Landscape of Statistics and Probability

1 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 $\boldsymbol{X}$，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 $\boldsymbol{\Sigma}$：

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho_{1,2}\sigma_1\sigma_2 \\ \rho_{1,2}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \tag{1}$$

什么条件下 $\boldsymbol{\Sigma}$ 是正定矩阵？

定义如下二元函数：

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

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 $\boldsymbol{\Sigma}$ 进行 Cholesky 分解：

$$\boldsymbol{\Sigma} = \boldsymbol{R}^{\mathrm{T}}\boldsymbol{R} \tag{3}$$

矩阵 $\boldsymbol{\Sigma}$ 能进行 Cholesky 分解的前提是什么？

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

特征值分解

对 $\boldsymbol{\Sigma}$ 特征值分解：

$$\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}} \tag{4}$$

为什么等式右侧第二个矩阵 $\boldsymbol{V}$ 对应转置运算？

矩阵 $\boldsymbol{V}$ 有什么特殊性质？如何从向量空间角度理解 $\boldsymbol{V}$？

矩阵 $\boldsymbol{\Lambda}$ 有什么特殊性质？

如果把 $\boldsymbol{V}$ 写成 $[\boldsymbol{v}_1, \boldsymbol{v}_2]$，(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 $\boldsymbol{X}$ 奇异值分解可以获得其奇异值 s_j，对 $\boldsymbol{X}$ 的协方差矩阵 $\boldsymbol{\Sigma}$ 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：

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

$(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$ 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。

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

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

X 的样本方差为：

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

X 的样本标准差为：

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

对于样本数据，随机变量 X 和 Y 的协方差为：

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

对于样本数据，随机变量 X 和 Y 的相关性系数为：

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

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：

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

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：

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

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：

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

贝叶斯定理为：

$$\Pr(A|B)\Pr(B) = \Pr(B|A)\Pr(A) = \Pr(A,B) \tag{14}$$

假设 $A_1, A_2, \ldots, A_n$ 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：

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

如果事件 A 和事件 B 独立，则：

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

如果事件 A 和事件 B 在 C 发生条件下条件独立，则：

$$\Pr(A,B|C) = \Pr(A|C) \cdot \Pr(B|C) \tag{17}$$

离散随机变量

离散随机变量 X 的概率质量函数满足：

$$\sum_{x} p_X(x) = 1, \quad 0 \le p_X(x) \le 1 \tag{18}$$

离散随机变量 X 的期望值为：

$$\mathrm{E}(X) = \sum_{x} x \cdot p_X(x) \tag{19}$$

离散随机变量 X 的方差为：

$$\mathrm{var}(X) = \sum_{x} (x - \mathrm{E}(X))^2 \cdot p_X(x) \tag{20}$$

二元离散随机变量 (X, Y) 的概率质量函数满足：

$$\sum_{x}\sum_{y} p_{X,Y}(x,y) = 1, \quad 0 \le p_{X,Y}(x,y) \le 1 \tag{21}$$

(X, Y) 的协方差定义为：

$$\begin{aligned} \mathrm{cov}(X,Y) &= \mathrm{E}\big((X - \mathrm{E}(X))(Y - \mathrm{E}(Y))\big) \\ &= \sum_{x}\sum_{y} p_{X,Y}(x,y)(x - \mathrm{E}(X))(y - \mathrm{E}(Y)) \end{aligned} \tag{22}$$

边缘概率 $p_X(x)$ 为：

$$p_X(x) = \sum_{y} p_{X,Y}(x,y) \tag{23}$$

边缘概率 $p_Y(y)$ 为：

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

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：

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

$p_{X|Y}(x|y)$ 对 x 求和等于 1：

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

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：

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

$p_{Y|X}(y|x)$ 对 y 求和等于 1：

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

如果离散变量 X 和 Y 独立，则：

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

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

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

离散分布

$[a, b]$ 上离散均匀分布的概率质量函数为：

$$p_x(x)=\frac{1}{b-a+1},\quad x=a,a+1,\cdots b-1,b \tag{31}$$

伯努利分布的概率质量函数为：

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

其中，p 的取值范围为 $[0, 1]$。

二项分布的概率质量函数为：

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

多项分布的概率质量函数为：

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

其中 x_i (i = 1, 2, …, 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：

$$p_x\left(x\right)=\frac{\exp\left(-\lambda\right)\lambda^x}{x!},\quad x=0,1,2,\ldots \tag{35}$$

