概率统计基础与 AI 应用

整理：May

2025 年 8 月 21 日

目录

1 概率论基础（Probability Theory）

1.1 概率公理（Kolmogorov 公理

概率论的基础由三个公理组成：

1. 非负性：
$$P(A) \geq 0, \quad \forall A \subseteq \Omega$$
2. 规范化：
$$P(\Omega) = 1$$
3. 可列可加性（Additivity）：对互斥事件 $A_1, A_2, \ldots$
$$P\left(\bigcup_i A_i\right) = \sum_i P(A_i)$$

推论：

- 空事件概率为 0: $P(\emptyset) = 0$
- 对任意事件 A: $P(A^c) = 1 - P(A)$

1.2 条件概率与独立性

- 条件概率：
$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad P(B) > 0$$
- 独立事件：
$$P(A \cap B) = P(A)P(B)$$

应用：

- 朴素贝叶斯分类器：假设特征条件独立，计算类别概率
- 生成模型：独立假设用于简化联合分布

1.3 全概率公式与贝叶斯定理

- 全概率公式：
$$P(A) = \sum_i P(A|B_i)P(B_i)$$

- 贝叶斯定理：

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

应用：

- 贝叶斯推断（后验概率计算）
- 最大后验估计 (MAP)：

$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta|X) = \arg\max_{\theta} P(X|\theta)P(\theta)$$

2 随机变量与分布(Random Variables & Distributions)

2.1 随机变量

- 离散：取有限或可数值，如 $X \in \{0, 1\}$
- 连续：有概率密度函数（PDF）

$$P(a \leq X \leq b) = \int_a^b f_X(x)\, dx$$

2.2 期望、方差、协方差

- 期望：

$$\mathbb{E}[X] = \sum x_i P(X = x_i) \quad \text{（离散）}$$

$$\mathbb{E}[X] = \int x f_X(x)\, dx \quad \text{（连续）}$$

- 方差：

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

- 协方差：

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

2.3 常用分布及公式

分布	PDF/PMF	期望	方差	AI 应用
伯努利	$P(X=1)=p, P(X=0)=1-p$	p	$p(1-p)$	二分类、Dropout 模拟
二项	$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$	np	$np(1-p)$	样本统计
高斯	$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/2\sigma^2}$	μ	σ^2	回归、VAE、生成模型
指数	$f(x)=\lambda e^{-\lambda x}$	$1/\lambda$	$1/\lambda^2$	等待时间建模
多项	$P(X_1,\ldots,X_k)=\frac{n!}{x_1!\ldots x_k!}\prod p_i^{x_i}$	np_i	$np_i(1-p_i)$	语言模型、词频统计

3 最大似然与贝叶斯估计

3.1 最大似然估计 (MLE)

给定观测数据 $X=\{x_1,\ldots,x_n\}$ 和参数 θ:

$$\hat{\theta}_{MLE}=\arg\max_{\theta}\prod_{i=1}^{n}p(x_i|\theta)$$

对数似然:

$$\mathcal{L}(\theta)=\sum_{i=1}^{n}\log p(x_i|\theta)$$

应用:

- 线性回归参数估计
- 逻辑回归训练

3.2 贝叶斯估计 (MAP)

$$\hat{\theta}_{MAP}=\arg\max_{\theta}P(\theta|X)=\arg\max_{\theta}P(X|\theta)P(\theta)$$

应用:

- 朴素贝叶斯
- 贝叶斯神经网络

4 概率统计在 AI 中的应用场景

场景	概率工具	公式/方法
分类	条件概率、贝叶斯	朴素贝叶斯：$P(C\|X) \propto P(X\|C)P(C)$
回归	高斯分布假设	$\hat{\theta} = (X^T X)^{-1} X^T y$
生成模型	MLE, KL 散度, 高斯	VAE ELBO: $\mathbb{E}_{q(z)}[\log p(x\|z)] - KL(q(z)\|\|p(z))$
强化学习	马尔可夫决策过程	$P(s_{t+1}\|s_t, a_t)$
NLP/LLM	多项分布, 交叉熵	$P(w_1, \ldots, w_n) = \prod_t P(w_t\|w_{<t})$
不确定性	方差, 协方差, 高斯	Bayesian NN: 后验方差估计

5 常见证明思路

1. MLE 推导对数似然取导 $\rightarrow$ 令 $\frac{\partial \mathcal{L}}{\partial \theta} = 0 \rightarrow$ 解参数示例：线性回归的 $\theta = (X^T X)^{-1} X^T y$
2. 贝叶斯公式推导从条件概率定义：

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

   联合概率分解：

$$P(A \cap B) = P(B|A)P(A) \quad \Rightarrow \quad \text{得贝叶斯定理}$$

3. 方差公式

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

   展开平方 $\rightarrow$ 分解期望 $\rightarrow$ 得公式

6 线性代数基础 (Linear Algebra Basics)

6.1 公理与定义

6.1.1 向量空间 (Vector Space)

