



机器学习与概率基础笔记

一、机器学习基础

1. 核心目标

机器学习的目标是从**数据 (Data)** 中自动学习出**规则 (Rule)** 或 **模式 (Pattern)**。

- **类比:** 不是直接告诉机器规则 (if-then), 而是给机器大量**例子 (Examples)**, 让其自行**归纳总结**。
- **例子:** 给机器看大量猫狗图片, 让其自己学习区分特征, 而非人为定义“猫有尖耳朵”。

2. 实现手段: 数学模型

机器通过**数学模型 (Mathematical Model)** 来描述学习到的规则。

- **模型:** 通常是一个函数或公式, 包含输入、输出和可调参数。

$$y = f(x, w)$$

其中:

- x : **输入 (Input)** - 提供给模型的数据 (如图片、文本)。
- w : **参数 (Parameters)** - 模型内部需要学习和调整的变量, 代表模型的“知识”或“权重”。
- f : **模型/函数 (Model/Function)** - 定义了如何根据输入 x 和参数 w 计算输出。
- y : **输出 (Output)** - 模型根据输入计算得到的结果 (如分类标签、预测值)。

3. “学习”的本质: 参数推断/估计

机器学习中的**学习 (Learning)** 过程, 本质上就是**调整模型参数 w** 的过程。

- **目标:** 通过分析大量**训练数据** (已知的输入 x 和对应的**正确输出 y_{true}**), 不断调整参数 w , 使得模型的输出 $y = f(x, w)$ 越来越接近 y_{true} 。
- **过程:**
 - i. 给定输入 x 。

- ii. 模型使用当前参数 w 计算输出 $y = f(x, w)$ 。
- iii. 比较 y 与真实输出 y_{true} 的差异 (计算损失 Loss)。
- iv. 根据差异调整参数 w , 以减小未来的差异。
- v. 重复此过程, 直到 w 足够好。
- **成功标志:** 当模型 (即确定下来的参数 w) 能够对未曾见过的新数据也做出准确预测时, 我们称机器“学会”了。

4. 简单示例 (分类问题)

- **输入 (Input x):** 不同的图形模式 (如红蓝方块组成)。
 - **输出 (Output y):** 图形对应的类别标签 (如 0 或 1)。
 - **学习任务:** 找到合适的参数 w , 使得模型 $f(x, w)$ 能够根据输入的图形 x 正确预测其类别 0 或 1。
-

二、机器学习过程图示

1. 学习样本 (使用模型表示)

将训练数据表示为模型需要满足的等式:

- $0 = f(\text{图形1}, w)$
 - $0 = f(\text{图形2}, w)$
 - $1 = f(\text{图形3}, w)$
 - $1 = f(\text{图形4}, w)$
- 这里的 w 是待学习的参数。

2. 学习过程 (箭头)

代表机器通过分析上述样本, 不断调整参数 w 的过程。

3. 学习结果 (Learned Rule / Decision Boundary)

学习完成后, 得到的参数 w 定义了一个具体的模型 $f(x, w)$ 。

- **可视化:** 可以看作机器在输入空间中画出的**决策边界**, 区分不同类别的区域。例如, 哪些模

式被归为 1, 哪些被归为 0。

- **简化表示:** 学习到的规则可能关注某些关键特征 (如蓝色中心 vs. 蓝色边缘)。

三、数据集 (Dataset)

1. 定义

用于机器学习模型学习的数据集合。

2. 数据种类与形式

数据可以是数字、文本、图像、声音等多种形式。

3. 按时间依赖性分类

一种重要的数据分类方式是看数据点之间是否存在**时间顺序依赖**。

- **时间依赖数据 (时序数据 - Time Series Data):**
 - 数据点的**顺序**和**时间间隔**非常重要。
 - 后续数据通常与前面数据相关。
 - **例子:** 股票价格、气温记录、心电图。横轴通常是时间。

$$(x_1, t_1), (x_2, t_2), \dots, (x_n, t_n) \quad \text{where } t_1 < t_2 < \dots < t_n$$

- **非时间依赖数据 (截面数据 - Cross-sectional Data):**
 - 每个数据点相对独立, 顺序通常不重要。
 - **例子:** 用户信息 (年龄、性别)、静态图像分类 (之前的红蓝方块)。

四、图像作为数据 (Images as Data)