连续随机变量

连续随机变量 X 的概率密度函数满足：

$$\int_{-\infty}^{+\infty}f_X\left(x\right)\mathrm{d}x=1,\quad f_X\left(x\right)\geq 0 \tag{36}$$

连续随机变量 X 期望为：

$$\mathrm{E}\left(X\right)=\int_x x\cdot f_X\left(x\right)\mathrm{d}x \tag{37}$$

连续随机变量 X 方差为：

$$\mathrm{var}\left(X\right)=\mathrm{E}\left[\left(X-\mathrm{E}\left(X\right)\right)^2\right]=\int_x\left(x-\mathrm{E}\left(X\right)\right)^2\cdot f_X\left(x\right)\mathrm{d}x \tag{38}$$

连续随机变量 X 的边缘概率密度函数 $f_X(x)$ 为：

$$f_X\left(x\right)=\int_y f_{X,Y}\left(x,y\right)\mathrm{d}y \tag{39}$$

连续随机变量 Y 的边缘概率密度函数 $f_Y(y)$ 为：

$$f_Y\left(y\right)=\int_x f_{X,Y}\left(x,y\right)\mathrm{d}x \tag{40}$$

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：

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

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：

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

利用贝叶斯定理，联合概率 $f_{X,Y}(x,y)$ 为：

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

如果连续随机变量 X 和 Y 独立，下则：

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

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：

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

连续分布

区间 $[a, b]$ 的连续均匀分布概率密度函数为：

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

一元学生 t-分布的概率密度函数为：

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

指数分布的概率密度函数为：

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

Beta(α, β) 分布的概率密度函数为：

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

Dirichlet 分布概率密度函数为：

$$f_{X_1,\ldots,X_K}(x_1,\ldots,x_K;\alpha_1,\ldots,\alpha_K) = \frac{1}{\mathrm{B}(\alpha_1,\ldots,\alpha_K)} \prod_{i=1}^{K} x_i^{\alpha_i-1}, \quad \sum_{i=1}^{K} x_i = 1 \tag{50}$$

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 $\mathrm{E}(Y|X = x)$ 为：

$$\mathrm{E}\left(Y|X=x\right)=\sum_{y} y\cdot p_{Y|X}\left(y|x\right) \tag{51}$$

E(Y) 的全期望定理为：

$$\mathrm{E}\left(Y\right)=\sum_{x}\mathrm{E}\left(Y\mid X=x\right)\cdot p_{X}\left(x\right) \tag{52}$$

给定 $Y=y$ 条件下，X 的条件期望 E($X|Y=y$) 定义为：

$$\mathrm{E}\left(X|Y=y\right)=\sum_{x} x\cdot p_{X|Y}\left(x|y\right) \tag{53}$$

E(Y) 的全期望定理为：

$$\mathrm{E}\left(X\right)=\sum_{y}\mathrm{E}\left(X\mid Y=y\right)\cdot p_{Y}\left(y\right) \tag{54}$$

给定 $X=x$ 条件下，Y 的条件方差 var($Y|X=x$) 为：

$$\mathrm{var}\left(Y|X=x\right)=\sum_{y}\left(y-\mathrm{E}\left(Y|X=x\right)\right)^{2}\cdot p_{Y|X}\left(y|x\right) \tag{55}$$

给定 $Y=y$ 条件下，X 的条件方差 var($X|Y=y$) 为：

$$\mathrm{var}\left(X|Y=y\right)=\sum_{x}\left(x-\mathrm{E}\left(X|Y=y\right)\right)^{2}\cdot p_{X|Y}\left(x\mid y\right) \tag{56}$$

对于 var(Y)，全方差定理为：

$$\mathrm{var}\left(Y\right)=\mathrm{E}\left(\mathrm{var}\left(Y\mid X\right)\right)+\mathrm{var}\left(\mathrm{E}\left(Y\mid X\right)\right) \tag{57}$$

对于 var(X)，全方差定理为：

$$\mathrm{var}\left(X\right)=\mathrm{E}\left(\mathrm{var}\left(X\mid Y\right)\right)+\mathrm{var}\left(\mathrm{E}\left(X\mid Y\right)\right) \tag{58}$$

