第七章 概率基础

第 20 讲 随机变量

黄定江

DaSE @ ECNU

djhuang@dase.ecnu.edu.cn

1. 20.1 随机变量
2. 20.2 离散型随机变量及其分布
3. 20.3 连续型随机变量及其分布
4. 20.4 多维随机变量及其分布
5. 20.5 边缘分布、条件分布和分布的混合
6. 20.6 随机变量的变换

1. 20.1 随机变量
2. 20.2 离散型随机变量及其分布
3. 20.3 连续型随机变量及其分布
4. 20.4 多维随机变量及其分布
5. 20.5 边缘分布、条件分布和分布的混合
6. 20.6 随机变量的变换

20.1.1 随机变量

统计学与数据挖掘都跟数据有关. 怎么将样本空间与事件同数据联系起来呢？
这条联系的纽带就是随机变量.

定义 1

随机变量即映射

$$X : \Omega \to R$$

该映射对每一个输出 ω 赋予实值 $X(\omega)$.

随机变量举例

- 给定一个查询 Q 和文档集 D，从文档集中随机抽取一篇文档 d_i，查看 d_i 是否与查询相关. 该试验结果的样本空间是 $\{$不相关, 相关$\}$. 构造映射 $X(\omega)$:

$$X(\omega) = \begin{cases} 0, & \omega = \text{不相关} \\ 1, & \omega = \text{相关} \end{cases}$$

- 用随机变量 X 表示股票 A 中午 12 点的股价

$$X(\omega) = x,\ \omega = \text{股票中午 12 点的股价是}x\text{元}$$

随机变量举例

- 从某部电影的 1000 个影评中抽取 5 个影评，随机变量 X 表示正类影评 (对电影持肯定态度) 的个数.

$$X(\omega) = \begin{cases} 0, & \omega = \text{没有正类影评} \\ 1 & \omega = 1 \text{ 个正类影评} \\ 2, & \omega = 2 \text{ 个正类影评} \\ 3, & \omega = 3 \text{ 个正类影评} \\ 4, & \omega = 4 \text{ 个正类影评} \\ 5, & \omega = 5 \text{ 个正类影评} \end{cases}$$

20.1.2 累积分布函数

给定随机变量 X，定义它的累积分布函数（分布函数）如下:

定义 2

累积分布函数表示函数 $F_X : R \to [0, 1]$，其定义为:

$$F_X(x) = P(X \leq x)$$

累积分布函数简记为 CDF(Cumulative Distribution Function)，它包含了随机变量的所有信息，有时用 F 代替 F_X 来表示 CDF。

累积分布函数举例

例 1

从包含 500 个正类与 500 个负类的影视评论集中有放回的抽取 2 次，随机变量 X 表示正类影评的个数，则累积分布函数 *CDF:*

$$F_X(x) = \begin{cases} 0, & x < 0 \\ \frac{1}{4}, & 0 \le x < 1 \\ \frac{3}{4}, & 1 \le x < 2 \\ 1, & x \ge 2 \end{cases}$$

注: 使用 1000 个影评中正类影评的频率 0.5 来近似估计抽取到正类影评的概率.

累积分布函数性质

定理 1

令 X 的 *CDF* 为 F，Y 的 *CDF* 为 G，如果对所有 x 有 $F(x) = G(x)$，则对所有 A 都有 $P(X \in A) = P(Y \in A)$

定理 2

从实直线映射到 $[0, 1]$ 的函数 F 是某个概率 P 的 *CDF* 当且仅当 F 满足以下三个条件：

- F 是非降的：若 $x_1 < x_2$，则 $F(x_1) \leq F(x_2)$
- F 是规范的：

$$\lim_{x \to -\infty} F(x) = 0 \quad \text{且} \quad \lim_{x \to \infty} F(x) = 1$$

- F 是右连续的：$F_x = F(x^+)$ 对所有 x 都成立，其中

$$F(x^+) = \lim_{y \to x, y > x} F(y)$$

1 20.1 随机变量

2 20.2 离散型随机变量及其分布

3 20.3 连续型随机变量及其分布

4 20.4 多维随机变量及其分布

5 20.5 边缘分布、条件分布和分布的混合

6 20.6 随机变量的变换

离散型随机变量和概率密度函数

定义 3

如果 X 取可数个值 $\{x_1, x_2, \cdots\}$, 则 X 是离散型随机变量，定义 X 的概率函数或概率密度函数为 $f_X(x) = P(X = x)$

概率密度函数 f_X 满足：对于 $x \in R$，有

$$f_X(x) \geq 0 \quad \text{并且} \sum_i f_X(x_i) = 1.$$

随机变量 X 的累积分布函数 $F_X(x)$ 和 f_X 的关系如下:

$$F_X(x) = P(X \leq x) = \sum_{x_i \leq x} f_X(x_i).$$

有时用 f 代替 f_X.

一维离散型概率密度函数举例

例 2

例1中随机变量 X 的概率密度函数是:

$$f_X(x) = \begin{cases} \frac{1}{4}, & x = 0, \\ \frac{1}{2}, & x = 1, \\ \frac{1}{4}, & x = 2, \\ 0, & \text{其他} \end{cases}$$

注: 使用 1000 个影评中正类影评的频率 0.5 来近似估计抽取到正类影评的概率.

常用的离散型随机变量和概率密度函数

符号说明：通常用 $X \sim F$ 表示 X 服从分布 F，读作“x 服从分布 F”，而并不是“x 与 F 近似”。

单点分布 仅在一个点 a 上有概率，记为 $X \sim \delta_a$，即 $P(X = a) = 1$，那么

$$F(x) = \begin{cases} 0, & x < a \\ 1, & x \geq a \end{cases}$$

概率密度函数在 $x = a$ 处 $f(x) = 1$，其他情形下为 0.

