

1

Landscape of Statistics and Probability

本书概率统计全景

公式连篇，丛书最无聊的一章



概率论作为数学学科，可以而且应该从公理开始建设，和几何、代数的路一样。

The theory of probability as mathematical discipline can and should be developed from axioms in exactly the same way as Geometry and Algebra.

—— 安德雷·柯尔莫哥洛夫 (Andrey Kolmogorov) | 概率论公理化之父 | 1903 ~ 1987



1.1 必备数学工具：一个小测验

本书前文提到，《统计至简》一册的核心特点是——多元。《矩阵力量》中介绍的线性代数工具是本书的核心数学工具。因此，在开始本书阅读之前，请大家完成本节这个小测验。

如果大家能够轻松完成这个测验，欢迎大家开始本书后续内容学习；否则，建议大家重温《矩阵力量》中重要数学工具。

数据矩阵

给定数据矩阵 \mathbf{X} ，如何求其中心化数据、标准化数据、质心、协方差矩阵、相关系数矩阵？

协方差矩阵

给定 2×2 协方差矩阵 Σ ：

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho_{1,2}\sigma_1\sigma_2 \\ \rho_{1,2}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \quad (1)$$

什么条件下 Σ 是正定矩阵？

定义如下二元函数：

$$f(x_1, x_2) = \mathbf{x}^T \Sigma \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} \sigma_1^2 & \rho_{1,2}\sigma_1\sigma_2 \\ \rho_{1,2}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad (2)$$

这个函数的图像是什么？二元函数的等高线形状有何特点？

Cholesky 分解

对协方差矩阵 Σ 进行 Cholesky 分解：

$$\Sigma = \mathbf{R}^T \mathbf{R} \quad (3)$$

矩阵 Σ 能进行 Cholesky 分解的前提是什么？

分解结果 \mathbf{R} 的特点是什么？如何从几何角度理解 \mathbf{R} ？

特征值分解

对 Σ 特征值分解：

$$\Sigma = \mathbf{V} \mathbf{\Lambda} \mathbf{V}^T \quad (4)$$

为什么等式右侧第二个矩阵 V 对应转置运算？

矩阵 V 有什么特殊性质？如何从向量空间角度理解 V ？

矩阵 A 有什么特殊性质？

如果把 V 写成 $[v_1, v_2]$, (4) 可以如何展开？

奇异值分解

奇异值分解有哪四种类型？每种类型之间存在怎样的关系？

数据矩阵 X 奇异值分解可以获得其奇异值 s_j ，对 X 的协方差矩阵 Σ 特征值分解可以得到特征值 λ_j 。奇异值 s_j 和特征值 λ_j 存在怎样的量化关系？

多元高斯分布

多元正态分布的概率密度函数为：

$$f_{\mathbf{x}}(\mathbf{x}) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)}{(2\pi)^{\frac{D}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \quad (5)$$

$(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$ 的含义是什么？ $|\boldsymbol{\Sigma}|^{\frac{1}{2}}$ 的含义是什么？

马氏距离的定义是什么？马氏距离和欧氏距离差别是什么？

测验题目到此结束。

从“生活常识”到“冷冰冰”的公式

概率统计的背后的思想更贴近“生活常识”。人们在谈论可能性的时候，大脑就不自觉进入“概率统计”模式。下面举几个例子。看着天上云层很厚，可能两小时就会下雨。昨晚淋了雨，估计今天要感冒。估计这次考试通过率 80% 以上。咱们剪刀石头布，三局两胜决胜负。

可惜的是，当数学家将这些生活常识“翻译成”数学语言之后，它们就变成了冷冰冰“火星文”。

本章下面就用数学手册、备忘录这种范式概括本书中核心公式。每一节对应本书一个板块。

而本章之后，我们就用丰富的图形给这些公式以色彩和温度。

1.2 统计描述

给定随机变量 X 的样本 $\{x^{(1)}, x^{(2)}, \dots, x^{(n)}\}$, X 的样本均值为:

$$\mu_X = \frac{1}{n} \left(\sum_{i=1}^n x^{(i)} \right) = \frac{x^{(1)} + x^{(2)} + x^{(3)} + \dots + x^{(n)}}{n} \quad (6)$$

X 的样本方差为:

$$\text{var}(X) = \sigma_X^2 = \frac{1}{n-1} \sum_{i=1}^n \left(x^{(i)} - \mu_X \right)^2 \quad (7)$$