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：

$$\mathrm{E}\left(Y|X=x\right)=\int_{y} y\cdot f_{Y|X}\left(y|x\right)\mathrm{d}y \tag{59}$$

条件方差 var($Y|X=x$) 为：

$$\mathrm{var}\left(Y|X=x\right)=\int_{y}\left(y-\mathrm{E}\left(Y|X=x\right)\right)^{2}\cdot f_{Y|X}\left(y|x\right)\mathrm{d}y \tag{60}$$

在给定 $Y=y$ 条件下，条件期望 E($X|Y=y$) 为：

$$\mathrm{E}\left(X|Y=y\right)=\int_{x} x\cdot f_{X|Y}\left(x|y\right)\mathrm{d}x \tag{61}$$

条件方差 var($X|Y=y$) 定义为：

$$\operatorname{var}\left(X|Y=y\right)=\int_{x}\left(X-\mathrm{E}\left(X|Y=y\right)\right)^{2}\cdot f_{X|Y}\left(x|y\right)\mathrm{d}x \tag{62}$$

1.4 高斯

一元高斯分布

一元高斯分布的概率密度函数为：

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

标准正态分布的概率密度函数为：

$$f_{Z}\left(z\right)=\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-z^{2}}{2}\right) \tag{64}$$

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：

$$f_{X,Y}\left(x,y\right)=\frac{1}{2\pi\sigma_{X}\sigma_{Y}\sqrt{1-\rho_{X,Y}^{2}}}\times\exp\left(\frac{-1}{2}\frac{1}{\left(1-\rho_{X,Y}^{2}\right)}\left(\left(\frac{x-\mu_{X}}{\sigma_{X}}\right)^{2}-2\rho_{X,Y}\left(\frac{x-\mu_{X}}{\sigma_{X}}\right)\left(\frac{y-\mu_{Y}}{\sigma_{Y}}\right)+\left(\frac{y-\mu_{Y}}{\sigma_{Y}}\right)^{2}\right)\right) \tag{65}$$

X 的边缘概率密度函数为：

$$f_{X}\left(x\right)=\frac{1}{\sigma_{X}\sqrt{2\pi}}\exp\left(\frac{-1}{2}\left(\frac{x-\mu_{X}}{\sigma_{X}}\right)^{2}\right) \tag{66}$$

Y 的边缘概率密度函数为：

$$f_{Y}\left(y\right)=\frac{1}{\sigma_{Y}\sqrt{2\pi}}\exp\left(\frac{-1}{2}\left(\frac{x-\mu_{Y}}{\sigma_{Y}}\right)^{2}\right) \tag{67}$$

多元高斯分布

多元高斯分布的概率密度函数为：

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

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：

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

条件期望 $\mathrm{E}(Y|X=x)$ 为：

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

条件方差 $\mathrm{var}(Y|X=x)$ 为：

$$\mathrm{var}\left(Y|X=x\right)=\left(1-\rho_{X,Y}^2\right)\sigma_Y^2 \tag{71}$$

如果随机变量向量 $\boldsymbol{\chi}$ 和 $\boldsymbol{\gamma}$ 服从多元高斯分布：

$$\begin{bmatrix}\boldsymbol{\chi}\\ \boldsymbol{\gamma}\end{bmatrix}\sim N\left(\begin{bmatrix}\boldsymbol{\mu}_{\chi}\\ \boldsymbol{\mu}_{\gamma}\end{bmatrix},\begin{bmatrix}\boldsymbol{\Sigma}_{\chi\chi} & \boldsymbol{\Sigma}_{\chi\gamma}\\ \boldsymbol{\Sigma}_{\gamma\chi} & \boldsymbol{\Sigma}_{\gamma\gamma}\end{bmatrix}\right) \tag{72}$$

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

$$\left\{\boldsymbol{\gamma}|\boldsymbol{\chi}=\boldsymbol{x}\right\}\sim N\left(\underbrace{\boldsymbol{\Sigma}_{\gamma\chi}\boldsymbol{\Sigma}_{\chi\chi}^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}_{\chi}\right)+\boldsymbol{\mu}_{\gamma}}_{\text{Expectation}},\quad\underbrace{\boldsymbol{\Sigma}_{\gamma\gamma}-\boldsymbol{\Sigma}_{\gamma\chi}\boldsymbol{\Sigma}_{\chi\chi}^{-1}\boldsymbol{\Sigma}_{\chi\gamma}}_{\text{Variance}}\right) \tag{73}$$

