数据科学基础

Foundations of Data Science

5.3 多维概率分布

陈振宇
南京大学智能软件工程实验室
www.iselab.cn

二维随机变量

【二维随机变量】设实验E的样本空间为$S=\{\mathrm{e}\}, X=X(\mathrm{e})$和$Y=Y(\mathrm{e})$是定义在$S$上的随机变量,由它们构成的一个变量$(X,Y)$叫做二维随机变量。

【联合分布】 设(X,Y)是二维随机变量, x,y是任意实数, 称二元函数

$$F(x,y)=P(X\le x, Y\le y)$$

为二维随机变量(X,Y)的联合分布函数.

请注意:

$$P\{x_1<X\le x_2, y_1<Y\le y_{2\}}$$
$$=F(x_2,y_2)-F(x_2,y_1)-F(x_1,y_2)+F(x_1,y_1)$$

二维离散随机变量

【二维离散随机变量】 若二维随机变量(X,Y)的可能值(x_i,y_i)只有有限对或可列无限对,则称(X,Y)是二维离散随机变量.

(X,Y)是离散型二维随机变量 $\Leftrightarrow$ X和Y都是离散型随机变量.

【联合分布】 称$P\{X=x_i,Y=y_j=p_{ij},i,j=1,2,\cdots$为(X,Y)的联合分布律. 且满足$\sum_i\sum_j p_{ij}=1$

【分布函数】二维随机变量(X,Y)的分布函数定义为$F(x,y)=\sum_{x_i\le x}\sum_{y_j\le y}p_{ij}$

二维连续随机变量

【二维连续型变量】对任意(X,Y),如果存在非负函数$f(x,y)$,使对任意实数对(x,y)有$F(x,y)=\int_{-\infty}^{y}\int_{-\infty}^{x}f(u,v)\,dudv$.则称$(X,Y)$为二维连续型随机变量,其中$f(x,y)$称为$(X,Y)$的联合概率密度函数.

二维随机变量具有以下性质:

1. F在(x,y)点连续, 则$\frac{\partial^2 F(x,y)}{\partial x\partial y}=f(x,y)$
2. 在任意平面G上的概率为

$$P\{(X,Y)\in G\}=\iint_G f(x,y)dxdy$$

二维随机变量计算

$$F(x,y) = 1 - e^{-0.01x} - e^{-0.01y} + e^{-0.01(x+y)}, x \geq 0, y \geq 0$$

其它$F(x,y) = 0$。

求

1. $P(X < 120, Y < 120)$

2. $P(X > 120, Y > 120)$

3. $P(Y \leq X)$

二维随机变量计算

$$(1)P(X<120,Y<120)=F(120,120)=1-e^{-1.2}-e^{-1.2}+e^{-2.4}=\left(1-e^{-1.2}\right)^2$$

$$(2)P(X>120,Y>120)=F(+\infty,+\infty)-F(120,+\infty)-F(+\infty,120)+$$

$$F(120,120)=1-(1-e^{-1.2})-(1-e^{-1.2})+\left(1-e^{-1.2}\right)^2=e^{-2.4}$$

$$(3)f(x,y)=\frac{\partial^2 F}{\partial x\partial y}=(0.01)^2\,e^{-0.01(x+y)}$$

$$P(Y\le X)=\int\int_{y\le x}f(x,y)dxdy=\int_{-\infty}^{+\infty}dx\int_{-\infty}^{x}f(x,y)dxdy$$

$$=\int_0^{+\infty}dx\int_0^{x}(0.01)^2e^{-0.01(x+y)}\,dy=\int_0^{+\infty}\left(-0.01e^{-0.02x}+0.01e^{-0.01x}\right)dx$$

$$=\left(0.5e^{-0.02x}-e^{-0.01x}\right)\Big|_0^{+\infty}=(0-0.5)-(0-1)=0.5$$

二维随机变量的边缘分布

【边缘分布】设(X,Y)为二维随机变量,则称随机变量X的概率分布为(X,Y)关于X的边缘分布;随机变量Y的概率分布为(X,Y)关于Y的边缘分布。

对于离散二维随机变量(X,Y), 有

- $F_X(x) = F(x,\infty) = \sum_{x_i \le x} \sum_{j=1}^{\infty} p_{ij}$, $F_Y(y) = F(\infty, y) = \sum_{i=1}^{\infty} \sum_{y_i \le y} p_{ij}$
- $p_{i\cdot} = P\{X = x_i\} = \sum_{j=1}^{\infty} p_{ij}$, $p_{\cdot j} = P\{Y = y_j\} = \sum_{i=1}^{\infty} p_{ij}$