离散均匀分布 令 $k > 1$ 为给定的整数，假设 X 具有如下概率密度函数:

$$f(x) = \begin{cases} \frac{1}{k}, & x = 1, 2, \cdots, k. \\ 0, & \text{其他}. \end{cases}$$

则称 X 在 $\{1, 2, \cdots, k\}$ 上服从均匀分布.

伯努利分布 随机变量 X 只取两个值，一般用 $0, 1$ 表示，则概率密度函数为:

$$P(X=x)=\begin{cases} p, & x=1, \\ 1-p, & x=0. \end{cases} \tag{1}$$

其中 $p \in [0,1]$，称 X 服从伯努利分布，记为 $X \sim Bernoulli(p)$，概率密度函数可简写为:

$$f(x)=p^x(1-p)^{1-x}$$

其中 $x \in [0,1]$. 在统计机器学习中的逻辑回归分类模型假设数据服从伯努利分布，进而对数据进行建模.

二项式分布 假设从若干影视评论中有放回的抽取 n 次，令随机变量 X 表示抽取到正类影评的次数. 假设每次取影评都是独立的且取到正类影评的概率是 p. 概率密度函数:

$$f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & x = 0, 1, \cdots, n \\ 0, & \text{其他.} \end{cases}$$

具有上述概率密度函数的随机变量称为伯努利随机变量，记 $X \sim Binomial(n, p)$. 若 $X_1 \sim Binomial(n_1, p)$，$X_2 \sim Binomial(n_2, p)$ 并且独立，则 $X_1 + X_2 \sim Binomial(n_1 + n_2, p)$.

注意：X 是随机变量；x 表示随机变量的一个特定的值；n 和 p 是参数，也即是固定实数。参数 p 通常未知，需要根据数据去估计，这就是统计推断要完成的事情。在多数统计模型中，既有随机变量，又有参数，因此不能把它们混淆了。

几何分布 如果在二项式分布的示例中，不是有放回的取 n 次，而是直到取到正类影评为止，令随机变量 X 表示第一次取得正类影评的次数，则 X 的密度函数为:

$$P(X=k)=p(1-p)^{k-1}, k=1,2,3,\cdots .$$

则 X 服从参数为 $p\in(0,1)$ 的几何分布，记为 $X\sim Geom(p)$. 对于几何分布有:

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

泊松分布 如果

$$f(x) = e^{-\lambda}\frac{\lambda^x}{x!},\ x \geq 0,$$

则随机变量 X 服从参数为 λ 的泊松分布，记为 $X \sim Poisson(\lambda)$，并且有:

$$\sum_{x=0}^{\infty} f(x) = e^{-\lambda}\sum_{x=0}^{\infty}\frac{\lambda^x}{x!} = e^{-\lambda}e^{\lambda} = 1$$

泊松分布常用于罕见事件的计数，如放射性元素的衰变与交通事故. 若 $X_1 \sim Poisson(\lambda_1)$，$X_2 \sim Poisson(\lambda_2)$ 且独立，则 $X_1 + X_2 \sim Poisson(\lambda_1 + \lambda_2)$.

1 20.1 随机变量

2 20.2 离散型随机变量及其分布

3 20.3 连续型随机变量及其分布

4 20.4 多维随机变量及其分布

5 20.5 边缘分布、条件分布和分布的混合

6 20.6 随机变量的变换

20.3.1 连续型随机变量和概率密度函数

定义 4

如果存在某个函数 f_X，对所有 x 有:

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

并且对任意 $a \leq b$ 有:

$$P(a < X < b) = \int_a^b f_X(x)\,dx,$$

则随机变量 X 是连续型随机变量。函数 f_X 称为**概率密度函数**(PDF)，且有 X 的概率累积分布函数 (CDF)

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt,$$

以及 $f_X(x) = F_X'(x)$ 在 F_X 可微的点均成立.

常用的连续型随机变量和概率密度函数

均匀分布 假设 X 的概率密度函数为:

$$f_X(x) = \begin{cases} 1, & 0 \le x \le 1 \\ 0, & \text{其他} \end{cases}$$

显然 $f_X(x) \ge 0$ 且 $\int_{-\infty}^{\infty} f(x) = 1$. 称具有这种概率密度函数的随机变量 X 服从 $(0,1)$ **均匀分布**，记为 $X \sim Uniform(0,1)$，其含义就是从 0 到 1 之间随机选取一点。*CDF* 为:

$$F_X(x) = \begin{cases} 0, & x < 0 \\ x, & 0 \le x \le 1 \\ 1, & x > 1, \end{cases}$$

- 在机器学习中，均匀分布可以用来模拟采样.
- 在深度学习中，均匀分布可以用来初始化神经网络模型的参数.

高斯分布

正态 (高斯) 分布 如果随机变量 X 的概率密度函数是:

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

则 X 服从参数为 μ 和 σ 的正态 (高斯) 分布，记为 $X \sim N(\mu, \sigma^2)$，其中 $\mu \in R$, $\sigma > 0$.

- 参数 μ 称为分布的“中心” (均值),σ 是分布的散步程度 (标准差)(均值和方差将在后面讲述). 如果 $\mu = 0, \sigma = 1$ 则称 X 服从**标准正态分布**.
- 正态分布在概率和统计中扮演重要角色，许多自然现象可以用正态分布来近似，如男性身高. 在深度学习中，正态分布也是常用的参数初始化方法

指数分布、伽马分布

指数分布 如果随机变量 X 的概率密度函数是:

$$f(x) = \frac{1}{\beta} e^{-x/\beta}$$

则 X 服从参数为 β 的指数分布, 记为 $X \sim Exp(\beta)$，其中 $\beta > 0$.

- 指数分布常用于电子元件的寿命和两次罕见事件的等待时间。

伽马分布 如果随机变量 X 的概率密度函数是:

$$f(x) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, x > 0$$

其中伽马函数 $\Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1} e^{-y} dx, \alpha > 0$. 则 X 服从参数为 α, β 的伽马分布, 记为 $X \sim Gamma(\alpha, \beta)$，其中 $\alpha, \beta > 0$.

