

1

Landscape of Statistics and Probability

本书概率统计全景

公式连篇，可能是“鸢尾花书”最枯燥无味的一章



概率论作为数学学科，可以且应该从公理开始建设，和几何、代数的思路一样。

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》中相关数学工具。

数据矩阵

给定数据矩阵 \mathbf{X} ，如何求其质心、中心化数据、标准化数据、格拉姆矩阵、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 ：

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

什么条件下 Σ 是正定矩阵？

定义如下二元函数：

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

相关性系数 $\rho_{1,2}$ 的取值范围是什么？这个二元函数的图像是什么？

当 σ_1 和 σ_2 均为 1 时，这个二元函数等高线形状随 $\rho_{1,2}$ 如何变化？

Cholesky 分解

对协方差矩阵 Σ 进行 Cholesky 分解：

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

矩阵 Σ 能进行 Cholesky 分解的前提是什么？

上三角矩阵 \mathbf{R} 的特点是什么？如何从几何角度理解 \mathbf{R} ？

特征值分解

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

对 Σ 特征值分解：

$$\Sigma = V \Lambda V^T \quad (\text{test.4})$$

等式右侧第二个矩阵 V 对应转置运算，为什么？

矩阵 V 有什么特殊性质？如何从向量空间角度理解 V ？

矩阵 Λ 有什么特殊性质？

如果把 V 写成 $[v_1, v_2]$ ，上式可以如何展开？

将 (test.4) 写成：

$$V^T \Sigma V = \Lambda \quad (\text{test.5})$$

把 V 写成 $[v_1, v_2]$ ，上式如何展开？

几何角度来看，上式代表什么？

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 的格拉姆矩阵 G 特征值分解可以得到特征值 $\lambda_{G,j}$ 。奇异值 s_j 和特征值 $\lambda_{G,j}$ 存在怎样的量化关系？

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 又存在怎样的量化关系？

奇异值分解和向量四个空间有怎样联系？

多元高斯分布

多元正态分布的概率密度函数 PDF 为：

$$f_X(\mathbf{x}) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} \quad (\text{test.6})$$

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

$(2\pi)^{\frac{D}{2}}$ 的作用是什么？ $|\Sigma|^{\frac{1}{2}}$ 的含义是什么？

什么情况下，上式不成立？

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。



本书不就上述题目给出具体答案，所有答案都在《矩阵力量》一册，请大家自行查阅。

本章下面先用数学手册、备忘录这种范式罗列本书中 100 个核心公式，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 n 个样本 $\{x^{(1)}, x^{(2)}, \dots, x^{(n)}\}$ ， X 的样本均值为：

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

X 的样本方差为：

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

X 的样本标准差为：

$$\sigma_X = \text{std}(X) = \sqrt{\text{var}(X)} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n \left(x^{(i)} - \mu_X \right)^2} \quad (3)$$

对于样本数据，随机变量 X 和 Y 的协方差为：

$$\text{cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^n \left(x^{(i)} - \mu_X \right) \left(y^{(i)} - \mu_Y \right) \quad (4)$$

对于样本数据，随机变量 X 和 Y 的相关性系数为：

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \quad (5)$$

▲ 注意，除非特殊说明，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：

$$\Pr(A) = \frac{n_A}{n} \quad (6)$$

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：

$$\Pr(A|B) = \frac{\Pr(A, B)}{\Pr(B)} \quad (7)$$

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：

$$\Pr(B|A) = \frac{\Pr(A, B)}{\Pr(A)} \quad (8)$$

贝叶斯定理为：

$$\Pr(A|B)\Pr(B) = \Pr(B|A)\Pr(A) = \Pr(A, B) \quad (9)$$

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：

$$\Pr(B) = \sum_{i=1}^n \Pr(A_i, B) \quad (10)$$

如果事件 A 和事件 B 独立，则：

$$\begin{aligned} \Pr(A|B) &= \Pr(A) \\ \Pr(B|A) &= \Pr(B) \\ \Pr(A, B) &= \Pr(A)\Pr(B) \end{aligned} \quad (11)$$