图像对于计算机而言也是一种数据, 需要转化为数字形式处理。

1. 向量表示 (Vector Representation)

- 忽略像素的空间结构，将图像像素按行（或列）展开成一个长向量。
- 二值图像 (Binary):** 每个像素用 0 或 1 表示 (如 黑/白, 或示例中的 蓝/红)。
一个 $H \times W$ 的二值图像可以表示为 $\mathbf{x} \in \{0, 1\}^{H \times W}$ 或拉直为 $\mathbf{v} \in \mathbb{R}^{H \times W}$ (实际为 $\{0, 1\}$)。
- 彩色图像 (RGB):** 每个像素由 红(R)、绿(G)、蓝(B) 三个通道的强度值表示 (通常 0-255)。
一个 $H \times W$ 的 RGB 图像表示为 $\mathbf{x} \in \mathbb{Z}^{H \times W \times 3}$ 或拉直为 $\mathbf{v} \in \mathbb{R}^{H \times W \times 3}$ (实际为 $\{0, \dots, 255\}$)。
 - 纯红: (255, 0, 0)
 - 纯蓝: (0, 0, 255)

2. 矩阵表示 (Matrix Representation)

- 更自然地保留图像的二维空间结构。
 - 一个 $H \times W$ 的灰度图或二值图像可以直接表示为一个 $H \times W$ 的矩阵。
 - 彩色图像可以表示为 $H \times W \times 3$ 的三维张量 (Tensor)。
-

五、学习的抽象定义

1. 机器学习流程

建立一个框架/机制，能根据输入数据 x 自动产生合适的输出数据 y 。

2. 数学表示

$$y = f(x)$$

- x : 输入 (Input)
- y : 输出 (Output)
- f : 函数/模型 (Function/Model), 代表输入到输出的映射规则。

3. “学习”的核心

学习的过程就是确定函数 f 的具体形式 (通常是通过确定其内部参数 w) 的过程。

Learning \equiv Finding the optimal $f(\cdot)$ (often by finding optimal w in $y = f(x, w)$)

通过分析训练数据 (x, y_{true}) 来找到能最好拟合数据并具有泛化能力的 f 。

六、机器学习类型：监督学习 (Supervised Learning)

1. 定义

像有老师指导一样学习。

2. 核心要素：带标签的训练数据

使用**训练数据 (Training Data)** 进行学习，训练数据包含**输入 x** 和其对应的**正确输出 y** (标签/答案) 的配对。

$$\mathcal{D}_{train} = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$$

3. 学习目标

找到一个模型 f (通过确定参数 w)，使得对于训练数据中的 x_i ，模型的输出 $f(x_i, w)$ 尽可能接近真实的 y_i 。

$$\min_w \sum_{i=1}^N \text{Loss}(f(x_i, w), y_i)$$

4. 名称来源

因为有“正确答案” y 作为监督信号来指导模型的学习，故名“监督学习”。

5. 例子

- 图像分类 (输入图片 x , 输出类别标签 y)
 - 回归预测 (输入房屋特征 x , 输出房价 y)
-

七、机器学习类型：无监督学习与强化学习

1. 无监督学习 (Unsupervised Learning)

- **特点:** 训练数据只有输入 x , 没有对应的标签 y 。

$$\mathcal{D}_{train} = \{x_1, x_2, \dots, x_N\}$$

- **目标:** 让机器自己发现数据中的**内在结构、模式或关系**。
- **典型例子:**
 - **聚类 (Clustering):** 将相似的数据点自动分组。
 - **降维 (Dimensionality Reduction):** 找到数据的更紧凑表示。

2. 强化学习 (Reinforcement Learning)

- **框架:** 不同于监督/无监督。
 - **核心思想:** 智能体 (Agent) 在环境 (Environment) 中通过**试错 (Trial and Error)** 来学习。
 - i. Agent 观察环境状态 s 。
 - ii. Agent 采取一个行动 a 。
 - iii. 环境根据行动 a 给出**奖励 (Reward) r** 或惩罚, 并转移到新状态 s' 。
 - iv. Agent 的目标是学习一个**策略 (Policy) $\pi(s) \rightarrow a$** , 使得长期累积奖励最大化。
 - **典型例子:**
 - 游戏 AI (如 AlphaGo)
 - 机器人控制
 - **老虎机问题 (Bandit Problems):** 在多个选项中找到回报最高的选项。
 - **特点:** 延迟奖励, 需要探索与利用的平衡。常与深度学习结合 (Deep Reinforcement Learning)。
-