X 的样本标准差为:

$$\sigma_X = \text{std}(X) = \sqrt{\text{var}(X)} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n \left(x^{(i)} - \mu_X \right)^2} \quad (8)$$

对于样本数据, 随机变量 X 和 Y 的协方差为:

$$\text{cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^n \left(x^{(i)} - \mu_X \right) \left(y^{(i)} - \mu_Y \right) \quad (9)$$

对于样本数据, 随机变量 X 和 Y 的相关性系数为:

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \quad (10)$$

注意, 本书一般不从符号上区分总体、样本的均值、方差、标准差等。

1.3 概率

古典概率模型

设样本空间 Ω 由 n 个等可能事件构成, 事件 A 的概率为:

$$\Pr(A) = \frac{n_A}{n} \quad (11)$$

其中, n_A 为含于事件 A 的试验结果数量。

A 和 B 为样本空间 Ω 中的两个事件, 其中 $\Pr(B) > 0$ 。那么, 事件 B 发生的条件下事件 A 发生的条件概率为:

$$\Pr(A|B) = \frac{\Pr(A, B)}{\Pr(B)} \quad (12)$$

其中， $\Pr(A, B)$ 为 A 和 B 事件的联合概率， $\Pr(B)$ 也叫 B 事件边缘概率。

类似地，如果 $\Pr(A) > 0$ ，事件 A 发生的条件下事件 B 发生的条件概率为：

$$\Pr(B|A) = \frac{\Pr(A, B)}{\Pr(A)} \quad (13)$$

贝叶斯定理为：

$$\Pr(A|B)\Pr(B) = \Pr(B|A)\Pr(A) = \Pr(A, B) \quad (14)$$

假设 A_1, A_2, \dots, A_n 互不相容，形成对样本空间 Ω 的分割。 $\Pr(A_i) > 0$ ，对于空间 Ω 中任意事件 B ，全概率定理为：

$$\Pr(B) = \sum_{i=1}^n \Pr(A_i, B) \quad (15)$$

如果事件 A 和事件 B 独立，则：

$$\begin{aligned} \Pr(A|B) &= \Pr(A) \\ \Pr(B|A) &= \Pr(B) \\ \Pr(A, B) &= \Pr(A)\Pr(B) \end{aligned} \quad (16)$$

如果事件 A 和事件 B 在 C 发生条件下条件独立，则：

$$\Pr(A, B|C) = \Pr(A|C) \cdot \Pr(B|C) \quad (17)$$

离散随机变量

离散随机变量 X 的概率质量函数满足：

$$\sum_x p_X(x) = 1, \quad 0 \leq p_X(x) \leq 1 \quad (18)$$

离散随机变量 X 的期望值为：

$$E(X) = \sum_x x \cdot p_X(x) \quad (19)$$

离散随机变量 X 的方差为：

$$\text{var}(X) = \sum_x (x - E(X))^2 \cdot p_X(x) \quad (20)$$

二元离散随机变量 (X, Y) 的概率质量函数满足：

$$\sum_x \sum_y p_{X,Y}(x, y) = 1, \quad 0 \leq p_{X,Y}(x, y) \leq 1 \quad (21)$$

(X, Y) 的协方差定义为：

$$\begin{aligned}\text{cov}(X, Y) &= E((X - E(X))(Y - E(Y))) \\ &= \sum_x \sum_y p_{X,Y}(x, y)(x - E(X))(y - E(Y))\end{aligned}\quad (22)$$

边缘概率 $p_X(x)$ 为：

$$p_X(x) = \sum_y p_{X,Y}(x, y) \quad (23)$$

边缘概率 $p_Y(y)$ 为：

$$p_Y(y) = \sum_x p_{X,Y}(x, y) \quad (24)$$

在给定事件 $\{Y = y\}$ 条件下， $p_Y(y) > 0$ ，事件 $\{X = x\}$ 发生的概率的条件概率质量函数 $p_{X|Y}(x|y)$ 为：

