

# 第七章 概率基础

## 第 21 讲 随机变量的数字特征

黄定江

DaSE @ ECNU  
djhuang@dase.ecnu.edu.cn

① 21.1 期望

② 21.2 方差

③ 21.3 矩和协方差矩阵

## 1 21.1 期望

## 2 21.2 方差

## 3 21.3 矩和协方差矩阵

## 21.1.1 期望：定义

## 定义 1

随机变量  $X$  的期望值或均值或一阶矩定义为

$$E(X) = \int_{-\infty}^{\infty} x dF(x) = \begin{cases} \sum_x x f(x), & X \text{ 为离散型随机变量,} \\ \int_x x f(x) dx, & X \text{ 为连续型随机变量.} \end{cases}$$

其中  $F(x)$  是概率累积函数. 也可以使用如下符号表示  $X$  的期望:

$$E(X) = EX = \int_{-\infty}^{\infty} x dF(x) = \mu = \mu_X.$$

需要说明的是, 若  $X$  是离散型随机变量且期望存在, 要求

$$\sum_x |x| f(x) < +\infty$$

若  $X$  是连续型随机变量且期望存在, 要求:

$$\int_{-\infty}^{\infty} |x| f(x) dx < +\infty$$

## 例 1

令  $X \sim \text{Bernoulli}(p)$ , 则

$$E(X) = \sum_{x=0}^1 xf(x) = (0 \times (1-p)) + (1 \times p) = p$$

## 例 2

令  $X \sim \text{Uniform}(-1, 3)$  则

$$E(X) = \int xf_X(x) = \int xf(x) dx = \frac{1}{4} \int_{-1}^3 x dx = 1$$

## 期望的求法：懒惰统计学家法则

## 定理 1

设  $Y$  是随机变量  $X$  的函数:  $Y = r(X)$ . 则

$$E(Y) = E(r(X)) = \int r(x) dF_X(x)$$

定理的证明超出了本课程的范围, 这里不再详述. 但是定理的重要意义在于当我们求  $E(Y)$  时, 不必算出  $Y$  的概率密度函数, 而只需利用  $X$  的概率密度函数就可以了.

## 例 3

令  $X \sim \text{Uniform}(0, 1)$ ,  $Y = r(X) = e^X$ , 则

$$E(Y) = \int_0^1 e^x f(x) dx = \int_0^1 e^x dx = e - 1$$

另一种方法就是先求出  $f_Y(y) = 1/y$ , 其中  $0 < y < e$ , 从而计算出  $E(Y) = \int_1^e y f(y) dy = e - 1$

## 期望的性质

### 性质

1. 设  $C$  是常数, 则有  $E(C) = C$
2. 设  $X$  是随机变量,  $C$  是常数, 则有

$$E(CX) = CE(X)$$

3. 设  $X$  和  $Y$  是两个随机变量, 则有:

$$E(X + Y) = E(X) + E(Y)$$

4. 设  $X$  和  $Y$  是相互独立的两个随机变量, 则有:

$$E(XY) = E(X)E(Y)$$

## 常见离散型随机变量的期望

- 随机变量  $X$  服从参数为  $p$  的伯努利分布的期望

$$E(X) = p \times 1 + (1 - p) \times 0 = p$$

- 随机变量  $X$  服从参数为  $n, p$  的二项式分布的期望

$$E(X) = \sum_{k=0}^n k \times P(X = k) = \sum_{k=0}^n k \times C_n^k p^k (1 - p)^{n-k} = np$$

- 随机变量  $X$  服从参数为  $p$  的几何分布 ( $0 < p < 1$ ) 的期望

$$E(X) = \sum_{k=1}^n k \times P(X = k) = \sum_{k=1}^n k \times (1 - p)^{k-1} p = \frac{1}{p}$$

- 随机变量  $X$  服从参数为  $\lambda$  的泊松分布的期望

$$E(X) = \sum_{k=0}^{\infty} k e^{-\lambda} \frac{\lambda^k}{k!} = \lambda$$



## 连续型随机变量的期望

- 随机变量  $X \sim U(a, b)$  的期望

$$E(X) = \int_{-\infty}^{+\infty} xf(x) dx = \int_a^b x \frac{1}{b-a} dx = \frac{b+a}{2}$$

- 随机变量  $X$  服从参数为  $\mu, \gamma$  的 Laplace 分布的期望

$$E(X) = \int_{-\infty}^{+\infty} xf(x) dx = \int_{-\infty}^{+\infty} x \frac{1}{2\gamma} \exp\left(-\frac{|x-\mu|}{\gamma}\right) dx = \mu$$