对于连续二维随机变量(X,Y), 有

- $F_X(x) = \int_{-\infty}^{x} \int_{-\infty}^{\infty} f(x,y)dxdy$, $F_Y(y) = \int_{-\infty}^{\infty} \int_{-\infty}^{y} f(x,y)dxdy$
- $f_X(x) = \int_{-\infty}^{\infty} f(x,y)dxdy$, $f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx$

离散随机变量的条件概率

设(X,Y)是二维离散型随机变量，其分布律为$P\{X=x_i,Y=y_j\}=p_{ij}$,其边缘概率分别为$p_{i\cdot},p_{\cdot j}$. 则条件概率定义为

$$P\{X=x_i \mid Y=y_j\}=\frac{P\{X=x_i,Y=y_j\}}{P\{Y=y_j\}}=\frac{p_{ij}}{p_{\cdot j}}$$

$$P\{Y=y_j \mid \mathrm{X}=x_j\}=\frac{P\{X=x_i,Y=y_j\}}{P\{X=x_i\}}=\frac{p_{ij}}{p_{i\cdot}}$$

离散随机变量概率计算示例

【例】对一群人进行吸烟X和身体健康Y调查. $X=1$健康,$X=0$一般, $X=-1$不健康; $Y=0$不吸烟, $Y=1$每天不多于15支,$Y=2$每天多于15支. (X,Y)的联合分布律如下:

$X\|Y$	0	1	2
1	0.35	0.04	0.025
0	0.025	0.15	0.04
-1	0.02	0.1	0.25

1. 试求X,Y的边缘分布律;
2. 试求$P(X=-1|Y=2)$的值.

离散随机变量概率计算示例

解: (1)X,Y的边缘分布律分别为:

X	1	0	-1	Y	0	1	2
$p_{i\cdot}$	0.415	0.215	0.37	$p_{\cdot j}$	0.395	0.290	0.315

(2) $P(X=-1|Y=2)=\frac{0.25}{0.315}=0.794$

连续随机变量条件概率

设(X,Y)是二维连续型随机变量，其概率密度为$f(x,y)$,其边缘概率密度分别为$f_X(x), f_Y(y)$. 则条件概率密度定义为

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

其条件概率分布定义为

$$F_{(X|Y)}(x|y) = P\{X \le x|Y = y\} = \int_{-\infty}^{x} \frac{f(x,y)}{f_Y(y)}dx$$

$$F_{(Y|X)}(y|x) = P\{Y \le y|X = x\} = \int_{-\infty}^{y} \frac{f(x,y)}{f_X(x)}dy$$

连续随机变量概率计算示例

【例】 设X在$(0,1)$上随机均匀取值. 对于给定$X = x (0 < x < 1)$时, Y在区间$(x, 1)$上均匀分布, 求Y的概率密度$f_Y(y)$.

连续随机变量概率计算示例

解: 对于任意$x(0<x<1)$, 在$X=x$的条件下,Y的条件概率密度为

$$f_{Y|X}(y|x)=\frac{1}{1-x}, x<y<1;\ f_{Y|X}(y|x)=0, \text{其它}$$

故(X,Y)的概率密度为

$$f(x,y)=f_X(x)f_{(Y|X)}(y|x)=1\cdot\frac{1}{1-x}$$

所以Y的边缘概率密度为

$$f_Y(y)=\int_{-\infty}^{\infty}f(x,y)dx=-\ln(1-y), 0<y<1$$

二维随机变量独立性

【独立】设$F(x,y)$及$F_X(x),F_Y(y)$分别是二维随机变量(X,Y)的分布函数及边缘分布函数.若对所有x,y有

$$P\{X \le x, Y \le y\} = P\{X \le x\}P\{Y \le y\}$$

则称随机变量X和Y是相互独立的. 等价命题有

$$F(x,y) = F_X(x)F_Y(y)$$

$$f(x,y) = f_X(x)f_Y(y)$$

$$P\{X = x_i, Y = y_j\} = P\{X = x_i\}P\{Y = y_j\}, i,j = 1,2,\cdots$$

对于连续, 只要条件几乎出出成立即可.

