基本概率知识

随机变量概率分布和贝叶斯公式

孔令晓

Notation	*Concept*
Ω	a sample space
ω	an atom
$p(\omega)$	a probability measure
X	a random variable
x	a realization of a random variable
$P(X, Y, Z)$	a joint distribution
$P(X \mid Y) = P(X, Y)/P(Y)$	conditional probability
$P(X, Y, Z) = P(X)P(Y \mid X)P(Z \mid X, Y)$	the chain rule
$P(X) = \sum_y P(X, Y = y)$	marginalization
$P(Y \mid X) \propto P(X \mid Y)P(Y)$	Bayes rule
$X \perp\!\!\!\perp Y \iff P(X, Y) = P(X)P(Y)$	independence
$X \perp\!\!\!\perp Y \mid Z \iff P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z)$	conditional independence
$\mathbb{E}[f(X)] = \sum_x f(x)p(x)$	expectation
$\mathbb{E}[f(X) \mid Y = y] = \sum_x f(x)p(x \mid y)$	conditional expectation

图 1 符号和概念

- 样本空间（Sample spaces）

样本空间是一个随机过程所有可能的结果组成的集合。随机变量是在一个样本空间中定义的，可以是离散的或连续的。

- 原子（atoms）

原子是样本空间中的“值”，如掷骰子{1，2，3，4，5，6}。

- 概率测度（probability measure）

概率测度是样本空间中每个原子的概率。

例如，对于一个不均匀硬币，一个概率函数是

p（heads）=0.7

p（tails）=0.3

概率是非负的且总和为 1

$$p(\omega) \geq 0 \quad \forall \omega \in \Omega$$

$$\sum_{\omega \epsilon \Omega} p(\omega) = 1$$

- 事件（events）

事件是样本空间的一个子集，原子是最小的事件

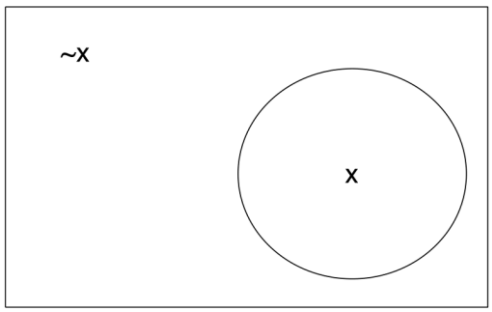

图 2 样本空间、原子、事件的关系

一个原子是盒子里的一个点，所有的原子组成样本空间。样本空间有一个概率测度 p(ω)，总和为 1。事件是原子的集合，这幅图中的两个事件 x 和~x。事件的概率是其原子的概率之和。

一个骰子的样本空间是{1，2，3，4，5，6}，一个事件 a={4，5，6}，它的概率 p（a）=p（4）+p（5）+p（6）。

- 随机变量（random variables）

随机变量是一个“概率结果”，如抛硬币的结果。一个随机变量是一组分隔样本空间的事件。最简单的随机变量将空间划分为单个原子。

随机变量 X 是一组事件，每一个事件 x 都是原子的一个集合，随机变量划分样本空间，所以$\Omega = \bigcup_{x \in X} x$。事件 x 的概率是$\mathrm{p(x)} = \sum_{\omega \in x} p(\omega)$。P（X）是随机变量的概率分布。当 X 是离散的，P（X）是一个总和为 1 的概率向量。

这个随机变量的定义有助于理解什么是分布，但这也似乎是一种直观的感觉，所以我们调整随机变量的定义同时表示两件事，将样本空间划分为事件，并将每个事件附加一个值。例如，X 是一个随机选择的人的高度，而 X 是这个高度的一个实现，比如 72 英寸。

- 联合分布（Joint distributions）

联合分布是同时对多个随机变量的概率分布。

假设三个骰子 X，Y，Z，每一个随机变量都有一个样本空间Ω_X，Ω_Y，Ω_Z，取这些样本空间的乘积空间$\Omega_{X,Y,Z} = \Omega_X \times \Omega_Y \times \Omega_Z$，定义一个概率测度 p(x,y,z)（x 是$\Omega_X$中的原子）。这个概率测度是一个联合概率测度。它描述了三个骰子的每一个可能的概率。