- 随机变量  $X$  服从参数为  $\mu, \sigma$  的高斯分布的期望

$$E(X) = \int_{-\infty}^{+\infty} xf(x) dx = \int_{-\infty}^{+\infty} x \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \mu$$

## 期望应用举例

## 例 4

期望在机器学习中随处可见。比如在衡量分类模型准确率时我们使用期望来定义：

$$accuracy = E(I(y = \hat{f}(x))) = \sum I(y = \hat{f}(x)) dF(x)$$

其中  $F(x)$  是随机变量  $x$  的累积分布函数， $\hat{f}(x)$  是训练好的模型，且：

$$I(y = \hat{f}(x)) = \begin{cases} 1 & \text{如果 } y = \hat{f}(x) \\ 0 & \text{如果 } y \neq \hat{f}(x) \end{cases}$$

但在实际中，我们并不知道数据真实的概率分布函数，常取的做法是采集  $n$  个未在训练集中出现过的样本作为测试集，使用：

$$\sum_i^n \frac{1}{n} I(y_i = \hat{f}(x_i))$$

来衡量模型的准确率。

此外强化学习中价值函数（状态价值函数和动作价值函数）也是由数学期望来定义的。

## 21.1.2 条件期望

假设  $X$  和  $Y$  为随机变量, 当  $Y = y$  时  $X$  的均值是多少? 方法跟前面计算  $X$  的均值一样, 只不过将期望定义中的  $f_X(x)$  用  $f_{X|Y}(x|y)$  代替就可以了.

## 定义 2

给定  $Y = y$  情况下  $X$  的条件期望为

$$E(X|Y = y) = \begin{cases} \sum xf_{X|Y}(x|y), & \text{离散情形,} \\ \int xf_{X|Y}(x|y)dx, & \text{连续情形} \end{cases}$$

如果  $r(x, y)$  为  $x, y$  的函数, 则

$$E(r(X, Y)|Y = y) = \begin{cases} \sum r(x, y)f_{X|Y}(x|y), & \text{离散情形,} \\ \int r(x, y)f_{X|Y}(x|y)dx, & \text{连续情形} \end{cases}$$

**注意!** 条件期望与期望有一些区别, 期望  $E(X)$  是一个值, 而  $E(X|Y=y)$  是  $y$  的函数. 在观察  $y$  之前, 并不知道  $E(X|Y=y)$  的值, 所以它是一个随机变量, 记为  $E(X|Y)$ . 换句话说,  $E(X|Y)$  是随机变量, 当  $Y=y$  时, 其值为  $E(X|Y=y)$ . 类似的,  $E(r(X, Y)|Y)$  是随机变量, 当  $Y=y$  时, 其值为  $E(r(X, Y)|Y=y)$ . 这一点很容易引起混淆, 下面举一个例子来说明.

### 例 5

假设  $X \sim \text{Uniform}(0, 1)$ , 当观察到  $X=x$  后, 假设  $Y|X=x \sim \text{Uniform}(x, 1)$ , 凭直觉  $E(Y|X=x) = (1+x)/2$ , 事实上  $f_{Y|X}(y|x) = 1/(1-x)$ , 其中  $x < y < 1$ , 故

$$E(Y|X=x) = \int_0^1 y f_{Y|X}(y|x) dy = \frac{1}{1-x} \int_x^1 y dy = \frac{1+x}{2}$$

因此  $E(Y|X) = (1+X)/2$ , 它是一个随机变量. 当观察到  $X=x$  后, 其值为  $E(Y|X=x) = (1+x)/2$

## 定理 2

(期望迭代法则) 对随机变量  $X$  和  $Y$ , 假设期望均存在, 则有

$$E[E(Y|X)] = E(Y), \quad E[E(X|Y)] = E(X)$$

更一般的, 对任意函数  $r(x, y)$  有

$$E[E(r(X, Y)|X)] = E(r(X, Y))$$

## 证明.

下面证明第一个等式, 利用条件期望的定义和  $f(x, y) = f(x)f(y|x)$

$$\begin{aligned} E[E(Y|X)] &= \int E(Y|X=x)f_X(x)dx = \iint yf(y|x)dyf(x)dx \\ &= \iint yf(y|x)f(x)dxdy = \iint yf(x, y)dxdy = E(Y) \end{aligned}$$



## 例 6

回到例5中, 试问怎么计算  $E(Y)$ ? 一种方法是求出联合密度函数  $f(x, y)$ , 然后计算  $E(Y) = \iint yf(x, y) dx dy$ . 另一种更简单的方式可以分两步来实现. 首先计算  $E(Y|X) = (1 + X)/2$ , 从而

$$\begin{aligned} E(Y) &= E[E(Y|X)] = E((1 + X)/2) \\ &= \frac{1 + E(X)}{2} = \frac{(1 + (1/2))}{2} = \frac{3}{4} \end{aligned}$$