如果事件 A 和事件 B 在 C 发生条件下条件独立，则：

$$\Pr(A, B|C) = \Pr(A|C) \cdot \Pr(B|C) \quad (12)$$

离散随机变量

离散随机变量 X 的概率质量函数满足：

$$\sum_x p_X(x) = 1, \quad 0 \leq p_X(x) \leq 1 \quad (13)$$

离散随机变量 X 的期望值为：

$$E(X) = \sum_x x \cdot p_X(x) \quad (14)$$

离散随机变量 X 的方差为：

$$\text{var}(X) = \sum_x (x - E(X))^2 \cdot p_X(x) \quad (15)$$

二元离散随机变量 (X, Y) 的概率质量函数满足：

$$\sum_x \sum_y p_{X,Y}(x, y) = 1, \quad 0 \leq p_{X,Y}(x, y) \leq 1 \quad (16)$$

(X, Y) 的协方差定义为：

$$\begin{aligned} \text{cov}(X, Y) &= E((X - E(X))(Y - E(Y))) \\ &= \sum_x \sum_y p_{X,Y}(x, y)(x - E(X))(y - E(Y)) \end{aligned} \quad (17)$$

边缘概率 $p_X(x)$ 为：

$$p_X(x) = \sum_y p_{X,Y}(x, y) \quad (18)$$

边缘概率 $p_Y(y)$ 为：

$$p_Y(y) = \sum_x p_{X,Y}(x, y) \quad (19)$$

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

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x, y)}{p_Y(y)} \quad (20)$$

$p_{X|Y}(x|y)$ 对 x 求和等于 1：

$$\sum_x p_{X|Y}(x|y) = 1 \quad (21)$$

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

$$p_{Y|X}(y|x) = \frac{p_{X,Y}(x, y)}{p_X(x)} \quad (22)$$

$p_{Y|X}(y|x)$ 对 y 求和等于 1：

$$\sum_y p_{Y|X}(y|x) = 1 \quad (23)$$

如果离散随机变量 X 和 Y 独立，则：

$$\begin{aligned} p_{X|Y}(x|y) &= p_X(x) \\ p_{Y|X}(y|x) &= p_Y(y) \\ p_{X,Y}(x, y) &= p_Y(y) \cdot p_X(x) \end{aligned} \quad (24)$$

离散分布

$[a, b]$ 上离散均匀分布的概率质量函数为：

$$p_X(x) = \frac{1}{b-a+1}, \quad x = a, a+1, \dots, b-1, b \quad (25)$$

伯努利分布的概率质量函数为：

$$p_X(x) = p^x (1-p)^{1-x} \quad x \in \{0, 1\} \quad (26)$$

其中， p 的取值范围为 $[0, 1]$ 。

二项分布的概率质量函数为：

$$p_X(x) = C_n^x p^x (1-p)^{n-x}, \quad x = 0, 1, \dots, n \quad (27)$$

多项分布的概率质量函数为：

$$p_{X_1, \dots, X_K}(x_1, \dots, x_K; n, p_1, \dots, p_K) = \begin{cases} \frac{n!}{(x_1!)(x_2!)\dots(x_K!)} \times p_1^{x_1} \times \dots \times p_K^{x_K} & \text{when } \sum_{i=1}^K x_i = n \\ 0 & \text{otherwise} \end{cases} \quad (28)$$

其中 $x_i (i = 1, 2, \dots, K)$ 为非负整数； p_i 取值范围为 $(0, 1)$ ，且 $\sum_{i=1}^K p_i = 1$ 。

泊松分布的概率质量函数为：

$$p_X(x) = \frac{\exp(-\lambda) \lambda^x}{x!}, \quad x = 0, 1, 2, \dots \quad (29)$$

其中， λ 大于 0。

连续随机变量

连续随机变量 X 的概率密度函数满足：

$$\int_{-\infty}^{+\infty} f_X(x) dx = 1, \quad f_X(x) \geq 0 \quad (30)$$

连续随机变量 X 期望为：

$$E(X) = \int_{-\infty}^{+\infty} x \cdot f_X(x) dx \quad (31)$$

连续随机变量 X 方差为：

$$\text{var}(X) = E\left[\left(X - E(X)\right)^2\right] = \int_{-\infty}^{+\infty} (x - E(X))^2 \cdot f_X(x) dx \quad (32)$$

