影响。

4. **难以处理高维非线性关系**对于图像、语音等复杂数据，通常需要更强模型（如神经网络）。

5. **类别不平衡问题明显**在正负样本差距较大时，模型容易偏向多数类。

12 总结

逻辑回归是一种经典的概率分类模型，通过 Sigmoid 函数实现概率映射，并利用梯度下降优化参数。其优点是简单高效、易解释，但在复杂问题中需要结合特征工程或更高级模型。

13 支持向量机（SVM）算法学习报告

13.1 什么是支持向量机？

支持向量机（Support Vector Machine，简称 SVM）是一种经典的监督学习分类模型，属于广义线性分类器。其核心目标是寻找一个最优决策超平面，使得两类样本之间的间隔最大化，从而获得最稳定、泛化能力最强的分类边界。

13.1.1 核心思想

- 找到位于样本集合边缘的关键数据点，称为**支持向量**。
- 仅依靠这些支持向量确定最终的决策超平面。
- 追求**间隔最大化**，让超平面距离两类样本都尽可能远。

13.2 SVM 核心数学表达

SVM 的目标是最大化分类间隔，等价于最小化权值向量的模长。

13.2.1 线性可分 SVM 目标函数

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2$$

$$\text{s.t.} \quad y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) \geq 1, \quad i = 1, 2, \ldots, N$$

其中 $\boldsymbol{w}$ 为法向量，b 为偏置，y_i 为样本标签（$+1$ 或 -1），$\boldsymbol{x}_i$ 为输入样本。

13.2.2 核心技巧（非线性分类）

当数据线性不可分时，使用核函数将低维数据映射到高维空间：

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x}_j)$$

常用核函数：线性核、多项式核、高斯核（RBF）、Sigmoid 核。

13.3 SVM 超平面示意图

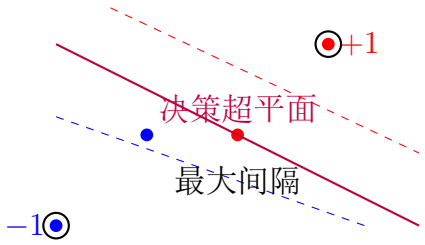


图 4: SVM 最大间隔超平面与支持向量示意图

13.4 SVM 适用场景

1. **高维小样本场景**：特征维度远大于样本数量（如文本分类、基因序列分析）。
2. **小样本学习**：数据量较少时仍能稳定训练，优于深度学习。
3. **非线性可分数据分类**：通过核函数将低维数据映射到高维空间实现线性可分。
4. **需要高可解释性与理论可靠性的任务**：数学基础严谨，边界清晰。

13.5 学习 SVM 的意义

- 机器学习经典算法，是理解分类、间隔、核方法的基础。
- 在高维小样本任务中不可替代。
- 与深度学习特征映射思想相通，能加深对 AI 模型的理解。

13.6 总结

- SVM 是一种监督学习分类模型，核心是找到间隔最大的决策超平面。
- 分类边界只由少数支持向量决定，抗干扰能力强。
- 擅长高维小样本、小样本、非线性数据分类。
- 核技巧是解决非线性问题的关键。

14 支持向量机：软间隔、松弛变量与超参数调优

摘要

本笔记基于课堂讲授内容，详细整理了支持向量机（SVM）在处理线性不可分数据时的关键技术——软间隔与松弛变量。同时结合大模型扩充，深入探讨了惩罚参数 C 与核参数 γ 对模型性能的影响及其调优策略。

14.1 核心概念：软间隔与松弛变量

14.1.1 背景与动机

在理想的“硬间隔”支持向量机中，我们假设数据是绝对线性可分的。然而在现实世界中，这种假设往往过于苛刻，主要面临以下挑战：

- **无法收敛**：当数据包含噪声或本质不可分时，优化算法找不到满足所有约束的解。
- **过拟合 (Overfitting)**：为了强行包容离群点（Outliers），分类面会变得极其扭曲，导致在测试集上的泛化能力极差。

14.1.2 软间隔 (Soft Margin)

软间隔机制允许部分样本点“违反”规则。即允许某些样本进入间隔带，甚至被错误分类。其核心思想是在**最大化间隔**与**最小化分类错误**之间寻求一种折中（Trade-off）。

14.1.3 松弛变量 (Slack Variable) ξ

为了量化“违反规则”的程度，我们为每个样本点 (x_i, y_i) 引入一个非负变量 $\xi_i \geq 0$。其几何意义如下：

- $\xi_i = 0$：样本点落在正确分类的间隔边界外（或边界上）。
- $0 < \xi_i \leq 1$：样本点在间隔带内部，但仍能被正确分类。
- $\xi_i > 1$：样本点跨过了决策边界，被错误分类。

14.2 数学优化模型

14.2.1 原始问题 (Primal Problem)

引入松弛变量后，SVM 的优化目标不再仅仅是最小化 $\frac{1}{2}\|w\|^2$，还需考虑所有松弛变量的和。目标函数定义为：

$$\min_{w,b,\xi} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i \tag{8}$$

满足约束条件：

$$\begin{cases} y_i(w^T x_i + b) \geq 1 - \xi_i \\ \xi_i \geq 0, \quad i = 1, 2, \ldots, m \end{cases} \tag{9}$$

其中，C 被称为**惩罚参数**（Penalty Parameter）。

14.2.2 对偶问题 (Dual Problem)

通过拉格朗日对偶性，我们得到其对偶形式。软间隔相比硬间隔，在数学上最显著的变化是拉格朗日乘子 α_i 的取值范围增加了一个上限 C（盒约束）：

$$\max_{\alpha} \quad \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \tag{10}$$

满足：

$$s.t. \quad 0 \leq \alpha_i \leq C, \quad \sum_{i=1}^{m} \alpha_i y_i = 0 \tag{11}$$

14.3 超参数的影响与调优

14.3.1 惩罚参数 C 的作用

C 决定了模型对“错分”的惩罚力度。

- **高 C 值**：极其重视准确性。决策边界会变得非常精细甚至复杂，旨在减少训练集的误差，极易导致**过拟合**。
- **低 C 值**：追求更平滑的决策面。它对误分类的宽容度较高，虽然训练精度可能下降，但通常能获得更好的**泛化性能**。

14.3.2 核参数 γ (针对 RBF 核)

γ 控制了单个支持向量的影响范围（即高斯核的宽度）。

- **γ 较大**：支持向量的影响力集中在局部，边界会围绕支持向量产生“岛状”隆起，容易过拟合。
- **γ 较小**：支持向量的影响力扩散到全局，边界会非常平滑，可能导致欠拟合。

14.3.3 实践建议

在实际科研中，推荐使用 **网格搜索 (Grid Search)** 在指数范围内寻找最优组合：$C \in \{2^{-5}, \dots, 2^{15}\}$, $\gamma \in \{2^{-15}, \dots, 2^{3}\}$.

14.4 结论

软间隔 SVM 的引入，标志着分类器从“绝对严谨”向“统计泛化”的转变。通过调节松弛变量和惩罚参数，我们可以构建出在现实噪声环境下依然稳健的分类模型。