- 指数分布即为 $Gamma(1, \beta)$ 分布。
- 如果 $X_i \sim Gamma(\alpha_i, \beta)$ 且相互独立，则 $\sum_{i=1}^{n} X_i \sim Gamma(\sum_{i=1}^{n} \alpha_i, \beta)$

贝塔分布

贝塔分布 如果随机变量 X 的概率密度函数是:

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, 0 < x < 1$$

则 X 服从参数为 $\alpha > 0$ 和 $\beta > 0$ 的贝塔分布, 记为 $X \sim Beta(\alpha, \beta)$.

t 分布、柯西分布

t 分布 如果随机变量 X 的概率密度函数是:

$$f(x) = \frac{\Gamma(\frac{v+1}{2})}{\Gamma(\frac{v}{2})} \frac{1}{(1+x^2/v)^{(v+1)/2}}$$

则 X 服从自由度为 v 的 t 分布, 记为 $X \sim t_v$.

- t 分布的概率密度函数图形与正态分布的概率密度函数图形类似，但前者尾部较重。
- 事实上，正态分布相当于 $v = \infty$ 的 t 分布.

柯西分布 如果随机变量 X 的概率密度函数是:

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

则 X 服从柯西分布.

- 柯西分布相当于自由度 $v = 1$ 的 t 分布.

$\mathcal{X}^2$ 分布

$\mathcal{X}^2$ 分布 如果随机变量 X 的概率密度函数是:

$$f(x) = \frac{1}{\Gamma(p/2)2^{p/2}} x^{(p/2)-1} e^{-x/2}, \qquad x > 0$$

则 X 服从自由度为 p 的 $\mathcal{X}^2$ 分布, 记为 $X \sim \mathcal{X}_p^2$。

- 如果 $Z_1, \ldots, Z_p$ 是独立标准正态随机变量，则 $\sum_{i=1}^{p} Z_i^2 \sim \mathcal{X}_p^2$

20.3.2 随机变量累积分布函数的性质

定理 3

令 F 为随机变量 X 的 *CDF*，则

- $P(X = x) = F(x) - F(x^-)$，其中，$F(x^-) = \lim_{y \to x^-} F(y)$
- $P(x < X \leq y) = F(y) - F(x)$
- $P(X > x) = 1 - F(x)$
- 如果 X 是连续的，则

$$\begin{aligned} F(b) - F(a) &= P(a < X < b) = P(a \leq X < b) \\ &= P(a < X \leq b) = P(a \leq X \leq b) \end{aligned}$$

随机变量累积分布函数的逆

定义 5

令 X 为一个随机变量，其 *CDF* 为 F，逆 *CDF* 或者分位数函数定义为

$$F^{-1}(q) = \inf\{x : F(x) > q\}$$

其中，$q \in [0,1]$. 如果 F 严格递增且连续则 $F^{-1}(q)$ 是满足 $F(x) = q$ 的唯一实数 x。

称 $F^{-1}(1/4)$ 为第一分位数，$F^{-1}(1/2)$ 为中位数，$F^{-1}(3/4)$ 为第三分位数。

如果 $F_X(x) = F_Y(x)$ 对所有 x 成立，则两个随机变量 X 和 Y 是同分布的，记为 $X \triangleq Y$。

这并不是表示 X 和 Y 是相等的，它表示所有关于 X 和 Y 的概率陈述是相同的。

1 20.1 随机变量

2 20.2 离散型随机变量及其分布

3 20.3 连续型随机变量及其分布

4 20.4 多维随机变量及其分布

5 20.5 边缘分布、条件分布和分布的混合

6 20.6 随机变量的变换

引言

- 在实际生产与理论研究中，常常会遇到这种情况：需要同时用几个随机变量才能较好地描绘某一试验或现象.
- 买西瓜时，根据 (色泽，敲声，根蒂) 来挑选西瓜. 航天飞船返航时的位置需要用 (经度，维度) 来确定.
- 在数据科学中使用多维随机变量来描述一个数据样本. 例如在金融反欺诈领域要决定是否给一人贷款时，要观察此人的 (收入，年龄，是否结婚，学历) 等等.
- 粗略地，称 n 个随机变量 $x_1, x_2, \cdots, x_n$ 的总体 $X = (x_1, x_2, \cdots, x_n)$ 为 n 元随机变量 (或 n 维随机变量). 由于 2 维随机变量与 n 维随即变量没有什么原则的区别，故着重讨论二维的情形.

20.4.1 二维随机变量

定义 6

设 E 是一个随机试验，它的样本空间是 $S=\{e\}$，设 $X=X(e), Y=Y(e)$ 是定义在 S 上的随机变量，由它们构成的一个向量 (X, Y) 叫做二维随机向量或二维随机变量。

二维随机变量 (X, Y) 的性质不仅与 X 及 Y 有关，而且还依赖于这两个随机变量的相互关系。因此逐个地研究 X 或 Y 的性质是不够的，还需将 (X, Y) 作为一个整体来进行研究。

定义 7

设 (X, Y) 是二维随机变量，对任意实数 x, y，二元函数：

$$F(x, y)=P\{(X \leq x) \cap (Y \leq y)\}=P\{X \leq x, Y \leq y\}$$

称为二维随机变量 (X, Y) 的分布函数，或称为随机变量 X 和 Y 的联合分布函数。

联合分布函数性质

性质 1

- $F(x,y)$ 是变量 x,y 的不减函数，即

  对于任意固定的 y，当 $x_2 > x_1$ 时，$F(x_2,y) \geq F(x_1,y)$;

  对于任意固定的 x，当 $y_2 > y_1$ 时 $F(x,y_2) \geq F(x,y_1)$
- 当变量 x 趋近 $-\infty$ 时，$F(x,y)=0$: $\lim\limits_{x\to-\infty} F(x,y)=0$