在乘积空间$\Omega_{X,Y,Z}$中，从简单样本空间中得到的简单随机变量 X、Y、Z 现在是复杂随机变量。随机变量 X 以第一个值划分空间。表达式 P（X=3）是第一个骰子转到 3 的概率，是所有 x=3 的原子（x，y，z）的概率之和。随机变量 Y 以第二个值划分空间，随机变量 Z 是第三个。

其他随机变量是骰子的集合，比如（X，Y）或（X，Z）。通过三个骰子的两个值划分样本空间。P（X，Y）和 P（X，Z）可以计算为联合概率测度的总和。全联合是 P（X，Y，Z）。

- 条件分布（Conditional distributions）

条件分布是一个随机变量在其他随机变量的值的影响下的分布。$\mathrm{P}(\mathrm{X} = \mathrm{x} | \mathrm{Y} = \mathrm{y})$表示 Y=y 时 X=x 的概率。

条件概率是直观的。假设两个舍友，小明喜欢看如懿传，小王不喜欢。现在考虑两个随机变量的联合分布 P（X，Y），X 代表小明是否在看如懿传，Y 代表小王是否在宿舍。

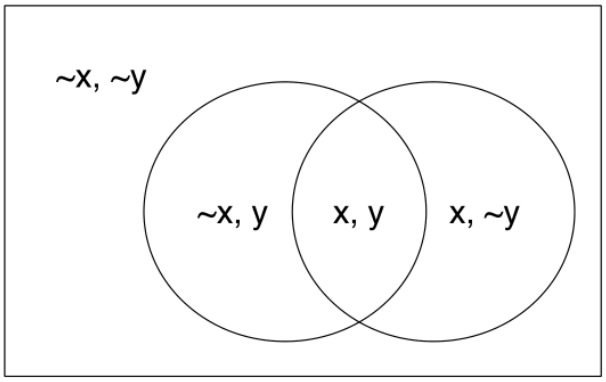

图 3 条件概率

我们可以考虑的分布是

$$\mathrm{P}(\text{Ming's TV is on} \mid \text{Wang is in dormitory}) = 0.1$$

$$\mathrm{P}(\text{Ming's TV is not on} \mid \text{Wang is in dormitory}) = 0.9$$

$$\mathrm{P}(\text{Ming's TV is on}\,|\,\text{Wang is not in dormitory}) = 0.7$$
$$\mathrm{P}(\text{Ming's TV is not on}\,|\,\text{Wang is not in dormitory}) = 0.3$$

现在$\mathrm{P}(\mathrm{X}=\mathrm{x}|\mathrm{Y}=\mathrm{y})$对于 y 的每一个值都是不同的分布。

$$\sum_x \mathrm{P}(\mathrm{X}=\mathrm{x}|\mathrm{Y}=\mathrm{y}) = 1$$

$$\sum_y \mathrm{P}(\mathrm{X}=\mathrm{x}|\mathrm{Y}=\mathrm{y}) \neq 1\text{（至少，不一定）}$$

条件概率定义为

$$\mathrm{P}(\mathrm{X}=\mathrm{x}|\mathrm{Y}=\mathrm{y}) = \frac{P(X=x,Y=y)}{P(Y=y)}(P(Y)>0)$$

- 链式法则

条件概率有助于推导链式法则。链式法则将联合分布定义为条件分布的乘积，

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

一般来说，对于任意一组 N 个变量

$$\mathrm{P}(X_1,\ldots,X_N) = \prod_{n=1}^{N} P(X_n|X_1,\ldots,X_{n-1})$$

- 边缘分布（Marginalization）

给定一个随机变量集合的联合分布，它们的子集的分布是什么？例如，给定 P（X，Y，Z），P（X）是什么？分布 P（X）叫做 X 的边缘分布。

$$\mathrm{P}(\mathrm{X}=\mathrm{x}) = \sum_{y\in Y}\sum_{z\in Z} P(X=x,Y=y,Z=z)$$