给定 (X, Y) 的联合概率分布 $f_{X,Y}(x,y)$ ， X 的边缘概率密度函数 $f_X(x)$ 为：

$$f_X(x) = \int_y f_{X,Y}(x, y) dy \quad (33)$$

连续随机变量 Y 的边缘概率密度函数 $f_Y(y)$ 为：

$$f_Y(y) = \int_x f_{X,Y}(x, y) dx \quad (34)$$

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

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} \quad (35)$$

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

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)} \quad (36)$$

利用贝叶斯定理，联合概率 $f_{X,Y}(x, y)$ 为：

$$f_{X,Y}(x, y) = f_{X|Y}(x|y) f_Y(y) = f_{Y|X}(y|x) f_X(x) \quad (37)$$

如果连续随机变量 X 和 Y 独立，则：

$$\begin{aligned} f_{X|Y}(x|y) &= f_X(x) \\ f_{Y|X}(y|x) &= f_Y(y) \\ f_{X,Y}(x, y) &= f_X(x) f_Y(y) \end{aligned} \quad (38)$$

连续分布

区间 $[a, b]$ 的连续均匀分布概率密度函数为：

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b, \\ 0 & \text{for } x < a \text{ or } x > b \end{cases} \quad (39)$$

一元学生 t -分布的概率密度函数为：

$$f_X(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi} \cdot \Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{(\nu+1)}{2}} \quad (40)$$

其中， ν 大于 0。

指数分布的概率密度函数为：

$$f_X(x) = \begin{cases} \lambda \exp(-\lambda x) & x \geq 0 \\ 0 & x < 0 \end{cases} \quad (41)$$

其中, λ 大于 0。

Beta(α, β) 分布的概率密度函数为:

$$f_X(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} \quad (42)$$

其中, α 和 β 均大于 0。这个 PDF 也可以写成:

$$f_X(x; \alpha, \beta) = \frac{x^{\alpha-1} (1-x)^{\beta-1}}{B(\alpha, \beta)} \quad (43)$$

其中, Beta 函数 $B(\alpha, \beta)$ 为:

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} \quad (44)$$

Dirichlet 分布概率密度函数为:

$$f_{X_1, \dots, X_K}(x_1, \dots, x_K; \alpha_1, \dots, \alpha_K) = \frac{1}{B(\alpha_1, \dots, \alpha_K)} \prod_{i=1}^K x_i^{\alpha_i-1}, \quad \sum_{i=1}^K x_i = 1 \quad (45)$$

其中, α_i 大于 0。注意, 对于 Dirichlet 分布, 本书后续常用 θ 代替 x 。

Beta 函数 $B(\alpha_1, \dots, \alpha_K)$ 为:

$$B(\alpha_1, \dots, \alpha_K) = \frac{\prod_{i=1}^K \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^K \alpha_i\right)} \quad (46)$$

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:

$$E(Y|X = x) = \sum_y y \cdot p_{Y|X}(y|x) \quad (47)$$

$E(Y)$ 的全期望定理为:

$$E(Y) = \sum_x E(Y|X = x) \cdot p_X(x) \quad (48)$$

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

$$E(X|Y = y) = \sum_x x \cdot p_{X|Y}(x|y) \quad (49)$$

$E(X)$ 的全期望定理为：

$$E(X) = \sum_y E(X|Y=y) \cdot p_Y(y) \quad (50)$$

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

$$\text{var}(Y|X=x) = \sum_y (y - E(Y|X=x))^2 \cdot p_{Y|X}(y|x) \quad (51)$$

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

$$\text{var}(X|Y=y) = \sum_x (x - E(X|Y=y))^2 \cdot p_{X|Y}(x|y) \quad (52)$$

对于 $\text{var}(Y)$ ，全方差定理为：

$$\text{var}(Y) = E(\text{var}(Y|X)) + \text{var}(E(Y|X)) \quad (53)$$

对于 $\text{var}(X)$ ，全方差定理为：

$$\text{var}(X) = E(\text{var}(X|Y)) + \text{var}(E(X|Y)) \quad (54)$$

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：

$$E(Y|X=x) = \int_y y \cdot f_{Y|X}(y|x) dy \quad (55)$$

条件方差 $\text{var}(Y|X=x)$ 为：

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

在给定 $Y=y$ 条件下，条件期望 $E(X|Y=y)$ 为：

$$E(X|Y=y) = \int_x x \cdot f_{X|Y}(x|y) dx \quad (57)$$

条件方差 $\text{var}(X|Y=y)$ 定义为：

$$\text{var}(X|Y=y) = \int_x (x - E(X|Y=y))^2 \cdot f_{X|Y}(x|y) dx \quad (58)$$

1.4 高斯

一元高斯分布

一元高斯分布的概率密度函数为：

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \quad (59)$$

标准正态分布的概率密度函数为：

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \quad (60)$$

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 ， (X, Y) 的概率密度函数为：

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{X,Y}^2}} \times \exp\left(\frac{-1}{2(1-\rho_{X,Y}^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho_{X,Y}\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right) \quad (61)$$

X 的边缘概率密度函数为：

$$f_X(x) = \frac{1}{\sigma_X\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right) \quad (62)$$

Y 的边缘概率密度函数为：

$$f_Y(y) = \frac{1}{\sigma_Y\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right) \quad (63)$$

多元高斯分布

多元高斯分布的概率密度函数为：

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)}{(2\pi)^{\frac{D}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \quad (64)$$

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：

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

条件期望 $E(Y|X=x)$ 为：

$$E(Y|X=x) = \mu_Y + \rho_{X,Y} \frac{\sigma_Y}{\sigma_X} (x - \mu_X) \quad (66)$$

条件方差 $\text{var}(Y|X=x)$ 为：

$$\text{var}(Y|X=x) = (1 - \rho_{X,Y}^2) \sigma_Y^2 \quad (67)$$

如果随机变量向量 χ 和 γ 服从多元高斯分布：

$$\begin{bmatrix} \chi \\ \gamma \end{bmatrix} \sim N \left(\begin{bmatrix} \mu_\chi \\ \mu_\gamma \end{bmatrix}, \begin{bmatrix} \Sigma_{\chi\chi} & \Sigma_{\chi\gamma} \\ \Sigma_{\gamma\chi} & \Sigma_{\gamma\gamma} \end{bmatrix} \right) \quad (68)$$

其中，

$$\chi = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_D \end{bmatrix}, \quad \gamma = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_M \end{bmatrix} \quad (69)$$

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

$$\{\gamma|\chi=\mathbf{x}\} \sim N \left(\underbrace{\Sigma_{\gamma\chi} \Sigma_{\chi\chi}^{-1} (\mathbf{x} - \mu_\chi) + \mu_\gamma}_{\text{Expectation}}, \underbrace{\Sigma_{\gamma\gamma} - \Sigma_{\gamma\chi} \Sigma_{\chi\chi}^{-1} \Sigma_{\chi\gamma}}_{\text{Variance}} \right) \quad (70)$$

给定 $\chi = \mathbf{x}$ 的条件下 γ 的条件期望为：

$$E(\gamma|\chi=\mathbf{x}) = \mu_{\gamma|\chi=\mathbf{x}} = \Sigma_{\gamma\chi} \Sigma_{\chi\chi}^{-1} (\mathbf{x} - \mu_\chi) + \mu_\gamma \quad (71)$$

协方差矩阵

随机变量向量 χ 的协方差矩阵为：

$$\begin{aligned} \text{var}(\chi) &= \text{cov}(\chi, \chi) = E \left[(\chi - E(\chi)) (\chi - E(\chi))^T \right] \\ &= E(\chi \chi^T) - E(\chi) E(\chi)^T \end{aligned} \quad (72)$$