  当变量 y 趋近 $-\infty$ 时，$F(x,y)=0$: $\lim\limits_{y\to-\infty} F(x,y)=0$

  当变量 x 趋近 $+\infty$，变量 y 趋近 $+\infty$ 时，$F(x,y)=1$: $\lim\limits_{\substack{x\to+\infty\\ y\to+\infty}} F(x,y)=1$

  当变量 x 趋近 $-\infty$，变量 y 趋近 $-\infty$ 时，$F(x,y)=0$: $\lim\limits_{\substack{x\to-\infty\\ y\to-\infty}} F(x,y)=0$
- F 对 x 是右连续的:$F_{x,y}=F(x^+,y)$ 对所有 x 都成立，$F(x^+,y)=\lim\limits_{t\to x,t>x} F(t,y)$

  F 对 y 是右连续的:$F_{x,y}=F(x,y^+)$ 对所有 y 都成立，$F(x,y^+)=\lim\limits_{t\to y,t>y} F(x,y)$
- 对于任意的四个实数 $x_1<x_2, y_1<y_2$, 有: $F(x_2,y_2)-F(x_1,y_2)-F(x_2,y_1)+F(x_1,y_1)\geq 0$

二维离散型分布函数

定义 8

如果二维随机变量 (X, Y) 只可能取到有限对或可列无穷多对值时，则称 (X, Y) 为二维**离散型**随机变量.

定义 9

设二维离散型随机变量 (X, Y) 所有可能的取值为 $(x_i, y_j), i, j = 1, 2, \cdots$，记 $P(X = x_i, Y = y_j) = p_{i,j}, i, j = 1, 2, \cdots$，由概率的定义有

$$p_{i,j} \geq 0, \quad \sum_{i=1}^{+\infty} \sum_{j=1}^{+\infty} p_{i,j} = 1,$$

我们称 $P(X = x_i, Y = y_j) = p_{ij}, i, j = 1, 2, \cdots$ 为二维离散型随机变量 (X, Y) 的概率密度，或随机变量 X 和 Y 的联合概率密度。

二离散型随机变量举例

例 3

一个袋子里有 4 个好瓜，1 个坏瓜，采用无放回挑选 2 次，设:

$$X=\begin{cases}1, & \text{第一次挑到好瓜}\\ 0, & \text{第一次挑到坏瓜}\end{cases} \qquad Y=\begin{cases}1, & \text{第二次挑到好瓜}\\ 0, & \text{第二次挑到坏瓜}\end{cases}$$

故联合概率分布:

Table 1: X, Y 联合概率分布

$X \backslash P \mid Y$	0	1
0	0	$\frac{1}{5}$
1	$\frac{1}{5}$	$\frac{3}{5}$

二维连续型随机变量

定义 10

对于二维随机变量的分布函数 $F(x,y)$，如果存在非负函数 $f(x,y)$ 使得对于任意 x,y 有

$$F(x,y)=\int_{-\infty}^{x}\int_{-\infty}^{y}f(u,v)\,dudv$$

则称 (X,Y) 是**连续型二维随机变量**，函数 $f(x,y)$ 称为二维随机变量 (X,Y) 的**概率密度**，或称为随机变量 X 和 Y 的**联合概率密度**.

二维连续性随机变量

二维随机变量的概率密度函数具有以下性质:

- $f(x, y) \geq 0$.
- 随机变量在 xOy 平面上的积分为 1
$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y)\, dxdy = F(+\infty, +\infty) = 1$$
- 设 G 是 xOy 平面上的区域，点 (X, Y) 落在 G 内的概率是:
$$P\{(X, Y) \in G\} = \iint_G f(x, y)\, dxdy$$
- 若 $f(x, y)$ 在点 (x, y) 连续，则有
$$\frac{\partial^2 F(x, y)}{\partial x \partial y} = f(x, y)$$

二维连续型随机变量举例：均匀分布

例 4

若二维随机变量 $(X\ Y)$ 的联合密度函数:

$$f(x,y)=\begin{cases}\frac{1}{(b_1-a_1)(b_2-a_1)} & a_1\le x\le b_1, a_1\le y\le b_2\\ 0 & otherelse\end{cases}$$

求 (1) 分布函数 $F(x,y)$.(2) 求 $P(Y\le X)$

解

(1)

$$F(x,y)=\int_{-\infty}^{y}\int_{-\infty}^{x}f(x,y)dxdy=\begin{cases}0 & x<a_1\ or\ y<a_1\\ \frac{(x-a_1)(y-a_1)}{(b_1-a_1)(b_2-a_1)} & a_1\le x<b_1\ and\ a_1\le y<b_2\\ \frac{x-a_1}{b_1-a_1} & a_1\le x<b_1\ and\ b_2\le y\\ \frac{y-a_1}{b_2-a_1} & b_1\le x\ and\ a_1\le y<b_2\\ 1 & b_1\le x\ and\ b_2\le y\end{cases}$$

(2) 设 $G=\{(x,y)|y\le x\}$ 则有

$$P(Y\le X)=\iint_G f(x,y)\,dxdy$$

当 $b_1<b_2$ 时

$$P(Y\le X)=\int_{a_1}^{b_1}dx\int_{a_1}^{x}\frac{1}{(b_1-a_1)(b_2-a_1)}dy=\frac{1}{2}\frac{b_1-a_1}{b_2-a_1}$$

当 $b_1\ge b_2$ 时

$$P(Y\le X)=1-\int_{a_1}^{b_2}dy\int_{a_1}^{y}\frac{1}{(b_1-a_1)(b_2-a_1)}dx=1-\frac{1}{2}\frac{b_2-a_1}{b_1-a_1}$$