八、概率论复习：基础概念

1. 随机变量 (Random Variables)

其值是随机现象结果的变量 (e.g., X, Y)。

- 离散随机变量: 取有限或可数个特定值 (e.g., x_1, x_2, \dots)。
- 连续随机变量: 取值在某个区间内连续。

2. 概率分布 (Probability Distribution)

描述随机变量取不同值的可能性。

- $p(X = x_i)$: 离散变量 X 取值为 x_i 的概率。
- $p(x)$: 连续变量 X 在值 x 处的概率密度函数 (PDF)。

3. 联合概率 (Joint Probability)

两个或多个随机变量**同时**取特定值的概率。

- $p(X = x_i, Y = y_j)$: X 取 x_i **并且** Y 取 y_j 的概率。

4. 边缘概率 (Marginal Probability)

只关心某个变量的概率，忽略其他变量。通过**加法规则 (Sum Rule)** 从联合概率计算得到。

- **加法规则 (Sum Rule)**:

$$p(X = x_i) = \sum_{j=1}^{N_Y} p(X = x_i, Y = y_j)$$

(对于离散变量)

$$p(x) = \int p(x, y) dy$$

(对于连续变量)

- **边缘化 (Marginalization)**: 上述通过对 Y 的所有可能值求和（或积分）来获得 X 的边缘概率的过程，称为将 Y **边缘化掉 (marginalizing out Y)**。
-

九、概率论复习：条件概率

1. 定义 (Conditional Probability)

在**某个事件已经发生**的条件下，另一个事件发生的概率。

- $p(Y = y_j | X = x_i)$: 在事件 $X = x_i$ **已经发生**的条件下，事件 $Y = y_j$ 发生的概率。
- 竖线 $|$ 读作 "given" 或 "在...条件下"。

2. 乘法规则 (Product Rule)

连接联合概率、条件概率和边缘概率。从条件概率定义导出：

$$p(X = x_i, Y = y_j) = p(Y = y_j | X = x_i)p(X = x_i)$$

也可以写成：

$$p(X = x_i, Y = y_j) = p(X = x_i | Y = y_j)p(Y = y_j)$$

3. 直观理解

要让两件事 A 和 B 同时发生，其概率等于 A 发生的概率乘以在 A 发生的条件下 B 发生的概率。

十、概率分布的简化记法

为了书写方便，常用简化记法：

- $p(X)$: 表示随机变量 X 的**概率分布** (函数本身)。
- $p(x)$: 表示随机变量 X 取**特定值** x 的**概率** (或概率密度)。即 $p(X = x)$ 的简写。
- $p(X, Y)$: 联合概率 $p(X = x, Y = y)$ 的简写。
- $p(Y|X)$: 条件概率 $p(Y = y | X = x)$ 的简写。

使用简化记法重写基本规则：

- 加法规则 (Sum Rule):

$$p(X) = \sum_Y p(X, Y) \quad (\text{Discrete } Y)$$

$$p(x) = \int p(x, y) dy \quad (\text{Continuous } Y)$$

(通过对 Y 求和/积分, 得到 X 的边缘概率)

- 乘法规则 (Product Rule):

$$p(X, Y) = p(Y|X)p(X) = p(X|Y)p(Y)$$

十一、贝叶斯定理 (Bayes' Theorem)

1. 推导

基于乘法规则的两种写法:

$$p(X, Y) = p(Y|X)p(X)$$

$$p(X, Y) = p(X|Y)p(Y)$$

令两式右边相等:

$$p(Y|X)p(X) = p(X|Y)p(Y)$$

整理可得贝叶斯定理 (假设 $p(X) > 0$):

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

2. 公式

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

3. 重要性

- 贝叶斯理论是现代机器学习和统计推断的核心之一。
- 它提供了一种**翻转条件概率**的方法，允许我们根据观察到的**证据 (Evidence X)** 来**更新**我们对某个事件或参数 (Y) 的**信念 (Belief)**。