独立性示例

【例】X,Y服从同一分布, 其分布律为

X,Y	-1	0	1
p	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

已知$P(X = Y) = 0$, 判断X,Y是否相关, 是否独立.

解: 求X,Y的联合概率和边缘概率

$X\|Y$	-1	0	1	$p_{\cdot j}$
-1	0	$1/4$	0	$1/4$
0	$1/4$	0	$1/4$	$1/2$
1	0	$1/4$	0	$1/4$
$p_{i\cdot}$	$1/4$	$1/2$	$1/4$	

独立性示例

$$E(X)=E(Y)=-1*\frac{1}{4}+0*\frac{1}{2}+1*\frac{1}{4}=0$$

$$E(XY)=(-1)*(0)*\frac{1}{4}+0*(-1)*\frac{1}{4}+1*0*\frac{1}{4}+1*0*\frac{1}{4}=0$$

所以$Cov(X,Y)=0, X,Y$不相关.

$$p_{-1,-1}=0\neq\ p_{-1\cdot}p_{\cdot-1}=\frac{1}{4}*\frac{1}{4}$$

所以X,Y不独立.

独立性示例

【例】(X,Y)的概率密度如下, 问X,Y是否独立?

$$f(x,y)=6e^{-(2x+3y)}, x>0, y>0$$

解: X,Y的边缘概率密度分别为:

$$f_X(x)=\int_{-\infty}^{\infty}6e^{-(2x+3y)}dy=2e^{-2x}, x>0$$

$$f_Y(y)=\int_{-\infty}^{\infty}6e^{-(2x+3y)}dx=3e^{-3y}, y>0$$

$$f(x,y)=f_X(x)f_Y(y)$$

所以X,Y相互独立.

概率计算

【例】设X,Y相互独立, 已知(X,Y)的联合分布律如表求未知概率值.

p_{ij}	0	1	2	$p_{i\cdot}$
1	0.01	0.2		
2	0.03			
$p_{\cdot j}$				

概率计算

【例子】　设X,Y是两个相互独立的随机变量,X在$(0,1)$上服从均匀分布,Y的概率密度为:

$$f_Y(y)=\frac{1}{2}e^{-\frac{y}{2}},y>0$$

1. 求$f(x,y)$

2. 设有a的二次方程$a^2+2aX+Y=0$,求此方程有实根的概率.

概率计算

解：

$$(1)\ f(x,y)=f_X(x)f_Y(y)=\frac{1}{2}e^{-\frac{y}{2}},0<x<1,y>0$$

$$(2)\ P\left(X^2\geq Y\right)=\iint_{X^2\geq Y}\frac{1}{2}e^{-\frac{y}{2}}\,dxdy$$

$$=\int_0^1 dx\int_0^{x^2}\frac{1}{2}e^{-\frac{y}{2}}\,dy=\int_0^1\left(1-e^{-\frac{x^2}{2}}\right)dx$$

$$=1-\int_0^1 e^{-\frac{x^2}{2}}\,dx=1-\sqrt{2\pi}\left(\Phi(1)-\Phi(0)\right)=0.1448$$

二维随机变量的矩

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

证明:

$$E(X+Y)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x+y)f(x,y)dxdy$$

$$=\int_{-\infty}^{+\infty}xf(x,y)dxdy+\int_{-\infty}^{+\infty}yf(x,y)dxdy=E(X)+E(Y)$$

二维随机变量的矩

X和Y独立

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

证明:

$$E(XY) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(xy)f(x,y)dxdy = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(xy)f_X(x)f_Y(y)dxdy$$

$$= \int_{-\infty}^{+\infty}xf_X(x)dx\int_{-\infty}^{+\infty}yf_Y(y)dy = E(X)E(Y)$$

二维随机变量的矩

X和Y独立

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

证明：

$$D(X+Y)=E\left(\left((X+Y)-E(X+Y)\right)^2\right)=E\left(\left(\left(X-E(X)\right)+\left(Y-E(Y)\right)\right)^2\right)$$

$$=E\left(\left(X-E(X)\right)^2\right)+E\left(\left(Y-E(Y)\right)^2\right)+2E\left(\left(X-E(X)\right)\left(Y-E(Y)\right)\right)$$

X和Y独立，则$X-E(X)$和$Y-E(Y)$独立，所以$E\left(\left(X-E(X)\right)\left(Y-E(Y)\right)\right)=$
$E(X-EX)E(Y-EY)=0, D(X+Y)=D(X)+D(Y)$

协方差

【协方差】 $E\{[(X-EX)][(Y-EY)]\}$称为随机变量X和Y的协方差$Cov(X,Y)$,

即$Cov(X,Y)=E\{[(X-EX)][(Y-EY)]\}$.