1 21.1 期望

2 21.2 方差

3 21.3 矩和协方差矩阵

## 21.2.1 方差：定义

## 定义 3

令随机变量  $X$  的均值为  $\mu$ ,  $X$  的方差记为  $\sigma^2$ , 定义为

$$\begin{aligned}\sigma^2 &= E(X - \mu)^2 \\ &= \int (x - \mu)^2 dF(X) \\ &= \begin{cases} \sum_x (x - \mu)^2 f(x), & X \text{ 为离散型随机变量,} \\ \int_x (x - \mu)^2 f(x) dx, & X \text{ 为连续型随机变量.} \end{cases}\end{aligned}$$

其中假设期望存在. 标准差定义为  $sd(X) = \sqrt{\sigma^2} = \sigma$



## 例 7

设随机变量  $X$  具有数学期望  $E(X) = \mu$ , 方差  $D(X) = \sigma^2 \neq 0$ , 记

$$X^* = \frac{X - \mu}{\sigma}$$

则

$$E(X^*) = \frac{1}{\sigma} E(X - \mu) = \frac{1}{\sigma} [E(X) - \mu] = 0$$

$$D(X^*) = E((X^*)^2) = \int \left(\frac{x - \mu}{\sigma}\right)^2 f(x) dx = \frac{1}{\sigma^2} \int (x - \mu)^2 f(x) dx = 1$$

因此对于任何一个具有均值和方差的分布, 我们总可以通过这样的变换将其变为均值为 0, 方差为 1 的分布。

## 方差的性质

### 性质

1. 设  $X$  是随机变量, 有

$$D(X) = E(X^2) - [E(X)]^2$$

2. 设  $C$  是常数, 则有  $D(C) = 0$

3. 设  $X$  是随机变量,  $C$  是常数, 则有

$$D(CX) = C^2 D(X) \quad D(X + C) = D(X)$$

4. 设  $X$  和  $Y$  是两个随机变量, 则有:

$$D(X + Y) = D(X) + D(Y) + 2E\{(X - E(X))(Y - E(Y))\}$$

若  $X$  与  $Y$  相互独立, 则有:

$$D(X + Y) = D(X) + D(Y)$$

## 定义 4

如果  $X_1, \dots, X_n$  为随机变量, 则定义样本均值为

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

样本方差为

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

## 定理 3

令  $X_1, X_2, \dots, X_n$  是独立同分布的随机变量且  $\mu = E(X_i), \sigma^2 = D(X_i)$ , 则

$$E(\bar{X}_n) = \mu, D(\bar{X}_n) = \frac{\sigma^2}{n}, E(S_n^2) = \sigma^2$$

## 常见离散型随机变量的方差

- 随机变量  $X$  服从参数为  $p$  的伯努利分布的方差

$$D(X) = (1-p)^2 \times p + (0-p)^2 \times (1-p) = p(1-p)$$

- 随机变量  $X$  服从参数为  $n, p$  的二项式分布的方差

$$\begin{aligned} D(X) &= \sum_{k=0}^n (k - E(X))^2 \times P(X = k) \\ &= \sum_{k=0}^n (k - np)^2 \times C_n^k p^k (1-p)^{n-k} \\ &= np(1-p) \end{aligned}$$

- 随机变量  $X$  服从参数为  $p$  的几何分布,  $0 < p < 1$  的方差

$$D(X) = \sum_{k=1}^n (k - E(k))^2 \times P(X = k) = \sum_{k=1}^n \left(k - \frac{1}{p}\right)^2 \times (1-p)^{k-1} p = \frac{1-p}{p^2}$$

- 随机变量  $X$  服从参数为  $\lambda$  的泊松分布的方差  $D(X) = \lambda$

## 常见连续型随机变量的方差

- 随机变量  $X \sim U(a, b)$  的方差

$$D(X) = \int_{-\infty}^{+\infty} (x - E(X))^2 f(x) dx = \int_a^b \left(x - \frac{b+a}{2}\right)^2 \frac{1}{b-a} dx = \frac{(b-a)^2}{12}$$

- 随机变量  $X$  服从参数为  $\mu, \gamma$  的 Laplace 分布的方差

$$D(X) = \int_{-\infty}^{+\infty} (x - E(X))^2 f(x) dx = \int_{-\infty}^{+\infty} (x - \mu)^2 \frac{1}{2\gamma} \exp\left(-\frac{|x - \mu|}{\gamma}\right) dx = 2\gamma^2$$