给定 $\boldsymbol{\chi}=\boldsymbol{x}$ 的条件下 $\boldsymbol{\gamma}$ 的条件期望为：

$$\mathrm{E}\left(\boldsymbol{\gamma}|\boldsymbol{\chi}=\boldsymbol{x}\right)=\boldsymbol{\mu}_{\gamma|\chi=x}=\boldsymbol{\Sigma}_{\gamma\chi}\boldsymbol{\Sigma}_{\chi\chi}^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}_{\chi}\right)+\boldsymbol{\mu}_{\gamma} \tag{74}$$

协方差矩阵

随机变量的列向量 $\boldsymbol{\chi}$ 的协方差矩阵为：

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

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

$$\boldsymbol{\Sigma} = \frac{\left(\boldsymbol{X} - \mathrm{E}(\boldsymbol{X})\right)^{\mathrm{T}}\left(\boldsymbol{X} - \mathrm{E}(\boldsymbol{X})\right)}{n-1} \tag{76}$$

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

$$\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}} = \sum_{j=1}^{D}\lambda_j \boldsymbol{v}_j \boldsymbol{v}_j^{\mathrm{T}} \tag{77}$$

协方差矩阵 $\boldsymbol{\Sigma}$ 的迹为特征值之和：

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

$\boldsymbol{\Sigma}$ 的行列式值为其特征值乘积：

$$|\boldsymbol{\Sigma}| = |\boldsymbol{\Lambda}| = \prod_{j=1}^{D}\lambda_j \tag{79}$$

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

$$\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^{\mathrm{T}} = \boldsymbol{R}^{\mathrm{T}}\boldsymbol{R} \tag{80}$$

合并协方差矩阵为：

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

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：

$$Y = h(X) = aX + b \tag{82}$$

其中，a 和 b 为常数。

(82) 中，Y 和 X 的期望、方差之间关系为：

$$\begin{aligned} \mathrm{E}(Y) &= a\,\mathrm{E}(X) + b \\ \mathrm{var}(Y) &= \mathrm{var}(aX + b) = a^2\,\mathrm{var}(X) \end{aligned} \tag{83}$$

如果 Y 和二元随机变量 (X_1, X_2) 存在如下关系：

$$Y = aX_1 + bX_2 \tag{84}$$

Y 的期望、方差为：

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

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

$$Y = \boldsymbol{v}^{\mathrm{T}}\boldsymbol{\chi} \tag{86}$$

Y 的期望、方差为：

$$\begin{aligned} \mathrm{E}(Y) &= \boldsymbol{v}^{\mathrm{T}}\boldsymbol{\mu}_{\chi} \\ \mathrm{var}(Y) &= \boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Sigma}_{\chi}\boldsymbol{v} \end{aligned} \tag{87}$$

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

$$\boldsymbol{\gamma} = \boldsymbol{V}^{\mathrm{T}}\boldsymbol{\chi} \tag{88}$$

$\boldsymbol{\gamma}$ 的期望、协方差矩阵为：

$$\begin{aligned} \mathrm{E}(\boldsymbol{\gamma}) &= \boldsymbol{V}^{\mathrm{T}}\boldsymbol{\mu}_{\chi} \\ \mathrm{var}(\boldsymbol{\gamma}) &= \boldsymbol{V}^{\mathrm{T}}\boldsymbol{\Sigma}_{\chi}\boldsymbol{V} \end{aligned} \tag{89}$$

蒙特卡洛模拟

多元随机数 $\boldsymbol{Z}$ 服从 $N(\boldsymbol{0}, \boldsymbol{I}_{D\times D})$，多元随机数 $\boldsymbol{X}$ 服从 $N(\mathrm{E}(\boldsymbol{X}), \boldsymbol{\Sigma}_{D\times D})$，两者关系为：

$$\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{R} + \mathrm{E}(\boldsymbol{X}) \tag{90}$$

其中，矩阵 $\boldsymbol{R}$ 来自 (80)。

1.6 频率派

频率派统计推断

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

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

这 n 个随机变量的平均值 $\bar{X}$ 近似服从如下正态分布：

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

最大似然估计的优化问题为：

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

概率密度估计

概率密度估计函数：

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

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

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

1.7 贝叶斯派

贝叶斯分类

利用贝叶斯定理分类：

$$f_{Y|X}\left(C_k\,|\,x\right) = \frac{f_{X|Y}\left(x|C_k\right)p_Y\left(C_k\right)}{f_X\left(x\right)} \tag{96}$$