20.4.2 n 维随机变量

定义 11

设 E 是一个随机试验，它的样本空间是 $S=\{e\}$，设 $X_1=X_1(e),X_2=X_2(e),\ldots,X_n=X_n(e)$ 是定义在 S 上的随机变量，由它们构成的一个 n 维向量 $(X_1,X_2,\ldots,X_n)$ 叫做 *n* 维随机向量或 *n* 维随机变量。

定义 12

设 $(X_1,X_2,\ldots,X_n)$ 是 *n* 维随机变量，对任意 *n* 个实数 $x_1,x_2,\ldots,x_n$，*n* 元函数：

$$F(x_1,x_2,\ldots,x_n)=P\{X_1\le x_1,X_2\le x_2,\ldots,X_n\le x_n\}$$

称为 *n* 维随机变量 $(X_1,X_2,\ldots,X_n)$ 的分布函数，或称为随机变量 $X_1,X_2,\ldots,X_n$ 的联合分布函数。

1 20.1 随机变量

2 20.2 离散型随机变量及其分布

3 20.3 连续型随机变量及其分布

4 20.4 多维随机变量及其分布

5 20.5 边缘分布、条件分布和分布的混合

6 20.6 随机变量的变换

20.5.1 边缘分布：离散型边缘分布

定义 13

如果 (X, Y) 具有联合密度函数 $f_{X,Y}$，则 X 的**边缘概率密度函数**定义为：

$$f_X(x) = P(X = x) = \sum_y P(X = x, Y = y) = \sum_y f(x, y)$$

Y 的**边缘概率密度函数**定义为： $f_Y(y) = P(Y = y) = \sum_x P(X = x, Y = y) = \sum_x f(x, y)$

相应的边缘分布函数可以由 (X, Y) 的分布函数 $F(x, y)$ 所确定，事实上，

$$F_X(x) = P\{X \le x\} = P\{X \le x, Y < \infty\} = F(x, \infty)$$

即 $F_X(x) = F(x, \infty)$ 就是说，只要在函数 $F(x, y)$ 中令 $y \to \infty$ 就能得到 $F_X(x)$。同理 $F_Y(y) = F(\infty, y)$

例 5

在例3中：$f_X(X = 0) = P(X = 0) = P(X = 0, Y = 0) + P(X = 0, Y = 1) = 0 + \frac{1}{5} = \frac{1}{5}$

$$f_Y(Y = 1) = P(Y = 1) = P(X = 0, Y = 1) + P(X = 1, Y = 1) = \frac{1}{5} + \frac{3}{5} = \frac{4}{5}$$

连续型边缘分布

定义 14

若二维随机变量 $(X\ Y)$ 的概率密度函数是 $f(x,y)$，称 $f_X(x)$ 和 $f_Y(y)$:

$$f_X(x)=\int_{-\infty}^{+\infty} f(x,y)dy \quad f_Y(y)=\int_{-\infty}^{+\infty} f(x,y)dx$$

分别是随机变量 X,Y 的边缘密度函数，相应的边缘分布函数记为 F_X 和 F_Y。

例 6

在例4中二维随机变量 $(X\ Y)$ 的联合密度函数: $f(x,y)=\begin{cases}\frac{1}{(b_1-a_1)(b_2-a_2)} & a_1\le x\le b_1, a_2\le y\le b_2\\ 0 & otherelse\end{cases}$

则称 X 和 Y 的边缘密度函数:

$$f_X(x)=\int_{-\infty}^{+\infty} f(x,y)dy=\int_{a_2}^{b_2}\frac{1}{(b_1-a_1)(b_2-a_2)}dy=\frac{1}{b_1-a_1}$$

$$f_Y(y)=\int_{-\infty}^{+\infty} f(x,y)dx=\int_{a_1}^{b_1}\frac{1}{(b_1-a_1)(b_2-a_2)}dy=\frac{1}{b_2-a_2}$$

20.5.2 独立随机变量

定义 15

如果对于任意 A 和 B 有

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

则称随机变量 X 和 Y 是独立的，记 $X \amalg Y$，否则称 X 和 Y 是相依的.

原则上，为检验两个随机变量 X, Y 是否独立，需要对所有的子集 A, B 检验。所幸，我们有以下定理：

定理 4

令 X, Y 具有联合 PDF$f_{X,Y}$，则 $X \amalg Y$ 当且仅当 $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ 对所有 x, y 成立。

例 7

假设 X, Y 独立并具有相同密度函数

$$f(x) = \begin{cases} 2x, & 0 \le x \le 1 \\ 0, & otherwise \end{cases}$$

从而

$$\begin{aligned} P(X+Y \le 1) &= \int\int_{x+y \le 1} f(x,y)\,dydx \\ &= 4\int_0^1 x(\int_0^{1-x} ydy)\,dx \\ &= 4\int_0^1 x\frac{(1-x)^2}{2}\,dx = \frac{1}{6} \end{aligned}$$

定理 5

假设 X, Y 的范围是矩形（可能无穷），如果对函数 g, h(不一定是概率密度函数) 有 $f(x, y) = g(x)h(y)$ 成立，则 X, Y 独立。

例 8

令 X, Y 具有概率密度函数

$$f(x, y) = \begin{cases} 2e^{-(x+2y)}, & x > 0, y > 0 \\ 0, & otherwise \end{cases}$$

X, Y 的范围是矩形 $(0, \infty) \times (0, \infty)$ 又 $f(x, y) = 2e^{-x} \times e^{-2y}$，从而 $X \amalg Y$

定理 6

设 $(X_1, X_2, \ldots, X_m)$ 和 $(Y_1, Y_2, \ldots, Y_n)$ 相互独立，则 $X_i (i = 1, 2, \ldots, m)$ 和 $Y_j (j = 1, 2, \ldots, n)$ 相互独立。且若 g, h 为连续函数，则 $h(X_1, X_2, \ldots, X_m), g(Y_1, Y_2, \ldots, Y_n)$ 相互独立。