- 随机变量  $X$  服从参数为  $\mu, \sigma$  的高斯分布的方差

$$D(X) = \int_{-\infty}^{+\infty} (x - E(X))^2 f(x) dx = \int_{-\infty}^{+\infty} (x - \mu)^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \sigma^2$$

### 21.2.2 方差的应用：过拟合与偏差-方差分解

- 在机器学习领域，我们将现有的数据划分为训练集和测试集，在训练集上训练一个模型  $\hat{f}$ .
- 如果模型在训练集上表现很好（如对于分类问题，表现好意味着分类准确率高），而在测试集上表现很差，则此时的模型输入处于过拟合状态. 为了避免过拟合，我们需要在模型的拟合能力与复杂度之间进行权衡.
- 拟合能力强的模型一般复杂度会比较高，容易导致过拟合. 相反，如果限制模型的复杂度，降低其拟合能力，又可能会导致欠拟合. 因此，如何在模型的拟合能力和复杂度之间取得一个较好的平衡，对一个机器学习算法来讲十分重要.
- 偏差-方差分解（Bias-Variance Decomposition）为我们提供一个很好的分析和指导工具.

## 例 8

以回归问题为例. 假设样本的真实分布是  $p_r(x, y)$ , 并采用平方损失函数, 模型  $f(x)$  的期望误差为:  $\mathcal{R}(f) = E_{(x,y) \sim p_r(x,y)} [(y - f(x))^2]$ . 那么最优模型为:  $f^*(x) = E_{y \sim p_r(y|x)}[y]$  其中  $p_r(y|x)$  为样本的真实条件分布,  $f^*(x)$  为使用平方损失作为优化目标的最优模型, 其损失为:  $\epsilon = E_{(x,y) \sim p_r(x,y)} [(y - f^*(x))^2]$ .

通常损失  $\epsilon$  是由样本分布以及噪声引起的, 无法通过优化模型来减少.

期望错误可以分解为:

$$\begin{aligned}\mathcal{R}(f) &= E_{(x,y) \sim p_r(x,y)} [(y - f^*(x) + f^*(x) - f(x))^2] \\ &= E_{(x,y) \sim p_r(x,y)} [(y - f^*(x))^2] + E_{(x,y) \sim p_r(x,y)} [(f^*(x) - f(x))^2] \\ &\quad + 2E_{(x,y) \sim p_r(x,y)} [(y - f^*(x))(f^*(x) - f(x))] \\ &= E_{x \sim p_r(x)} [(f^*(x) - f(x))^2] + \epsilon\end{aligned}\tag{1}$$

其中

$$\begin{aligned} 2E_{(x,y) \sim p_r(x,y)} [(y - f^*(x))(f^*(x) - f(x))] &= 2 \int_x \int_y p_r(x, y)(y - f^*(x))(f^*(x) - f(x)) dx dy \\ &= 2 \int_x (f^*(x) - f(x)) dx \int_y p_r(x, y)(y - f^*(x)) dy \end{aligned}$$

对于给定的  $x_0$ :

$$\begin{aligned} \int_y p_r(x_0, y)((y - f^*(x_0))) dy &= \int_y p_r(x_0, y)(y - f^*(x_0)) dy \\ &= p_r(x_0) \int_y p_r(y|x_0)(y - f^*(x_0)) dy \\ &= p_r(x_0) \left( \int_y p_r(y|x_0)y dy - f^*(x_0) \int_y p_r(y|x_0) dy \right) \\ &= 0 \end{aligned}$$

故

$$2E_{(x,y) \sim p_r(x,y)} [(y - f^*(x))(f^*(x) - f(x))] = 0$$



- 式(1)  $\mathcal{R}(f) = E_{x \sim p_r(x)} [(f^*(x) - f(x))^2] + \epsilon$  中的第一项是当前训练出的模型与最优模型之间的差距，是机器学习算法可以优化的真实目标。
- 在实际训练一个模型  $f(x, y)$  时，训练集  $D$  是从真实分布  $p_r(x, y)$  上独立同分布地采样出来的有限样本集合。不同的训练集会得到不同的模型。
- 令  $f_D(x)$  表示在训练集  $D$  学习到的模型，一个机器学习算法（包括模型以及优化算法）的能力可以用不同训练集上的模型的平均性能来评价。
- 对于单个样本  $x$ ，不同训练集  $D$  得到模型  $f_D(x)$  和最优模型  $f^*(x)$  的期望差距为

$$\begin{aligned} E_D [(f_D(x) - f^*(x))^2] &= E_D [(f_D(x) - E_D [f_D(x)] + E_D [f_D(x)] - f^*(x))^2] \\ &= (E_D [f_D(x)] - f^*(x))^2 + E_D [(f_D(x) - E_D [f_D(x)])^2] \end{aligned} \quad (2)$$

