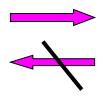
§ 4.3 协方差和相关系数

前面我们介绍了随机变量的数学期望和方差,对于多维随机变量,反映分量之间关系的数字特征中,最重要的,就是现在要讨论的

协方差Covariance和相关系数Correlation

问题 对于二维随机变量(X,Y):

已知联合分布」。 边缘分布



对二维随机变量,除每个随机变量各自 的概率特性外,相互之间可能还有某种联系 问题是用一个怎样的数去反映这种联系.

数
$$E([X-E(X)][Y-E(Y)])$$

反映了随机变量 X, Y 之间的某种关系

1. 协方差和相关系数的概念





定义
$$E([X-E(X)][Y-E(Y)])$$

称为 X, Y 的<mark>协方差Covariance</mark>.

$$cov(X,Y) = E([X - E(X)][Y - E(Y)])$$

若D(X) > 0, D(Y) > 0, 称



$$\frac{\text{cov}(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}$$

为X,Y的相关系数Correlation,记为

$$\rho_{XY} = \frac{\text{cov}(X, Y)}{\sqrt{D(X)}\sqrt{D(Y)}}$$

若 $\rho_{XY} = 0$, 称 X, Y 不相关.

更准确地说叫作"线性无关"、"线性不相关",这仅仅表明X 与Y 两随机变量之间没有线性相关性,并非表示它们之间一定没有任何内在的(非线性)函数关系。

计算协方差的一个简单公式



由协方差的定义及期望的性质,可得

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

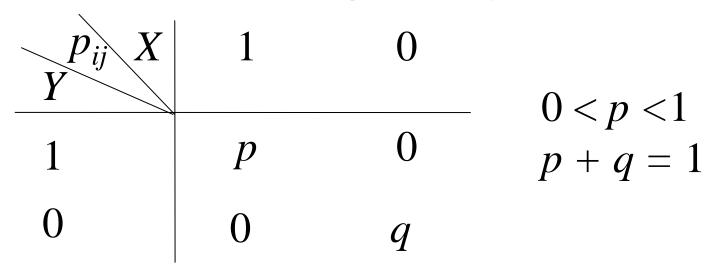
$$=E(XY)-E(X)E(Y)-E(X)E(Y)+E(X)E(Y)$$

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

$$\mathbb{E} \Gamma \quad Cov(X,Y) = E(XY) - E(X)E(Y)$$

可见,若X与Y独立,Cov(X,Y)=0.

例 已知X,Y的联合分布为



求 $cov(X,Y), \rho_{XY}$

解

X	1 0	Y	1 0	XY	1	0
\overline{P}	p q	\overline{P}	p q	\overline{P}	p	\overline{q}

$$E(X) = p, E(Y) = p,$$

$$D(X) = pq, D(Y) = pq,$$

$$E(XY) = p,$$

$$cov(X,Y) = pq, \ \rho_{XY} = 1$$

例 设(X,Y) ~ $N(\mu_1,\sigma_1^2;\mu_2,\sigma_2^2;\rho)$, 求 ρ_{XY}

解
$$cov(X,Y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x - \mu_1)(y - \mu_2) f(x,y) dx dy$$

$$\frac{\frac{x-\mu_{1}}{\sigma_{1}} = s}{\frac{y-\mu_{2}}{\sigma_{2}} = t} \frac{\sigma_{1}\sigma_{2}}{2\pi\sqrt{1-\rho^{2}}} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} ste^{-\frac{1}{2(1-\rho^{2})}(s-\rho t)^{2} - \frac{1}{2}t^{2}} dsdt$$

$$\stackrel{\diamondsuit}{=} \frac{s - \rho t = u}{2\pi \sqrt{1 - \rho^2}} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} t(\rho t + u) e^{-\frac{u^2}{2(1 - \rho^2)} - \frac{1}{2}t^2} du dt$$

$$= \sigma_1 \sigma_2 \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (\rho t^2 + tu) f(u) f(t) du dt$$

 $u \sim N(0, \sqrt{1-\rho^2}), t \sim N(0, 1)$

$$= \sigma_1 \sigma_2(\rho \int_{-\infty}^{+\infty} t^2 f(t) dt \int_{-\infty}^{+\infty} f(u) du + \int_{-\infty}^{+\infty} t f(t) dt \int_{-\infty}^{+\infty} u f(u) du) = \sigma_1 \sigma_2(\rho \mathsf{D(t)} + \mathsf{E(t)E(u)})$$

$$= \boldsymbol{\sigma}_1 \boldsymbol{\sigma}_2 \boldsymbol{\rho}$$

$$\rho_{XY} = \frac{\text{cov}(X, Y)}{\sqrt{D(X)}\sqrt{D(Y)}}$$

$$\rho_{XY} = \rho$$

若 $(X,Y) \sim N(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, \rho),$

则X,Y相互独立 $\longrightarrow X,Y$ 不相关

3-4讲过二维随机变量独立的充要条件 $f(x,y)=f_X(x)f_Y(y)$

2. 协方差和相关系数的性质



- $(1) \quad \operatorname{cov}(X,Y) = \operatorname{cov}(Y,X)$
- (2) cov(aX,bY) = ab cov(X,Y)
- (3) cov(X + Y, Z) = cov(X, Z) + cov(Y, Z)
- $(4) \quad \operatorname{cov}(X, X) = D(X)$
- (5) $D(X \pm Y) = D(X) + D(Y) \pm 2 \operatorname{cov}(X, Y)$