20.5.3 条件分布

如果 X, Y 是离散的，则可以计算假设已经观察到 $Y = y$ 情况下 X 的条件分布。特别地，$P(X = x|Y = y) = P(X = x, Y = y)/P(Y = y)$, 从而有如下条件密度函数的定义。

定义 16

假设随机变量 X, Y 是离散的，如果 $f_Y(y) > 0$ 则条件概率密度函数为

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

对于连续情形则有定义:

定义 17

假设随机变量 X, Y 是连续的，如果 $f_Y(y) > 0$ 则条件概率密度函数为

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

从而

$$P(X \in A|Y = y) = \int_A f_{X|Y}(x, y)\,dx$$

例 9

令 X, Y 服从单位正方形上的联合均匀分布，从而当 $0 \le x \le 1$ 时，$f_{X,Y}(x,y) = 1$，其他情况下 $f_{X,Y}(x,y) = 0$，即给定 $Y = y$，$X \sim U(0,1)$ 记作 $X|\,Y = y \sim U(0,1)$

例 10

令

$$f(x,y) = \begin{cases} x+y, & 0 \le x \le 1, 0 \le y \le 1 \\ 0, & otherwise \end{cases}$$

试求 $P(X < 1/4|\,Y = 1/3)$ 的值，易知 $f_Y(y) = y + 1/2$ 因此

$$f_{X|\,Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{x+y}{y+1/2}$$

所以

$$P(X < \frac{1}{4}|\,Y = \frac{1}{3}) = \int_0^{\frac{1}{4}} \frac{x+1/3}{1/3+1/2}\,dx = \frac{11}{80}$$

条件分布和联合分布的应用

监督学习的任务就是学习一个模型，应用这一模型，对给定的输入预测相应的输出。这个模型的一般形式为决策函数：$Y = f(X)$ 或条件概率分布 $P(Y|X)$。
监督学习方法又可以分成生成方法和判别方法。所学到的模型分别称为生成模型和判别模型。

- 生成方法由数据学习联合分布 $P(X, Y)$，然后求出条件概率分布 $P(Y|X)$ 作为预测的模型，即生成模型

$$P(Y|X) = \frac{P(X, Y)}{P(X)}$$

  这样的方法之所以称为生成方法，是因为模型表示了给定输入 X 产生输出 Y 的生成关系。典型的生成模型有朴素贝叶斯和隐马尔可夫模型。

- 判别方法由数据直接学习决策函数 $f(X)$ 或条件概率分布 $P(Y|X)$ 作为预测的模型，即判别模型。判别方法关心的是给定的输入 X，应该预测什么样的输出 Y。典型的判别模型包括 k 近邻、感知机、决策树、逻辑回归、最大熵模型、支持向量机、提升方法和条件随机场

生成方法和判别方法的特点

在监督学习中，生成方法和判别方法各有优缺点，适合于不同条件下的学习问题。

- **生成方法的特点：**
  - 生成方法可以还原出联合概率分布 $P(X, Y)$，而判别方法不能；
  - 生成方法的学习收敛速度更快，即当样本容量增加时，学到的模型可以更快地收敛于真实模型；
  - 当存在隐变量时，仍可以用生成方法学习，此时判别方法就不能用。
- **判别方法的特点：**
  - 判别方法直接学习的是条件概率 $P(Y|X)$ 或决策函数 $f(X)$，直接面对预测，往往学习的准确率更高；
  - 由于直接学习 $P(Y|X)$ 或 $f(X)$，可以对数据进行各种程度上的抽象、定义特征并使用特征，因此可以简化学习问题。

20.5.4 两个重要的多元分布：多项分布

定义 18

多项分布 二项分布的多元形式称为多项分布。假设一个坛子里装有 k 种颜色的球，编号为 '颜色 1，颜色 2，……，颜色 k'，随机从坛子中取一个球。令 $p=(p_1,p_2,\ldots,p_k)$，其中 $p_j\geq 0,\sum_{j=1}^k p_j=1$，假设 p_j 表示选取的球的颜色为颜色 j 的概率。抽取 n 次（独立重复抽取）并令 $X=(X_1,\ldots,X_k)$，其中 X_j 表示颜色 j 出现的次数。因此 $n=\sum_{j=1}^k X_j$，则 X 服从多元分布 (n,p)，记作 $X\sim Multinomail(n,p)$，其概率函数为

$$f(x)=\binom{n}{x_1,\ldots,x_n}p_1^{x_1}\ldots p_k^{x_k},\qquad \text{其中}\qquad \binom{n}{x_1,\ldots,x_n}=\frac{n!}{x_1!\ldots x_k!}.$$

定理 7

假设 $X\sim Multinomail(n,p)$，其中，$X=(X_1,\ldots,X_k),p=(p_1,p_2,\ldots,p_k)$，则 X_j 的边缘分布为二项分布 $Binomial(n,p_j)$.

多元正态分布

定义 19

多元正态分布 一元正态分布有两个参数 μ, σ，在多元情况下，μ 是一个向量，σ 被矩阵 Σ 取代，首先令 $Z = (Z_1, \ldots, Z_k)^T$，其中 $Z_1, \ldots, Z_k \sim N(0,1)$ 且独立，则 Z 的密度函数为

$$f(\boldsymbol{z}) = \prod_{i=1}^{k} f(z_i) = \frac{1}{(2\pi)^{k/2}} e^{-\frac{1}{2}\sum_{j=1}^{k} z_j^2} = \frac{1}{(2\pi)^{k/2}} e^{-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{z}}$$

称 Z 服从标准多元正态分布，记作 $Z \sim N(\mathbf{0}, \boldsymbol{I})$，其中 $\mathbf{0}$ 表示有 k 个 0 元素的向量，$\boldsymbol{I}$ 为 $k \times k$ 的单位矩阵。更一般地，如果 X 具有密度函数

$$f(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{k/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-1/2(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