样本数据矩阵 X 的协方差矩阵 Σ 为：

$$\boldsymbol{\Sigma} = \frac{(\mathbf{X} - \mathbf{E}(\mathbf{X}))^T (\mathbf{X} - \mathbf{E}(\mathbf{X}))}{n-1} \quad (73)$$

合并协方差矩阵为：

$$\boldsymbol{\Sigma}_{\text{pooled}} = \frac{1}{\sum_{k=1}^K (n_k - 1)} \sum_{k=1}^K (n_k - 1) \boldsymbol{\Sigma}_k = \frac{1}{n-K} \sum_{k=1}^K (n_k - 1) \boldsymbol{\Sigma}_k \quad (74)$$

其中， $\sum_{k=1}^K n_k = n$ 。

1.5 随机

随机变量的函数

如果 Y 和二元随机变量 (X_1, X_2) 存在如下关系：

$$Y = aX_1 + bX_2 = \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \quad (75)$$

Y 的期望、方差为：

$$\mathbf{E}(Y) = \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} \mathbf{E}(X_1) \\ \mathbf{E}(X_2) \end{bmatrix}, \quad \text{var}(Y) = \begin{bmatrix} a & b \end{bmatrix} \underbrace{\begin{bmatrix} \text{var}(X_1) & \text{cov}(X_1, X_2) \\ \text{cov}(X_1, X_2) & \text{var}(X_2) \end{bmatrix}}_{\boldsymbol{\Sigma}} \begin{bmatrix} a \\ b \end{bmatrix} \quad (76)$$

如果 $\boldsymbol{\chi} = [X_1, X_2, \dots, X_D]^T$ 服从 $N(\boldsymbol{\mu}_{\boldsymbol{\chi}}, \boldsymbol{\Sigma}_{\boldsymbol{\chi}})$ ， $\boldsymbol{\chi}$ 在单位向量 \mathbf{v} 方向上投影得到 Y ：

$$Y = \mathbf{v}^T \boldsymbol{\chi} \quad (77)$$

Y 的期望、方差为：

$$\begin{aligned} \mathbf{E}(Y) &= \mathbf{v}^T \boldsymbol{\mu}_{\boldsymbol{\chi}} \\ \text{var}(Y) &= \mathbf{v}^T \boldsymbol{\Sigma}_{\boldsymbol{\chi}} \mathbf{v} \end{aligned} \quad (78)$$

$\boldsymbol{\chi}$ 在规范正交系 \mathbf{V} 投影得到 $\boldsymbol{\gamma}$ ：

$$\boldsymbol{\gamma} = \mathbf{V}^T \boldsymbol{\chi} \quad (79)$$

$\boldsymbol{\gamma}$ 的期望、协方差矩阵为：

$$\begin{aligned} \mathbf{E}(\boldsymbol{\gamma}) &= \mathbf{V}^T \boldsymbol{\mu}_{\boldsymbol{\chi}} \\ \text{var}(\boldsymbol{\gamma}) &= \mathbf{V}^T \boldsymbol{\Sigma}_{\boldsymbol{\chi}} \mathbf{V} \end{aligned} \quad (80)$$

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：

$$E(X_k) = \mu, \quad \text{var}(X_k) = \sigma^2 \quad (81)$$

这 n 个随机变量的平均值 \bar{X} 近似服从如下正态分布：

$$\bar{X} = \frac{1}{n} \sum_{k=1}^n X_k \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad (82)$$

最大似然估计的优化问题为：

$$\hat{\theta}_{\text{MLE}} = \arg \max_{\theta} \prod_{i=1}^n f_{X_i}(x_i; \theta) = \arg \max_{\theta} \sum_{i=1}^n \ln f_{X_i}(x_i; \theta) \quad (83)$$

概率密度估计

概率密度估计函数为：

$$\hat{f}_X(x) = \frac{1}{n} \sum_{i=1}^n K_h(x - x^{(i)}) = \frac{1}{n} \frac{1}{h} \sum_{i=1}^n K\left(\frac{x - x^{(i)}}{h}\right), \quad -\infty < x < +\infty \quad (84)$$