- 式(2)中第一项  $(E_D [f_D(x)] - f^*(x))^2$  称为偏差 (*Bias*) 的平方，记为  $(bias.x)^2$ ，是指一个模型在不同训练集上的平均性能和最优模型的差异。
- 第二项  $E_D [(f_D(x) - E_D [f_D(x)])^2]$  称为方差 (*Variance*)，记为  $variance.x$ ，是指一个模型在不同训练集上的差异，可以用来衡量一个模型是否容易过拟合。

用  $E_D [(f_D(x) - f^*(x))^2]$  来代替式(1)中的  $(f(x) - f^*(x))^2$ , 则期望错误可写成:

$$\begin{aligned}\mathcal{R}(f) &= E_{x \sim p_r(x)} [E_D [(f_D(x) - f^*(x))^2]] + \epsilon \\ &= (bias)^2 + variance + \epsilon\end{aligned}\tag{3}$$

其中:

$$(bias)^2 = E_x [(E_D[f_D(x)] - f^*(x))^2]$$

$$variance = E_x [E_D[(f_D(x) - E_D[f_D(x)])^2]]$$

所以最小化期望误差等价于最小化偏差与方差之和.

- 图1给出了机器学习模型的四种偏差和方差组合情况.
- 每个图的中心点为最优模型  $f^*(x)$ , 蓝点为不同训练集  $D$  上得到的模型  $f_D(x)$ .
- 图1a 给出了一种理想情况, 方差和偏差都比较小.
- 图1b 为高偏差低方差的情况, 表示模型的泛化能力很好, 但拟合能力不足.
- 图1c 为低偏差高方差的情况, 表示模型的拟合能力很好, 但泛化能力比较差. 当训练数据比较少时会导致过拟合.
- 图1d 为高偏差高方差的情况, 是一种最差的情况.

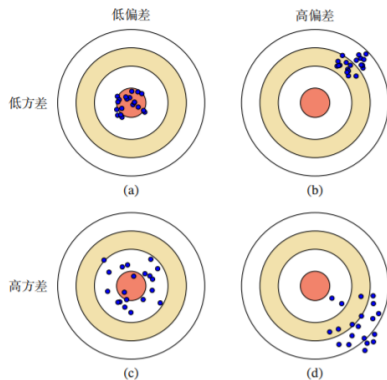


图 1: 偏差与方差的组合

方差一般会随着训练样本的增加而减少。当样本比较多时，方差比较少，这时可以选择能力强的模型来减少偏差。然而在很多机器学习任务上，训练集往往都比较有限，最优的偏差和最优的方差就无法兼顾。

- 以结构错误最小化为例，我们可以调整正则化系数  $\lambda$  来控制模型的复杂度。
- 当  $\lambda$  变大时，模型复杂度会降低，可以有效地减少方差，避免过拟合，但偏差会上升。
- 当  $\lambda$  过大时，总的期望错误反而会上升。
- 因此，一个好的正则化系数  $\lambda$  需要在偏差和方差之间取得比较好的平衡。

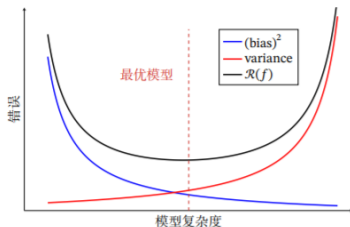


图 2: 机器学习模型的期望错误、偏差和方差随复杂度的变化情况

图2给出了机器学习模型的期望错误、偏差和方差随复杂度的变化情况，其中红色虚线表示最优模型。最优模型并不一定是偏差曲线和方差曲线的交点。

- 偏差和方差分解给机器学习模型提供了一种分析途径，但在实际操作中难以直接衡量。
- 一般来说，当一个模型在训练集上的错误率比较高时，说明模型的拟合能力不够，偏差比较高。
- 这种情况可以通过增加数据特征、提高模型复杂度、减少正则化系数等操作来改进模型。
- 当模型在训练集上的错误率比较低，但验证集上的错误率比较高时，说明模型过拟合，方差比较高。
- 这种情况可以通过降低模型复杂度、加大正则化系数、引入先验等方法来缓解。
- 此外，还有一种有效降低方差的方法为集成模型，即通过多个高方差模型的平均来降低方差。

### 21.2.3 协方差和相关系数

对于二维随机变量  $(X, Y)$ ，除了讨论期望与方差之外，还需讨论  $X$  和  $Y$  之间的相关关系的数字特征.