$f_{Y|X}(C_k\,|\,x)$ 叫后验概率，又叫成员值。$f_X(x)$ 为证据因子，也叫证据。$p_Y(C_k)$ 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可以是最大化后验概率：

$$\hat{y} = \arg\max_{C_k} f_{Y|X}\left(C_k\,|\,x\right) \tag{97}$$

其中，$k = 1,2,\ldots K$。

贝叶斯统计推断

模型参数的后验分布为：

$$f_{\Theta|X}\left(\theta\,|\,x\right) = \frac{f_{X|\Theta}\left(x\,|\,\theta\right)f_{\Theta}\left(\theta\right)}{\int_{\vartheta} f_{X|\Theta}\left(x\,|\,\vartheta\right)f_{\Theta}\left(\vartheta\right)\mathrm{d}\vartheta} \tag{98}$$

最大化后验估计的优化问题为：

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

1.8 椭圆三部曲

马氏距离

马氏距离的定义为：

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

D 维马氏距离的平方则服从自由度为 D 的卡方分布：

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

线性回归

一元线性回归表达式为：

$$y = \hat{y} + \varepsilon = b_0 + b_1 x + \varepsilon \tag{102}$$

最小二乘法优化问题为：

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

多元线性回归可以写成超定方程组：

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} \tag{104}$$

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

$$\boldsymbol{b} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{y} \tag{105}$$

主成分分析

对原始矩阵 $\boldsymbol{X}$ 进行经济型 SVD 分解：

$$\boldsymbol{X} = \boldsymbol{U}_{\boldsymbol{X}} \boldsymbol{S}_{\boldsymbol{X}} \boldsymbol{V}_{\boldsymbol{X}}^{\mathrm{T}} \tag{106}$$

其中，$\boldsymbol{S}_{\boldsymbol{X}}$ 为对角方阵。

利用 (106)，$\boldsymbol{X}$ 的格拉姆矩阵可以展开为：

$$\boldsymbol{G} = \boldsymbol{V}_{\boldsymbol{X}} \boldsymbol{S}_{\boldsymbol{X}}^{2} \boldsymbol{V}_{\boldsymbol{X}}^{\mathrm{T}} \tag{107}$$

上式便是格拉姆 $\boldsymbol{G}$ 的特征值分解。

对中心化数据矩阵 $\boldsymbol{X}_c$ 经济型 SVD 分解：

$$\boldsymbol{X}_c = \boldsymbol{U}_c \boldsymbol{S}_c \boldsymbol{V}_c^{\mathrm{T}} \tag{108}$$

而协方差矩阵 $\boldsymbol{\Sigma}$ 则可以写成：

$$\boldsymbol{\Sigma} = \boldsymbol{V}_c \frac{\boldsymbol{S}_c^2}{n-1} \boldsymbol{V}_c^{\mathrm{T}} \tag{109}$$

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

$$\lambda_{c\_j} = \frac{s_{c\_j}^2}{n-1} \tag{110}$$

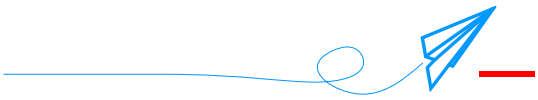
同样，对标准化数据矩阵 $\boldsymbol{Z}_{\boldsymbol{X}}$ 进行经济型 SVD 分解：

$$\boldsymbol{Z}_{\boldsymbol{X}} = \boldsymbol{U}_{\boldsymbol{Z}} \boldsymbol{S}_{\boldsymbol{Z}} \boldsymbol{V}_{\boldsymbol{Z}}^{\mathrm{T}} \tag{111}$$

相关性系数矩阵 $\boldsymbol{P}$ 则可以写成：

$$\boldsymbol{P} = \boldsymbol{V}_{\boldsymbol{Z}} \frac{\boldsymbol{S}_{\boldsymbol{Z}}^2}{n-1} \boldsymbol{V}_{\boldsymbol{Z}}^{\mathrm{T}} \tag{112}$$

上式相当于对 $\boldsymbol{P}$ 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。