一个集合 V 对向量加法和标量乘法封闭，并满足：

1. 加法交换律：$\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$
2. 加法结合律：$(\mathbf{u} + \mathbf{v}) + \mathbf{w} = \mathbf{u} + (\mathbf{v} + \mathbf{w})$
3. 存在零向量：$\mathbf{v} + \mathbf{0} = \mathbf{v}$
4. 存在负向量：$\mathbf{v} + (-\mathbf{v}) = \mathbf{0}$

5. 标量乘法结合律：$a(b\mathbf{v}) = (ab)\mathbf{v}$

6. 单位元：$1\mathbf{v} = \mathbf{v}$

7. 分配律：$a(\mathbf{u} + \mathbf{v}) = a\mathbf{u} + a\mathbf{v}$

6.1.2 矩阵 (Matrix)

- $A \in \mathbb{R}^{m \times n}$ 表示 $m \times n$ 的矩阵
- 基本运算：

$$
\begin{aligned}
A + B &= [a_{ij} + b_{ij}] \\
cA &= [ca_{ij}] \\
(AB)_{ij} &= \sum_k A_{ik} B_{kj} \\
A^T &= \text{转置矩阵} \\
AA^{-1} &= I \quad (\text{若 } A \text{ 可逆})
\end{aligned}
$$

6.1.3 内积与范数

$$
\begin{aligned}
\mathbf{u} \cdot \mathbf{v} &= \sum_i u_i v_i \\
\|\mathbf{v}\|_2 &= \sqrt{\mathbf{v} \cdot \mathbf{v}}
\end{aligned}
$$

6.2 线性代数在 AI 中的公式与应用

概念	公式	AI/ML 应用
向量表示	$\mathbf{x} \in \mathbb{R}^n$	特征向量, 词向量 (Word Embeddings)
矩阵运算	$Z = WX + b$	神经网络前向传播
矩阵转置	A^T	卷积核矩阵，梯度计算
矩阵逆	A^{-1}	线性回归闭式解: $\theta = (X^T X)^{-1} X^T y$
内积	$\mathbf{u} \cdot \mathbf{v}$	相似度计算（Cosine similarity）
范数	$\|\mathbf{v}\|_2$	正则化（L2）
特征值/特征向量	$A\mathbf{v} = \lambda\mathbf{v}$	PCA、SVD、降维
奇异值分解 (SVD)	$A = U\Sigma V^T$	低秩近似、矩阵补全
正交性	$U^T U = I$	正交初始化、QR 分解
线性方程组	$Ax = b$	参数求解、最小二乘

6.3 常用线性代数操作在 AI 模型中的公式

1. 线性回归:
$$\hat{\theta} = (X^T X)^{-1} X^T y$$

2. PCA 主成分分析:
$$\max_w \text{Var}(Xw), \quad s.t.\ \|w\|_2 = 1$$
解为协方差矩阵 $C = X^T X$ 的特征向量。

3. 神经网络前向传播:
$$\mathbf{h}^{(l)} = f(W^{(l)}\mathbf{h}^{(l-1)} + b^{(l)})$$

4. 梯度下降中的矩阵求导:
$$\frac{\partial \mathcal{L}}{\partial W} = \frac{\partial \mathcal{L}}{\partial \mathbf{h}} \frac{\partial \mathbf{h}}{\partial W}$$

5. SVD 低秩近似:
$$A \approx U_k \Sigma_k V_k^T$$
用于推荐系统、矩阵压缩。

6.4 常见证明思路

1. 逆矩阵存在性：若 $\det(A) \neq 0 \to A$ 可逆。
2. 特征值分解：对对称矩阵 $A = A^T$，特征值为实数，特征向量正交。
3. PCA 最大方差问题：
$$\max_w w^T C w, \quad s.t.\ \|w\|_2 = 1$$
拉格朗日乘子法 $\to$ 解为协方差矩阵特征向量。
4. 线性方程最小二乘：
$$\min_x \|Ax - b\|_2^2 \quad \Rightarrow \quad A^T A x = A^T b$$

6.5 总结

- 向量和矩阵是 AI 数据表示和计算的核心
- 特征分解和奇异值分解是降维和特征提取的基础
- 矩阵运算公式贯穿回归、神经网络、嵌入表示、推荐系统
- 线性代数证明方法多用拉格朗日法、矩阵运算规则、特征值性质

7 微积分基础（Calculus Fundamentals）

7.1 基本概念

函数：$y = f(x)$
极限：
$$\lim_{x \to a} f(x) = L$$
连续性：函数在 $x = a$ 连续当且仅当
$$\lim_{x \to a} f(x) = f(a)$$

7.2 导数（Derivatives）

定义：
$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$
多变量偏导：
$$\frac{\partial f}{\partial x_i} = \lim_{h \to 0} \frac{f(x_1, \ldots, x_i + h, \ldots, x_n) - f(x_1, \ldots, x_n)}{h}$$

梯度（Gradient）：

$$\nabla f = \left[\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n}\right]^T$$