图 2 中的图表显示了这是如何表示的，假设图表是乘积空间，每个原子都是（x,y,z）元组。$\mathrm{P}(\mathrm{X}=\mathrm{x})$就是 X=x 的集合的分布。

边缘分布也可以从链式法则中得到，

$$\sum_y\sum_z P(X=x,Y=y,Z=z)$$
$$= \sum_y\sum_z P(X=x)P(Y=y,Z=z|X=x)$$
$$= P(X=x)\sum_y\sum_z p(Y=y,Z=z|X=x)$$
$$= \mathrm{P}(\mathrm{X}=\mathrm{x})$$

- 贝叶斯法则（Bayes rule）

从链式法则和边缘分布可以得到贝叶斯法则。

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

假设 Y 是一种疾病而 X 是一种症状，一个病人带着症状走进一家诊所，她患有这种病吗？从人群中，我们可以估计出患病有症状的概率 P（X|Y）和疾病发生的概率 P（Y）。然后使用贝叶斯规则我们可以计算出有症状时疾病的分布 P（Y|X）。

贝叶斯规则对于贝叶斯统计和生成概率建模至关重要。领域知识和假设提供了一个简单的隐藏变量的先验分布 P（Y）以及 P（X|Y），然后可以从贝叶斯公式得到隐藏变量 P（Y|X）。

- 独立性（Independence）

  如果 X 不能告诉我们关于 Y 的信息，则随机变量 X 和 Y 是独立的。

$$\mathrm{P(Y|X)} = \mathrm{P(Y)}$$

  独立性意味着它们的联合分布因子

$$\mathrm{X} \perp \mathrm{Y} \Leftrightarrow \mathrm{P(X,Y)} = \mathrm{P(X)P(Y)}$$

  为什么？因为链式法则，

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

  下面是一些独立随机变量的例子：（1）第一次抛硬币，第二次抛硬币；（2）你是否使用电牙刷，蓝色是否是你最喜欢的颜色。

  下面是一些不独立的随机变量的例子：（1）你是否登记为共和党人，以及你 2016 年是否投票支持特朗普（2）天空的颜色，和白天的时间。

  请注意，独立性是样本空间上的联合概率测度的一个属性。一些联合分布满足因数分解，一些不。联合分布的分解是生成模型的一个关键概念。

- 条件独立性（Conditionalindependence）

  假设有两枚硬币，一个是不均匀的一个是均匀的，

$$\mathrm{P}(\mathrm{C}_1 = \mathrm{HEADS}) = 0.5 \quad \mathrm{P}(\mathrm{C}_2 = \mathrm{HEADS}) = 0.7$$

  我们随机选择其中一枚硬币，$\mathrm{Z} \in \{1,2\}$，翻转硬币 Cz 两次，记录输出（X，Y）。X 和 Y 独立吗？假设我们知道 Z 的值，哪个硬币会被翻转？他们是独立的吗？

  给定 Z，如果$\mathrm{P}(Y|X,\ Z = z) = \mathrm{P(Y|Z = z)}$成立，变量 X 和 Y 条件独立。

  同样，这意味着因子分解

$$\mathrm{X} \perp \mathrm{Y|Z} \Leftrightarrow \mathrm{P(X,Y|Z = z)} = \mathrm{P(X|Z = z)P(Y|Z = z)}$$(对于所有的 z)

  条件独立性又是联合概率测度的一个属性，它是如何在每个事件 z 中分解的。

- 连续随机变量（Continuous random variables）

  到目前为止，我们只使用了离散随机变量。随机变量可以是连续的，在一个连续随机变量中，我们有一个总和为 1 的密度 p(x)。如果$\mathrm{x} \in \mathbb{R}$，$\int_{-\infty}^{\infty} p(x)dx = 1$。

