

本文是斯坦福大学CS229机器学习课程的基础材料，[原始文件下载](#)

原文作者：Arian Maleki , Tom Do

翻译：[石振宇](#)

审核和修改制作：[黄海广](#)

备注：请关注[github](#)的更新。

CS229 机器学习课程复习材料-概率论

CS229 机器学习课程复习材料-概率论

概率论复习和参考

1. 概率的基本要素
 - 1.1 条件概率和独立性
2. 随机变量
 - 2.1 累积分布函数
 - 2.2 概率质量函数
 - 2.3 概率密度函数
 - 2.4 期望
 - 2.5 方差
 - 2.6 一些常见的随机变量
3. 两个随机变量
 - 3.1 联合分布和边缘分布
 - 3.2 联合概率和边缘概率质量函数
 - 3.3 联合概率和边缘概率密度函数
 - 3.4 条件概率分布
 - 3.5 贝叶斯定理
 - 3.6 独立性
 - 3.7 期望和协方差
4. 多个随机变量
 - 4.1 基本性质
 - 4.2 随机向量
 - 4.3 多元高斯分布
5. 其他资源

概率论复习和参考

概率论是对不确定性的研究。通过这门课，我们将依靠概率论中的概念来推导机器学习算法。这篇笔记试图涵盖适用于**CS229**的概率论基础。概率论的数学理论非常复杂，并且涉及到“分析”的一个分支：测度论。在这篇笔记中，我们提供了概率的一些基本处理方法，但是不会涉及到这些更复杂的细节。

1. 概率的基本要素

为了定义集合上的概率，我们需要一些基本元素，

- 样本空间 Ω ：随机实验的所有结果的集合。在这里，每个结果 $w \in \Omega$ 可以被认为是实验结束时现实世界状态的完整描述。
- 事件集（事件空间） \mathcal{F} ：元素 $A \in \mathcal{F}$ 的集合（称为事件）是 Ω 的子集（即每个 $A \subseteq \Omega$ 是一个实验可能结果的集合）。

备注： \mathcal{F} 需要满足以下三个条件：

(1) $\emptyset \in \mathcal{F}$

(2) $A \in \mathcal{F} \implies \Omega \setminus A \in \mathcal{F}$

(3) $A_1, A_2, \dots, A_i \in \mathcal{F} \implies \cup_i A_i \in \mathcal{F}$

• 概率度量 P ：函数 P 是一个 $\mathcal{F} \rightarrow \mathbb{R}$ 的映射，满足以下性质：

- 对于每个 $A \in \mathcal{F}$, $P(A) \geq 0$,
- $P(\Omega) = 1$
- 如果 A_1, A_2, \dots 是互不相交的事件(即当 $i \neq j$ 时, $A_i \cap A_j = \emptyset$), 那么:

$$P(\cup_i A_i) = \sum_i P(A_i)$$

以上三条性质被称为**概率公理**。

举例：

考虑投掷六面骰子的事件。样本空间为 $\Omega = \{1, 2, 3, 4, 5, 6\}$ 。最简单的事件空间是平凡事件空间 $\mathcal{F} = \{\emptyset, \Omega\}$ 。另一个事件空间是 Ω 的所有子集的集合。对于第一个事件空间，满足上述要求的唯一概率度量由 $P(\emptyset) = 0$, $p(\Omega) = 1$ 给出。对于第二个事件空间，一个有效的概率度量是将事件空间中每个事件的概率分配为 $i/6$ ，这里 i 是这个事件集中元素的数量；例如 $P(\{1, 2, 3, 4\}) = 4/6$, $P(\{1, 2, 3\}) = 3/6$ 。

性质：

- 如果 $A \subseteq B$ ，则： $P(A) \leq P(B)$
- $P(A \cap B) \leq \min(P(A), P(B))$
- (布尔不等式)： $P(A \cup B) \leq P(A) + P(B)$
- $P(\Omega|A) = 1 - P(A)$
- (全概率定律)：如果 A_1, \dots, A_k 是一些互不相交的事件并且它们的并集是 Ω ，那么它们的概率之和是1

1.1 条件概率和独立性

假设 B 是一个概率非0的事件，我们定义在给定 B 的条件下 A 的条件概率为：

$$P(A|B) \triangleq \frac{P(A \cap B)}{P(B)}$$

换句话说， $P(A|B)$ 是度量已经观测到 B 事件发生的情况下 A 事件发生的概率，两个事件被称为独立事件当且仅当 $P(A \cap B) = P(A)P(B)$ (或等价地, $P(A|B) = P(A)$)。因此，独立性相当于是说观察到事件 B 对于事件 A 的概率没有任何影响。

2. 随机变量

考虑一个实验，我们翻转10枚硬币，我们想知道正面硬币的数量。这里，样本空间 Ω 的元素是长度为10的序列。例如，我们可能有 $w_0 = \{H, H, T, H, T, H, H, T, T, T\} \in \Omega$ 。然而，在实践中，我们通常不关心获得任何特定正反序列的概率。相反，我们通常关心结果的实值函数，比如我们10次投掷中出现的正面数，或者最长的背面长度。在某些技术条件下，这些函数被称为**随机变量**。

更正式地说，随机变量 X 是一个的 $\Omega \rightarrow \mathbb{R}$ 函数。通常，我们将使用大写字母 $X(\omega)$ 或更简单的 X (其中隐含对随机结果 ω 的依赖)来表示随机变量。我们将使用小写字母 x 来表示随机变量的值。