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x, y)}{p_Y(y)} \quad (25)$$

$p_{X|Y}(x|y)$ 对 x 求和等于 1：

$$\sum_x p_{X|Y}(x|y) = 1 \quad (26)$$

在给定事件 $\{X = x\}$ 条件下， $p_X(x) > 0$ ，事件 $\{Y = y\}$ 发生的概率的条件概率质量函数 $p_{Y|X}(y|x)$ 为：

$$p_{Y|X}(y|x) = \frac{p_{X,Y}(x, y)}{p_X(x)} \quad (27)$$

$p_{Y|X}(y|x)$ 对 y 求和等于 1：

$$\sum_y p_{Y|X}(y|x) = 1 \quad (28)$$

如果离散变量 X 和 Y 独立，则：

$$\begin{aligned}p_{X|Y}(x|y) &= p_X(x) \\ p_{Y|X}(y|x) &= p_Y(y)\end{aligned}\quad (29)$$

如果离散随机变量 X 和 Y 独立，联合概率 $p_{X,Y}(x, y)$ 为：

$$p_{X,Y}(x, y) = p_Y(y) \cdot p_X(x) \quad (30)$$

离散分布

$[a, b]$ 上离散均匀分布的概率质量函数为：

$$p_X(x) = \frac{1}{b-a+1}, \quad x = a, a+1, \dots, b-1, b \quad (31)$$

伯努利分布的概率质量函数为：

$$p_X(x) = p^x (1-p)^{1-x} \quad x \in \{0, 1\} \quad (32)$$

其中， p 的取值范围为 $[0, 1]$ 。

二项分布的概率质量函数为：

$$p_X(x) = C_n^x p^x (1-p)^{n-x}, \quad x = 0, 1, \dots, n \quad (33)$$

多项分布的概率质量函数为：

$$p_{X_1, \dots, X_K}(x_1, \dots, x_K; n, p_1, \dots, p_K) \begin{cases} \frac{n!}{(x_1!) \times (x_2!) \cdots (x_K!)} \times p_1^{x_1} \times \cdots \times p_K^{x_K} & \text{when } \sum_{i=1}^K x_i = n \\ 0 & \text{otherwise} \end{cases} \quad (34)$$

其中 $x_i (i = 1, 2, \dots, K)$ 为非负整数，且 $\sum_{i=1}^K p_i = 1$ 。

泊松分布的概率质量函数为：

$$p_X(x) = \frac{\exp(-\lambda) \lambda^x}{x!}, \quad x = 0, 1, 2, \dots \quad (35)$$

连续随机变量

连续随机变量 X 的概率密度函数满足：

$$\int_{-\infty}^{+\infty} f_X(x) dx = 1, \quad f_X(x) \geq 0 \quad (36)$$

连续随机变量 X 期望为：

$$E(X) = \int_{-\infty}^{+\infty} x \cdot f_X(x) dx \quad (37)$$

连续随机变量 X 方差为：

$$\text{var}(X) = E[(X - E(X))^2] = \int_{-\infty}^{+\infty} (x - E(X))^2 \cdot f_X(x) dx \quad (38)$$

连续随机变量 X 的边缘概率密度函数 $f_X(x)$ 为：

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x, y) dy \quad (39)$$

连续随机变量 Y 的边缘概率密度函数 $f_Y(y)$ 为：

$$f_Y(y) = \int_x f_{X,Y}(x, y) dx \quad (40)$$

在给定的 $Y = y$ 条件下，且 $f_Y(y) > 0$ ，条件概率密度函数 $f_{X|Y}(x|y)$ 为：

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} \quad (41)$$

给定的 $X = x$ 条件下，且 $f_X(x) > 0$ ，条件概率密度函数 $f_{Y|X}(y|x)$ 为：

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)} \quad (42)$$

利用贝叶斯定理，联合概率 $f_{X,Y}(x, y)$ 为：

$$f_{X,Y}(x, y) = f_{X|Y}(x|y)f_Y(y) = f_{Y|X}(y|x)f_X(x) \quad (43)$$

如果连续随机变量 X 和 Y 独立，下则：

$$\begin{aligned} f_{X|Y}(x|y) &= f_X(x) \\ f_{Y|X}(y|x) &= f_Y(y) \end{aligned} \quad (44)$$

如果连续随机变量 X 和 Y 独立，则联合概率密度函数 $f_{X,Y}(x, y)$ 为：

$$f_{X,Y}(x, y) = f_X(x)f_Y(y) \quad (45)$$

连续分布

区间 $[a, b]$ 的连续均匀分布概率密度函数为：

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b, \\ 0 & \text{for } x < a \text{ or } x > b \end{cases} \quad (46)$$