  对于连续随机变量，概率是更小的区间的积分。

  高斯函数是一个连续分布。它的密度是

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

  x 越接近 μ，密度越高。高斯密度是一个钟形凸起，μ 是正态分布的位置参数，描述正态分布的集中趋势位置。σ 描述正态分布数据分布的离散程度，σ 越大，数据分布越分散，σ 越小，数据分布越集中。

- 期望（Expectation）

  期望是 f 的加权平均值，x 提供权重，

$$\mathbb{E}[\mathrm{f(X)}] = \sum_x p(x)f(x)$$

  在连续的情况下，期望是一个积分

$$\mathbb{E}[\mathrm{f(X)}] = \int p(x)f(x)dx$$

  条件期望是相似的，

$$\mathbb{E}[\mathrm{f(X)}|Y = y] = \sum_x p(x|y)f(x)$$

$[f(X)|Y=y]$是一个标量，但是$[f(X)|Y]$是随机变量，每个 y 值的标量，它的分布是 P(Y)。

有两种特殊的期望。$\mathbb{E}[X]$是 X 的均值。如果 X 是随机选择的人的身高，$\mathbb{E}[X]$就是人口的平均身高。$\mathbb{E}[(X-\mathbb{E}[X])^2]$是方差。它衡量的是平均值（平方）距离均值的距离，这就解释了 X 的分布是如何分布在平均水平上的。

假设我们有两个离散随机变量 X、Y，$f: \mathbb{R}^2 \to \mathbb{R}$是这两个变量的函数，f 的期望为：

$$\mathbb{E}[f(X,Y)] = \sum_x \sum_y p(x,y) f(x,y)$$

连续随机变量的期望为：

$$\mathbb{E}[f(X,Y)] = \int \int p(x,y) f(x,y) dx dy$$

- 方差（Variance）

方差是衡量随机变量或一组数据离散程度的度量。概率论中方差用来度量随机变量和其数学期望（即均值）之间的偏离程度。

$$\mathrm{Var}[X] = \mathbb{E}[(X-\mathbb{E}[X])^2]$$

其中 E（X）是 X 的期望值，X 是变量值，公式中的 E 是期望值 expected value 的缩写，意为“变量值与其期望值之差的平方和”的期望值。

$$\mathbb{E}\left[\left(X-\mathbb{E}(X)\right)^2\right] = \mathbb{E}[X^2 - 2\mathbb{E}[X]X + \mathbb{E}[X]^2]$$

$$= \mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + \mathbb{E}[X]^2$$

$$= \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

如果常数 a 属于 R，Var[a]=0，$\mathrm{Var}[af(X)]=a^2\,\mathrm{Var}[f(X)]$，Var[f(X)+a]= Var[f(X)]。

- 协方差（Covariance）

协方差在概率论和统计学中用于衡量两个变量的总体误差。而方差是协方差的一种特殊情况，即当两个变量是相同的情况。

我们可以用期望的概念来研究两个随机变量之间的关系。两个随机变量 X 和 Y 的协方差被定义为

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

$$= \mathbb{E}[XY - X\mathbb{E}[Y] - Y\mathbb{E}[X] + \mathbb{E}[X]\mathbb{E}[Y]]$$

$$= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[Y]\mathbb{E}[X] + \mathbb{E}[X]\mathbb{E}[Y]$$

$$= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

在这里，显示协方差的两种形式的平等的关键步骤是在第三个等式中，E[X]和 E[Y]实际上是可以从期望中拉出来的常数。当 Cov[X，Y]=0 时，我们说 X 和 Y 是不相关的（uncorrelated）。

（线性）E[f(X,Y) + g(X,Y)] = E[f(X,Y)] + E[g(X,Y)]

Var[X + Y] = Var[X] + V ar[Y] + 2Cov[X,Y]

如果 X，Y 是相互独立的，Cov[X,Y] = 0，E[f(X)g(Y)] = E[f(X)]E[g(Y)]

- 多维随机变量

假设我们有 n 个随机变量。当我们一起处理这些随机变量时，我们通常会发现把它们放在一个向量上很方便$X = [X_1\ X_2\ \ldots\ X_n]^T$. 。我们把得到的向量称为一个随机向量（一个随机向量是从Ω映射到 Rn）。