应用：

- 优化目标函数（线性回归、神经网络）
- 反向传播（Deep Learning）
- 梯度下降法：

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta L(\theta)$$

7.3 高阶导数与 Hessian 矩阵

二阶导数：

$$f''(x) = \frac{d^2 f}{dx^2}$$

Hessian 矩阵：

$$H = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

应用：

- 二阶优化算法（Newton 方法）
- 凸性判断：若 $H \succeq 0 \Rightarrow f$ 是凸函数

7.4 积分（Integration）

定积分：

$$\int_a^b f(x)\, dx$$

不定积分：

$$F(x) = \int f(x)\, dx, \quad F'(x) = f(x)$$

应用：

- 概率分布函数：

$$P(a \le X \le b) = \int_a^b f_X(x)\, dx$$

- 期望：

$$\mathbb{E}[X] = \int x f_X(x)\, dx$$

- 损失函数累积（RL 或连续时间优化）

7.5 常用公式

链式法则：

$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}$$

多元链式法则：

$$\frac{\partial z}{\partial x_i} = \sum_j \frac{\partial z}{\partial y_j} \frac{\partial y_j}{\partial x_i}$$

泰勒展开：

$$f(x) \approx f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \dots$$

应用：

- 梯度近似
- 激活函数线性化
- 优化算法推导

7.6 多重积分与卷积

二重积分：

$$\iint_R f(x,y)\,dxdy$$

卷积公式（连续）：

$$(f * g)(t) = \int_{-\infty}^{\infty} f(\tau) g(t-\tau)\,d\tau$$

应用：

- CNN 中卷积核运算
- 概率密度函数卷积
- 信号处理 / LLM 特征提取

7.7 证明思路示例

1. 导数定义推导梯度下降更新公式

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta L(\theta)$$

来源：一阶泰勒展开 + 梯度负方向下降损失。

2. 链式法则在反向传播

$$\frac{\partial L}{\partial W} = \frac{\partial L}{\partial y} \frac{\partial y}{\partial W}$$

3. Hessian 判定凸性若 $H(x) \succeq 0$，则 $\forall v, v^T H v \geq 0$ 二阶泰勒展开：

$$f(x+v) \approx f(x) + \nabla f^T v + \tfrac{1}{2} v^T H v$$

7.8 微积分在 AI 的主要应用场景

场景	工具/公式	应用示例
线性回归	导数、梯度	最小二乘法求参数
逻辑回归	链式法则、梯度	Sigmoid 损失反向传播
神经网络	偏导数、链式法则、Hessian	反向传播、二阶优化
优化算法	梯度、Hessian、泰勒展开	GD / SGD / Newton 方法
卷积神经网络	多重积分、卷积公式	图像特征提取
强化学习	定积分、期望	累积奖励、策略梯度
LLM & NLP	多变量微积分	注意力机制梯度、优化损失函数

8 信息论在 AI 中的作用

作用	应用场景	公式 / 方法
测量不确定性	生成模型、NLP	熵 $H(X)$
衡量依赖	特征选择、嵌入	互信息 $I(X;Y)$
分布对齐	VAE、RLHF	KL 散度 $D_{KL}(P\|\|Q)$
模型训练目标	LLM、语言建模	交叉熵损失 $H(P,Q)$
模型压缩 / 信息瓶颈	表示学习	$\max I(Z;Y) - \beta I(X;Z)$

9 信息论与生成型模型 / NLP 的联系

1. 语言模型训练：

$$\mathcal{L}_{CE} = -\sum_t \log P_\theta(w_t|w_{<t})$$

2. 变分自编码器（VAE）：

$$\text{ELBO} = \mathbb{E}_{q(z|x)}[\log p(x|z)] - D_{KL}[q(z|x)||p(z)]$$

3. 信息瓶颈 (Information Bottleneck)：

$$\max I(Z;Y) - \beta I(Z;X)$$

4. RLHF (Reinforcement Learning from Human Feedback)：

$$\mathcal{L}_{RLHF} = D_{KL}(P_\theta||P_r) - \mathbb{E}[R]$$

10 常见公式总结

$$
\begin{aligned}
I(x) &= -\log P(x) \\
H(X) &= -\sum_x P(x)\log P(x) \\
H(Y|X) &= -\sum_{x,y} P(x,y)\log P(y|x) \\
I(X;Y) &= \sum_{x,y} P(x,y)\log \frac{P(x,y)}{P(x)P(y)} \\
D_{KL}(P||Q) &= \sum_x P(x)\log \frac{P(x)}{Q(x)} \\
H(P,Q) &= -\sum_x P(x)\log Q(x)
\end{aligned}
$$

11 信息论公理与证明思路

1. **非负性**：

$$H(X) \geq 0, \quad D_{KL}(P||Q) \geq 0$$

证明：由 Jensen 不等式推出

2. **链式法则（Chain Rule）**：

$$H(X,Y) = H(X) + H(Y|X)$$

证明思路：联合概率分解 $\rightarrow$ 信息量期望

3. **互信息对称性**：

$$I(X;Y) = I(Y;X) = H(X) - H(X|Y)$$

4. **KL 与交叉熵关系**：

$$D_{KL}(P||Q) = H(P,Q) - H(P)$$

证明：代入交叉熵定义即可