#### 定义 5

令  $X$  和  $Y$  是均值分别为  $\mu_X$  和  $\mu_Y$ ，标准差分别是  $\sigma_X$  和  $\sigma_Y$  的随机变量，定义  $X$  和  $Y$  的协方差：

$$\text{Cov}(X, Y) = E\{(X - \mu_X)(Y - \mu_Y)\}$$

相关系数：

$$\rho = \rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

## 协方差的性质

由协方差定义可知

性质

1.

$$\text{Cov}(X, Y) = \text{Cov}(Y, X) \quad \text{Cov}(X, X) = D(X)$$

2. 对于任意两个随机变量  $X$  和  $Y$ , 下列等式成立:

$$D(X + Y) = D(X) + D(Y) + 2\text{Cov}(X, Y)$$

将  $\text{Cov}(X, Y)$  的定义展开, 易得

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$$

3.

$$\text{Cov}(aX, bY) = ab\text{Cov}(X, Y)$$

4.

$$\text{Cov}(X_1 + X_2, Y) = \text{Cov}(X_1, Y) + \text{Cov}(X_2, Y)$$

## 相关系数的性质

同时相关系数具有以下性质:

性质

1.

$$|\rho_{XY}| \leq 1$$

当  $\rho = 0$  时, 称随机变量  $X$  和  $Y$  不相关.

2.  $|\rho_{XY}| = 1$  的充要条件是存在  $a, b$  使  $P(Y = a + bX) = 1$



## 例 9

设  $(X, Y)$  服从二维正态分布, 它的概率密度函数

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp \left\{ \frac{-1}{2(1-\rho^2)} \left[ \frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho \frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2} \right] \right\}$$

我们可以分别计算出  $X$  和  $Y$  的边缘概率分布, 然后分别求出  $X$  和  $Y$  的期望, 方差以及相关系数. 由于计算太复杂, 这直接给出结

果:  $E(X) = \mu_1, E(Y) = \mu_2, D(X) = \sigma_1^2, D(Y) = \sigma_2^2, Cov(X, Y) = \rho$

这也就是说, 二维正态随机变量  $(X, Y)$  的概率密度中的参数  $\rho$  就是  $X$  和  $Y$  的相关系数, 因而二维正态随机变量的分布完全可由  $X, Y$  各自的数学期望, 方差以及它们的相关系数所确定.

## 定理 4

若  $(X, Y)$  服从二维正态分布, 那么  $X$  和  $Y$  相互独立的充要条件是  $\rho = 0$

证明.

必要性: 若  $X$  和  $Y$  相互独立, 则:

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y) = E(X)E(Y) - E(X)E(Y) = 0$$

故  $\rho = 0$

充分性: 若  $\rho = 0$ , 则:

$$\begin{aligned} f(x, y) &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp \left\{ \frac{-1}{2(1-\rho^2)} \left[ \frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho \frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2} \right] \right\} \\ &= \frac{1}{2\pi\sigma_1\sigma_2} \exp \left\{ \frac{-1}{2} \left[ \frac{(x-\mu_1)^2}{\sigma_1^2} + \frac{(y-\mu_2)^2}{\sigma_2^2} \right] \right\} \\ &= \frac{1}{\sqrt{2\pi}\sigma_1} \exp \left\{ \frac{-1}{2} \left[ \frac{(x-\mu_1)^2}{\sigma_1^2} \right] \right\} \times \frac{1}{\sqrt{2\pi}\sigma_2} \exp \left\{ \frac{-1}{2} \left[ \frac{(y-\mu_2)^2}{\sigma_2^2} \right] \right\} \\ &= f(x)f(y) \end{aligned}$$

所以随机变量  $(X, Y)$  相互独立. □

## 21.2.4 随机变量的内积

- 如果两个随机变量  $x$  和  $y$  是不相关的, 那么:  $D(x+y) = D(x) + D(y)$ 。
- 因为方差是以平方来衡量的, 这和勾股定理很像:  $c^2 = a^2 + b^2$  其中  $a, b$  是直角三角形的直角边,  $c$  是直角三角形的斜边。
- 接下来, 我们尝试从几何学的角度来解释不相关随机变量之间的关系。随机变量能被看成向量空间中的向量, 我们能定义内积来获取随机变量的几何性质, 如果我们定义:  $\langle x, y \rangle := Cov(x, y)$ 。
- 随机变量的长度是:  $\|x\| = \sqrt{Cov(x, x)} = \sqrt{D(x)} = \sigma(x)$ 。我们知道如果两个向量  $x \perp y \Leftrightarrow \langle x, y \rangle = 0$ 。对于我们这种情况意味着  $x$  与  $y$  正交当且仅当  $Cov(x, y) = 0$  或者说  $x$  与  $y$  不相关。