一元学生 t -分布的概率密度函数为：

$$f_X(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi} \cdot \Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{(\nu+1)}{2}} \quad (47)$$

指数分布的概率密度函数为：

$$f_X(x) = \begin{cases} \lambda \exp(-\lambda x) & x \geq 0 \\ 0 & x < 0 \end{cases} \quad (48)$$

Beta(α, β) 分布的概率密度函数为：

$$f_x(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} \quad (49)$$

Dirichlet 分布概率密度函数为：

$$f_{x_1, \dots, x_K}(x_1, \dots, x_K; \alpha_1, \dots, \alpha_K) = \frac{1}{B(\alpha_1, \dots, \alpha_K)} \prod_{i=1}^K x_i^{\alpha_i-1}, \quad \sum_{i=1}^K x_i = 1 \quad (50)$$

条件概率

如果 X 和 Y 均为离散随机变量，给定 $X = x$ 条件下， Y 的条件期望 $E(Y|X = x)$ 为：

$$E(Y|X = x) = \sum_y y \cdot p_{Y|X}(y|x) \quad (51)$$

$E(Y)$ 的全期望定理为：

$$E(Y) = \sum_x E(Y|X = x) \cdot p_X(x) \quad (52)$$

给定 $Y = y$ 条件下， X 的条件期望 $E(X|Y = y)$ 定义为：

$$E(X|Y = y) = \sum_x x \cdot p_{X|Y}(x|y) \quad (53)$$

$E(Y)$ 的全期望定理为：

$$E(X) = \sum_y E(X|Y = y) \cdot p_Y(y) \quad (54)$$

给定 $X = x$ 条件下， Y 的条件方差 $\text{var}(Y|X = x)$ 为：

$$\text{var}(Y|X = x) = \sum_y (y - E(Y|X = x))^2 \cdot p_{Y|X}(y|x) \quad (55)$$

给定 $Y = y$ 条件下， X 的条件方差 $\text{var}(X|Y = y)$ 为：

$$\text{var}(X|Y = y) = \sum_x (x - E(X|Y = y))^2 \cdot p_{X|Y}(x|y) \quad (56)$$

对于 $\text{var}(Y)$ ，全方差定理为：

$$\text{var}(Y) = E(\text{var}(Y|X)) + \text{var}(E(Y|X)) \quad (57)$$

对于 $\text{var}(X)$ ，全方差定理为：

$$\text{var}(X) = E(\text{var}(X|Y)) + \text{var}(E(X|Y)) \quad (58)$$

如果 X 和 Y 均为连续随机变量，在给定 $X = x$ 条件下，条件期望 $E(Y|X = x)$ 为：

$$E(Y|X = x) = \int_y y \cdot f_{Y|X}(y|x) dy \quad (59)$$

条件方差 $\text{var}(Y|X=x)$ 为：

$$\text{var}(Y|X=x) = \int_y (y - E(Y|X=x))^2 \cdot f_{Y|X}(y|x) dy \quad (60)$$

在给定 $Y=y$ 条件下，条件期望 $E(X|Y=y)$ 为：

$$E(X|Y=y) = \int_x x \cdot f_{X|Y}(x|y) dx \quad (61)$$

条件方差 $\text{var}(X|Y=y)$ 定义为：

$$\text{var}(X|Y=y) = \int_x (X - E(X|Y=y))^2 \cdot f_{X|Y}(x|y) dx \quad (62)$$

1.4 高斯

一元高斯分布

一元高斯分布的概率密度函数为：

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \quad (63)$$

标准正态分布的概率密度函数为：

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \quad (64)$$

二元高斯分布

如果 (X, Y) 服从二元高斯分布，且相关性系数不为 ± 1 ，其概率密度函数为：

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{X,Y}^2}} \times \exp\left(\frac{-1}{2(1-\rho_{X,Y}^2)} \left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho_{X,Y} \left(\frac{x-\mu_X}{\sigma_X}\right) \left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 \right)\right) \quad (65)$$

X 的边缘概率密度函数为：

$$f_X(x) = \frac{1}{\sigma_X\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right) \quad (66)$$

Y 的边缘概率密度函数为：

$$f_Y(y) = \frac{1}{\sigma_Y \sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{y - \mu_Y}{\sigma_Y}\right)^2\right) \quad (67)$$

多元高斯分布

多元高斯分布的概率密度函数为：

$$f_{\mathbf{x}}(\mathbf{x}) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)}{(2\pi)^{\frac{D}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \quad (68)$$