期望：考虑一个任意的函数 f：$\mathbb{R}^n \to \mathbb{R}$. 这个函数的期望值被定义为

$$\mathbb{E}[\mathrm{f(X)}] = \int_{\mathbb{R}^n} f(x_1, x_2, \ldots, x_n) p(x_1, x_2, x_n) dx_1 dx_2 \ldots dx_n$$

$\int_{\mathbb{R}^n}$ 是从负无穷到正无穷的 n 个连续的集成。如果函数 f：$\mathbb{R}^n \to \mathbb{R}^m$，f 的期望值是输出向量的元素期望值，如 f 为

$$\mathrm{f(x)} = \begin{bmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_m(x) \end{bmatrix}$$

则

$$\mathbb{E}[\mathrm{f(x)}] = \begin{bmatrix} \mathbb{E}[f_1(x)] \\ \mathbb{E}[f_2(x)] \\ \vdots \\ \mathbb{E}[f_m(x)] \end{bmatrix}$$

协方差矩阵（Covariance matrix）：

对于一个给定的随机向量 X：$\Omega \to \mathbb{R}^n$，它的协方差矩阵是 n×n 的平方矩阵，

$$\Sigma_{ij} = \mathrm{Cov}[X_i, X_j]$$

根据协方差的定义，我们有

$$\begin{aligned}
\Sigma &= \begin{bmatrix} Cov[X_1, X_1] & \cdots & Cov[X_1, X_n] \\ \vdots & \ddots & \vdots \\ Cov[X_n, X_1] & \cdots & Cov[X_n, X_n] \end{bmatrix} \\
&= \begin{bmatrix} E[X_1^2] - E[X_1]E[X_1] & \cdots & E[X_1 X_n] - E[X_1]E[X_n] \\ \vdots & \ddots & \vdots \\ E[X_n X_1] - E[X_n]E[X_1] & \cdots & E[X_n^2] - E[X_n]E[X_n] \end{bmatrix} \\
&= \begin{bmatrix} E[X_1^2] & \cdots & E[X_1 X_n] \\ \vdots & \ddots & \vdots \\ E[X_n X_1] & \cdots & E[X_n^2] \end{bmatrix} - \begin{bmatrix} E[X_1]E[X_1] & \cdots & E[X_1]E[X_n] \\ \vdots & \ddots & \vdots \\ E[X_n]E[X_1] & \cdots & E[X_n]E[X_n] \end{bmatrix} \\
&= E[XX^T] - E[X]E[X]^T = \ldots = E[(X - E[X])(X - E[X])^T].
\end{aligned}$$

协方差矩阵有许多有用的性质：

$\Sigma \succeq 0$：Σ 是半正定。

$\Sigma = \Sigma^T$：Σ 是对称的。

多元高斯分布（The multivariate Gaussian distribution）：

一个关于随机向量的概率分布特别重要的例子就是多元高斯或多元正态分布。随机向量 $X \in R^n$ 有一个多变量的正态（或高斯）分布，均值 $\mu \in R^n$，协方差矩阵 $\Sigma \in \mathbb{S}_{++}^n$，$\mathbb{S}_{++}^n$ 指的是对称正定 n×n 矩阵的空间。

$$f_{x_1, x_2, \ldots, x_n}(x_1, x_2, \ldots, x_n; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right)$$

我们把这个写出 $X \sim \mathcal{N}(\mu, \Sigma)$。μ 相当于每个正态分布的对称轴，是一个一维向量。一般来说，高斯随机变量在机器学习和统计学中非常有用，主要有两个原因。首先，在统计算法中建模“噪声”时，它们是非常常见的。通常，噪声可以

认为是大量的小的独立随机扰动的积累，影响测量过程；根据中心极限定理（Central Limit Theorems）独立随机变量的求和会趋向于“看似高斯函数”。其次，高斯随机变量对于许多分析操作来说是很方便的，因为在实践中出现的许多关于高斯分布的积分都有简单的闭合形式解。