**注意** 虽然试图使用欧几里得距离 (由上面的内积的定义构造) 来比较概率分布, 但不幸的是, 这并不是获得分布之间距离的最佳方法。由于概率质量 (或密度) 需要加起来为 1 的事实, 分布存在于一个流形中。对这种概率分布空间的研究称为信息几何。

1 21.1 期望

2 21.2 方差

3 21.3 矩和协方差矩阵

### 21.3.1 矩的定义

#### 定义 6

设  $X$  和  $Y$  是随机变量,

若  $E(X^k), k = 1, 2, \dots$  存在, 称它为  $X$  的  $k$  阶原点矩, 简称  $k$  阶矩.

若  $E\{[X - E(X)]^k\}, k = 2, 3, \dots$  存在, 称它为  $X$  的  $k$  阶中心矩.

若  $E\{X^k Y^l\}, k, l = 1, 2, 3, \dots$  存在, 称它为  $X$  和  $Y$  的  $k + l$  阶混合矩.

若  $E\{[X - E(X)]^k [Y - E(Y)]^l\}, k, l = 1, 2, 3, \dots$  存在, 称它为  $X$  和  $Y$  的  $k + l$  阶混合中心矩.

显然  $X$  的数学期望  $E(X)$  是  $X$  的一阶原点矩, 方差  $D(X)$  是  $X$  的二阶中心矩, 协方差  $Cov(X, Y)$  是  $X$  和  $Y$  的二阶混合中心矩.

## 协方差矩阵的定义

下面介绍  $n$  维随机变量的协方差矩阵. 先从二维随机变量讲起.

### 二维随机变量的协方差矩阵

二维随机变量  $(X_1, X_2)$  有 4 个二阶中心矩 (假设它们都存在), 分别记为

$$c_{11} = E\{[X_1 - E(X_1)]^2\}$$

$$c_{12} = E\{[X_1 - E(X_1)][X_2 - E(X_2)]\}$$

$$c_{21} = E\{[X_2 - E(X_2)][X_1 - E(X_1)]\}$$

$$c_{22} = E\{[X_2 - E(X_2)]^2\}$$

将它们排成矩阵的形式

$$\begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}$$

这个矩阵称为二维随机变量  $(X, Y)$  的协方差矩阵.

## 定义 7

设  $n$  维随机变量  $(X_1, X_2, \dots, X_n)$  的二阶混合中心矩

$$c_{ij} = \text{Cov}(X_i, X_j) = E[X_i - E(X_i)][X_j - E(X_j)], i, j = 1, 2, \dots, n$$

都存在, 则称矩阵:

$$\mathbf{C} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{bmatrix}$$

为  $n$  维随机变量  $(X_1, X_2, \dots, X_n)$  的协方差矩阵.

由于  $c_{ij} = c_{ji} (i \neq j; i, j = 1, 2, \dots, n)$ , 因此上述矩阵是一个对阵矩阵.

一般情况下  $n$  维随机变量的分布是不知道的, 或者太过复杂, 以致在数学上不易处理, 因此在实际应用中协方差矩阵就显得重要了.

### 例 10

我们以  $n$  维正态分布为例来介绍  $n$  维随机变量. 在介绍  $n$  维正态分布的概率密度函数之前, 我们先将二维正态分布的概率密度函数改成另外一种形式, 以便将它推广到  $n$  维随机变量的场合中去. 二维正态随机变量  $(X_1, X_2)$  的概率密度函数为

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp \left\{ \frac{-1}{2(1-\rho^2)} \left[ \frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho \frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2} \right] \right\}$$

现将上式中花括号内的式子写成矩阵形式, 为此引入下面的列矩阵

$$\mathbf{X} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

$(X_1, X_2)$  的协方差矩阵为

$$\mathbf{C} = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$



它的行列式  $|C| = \sigma_1^2 \sigma_2^2 (1 - \rho^2)$ ,  $C$  的逆矩阵为

$$C^{-1} = \frac{1}{|C|} \begin{bmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{bmatrix}$$