将定义式展开, 易得: $Cov(X,Y)=E(XY)-E(X)E(Y)$.

协方差的性质:

1. $Cov(X,Y)=Cov(Y,X)$

2. $Cov(aX,bY)=Cov(X,Y),a,b$为常数

3. $Cov(X_1+X_2,Y)=Cov(X_1,Y)+Cov(X_2,Y)$

相关系数

【相关系数】 设随机变量X,Y的数学期望、方差都存在, 称

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

为随机变量X,Y的相关系数.

相关系数的两条重要性质:

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

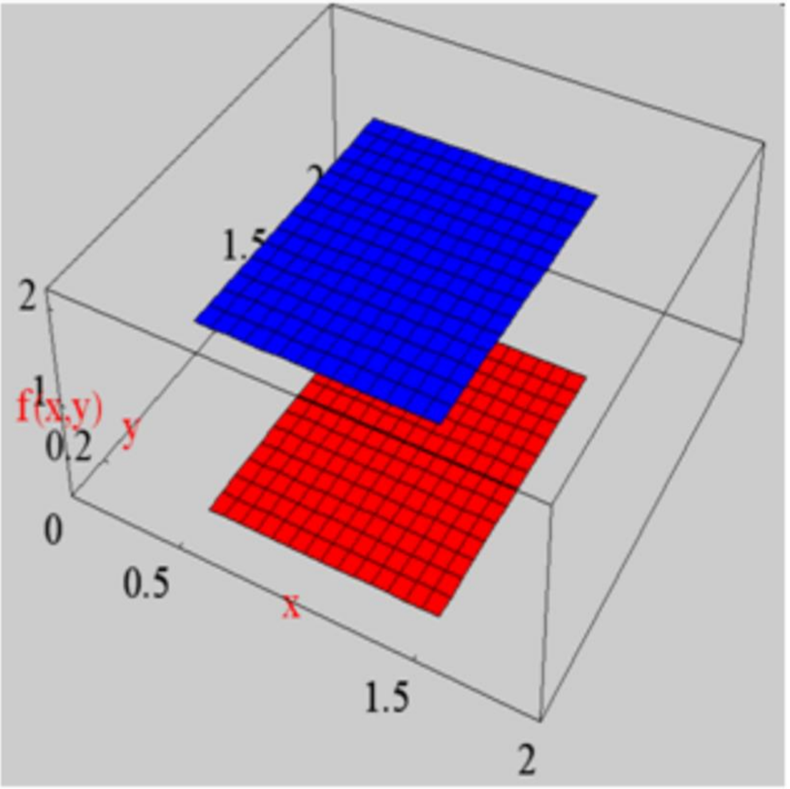
2. $|\rho_{XY}|=1$的充要条件为两个随机变量X和Y有线性关系

二维均匀分布

【均匀分布】设G是平面上的有界区域,其面积为A.
若二维随机变量(X,Y)具有概率密度

$$f(x,y)=\frac{1}{A},(x,y)\in G$$

其它$f(x,y)=0$, 称(X,Y)在G上的二维均匀分布。



二维均匀分布

【例】二维随机变量(X,Y)是在$x^2+y^2\le 1$上的均匀分布,即

$$f(x,y)=\frac{1}{\pi},x^2+y^2\le 1$$

其它$f(x,y)=0$, 求$f_{X|Y}(x,y)$。

二维均匀分布

先求边缘密度函数。因$x^2+y^2\le 1$时，$f(x,y)=\frac{1}{\pi}$，所以

$$f_Y(y)=\int_{-\infty}^{+\infty}\frac{1}{\pi}dx=\int_{x^2<1-y^2}\frac{1}{\pi}dx=\frac{2}{\pi}\sqrt{1-y^2}$$