则向量 X 服从多元正态分布，记作 $X \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$，其中 $|\boldsymbol{\Sigma}|$ 表示 $\boldsymbol{\Sigma}$ 的行列式，$\boldsymbol{\mu}$ 为长度为 k 的向量，$\boldsymbol{\Sigma}$ 为 $k \times k$ 的正定对称矩阵。当 $\boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma} = \boldsymbol{I}$ 时就是标准多元正态分布的情形。

性质 2

由于 $\boldsymbol{\Sigma}$ 是对称正定矩阵，可证明存在矩阵 $\boldsymbol{\Sigma}^{1/2}$，称为 $\boldsymbol{\Sigma}$ 的平方根，具有以下性质：

- $\boldsymbol{\Sigma}^{1/2}$ 是对称的
- $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2}$
- $\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{1/2} = \boldsymbol{I}$，其中 $\boldsymbol{\Sigma}^{-1/2} = (\boldsymbol{\Sigma}^{1/2})^{-1}$

定理 8

如果 $Z \sim N(\mathbf{0}, \boldsymbol{I})$ 且 $X = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}Z$，则 $X \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$，相反地，如果 $X \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$，则 $\boldsymbol{\Sigma}^{-1/2}(X - \boldsymbol{\mu}) \sim N(\mathbf{0}, \boldsymbol{I})$

假设将随机向量 X 划分为 $X=(X_a,X_b)$，则类似的有 $\boldsymbol{\mu}=(\mu_a,\mu_b)$ 和

$$\boldsymbol{\Sigma}=\begin{pmatrix}\Sigma_{aa} & \Sigma_{ab}\\ \Sigma_{ba} & \Sigma_{bb}\end{pmatrix}$$

定理 9

令 $X\sim N(\boldsymbol{\mu},\boldsymbol{\Sigma})$，则

1. X_a 的边缘分布为 $X_a\sim N(\mu_a,\Sigma_{aa})$
2. 给定 $X_a=x_a$ 的条件下 X_b 的条件分布为
$$X_b|X_a=x_a\sim N(\mu_b+\Sigma_{ba}\Sigma_{aa}^{-1}(x_a-\mu_a),\Sigma_{bb}-\Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab})$$
3. 如果 $\boldsymbol{a}$ 是向量，则 $\boldsymbol{a}^TX\sim N(\boldsymbol{a}^T\mu,\boldsymbol{a}^T\Sigma\boldsymbol{a})$
4. $V=(X-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(X-\boldsymbol{\mu})\sim\mathcal{X}_k^2$

20.5.5 分布的混合

- 组合一些简单的概率分布来定义新的概率分布也是一种比较常见的做法. 一种通用的组合方法是构造混合分布 (mixture distribution). 混合分布由一些组件 (component) 分布构成. 每次实验，样本是由哪个组件分布产生的取决于从一个多元伯努利分布采样的结果.

$$P(x) = \sum_i P(c = i)P(x|c = i)$$

  这里 $P(c = i)$ 是对各组件的一个多项分布。

- 例如已知中国男性和女性的身高分别服从高斯分布 $N1(x; \mu_1, \sigma_1)$，$N_2(x; \mu_2, \sigma_2)$. 则可以使用 2 个高斯分布的混合分布来描述中国人身高分布.

$$N(x) = k_1 N_1(x; \mu_1, \sigma_1) + k_2 N_2(x; \mu_2, \sigma_2)$$

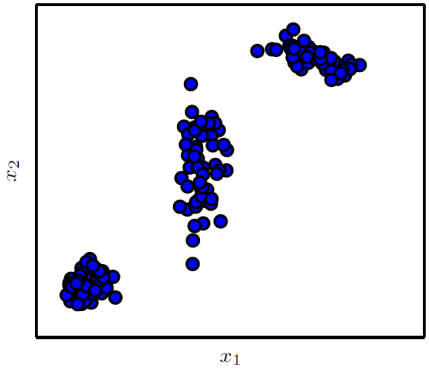
  这里 $k_1 + k_2 = 1$.

多元高斯混合模型

- **多元高斯混合模型** (Gaussian Mixture Model) 是一种非常常见的混合模型，它的组件 $p(x|c=i)$ 是多元高斯分布. 每个组件都有各自的均值 $\mu^{(i)}$ 和协方差矩阵 $\Sigma^{(i)}$.
- 高斯混合分布是概率密度的万能近似器 (universal approximator). 任何平滑的概率密度都可以用具有足够多组件的高斯混合模型以任意精度来逼近.
- 实际中使用高斯分布对数据分布进行建模时，有时会假设高斯分布组件有相同的方差，即 $\Sigma^{(i)}=\Sigma$.

多元高斯混合模型

- 某个高斯混合模型生成的样本



有三个组件. 从左到右，第一个组件具有各向同性的协方差矩阵，在每个方向上具有相同的方差. 第二个具有对角的协方差矩阵，沿着每个轴的对齐方向单独控制方差. 该示例中，沿着 x_2 轴的方差要比沿着 x_1 轴的方差大. 第三个组件具有满秩的协方差矩阵，使它能够沿着任意基的方向单独地控制方差.