其中，协方差矩阵 $\boldsymbol{\Sigma}$ 为正定矩阵。

条件高斯分布

如果 (X, Y) 服从二元高斯分布，且相关性系数不为 ± 1 ， $f_{Y|X}(y|x)$ 为：

$$f_{Y|X}(y|x) = \frac{1}{\sigma_Y \sqrt{1 - \rho_{X,Y}^2} \sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{y - \left(\mu_Y + \rho_{X,Y} \frac{\sigma_Y}{\sigma_X}(x - \mu_X)\right)}{\sigma_Y \sqrt{1 - \rho_{X,Y}^2}}\right)^2\right) \quad (69)$$

条件期望 $E(Y|X=x)$ 为：

$$E(Y|X=x) = \mu_Y + \rho_{X,Y} \frac{\sigma_Y}{\sigma_X}(x - \mu_X) \quad (70)$$

条件方差 $\text{var}(Y|X=x)$ 为：

$$\text{var}(Y|X=x) = (1 - \rho_{X,Y}^2) \sigma_Y^2 \quad (71)$$

如果随机变量向量 \mathbf{x} 和 \mathbf{y} 服从多元高斯分布：

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy} \end{bmatrix}\right) \quad (72)$$

给定 $\mathbf{x} = \mathbf{x}$ 的条件下， \mathbf{y} 服从如下多元高斯分布：

$$\{\mathbf{y}|\mathbf{x}=\mathbf{x}\} \sim N\left(\underbrace{\boldsymbol{\Sigma}_{yx} \boldsymbol{\Sigma}_{xx}^{-1}(\mathbf{x} - \boldsymbol{\mu}_x) + \boldsymbol{\mu}_y}_{\text{Expectation}}, \underbrace{\boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}_{yx} \boldsymbol{\Sigma}_{xx}^{-1} \boldsymbol{\Sigma}_{xy}}_{\text{Variance}}\right) \quad (73)$$

给定 $\mathbf{x} = \mathbf{x}$ 的条件下 \mathbf{y} 的条件期望为：

$$\mathbb{E}(y|\mathbf{x}) = \boldsymbol{\mu}_{y|\mathbf{x}} = \boldsymbol{\Sigma}_{yx} \boldsymbol{\Sigma}_{xx}^{-1} (\mathbf{x} - \boldsymbol{\mu}_x) + \boldsymbol{\mu}_y \quad (74)$$

协方差矩阵

随机变量的列向量 $\boldsymbol{\chi}$ 的协方差矩阵为：

$$\begin{aligned} \text{var}(\boldsymbol{\chi}) &= \text{cov}(\boldsymbol{\chi}, \boldsymbol{\chi}) = \mathbb{E} \left[(\boldsymbol{\chi} - \mathbb{E}(\boldsymbol{\chi})) (\boldsymbol{\chi} - \mathbb{E}(\boldsymbol{\chi}))^T \right] \\ &= \mathbb{E}(\boldsymbol{\chi} \boldsymbol{\chi}^T) - \mathbb{E}(\boldsymbol{\chi}) \mathbb{E}(\boldsymbol{\chi})^T \end{aligned} \quad (75)$$

样本数据矩阵 \mathbf{X} 的协方差矩阵 $\boldsymbol{\Sigma}$ 为：

$$\boldsymbol{\Sigma} = \frac{(\mathbf{X} - \mathbb{E}(\mathbf{X}))^T (\mathbf{X} - \mathbb{E}(\mathbf{X}))}{n-1} \quad (76)$$

协方差矩阵 $\boldsymbol{\Sigma}$ 谱分解后可以展开为：

$$\boldsymbol{\Sigma} = \mathbf{V} \boldsymbol{\Lambda} \mathbf{V}^T = \sum_{j=1}^D \lambda_j \mathbf{v}_j \mathbf{v}_j^T \quad (77)$$

协方差矩阵 $\boldsymbol{\Sigma}$ 的迹为特征值之和：

$$\text{trace}(\boldsymbol{\Sigma}) = \sum_{j=1}^D \sigma_j^2 = \sum_{j=1}^D \lambda_j \quad (78)$$