十二、贝叶斯定理的解读

$$\underbrace{p(Y|X)}_{\text{后验 Posterior}} = \frac{\overbrace{p(X|Y)}^{\text{似然 Likelihood}} \overbrace{p(Y)}^{\text{先验 Prior}}}{\underbrace{p(X)}_{\text{证据 Evidence}}}$$

- $p(Y)$: **先验概率 (Prior Probability)**
 - 在观察到任何数据 X 之前，我们对 Y 的初始信念或假设。
- $p(X|Y)$: **似然 (Likelihood)**
 - **假设 Y 为真 (或给定特定的 Y)**，我们观察到数据 X 的可能性有多大。
 - 它衡量了 Y 对数据 X 的解释程度。
 - **注意:** 当 X 固定时，似然是关于 Y 的函数 $L(Y) = p(X|Y)$ ，但它本身**不是** Y 的概率分布 (对 Y 积分不一定为 1)。
- $p(X)$: **证据 (Evidence) / 边缘似然 (Marginal Likelihood)**
 - 观察到数据 X 本身的总概率，与 Y 无关。
 - $p(X) = \sum_Y p(X|Y)p(Y)$ (离散) 或 $p(X) = \int p(X|Y)p(Y)dY$ (连续)。
 - 主要作用是**归一化**，确保后验概率对所有 Y 求和/积分为 1。
- $p(Y|X)$: **后验概率 (Posterior Probability)**
 - 在观察到数据 X 之后，我们对 Y 更新后的信念。
 - 这是贝叶斯推断的主要目标。

核心思想: 贝叶斯定理描述了一个**信念更新 (Belief Update)** 的过程:

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

(因为 $p(X)$ 对于给定的数据 X 是常数)

十三、概率的思考方式 (Philosophies of Probability)

1. 频率主义 (Frequentism)

- 概率定义:** 事件发生的**长期频率**，通过大量**重复的随机试验**得到。
- 特点:** 概率是客观存在的、描述事件固有属性的值。关注点：“在很多次试验中，这件事会发生多少次？”

2. 贝叶斯主义 (Bayesianism)

- 概率定义:** 对事件发生**不确定性 (Uncertainty)** 的度量，即**信念程度 (Degree of Belief)**。
- 特点:**
 - 概率可以是**主观的** (基于个人信息和判断) 或**客观的** (基于公认的先验知识)。
 - 关注点：“根据我目前的信息，我对这件事发生的相信程度是多少？”
 - 贝叶斯定理**描述了信念如何根据新证据进行更新。

十四、补充：现代概率论 (Kolmogorov Axioms)

1. 公理化基础

现代概率论由 Kolmogorov 基于**测度论 (Measure Theory)** 建立。

- 样本空间 (Sample Space) Ω :** 所有可能基本结果的集合。
- 事件域 (Sigma-Algebra) \mathcal{F} :** Ω 的子集构成的集合，包含我们关心的、可以赋予概率的事件。 \mathcal{F} 必须满足特定条件 (闭包性质)。
- 概率测度 (Probability Measure) P :** 定义在 \mathcal{F} 上的函数，给每个事件 $A \in \mathcal{F}$ 赋予一个 $[0, 1]$ 区间的概率值 $P(A)$ ，并满足：
 - 非负性: $P(A) \geq 0$ for all $A \in \mathcal{F}$.
 - 规范性: $P(\Omega) = 1$.
 - 可数可加性: 对于互不相交的事件 $A_1, A_2, \dots \in \mathcal{F}$, $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

2. 意义

- 提供了一个**统一的数学框架**。

- 该框架本身不偏袒频率主义或贝叶斯主义解释，只要计算出的概率满足公理，数学上都是有效的。
 - 两种哲学观点是对这个数学框架的不同应用和解读。
-

十五 & 十六、补充：可测空间 (Measurable Space)

1. σ -代数 (Sigma-Algebra) \mathcal{F}

给定一个非空集合 Ω (样本空间)，它的一个子集族 (集合的集合) \mathcal{F} 如果满足以下条件，则称为 Ω 上的 σ -代数：

1. $\Omega \in \mathcal{F}$ (全集/必然事件在族中)。 (Note: 原文写 $\emptyset \in M$ ，等价于 $\Omega \in M$ 因为有补集)
2. 若 $A \in \mathcal{F}$ ，则其补集 $A^c = \Omega \setminus A$ 也在 \mathcal{F} 中 (对补集运算封闭)。
3. 若 $A_1, A_2, \dots \in \mathcal{F}$ (可数个)，则它们的并集 $\bigcup_{i=1}^{\infty} A_i$ 也在 \mathcal{F} 中 (对可数并集运算封闭)。