1. 20.1 随机变量
2. 20.2 离散型随机变量及其分布
3. 20.3 连续型随机变量及其分布
4. 20.4 多维随机变量及其分布
5. 20.5 边缘分布、条件分布和分布的混合
6. 20.6 随机变量的变换

20.6.1 随机变量的变换：离散随机变量的变换

- 假设随机变量 X 有 PDFf_X 和 CDFF_X，令 $Y = r(X)$ 为 X 的函数，例如 $Y = X^2, Y = e^X$, 称 $Y = r(X)$ 为 X 的变换. 怎么去求 Y 的 PDF 和 CDF 呢？
- 在离散情形下，很容易求得 Y 的密度函数

$$f_Y(y) = P(Y = y) = P(r(X) = Y) = P(x : r(x) = y) = P(X \in r^{-1}(y))$$

例 11

假设 $P(X = -1) = P(X = 1) = 1/4, P(X = 0) = 1/2$, 令 $Y = X^2$，则
$P(Y = 0) = P(X = 0) = 1/2, P(Y = 1) = P(X = 1) + P(X = -1) = 1/2$，即 Y 的取值比 X 少，因为该变换不是一一变换。

x	$f_X(x)$
-1	$\frac{1}{4}$
0	$\frac{1}{4}$
1	$\frac{1}{2}$

y	$f_Y(y)$
0	$\frac{1}{2}$
1	$\frac{1}{2}$

连续随机变量的变换

求变换的连续随机变量的分布函数的一般步骤

连续情形下求 Y 的分布要难一些，主要有 3 步:

1. 对于每个 y，求集合 $A_y = \{x : r(x) \leq y\}$
2. 求 CDF

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(r(X) \leq y) \\ &= P(x : r(x) \leq y) \\ &= \int_{A_y} f_X(x)\,dx \end{aligned}$$

3. PDF 为 $f_Y(y) = F_Y^{'}(y)$

例 12

令 $f_X(x) = e^{-x}, x > 0$，从而 $F_X(x) = \int_0^x f_X(s)ds = 1 - e^{-x}$. 令 $Y = r(X) = \ln X$, 则 $A_y = \{x : x \le e^y\}$, 且

$$\begin{aligned} F_Y(y) &= P(Y \le y) = P(\ln X \le Y) \\ &= P(X \le e^y) = F_X(e^y) = 1 - e^{-e^y} \end{aligned}$$

因此 $f_Y(y) = e^y e^{-e^y}$，其中 $y \in R$

定理 10

对于连续型随机变量 X，其累积分布函数 $F_x(\cdot)$，则随机变量 y:

$$y = F_x(X)$$

服从均匀分布。

证明.

由于累积分布函数 $F_x(X)$ 是单调递增的，且值域为 $(0,1)$, 那么 $F_x(X)$ 的逆函数 $F_x^{-1}(\cdot)$ 存在，故:

$$F_y(Y) = P(y \leq Y) = P(F_x(X) \leq Y) = P(x \leq F_x^{-1}(Y)) = F_x(F_x^{-1}(Y)) = Y$$

故:

$$p_y(Y) = \frac{d}{dY} F_y(Y) = 1 \quad 0 \leq y \leq 1$$

□

20.6.2 多个随机变量的变换

在有些情况下，更关心多个随机变量的变换. 例如，如果 X 和 Y 为给定的随机变量，可能想知道 $X/Y, X+Y, max\{X, Y\}$ 或 $min\{X, Y\}$ 的分布。令 $Z = r(X, Y)$ 为所关注的函数，求 f_Z 的步骤与上一节相同.

求多个随机变量的变换分布的步骤

1. 对每个 z，求集合 $A_z = \{(x, y) : r(x, y) \leq z\}$

2 求 CDF

$$F_Y(y) = P(Z \leq z) = P(r(X, Y) \leq z)$$
$$= P(\{(x, y) : r(x, y) \leq z\}) = \int\int_{A_z} f_{X,Y}(x, y)\,dxdy$$

3. PDF 为 $f_Z(z) = F_Z^{'}(z)$

例 13

令 $X_1, X_2 \sim Uniform(0,1)$ 且独立, 求 $Y = X_1 + X_2$ 的分布函数.(X_1, X_2) 的联合密度函数为

$$f(x_1, x_2) = \begin{cases} 1, & 0 < x_1 < 1, 0 < x_2 < 1 \\ 0, & \text{其他.} \end{cases}$$

令 $r(x_1, x_2) = x_1 + x_2$，则有

$$F_Y(y) = P(Y \le y) = P(r(X_1, X_2) \le y) \quad = P(\{(x_1, x_2) : r(x_1, x_2) \le y\}) = \int\int_{A_y} f(x_1, x_2)\, dx_1\, dx_2$$

接下来求 A_y 是一个困难的环节，首先假设 $0 < y \le 1$，则 A_y 为由顶点 $(0,0), (y,0), (0,y)$ 组成的三角形区域, 见下图.

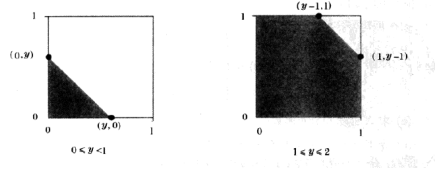


图 1: 集合 A_y 以及 A_y 包括正方形下方位于直线 $x_2 = y - x_1$ 下的所有点 (x_1, x_2)

例. 续

在这种情况下，$\int\int_{A_y} f(x_1,x_2)dx_1\,dx_2$ 为该区域的面积即 $y^2/2$. 如果 $1<y<2$，则 A_y 为单位正方形区域排除由定点 $(1,y-1),(1,1),(y-1,1)$ 组成的三角形. 该区域的面积为 $1-(2-y)^2/2$，因此

$$F_Y(y)=\begin{cases}0, & y<0\\ \frac{y^2}{2}, & 0\le y<1\\ 1-\frac{(2-y)^2}{2}, & 1\le y<2\\ 1, & y\ge 2.\end{cases}$$

微分得 PDF 为

$$f_Y(y)=\begin{cases}y, & 0\le y\le 1,\\ 2-y, & 1\le y\le 2,\\ 0, & \text{其他}\end{cases}$$

本讲小结

随机变量

- 离散随机变量
- 连续随机变量
- 累积分布函数
- 概率密度函数
- 常见的离散和连续概率分布

多元随机变量

- 离散和连续的多元随机变量
- 联合分布、联合概率密度函数
- 边缘分布、条件分布
- 独立的随机变量
- 随机变量的变换

概率分布在机器学习中应用：生成方法（生成模型）和判别方法（判别模型）!