$\boldsymbol{\Sigma}$ 的行列式值为其特征值乘积：

$$|\boldsymbol{\Sigma}| = |\mathbf{A}| = \prod_{j=1}^D \lambda_j \quad (79)$$

如果协方差矩阵 $\boldsymbol{\Sigma}$ 正定，对其 Cholesky 分解得到：

$$\boldsymbol{\Sigma} = \mathbf{L} \mathbf{L}^T = \mathbf{R}^T \mathbf{R} \quad (80)$$

合并协方差矩阵为：

$$\boldsymbol{\Sigma}_{\text{pooled}} = \frac{1}{\sum_{k=1}^K n_k - 1} \sum_{k=1}^K (n_k - 1) \boldsymbol{\Sigma}_k = \frac{1}{n - K} \sum_{k=1}^K (n_k - 1) \boldsymbol{\Sigma}_k \quad (81)$$

1.5 随机

随机变量的函数

形如下式的随机变量变换叫做 X 到 Y 的线性变换：

$$Y = h(X) = aX + b \quad (82)$$

其中, a 和 b 为常数。

(82) 中, Y 和 X 的期望、方差之间关系为:

$$\begin{aligned} E(Y) &= aE(X) + b \\ \text{var}(Y) &= \text{var}(aX + b) = a^2 \text{var}(X) \end{aligned} \quad (83)$$

如果 Y 和二元随机变量 (X_1, X_2) 存在如下关系:

$$Y = aX_1 + bX_2 \quad (84)$$

Y 的期望、方差为:

$$\begin{aligned} E(Y) &= aE(X_1) + bE(X_2) \\ \text{var}(Y) &= a^2 \text{var}(X_1) + b^2 \text{var}(X_2) + 2ab \text{cov}(X_1, X_2) \end{aligned} \quad (85)$$

如果 $\chi = [X_1, X_2, \dots, X_D]^T$ 服从 $N(\mu_\chi, \Sigma_\chi)$, χ 在单位向量 v 方向上投影得到 Y :

$$Y = v^T \chi \quad (86)$$

Y 的期望、方差为:

$$\begin{aligned} E(Y) &= v^T \mu_\chi \\ \text{var}(Y) &= v^T \Sigma_\chi v \end{aligned} \quad (87)$$

χ 在规范正交系 V 投影得到 γ :

$$\gamma = V^T \chi \quad (88)$$

γ 的期望、协方差矩阵为:

$$\begin{aligned} E(\gamma) &= V^T \mu_\chi \\ \text{var}(\gamma) &= V^T \Sigma_\chi V \end{aligned} \quad (89)$$

蒙特卡洛模拟

多元随机数 Z 服从 $N(0, I_{D \times D})$, 多元随机数 X 服从 $N(E(X), \Sigma_{D \times D})$, 两者关系为:

$$X = ZR + E(X) \quad (90)$$

其中, 矩阵 R 来自 (80)。

1.6 频率派

频率派统计推断

随机变量 X_1, X_2, \dots, X_n 独立同分布。 $X_k (k = 1, 2, \dots, n)$ 的期望和方差为：

$$E(X_k) = \mu, \quad \text{var}(X_k) = \sigma^2 \quad (91)$$

这 n 个随机变量的平均值 \bar{X} 近似服从如下正态分布：

$$\bar{X} = \frac{1}{n} \sum_{k=1}^n X_k \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad (92)$$

最大似然估计的优化问题为：

$$\hat{\theta}_{\text{MLE}} = \arg \max_{\theta} \prod_{i=1}^n f_{X_i}(x_i; \theta) \quad (93)$$

概率密度估计

概率密度估计函数：

$$\hat{f}_X(x) = \frac{1}{n} \sum_{i=1}^n K_h(x - x^{(i)}) = \frac{1}{n} \underbrace{\frac{1}{h} \sum_{i=1}^n K\left(\frac{x - x^{(i)}}{h}\right)}_{\text{Area} = n}, \quad -\infty < x < +\infty \quad (94)$$

核函数 $K(x)$ 满足两个重要条件：(1) 对称性；(2) 面积为 1：

$$\begin{aligned} K(x) &= K(-x) \\ \int_{-\infty}^{+\infty} K(x) dx &= \frac{1}{h} \int_{-\infty}^{+\infty} K\left(\frac{x}{h}\right) dx = 1 \end{aligned} \quad (95)$$