2. 可测集 (Measurable Set)

属于 σ -代数 \mathcal{F} 中的集合 A 称为**可测集**或**事件 (Event)**。只有这些集合才能被赋予概率。

3. 可测空间 (Measurable Space)

配对 (Ω, \mathcal{F}) 称为一个可测空间。它定义了基础空间和哪些事件是可以测量的。

- **例子 (对于 $\Omega = \{A, B, C\}$):**
 - 最小 σ -代数: $\mathcal{F}_0 = \{\emptyset, \Omega\}$
 - 最大 σ -代数 (幂集): $\mathcal{F}_1 = 2^{\Omega} = \{\emptyset, \{A\}, \{B\}, \{C\}, \{A, B\}, \{A, C\}, \{B, C\}, \Omega\}$
 - (Ω, \mathcal{F}_0) 和 (Ω, \mathcal{F}_1) 都是合法的可测空间。
-

十七 & 十八、补充：测度空间与概率空间

1. 测度 (Measure) μ

在可测空间 (Ω, \mathcal{F}) 上定义的函数 $\mu : \mathcal{F} \rightarrow [0, \infty]$, 满足:

- 非负性: $\mu(A) \geq 0$ for all $A \in \mathcal{F}$, and $\mu(\emptyset) = 0$.
- 可数可加性: 对于互不相交的事件 $A_1, A_2, \dots \in \mathcal{F}$, $\mu(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$.

2. 测度空间 (Measure Space)

三元组 $(\Omega, \mathcal{F}, \mu)$ 称为测度空间。

3. 概率测度 (Probability Measure) P

一种特殊的测度, 其**总测度**为 1。即 $P(\Omega) = 1$ 。

4. 概率空间 (Probability Space)

满足 $P(\Omega) = 1$ 的测度空间 (Ω, \mathcal{F}, P) 称为概率空间。这是现代概率论的数学基础。

- Ω : 样本空间 (所有可能结果)
- \mathcal{F} : 事件域 (可以赋予概率的事件集合)
- P : 概率测度 (为每个事件分配概率的规则)

十九、参数估计与贝叶斯理论

1. 机器学习与参数估计

机器学习的核心任务之一是**参数估计 (Parameter Estimation)**: 根据数据 D 找到模型 $y = f(x, w)$ 中最优的参数 w 。

2. 贝叶斯参数估计视角

- 将参数 w 视为**随机变量**，具有不确定性。
- 使用**概率分布**来表达关于 w 的信念。
- **先验分布** $p(w)$: 在看到数据 D **之前**，对参数 w 的初始信念。
- **似然函数** $p(D|w)$: 给定参数 w 时，观察到数据 D 的可能性。
- **后验分布** $p(w|D)$: 在看到数据 D **之后**，对参数 w 更新后的信念。
- **贝叶斯定理**连接三者：

$$p(w|D) = \frac{p(D|w)p(w)}{p(D)}$$

其中 $p(D) = \int p(D|w)p(w)dw$ 是证据项 (归一化常数)。

3. 目标

贝叶斯推断的目标是计算**后验分布** $p(w|D)$ ，它包含了数据和先验信息中关于 w 的所有知识。

二十、似然函数 (Likelihood Function)

1. 定义

似然函数 $L(w; D) = p(D|w)$ 是参数 w 的函数，表示在**不同参数** w 下，我们**实际观测到的固定数据** D 出现的可能性。

2. 与概率分布的区别 (重要!)

- $p(D|w)$ 当 w 固定, D 变化时, 是**数据 D 的概率分布** (对 D 积分/求和为 1)。
- $p(D|w)$ 当 D 固定, w 变化时, 是**参数 w 的似然函数**。它**不是 w 的概率分布** (对 w 积分/求和通常不为 1)。

3. 作用

衡量不同参数值 w 对观测数据 D 的**解释程度**或**拟合程度**。似然值越大, 表示该参数 w 越能“解

释”观测到的数据 D 。

二十一、最大似然估计 (Maximum Likelihood Estimation, MLE)