经计算可知

$$(\mathbf{X} - \boldsymbol{\mu})^T C^{-1} (\mathbf{X} - \boldsymbol{\mu}) = \frac{1}{|C|} (x_1 - \mu_1, x_2 - \mu_2) \begin{bmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{bmatrix} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}$$

$$(\mathbf{X} - \boldsymbol{\mu})^T C^{-1} (\mathbf{X} - \boldsymbol{\mu}) = \frac{1}{1 - \rho^2} \left[ \frac{(x - \mu_1)^2}{\sigma_1^2} - 2\rho \frac{(x - \mu_1)(y - \mu_2)}{\sigma_1\sigma_2} + \frac{(y - \mu_2)^2}{\sigma_2^2} \right]$$

于是  $(X_1, X_2)$  的概率密函可写成

$$f(x_1, x_2) = \frac{1}{(2\pi)^{2/2} (|C|)^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{X} - \boldsymbol{\mu})^T C^{-1} (\mathbf{X} - \boldsymbol{\mu}) \right\}$$

### 推广到 $n$ 维正态随变量

上式容易推广到  $n$  维正态随变量  $(X_1, X_2, \dots, X_n)$  的情况. 引入列矩阵

$$\mathbf{X} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad \boldsymbol{\mu} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} = \begin{pmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_n) \end{pmatrix}$$

$n$  维正态随机变量  $(X_1, X_2, \dots, X_n)$  的概率密度函数定义为:

$$f(x_1, x_2, \dots, x_n) = \frac{1}{(2\pi)^{n/2}(|\mathbf{C}|)^{1/2}} \exp \left\{ -\frac{1}{2}(\mathbf{X} - \boldsymbol{\mu})^T \mathbf{C}^{-1}(\mathbf{X} - \boldsymbol{\mu}) \right\}$$

其中  $\mathbf{C}$  是  $(X_1, X_2, \dots, X_n)$  的协方差矩阵.

## $n$ 维正态随机变量的性质

$n$  维正态随机变量具有以下四条重要的性质:

### 性质 1

$n$  维正态随机变量  $(X_1, X_2, \dots, X_n)$  的每一个分量  $X_i, i = 1, 2, \dots, n$  都是  $n$  维正态随机变量; 反之, 若  $X_1, X_2, \dots, X_n$  都是正态随机变量, 且相互独立, 则  $X_1, X_2, \dots, X_n$  是  $n$  维正态随机变量.

### 性质 2

$n$  维随机变量  $X_1, X_2, \dots, X_n$  服从  $n$  维正态分布的充要条件是  $X_1, X_2, \dots, X_n$  的任意线性组合

$$l_1 X_1 + l_2 X_2 + \dots + l_n X_n$$

服从一维正态分布 (其中  $l_1, l_2, \dots, l_n$  不全为 0)

### 性质 3

若  $X_1, X_2, \dots, X_n$  服从  $n$  维正态分布, 设  $Y_1, Y_2, \dots, Y_k$  是  $X_j (j = 1, 2, \dots, n)$  的线性函数, 则  $(Y_1, Y_2, \dots, Y_k)$  也服从多维正态分布.

这一性质称为正态变量的线性变换不变性.

### 性质 4

设  $(X_1, X_2, \dots, X_n)$  服从  $n$  维正态分布, 则 " $X_1, X_2, \dots, X_n$ " 相互独立与 " $X_1, X_2, \dots, X_n$ " 两两不相关是等价的.

## 协方差矩阵性质

### 性质 5

离散型随机变量的协方差矩阵是半正定矩阵。

### 性质 6

对于二元离散型随机变量  $(x, y)$ ，其协方差矩阵  $V$  等于多个矩阵之和，即：

$$V = \sum_{i,j} p_{ij} V_{i,j} = \sum_{i,j} p_{ij} U U^T = \sum_{i,j} p_{ij} \begin{bmatrix} x - \mu_x \\ y - \mu_y \end{bmatrix} \begin{bmatrix} x - \mu_x & y - \mu_y \end{bmatrix}$$

其中  $p_{ij}$  是二元离散随机变量的概率分布函数， $\mu_x$  和  $\mu_y$  分别是  $x$  和  $y$  的均值，且：

$$U = \begin{bmatrix} x - \mu_x \\ y - \mu_y \end{bmatrix}$$

上述性质对于  $n$  为随机变量也成立。

## 本讲小结

随机变量的数字特征：期望、方差、协方差、相关系数、矩等

### 期望

- 定义和性质
- 求法：懒惰统计学家法则
- 常见的离散和连续随机变量的期望
- 条件期望

### 方差

- 定义和性质
- 常见离散和连续型随机变量的性质
- 协方差、相关系数和内积
- 矩和协方差矩阵

期望和方差在机器学习中具有重要应用：衡量分类准确率，统计决策规则，过拟合与偏差-方差分解等！