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

$$\begin{aligned} K(x) &= K(-x) \\ \int_{-\infty}^{+\infty} K(x) dx &= \frac{1}{h} \int_{-\infty}^{+\infty} K\left(\frac{x}{h}\right) dx = 1 \end{aligned} \quad (85)$$

1.7 贝叶斯派

贝叶斯分类

利用贝叶斯定理分类：

$$f_{Y|X}(C_k | x) = \frac{f_{X|Y}(x | C_k) p_Y(C_k)}{f_X(x)} \quad (86)$$

$f_{Y|X}(C_k | x)$ 叫后验概率，又叫成员值。

$f_X(x)$ 为证据因子，也叫证据。

$p_Y(C_k)$ 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：

$$\hat{y} = \arg \max_{C_k} f_{Y|X}(C_k | x) = \arg \max_{C_k} f_{X|Y}(x | C_k) p_Y(C_k) \quad (87)$$

其中， $k = 1, 2, \dots, K$ 。

贝叶斯统计推断

模型参数的后验分布为：

$$f_{\Theta|X}(\theta | x) = \frac{f_{X|\Theta}(x | \theta) f_{\Theta}(\theta)}{\int_{\mathcal{G}} f_{X|\Theta}(x | g) f_{\Theta}(g) dg} \quad (88)$$

最大化后验估计的优化问题为：

$$\hat{\theta}_{\text{MAP}} = \arg \max_{\theta} f_{\Theta|X}(\theta | x) = \arg \max_{\theta} f_{X|\Theta}(x | \theta) f_{\Theta}(\theta) \quad (89)$$

1.8 椭圆三部曲

马氏距离

马氏距离的定义为：

$$d = \sqrt{(x - \mu)^T \Sigma^{-1} (x - \mu)} \quad (90)$$

D 维马氏距离的平方则服从自由度为 D 的卡方分布：

$$d^2 = (x - \mu)^T \Sigma^{-1} (x - \mu) \sim \chi^2_{(\text{df}=D)} \quad (91)$$

线性回归

多元线性回归可以写成超定方程组：

$$y = Xb \quad (92)$$

如果 $X^T X$ 可逆，则 b 为：

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (93)$$

主成分分析

注意，这部分公式实际上来自《矩阵力量》；此外，我们将会在《数据有道》也会用到这些公式。

对原始矩阵 \mathbf{X} 进行经济型 SVD 分解：

$$\mathbf{X} = \mathbf{U}_X \mathbf{S}_X \mathbf{V}_X^T \quad (94)$$

其中， \mathbf{S}_X 为对角方阵。

利用 (94)， \mathbf{X} 的格拉姆矩阵可以展开为：

$$\mathbf{G} = \mathbf{V}_X \mathbf{S}_X^2 \mathbf{V}_X^T \quad (95)$$

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：

$$\mathbf{X}_c = \mathbf{U}_c \mathbf{S}_c \mathbf{V}_c^T \quad (96)$$

而协方差矩阵 $\mathbf{\Sigma}$ 则可以写成：

$$\mathbf{\Sigma} = \mathbf{V}_c \frac{\mathbf{S}_c^2}{n-1} \mathbf{V}_c^T \quad (97)$$

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 (96) 中奇异值和 (97) 中特征值关系：

$$\lambda_{c-j} = \frac{s_{c-j}^2}{n-1} \quad (98)$$

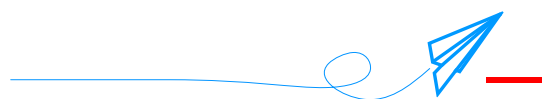
同样，对标准化数据矩阵 \mathbf{Z}_X 进行经济型 SVD 分解：

$$\mathbf{Z}_X = \mathbf{U}_Z \mathbf{S}_Z \mathbf{V}_Z^T \quad (99)$$

相关性系数矩阵 \mathbf{P} 则可以写成：

$$\mathbf{P} = \mathbf{V}_Z \frac{\mathbf{S}_Z^2}{n-1} \mathbf{V}_Z^T \quad (100)$$

上式相当于对 \mathbf{P} 特征值分解。



学完本册《统计至简》后，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像。