1. 核心思想

选择能够**最大化似然函数**的参数值 w^* 作为参数的估计值。即，找到最能解释我们观测到的数据的参数。

2. 数学公式

$$w_{\text{MLE}}^* = \operatorname{argmax}_w p(D|w)$$

或等价地，最大化对数似然 (Log-Likelihood)，因为 \log 是单调函数：

$$w_{\text{MLE}}^* = \operatorname{argmax}_w \log p(D|w)$$

(对数似然通常在计算上更方便，尤其是当 D 包含多个独立样本时，联合概率 $p(D|w)$ 是连乘形式，取对数后变成连加形式)。

3. 图示理解

给定观测数据 D (图中的红线)，比较不同参数 w_1, w_2, w_3 对应的概率分布 $p(D|w)$ 在 D 处的值（似然值）。选择使得 $p(D|w)$ 最大的那个 w (如图中的 w_2) 作为 w_{MLE}^* 。

4. 特点

- 是一种常用的点估计方法。
 - 结果完全由数据驱动。
 - 对于小数据集或极端数据，结果可能不稳定或不合理。
-

二十二、例题 1.1：抛硬币的 MLE

场景：抛硬币 3 次。设 X 为正面朝上的次数。

(1) 公平硬币 ($p = 0.5$) 的概率分布

使用二项分布 $P(X = k) = C(n, k)p^k(1 - p)^{n-k}$, 其中 $n = 3, p = 0.5$ 。

- $P(X = 0) = C(3, 0)(0.5)^0(0.5)^3 = 1 \times 1 \times (1/8) = 1/8$
- $P(X = 1) = C(3, 1)(0.5)^1(0.5)^2 = 3 \times (1/2) \times (1/4) = 3/8$
- $P(X = 2) = C(3, 2)(0.5)^2(0.5)^1 = 3 \times (1/4) \times (1/2) = 3/8$
- $P(X = 3) = C(3, 3)(0.5)^3(0.5)^0 = 1 \times (1/8) \times 1 = 1/8$

(2) 未知参数 p 的似然函数

设正面概率为未知参数 p 。观测数据 D 是具体的抛掷结果 (e.g., 观察到 k 次正面)。似然函数 $L(p; D) = p(D|p)$ 是关于 p 的函数。

- 若观测到 $k = 3$ 正面 ($D=(3,0)$): $L(p) = p(X = 3|p) = C(3, 3)p^3(1 - p)^0 = p^3$
- 若观测到 $k = 2$ 正面 ($D=(2,1)$): $L(p) = p(X = 2|p) = C(3, 2)p^2(1 - p)^1 = 3p^2(1 - p)$
- 若观测到 $k = 1$ 正面 ($D=(1,2)$): $L(p) = p(X = 1|p) = C(3, 1)p^1(1 - p)^2 = 3p(1 - p)^2$
- 若观测到 $k = 0$ 正面 ($D=(0,3)$): $L(p) = p(X = 0|p) = C(3, 0)p^0(1 - p)^3 = (1 - p)^3$

(3) 最大似然估计 p^*

找到使似然函数 $L(p)$ 最大的 p 值 ($p^* = \operatorname{argmax}_p L(p)$)。

- 对于二项分布, 观测到 n 次试验中 k 次成功, MLE 结果为 $\hat{p}_{\text{MLE}} = k/n$ 。
- 若观测到 $k = 3$: $p^* = \operatorname{argmax}_p p^3 = 1$. (MLE: $\hat{p} = 3/3 = 1$)
- 若观测到 $k = 2$: $p^* = \operatorname{argmax}_p 3p^2(1 - p)$. 求导或直接用公式 $\hat{p} = 2/3$.
- 若观测到 $k = 1$: $p^* = \operatorname{argmax}_p 3p(1 - p)^2$. 求导或直接用公式 $\hat{p} = 1/3$.
- 若观测到 $k = 0$: $p^* = \operatorname{argmax}_p (1 - p)^3 = 0$. (MLE: $\hat{p} = 0/3 = 0$)

讨论:

- MLE 结果非常依赖于观测数据。少量或极端数据会导致极端估计 (如 $p^* = 0$ 或 $p^* = 1$)。
- **贝叶斯估计**可以通过引入**先验分布** $p(p)$ (如假设 p 可能接近 0.5) 来缓和 MLE 的极端性, 得到一个结合先验和数据的更稳健的估计结果。