$$f_{X|Y}(x|y)=\frac{f(x,y)}{f_Y(y)}=\frac{\frac{1}{\pi}}{\frac{2}{\pi}\sqrt{1-y^2}}$$

$$f_{X|Y}(x|y)=\frac{1}{2\sqrt{1-y^2}},-\sqrt{1-y^2}\le x\le\sqrt{1-y^2}$$

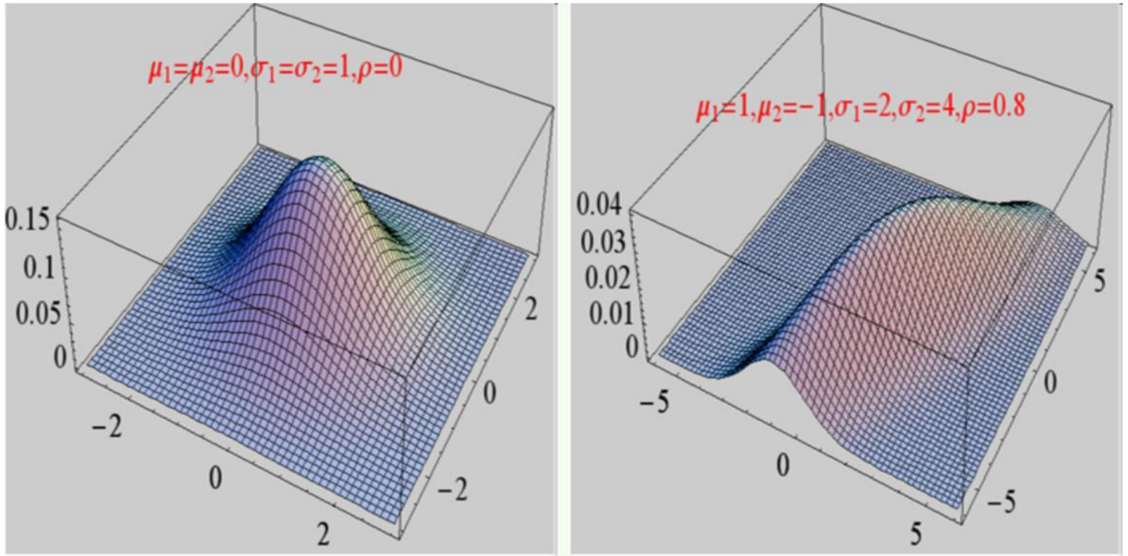
二维正态分布

【二维正态分布】如果随机变量(X,Y)的概率密度为:

$$f(x,y)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}e^{-\frac{1}{2(1-\rho)^2}[\frac{(x-\mu_1)^2}{\sigma_1^2}-2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}+\frac{(y-y_2)^2}{\sigma_2^2}]}$$

$-\infty<x,y<\infty$.则称(X,Y)服从参数为$\mu_1,\mu_2,\sigma_1,\sigma_2,\rho$的二维正态分布,记为

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



二维正态分布

求X,Y的边缘密度函数.

$$f_X(x)=\int_{-\infty}^{\infty}f(x,y)dy=\int_{-\infty}^{\infty}\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}e^{-\frac{1}{2(1-\rho)^2}\left[\frac{(x-\mu_1)^2}{\sigma^2}-2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}+\frac{(y-y_2)^2}{\sigma_2^2}\right]}dy$$

$$=\int_{-\infty}^{\infty}\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}e^{-\frac{1}{2(1-\rho)^2}\left[\frac{y-\mu_2}{\sigma_2}-\rho\frac{x-\mu_1}{\sigma_1}\right]^2}dy$$

$$=\frac{1}{\sqrt{2\pi}\sigma_1}e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1-\rho^2}}e^{-frac\left[y-\left(\mu_2+\rho\frac{\sigma_2}{\sigma_1}(x-\mu_1)\right)\right]^2 2\sigma_2^2(1-\rho^2)}dy$$

$$=\frac{1}{\sqrt{2\pi}\sigma_1}e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}$$

同理可得

$$f_Y(y)=\frac{1}{\sqrt{2\pi}\sigma_1}e^{-\frac{(y-\mu_1)^2}{2\sigma_1^2}}$$