1.7 贝叶斯派

贝叶斯分类

利用贝叶斯定理分类：

$$f_{Y|X}(C_k | x) = \frac{f_{X|Y}(x | C_k) p_Y(C_k)}{f_X(x)} \quad (96)$$

$f_{Y|X}(C_k|x)$ 叫后验概率，又叫成员值。 $f_X(x)$ 为证据因子，也叫证据。 $p_X(C_k)$ 为先验概率，表达样本集合中 C_k 类样本占比。 $f_{X|Y}(x|C_k)$ 为似然概率。

贝叶斯分类优化问题可以是最大化后验概率：

$$\hat{y} = \arg \max_{C_k} f_{Y|X}(C_k|x) \quad (97)$$

其中， $k = 1, 2, \dots, K$ 。

贝叶斯统计推断

模型参数的后验分布为：

$$f_{\Theta|X}(\theta|x) = \frac{f_{X|\Theta}(x|\theta)f_{\Theta}(\theta)}{\int_{\mathcal{G}} f_{X|\Theta}(x|\vartheta)f_{\Theta}(\vartheta)d\vartheta} \quad (98)$$

最大化后验估计的优化问题为：

$$\hat{\theta}_{\text{MAP}} = \arg \max_{\theta} f_{\Theta|X}(\theta|x) \quad (99)$$

1.8 椭圆三部曲

马氏距离

马氏距离的定义为：

$$d = \sqrt{(x - \mu)^T \Sigma^{-1} (x - \mu)} \quad (100)$$

D 维马氏距离的平方则服从自由度为 D 的卡方分布：

$$d^2 = (x - \mu)^T \Sigma^{-1} (x - \mu) \sim \chi^2_{(\text{df}=D)} \quad (101)$$

线性回归

一元线性回归表达式为：

$$y = \hat{y} + \varepsilon = b_0 + b_1 x + \varepsilon \quad (102)$$

最小二乘法优化问题为：

$$\arg \min_{b_0, b_1} \sum_{i=1}^n (\varepsilon^{(i)})^2 \quad (103)$$

多元线性回归可以写成超定方程组：

$$\mathbf{y} = \mathbf{X}\mathbf{b} \quad (104)$$

如果 $\mathbf{X}^T \mathbf{X}$ 可逆，则 \mathbf{b} 为：

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (105)$$

主成分分析

对原始矩阵 \mathbf{X} 进行经济型 SVD 分解：

$$\mathbf{X} = \mathbf{U}_X \mathbf{S}_X \mathbf{V}_X^T \quad (106)$$

其中， \mathbf{S}_X 为对角方阵。

利用 (106)， \mathbf{X} 的格拉姆矩阵可以展开为：

$$\mathbf{G} = \mathbf{V}_X \mathbf{S}_X^2 \mathbf{V}_X^T \quad (107)$$

上式便是格拉姆 \mathbf{G} 的特征值分解。

对中心化数据矩阵 \mathbf{X}_c 经济型 SVD 分解：

$$\mathbf{X}_c = \mathbf{U}_c \mathbf{S}_c \mathbf{V}_c^T \quad (108)$$

而协方差矩阵 $\mathbf{\Sigma}$ 则可以写成：

$$\mathbf{\Sigma} = \mathbf{V}_c \frac{\mathbf{S}_c^2}{n-1} \mathbf{V}_c^T \quad (109)$$

相信大家在上式中能够看到协方差矩阵 $\mathbf{\Sigma}$ 的特征值分解。请大家注意 (108) 中奇异值和 (109) 中特征值关系：

$$\lambda_{c-j} = \frac{s_{c-j}^2}{n-1} \quad (110)$$

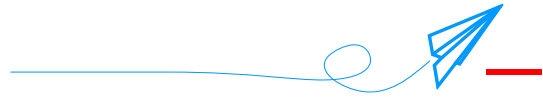
同样，对标准化数据矩阵 \mathbf{Z}_X 进行经济型 SVD 分解：

$$\mathbf{Z}_X = \mathbf{U}_Z \mathbf{S}_Z \mathbf{V}_Z^T \quad (111)$$

相关性系数矩阵 \mathbf{P} 则可以写成：

$$\mathbf{P} = \mathbf{V}_Z \frac{\mathbf{S}_Z^2}{n-1} \mathbf{V}_Z^T \quad (112)$$

上式相当于对 \mathbf{P} 特征值分解。



等大家学完本书，再回过头来看本章罗列的这些公式时，希望大家看到的不再是冷冰冰的符号，而是一幅幅彩色的图画。