(6)
$$-1 \le \rho_{xy} \le 1$$



(7) 如果 Y = aX + b 存在常数 $a, b(a \neq 0)$,

如果a>0,则 $\rho(X,Y)=1$

如果a<0,则 $\rho(X,Y) = -1$

即X和Y以概率1线性相关。

X,Y相互独立 $\stackrel{}{>}$ X,Y不相关 $\stackrel{\mathbb{Z}}{=}$



$$\rho_{XY} = 0 \longrightarrow X, Y 不相关$$

$$\longleftrightarrow cov(X,Y) = 0$$

若(X,Y)服从二维正态分布, X, Y相互独立 $\longrightarrow X, Y$ 不相关

X与Y之间没有线性关系并不表示没有关系!

例
$$X \sim U[-1,1], \quad Y = X^2$$

$$E(X) = E(X^3) = 0$$

$$cov(X,Y) = E(XY) - E(X)E(Y) = E(X^3) = 0$$

$$\rho_{XY} = \frac{cov(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}} = 0$$

X, Y不相关

$$Y = X^2$$

显然是不相互独立的

例 设
$$X$$
, Y 是两个随机变量,已知 $DX = 1$, $DY = 4$, $cov(X, Y)$ $= 1$, $\xi = X - 2Y$, $\eta = 2X - Y$ 试求: $\rho_{\xi,\eta}$

解:
$$D\xi = D(X - 2Y)$$

 $= DX + D(2Y) - 2\operatorname{cov}(X, 2Y)$
 $= DX + 4DY - 4\operatorname{cov}(X, Y)$
 $= 1 + 4 \times 4 - 4 \times 1 = 13$
 $D\eta = D(2X - Y)$
 $= 4DX + DY - 4\operatorname{cov}(X, Y)$
 $= 4 \times 1 + 4 - 4 \times 1 = 4$

$$cov(\xi, \eta) = cov(X - 2Y, 2X - Y)$$

$$= cov(X - 2Y, 2X) - cov(X - 2Y, Y)$$

$$= 2cov(X, X) - 4cov(Y, X)$$

$$- cov(X, Y) + 2cov(Y, Y)$$

$$= 2DX - 5cov(X, Y) + 2DY$$

$$= 2X - 5x + 2x + 4 = 5$$

$$\rho_{\xi, \eta} = \frac{cov(\xi, \eta)}{\sqrt{D\xi}\sqrt{D\eta}} = \frac{5}{\sqrt{13}\sqrt{4}} = \frac{5\sqrt{13}}{26}$$

相关系数实际意义

在数据挖掘中,相关系数可以分析**冗余**(相关、包含)问题。比如一个属性如果可能由其它属性包含,那么该属性就是冗余的。

- $\rho > 0$,X和Y是正相关,即X随着Y的值增加而增加。 ρ 越大,X与Y的相关性越强(即每个属性蕴含另一个的可能性越大)。因此如果 ρ 很大,表示X(或Y)可以作为冗余而被去掉。
- *ρ* =0, X和Y不相关
- $\rho < 0$,X和Y是负相关,即一个值随着另一个的减少而增加。意味着一个属性阻止另一个属性的出现。
- 注意: 相关并不意味因果关系。也就是说X和Y是相关,并不意味着X导致Y或反之。

3. 矩和协方差矩阵

$$E(X^k)$$
 — X 的 k 阶原点矩

$$E((X-E(X))^k)$$
— X 的 k 阶中心矩

$$E(X^kY^l)$$
 — X,Y 的 $k+l$ 阶混合原点矩

$$E((X-E(X))^{k}(Y-E(Y))^{l})$$

-X,Y的 k+l 阶混合中心矩

n维随机变量 $(X_1,X_2,...,X_n)$ 的协方差矩阵

若
$$c_{ij} = Cov(X_i, X_j)$$
 $i, j=1,2,...,n$

都存在, 称矩阵

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

为 $(X_1, X_2, ..., X_n)$ 的协方差矩阵

Summary

- 1 协方差和相关系数定义
 - 2 协方差的简化公式
 - 3 协方差性质
 - 4矩和协方差矩阵
- 5 协方差和相关系数的计算