举例：

在我们上面的实验中，假设 $X(\omega)$ 是在投掷序列 ω 中出现的正面的数量。假设投掷的硬币只有10枚，那么 $X(\omega)$ 只能取有限数量的值，因此它被称为**离散随机变量**。这里，与随机变量 X 相关联的集合取某个特定值 k 的概率为：

$$P(X = k) := P(\{\omega : X(\omega) = k\})$$

举例：

假设 $X(\omega)$ 是一个随机变量，表示放射性粒子衰变所需的时间。在这种情况下， $X(\omega)$ 具有无限多的可能值，因此它被称为**连续随机变量**。我们将 X 在两个实常数 a 和 b 之间取值的概率(其中 $a < b$)表示为：

$$P(a \leq X \leq b) := P(\{\omega : a \leq X(\omega) \leq b\})$$

2.1 累积分布函数

为了指定处理随机变量时使用的概率度量，通常可以方便地指定替代函数(**CDF**、**PDF**和**PMF**)，在本节和接下来的两节中，我们将依次描述这些类型的函数。

累积分布函数(CDF)是函数 $F_X : \mathbb{R} \rightarrow [0, 1]$ ，它将概率度量指定为：

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

通过使用这个函数，我们可以计算任意事件发生的概率。图1显示了一个样本**CDF**函数。

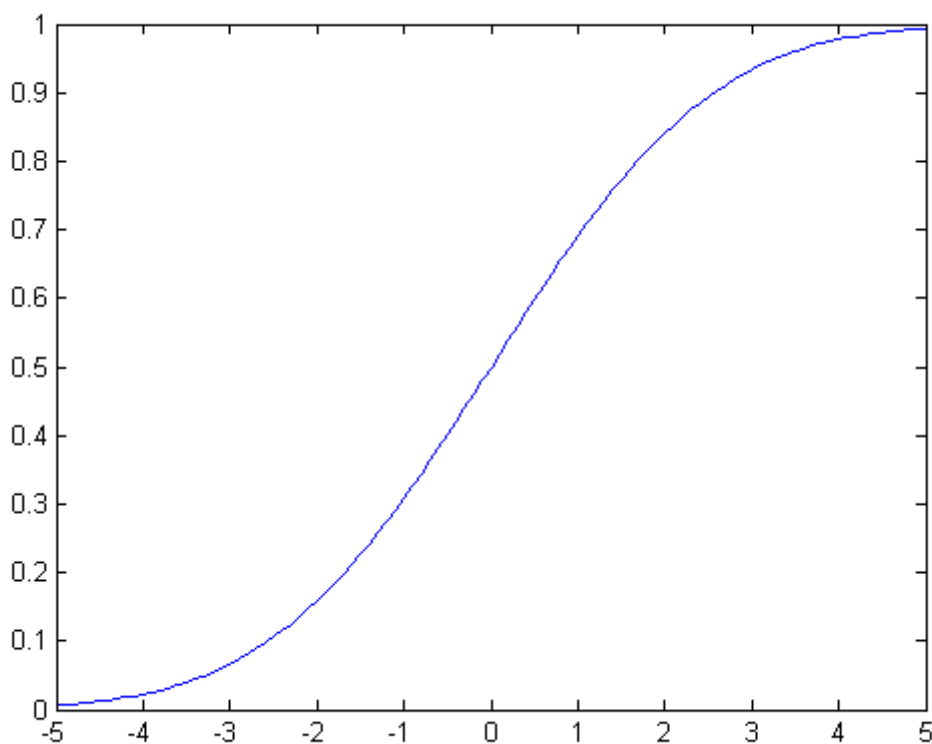


图1：一个累积分布函数(CDF)

性质：

- $0 \leq F_X(x) \leq 1$
- $\lim_{x \rightarrow -\infty} F_X(x) = 0$
- $\lim_{x \rightarrow \infty} F_X(x) = 1$
- $x \leq y \implies F_X(x) \leq F_X(y)$

2.2 概率质量函数

当随机变量 X 取有限种可能值(即， X 是离散随机变量)时，表示与随机变量相关联的概率度量的更简单的方法是直接指定随机变量可以假设的每个值的概率。特别地，**概率质量函数(PMF)**是函数 $p_X : \Omega \rightarrow \mathbb{R}$ ，这样：

$$p_X(x) \triangleq P(X = x)$$

在离散随机变量的情况下，我们使用符号 $Val(X)$ 表示随机变量 X 可能假设的一组可能值。例如，如果 $X(\omega)$ 是一个随机变量，表示十次投掷硬币中的正面数，那么 $Val(X) = \{0, 1, 2, \dots, 10\}$ 。

性质：

- $0 \leq p_X(x) \leq 1$
- $\sum_{x \in Val(X)} p_X(x) = 1$
- $\sum_{x \in A} p_X(x) = P(X \in A)$

2.3 概率密度函数

对于一些连续随机变量，累积分布函数 $F_X(x)$ 处可微。在这些情况下，我们将**概率密度函数(PDF)**定义为累积分布函数的导数，即：

$$f_X(x) \triangleq \frac{dF_X(x)}{dx}$$

请注意，连续随机变量的概率密度函数可能并不总是存在的(即，如果它不是处处可微)。

根据微分的性质，对于很小的 Δx ,

$$P(x \leq X \leq x + \Delta x) \approx f_X(x)\Delta x$$

CDF和**PDF**(当它们存在时!)都可用于计算不同事件的概率。但是应该强调的是，任意给定点的**概率密度函数(PDF)**的值不是该事件的概率，即 $f_X(x) \neq P(X = x)$ 。例如， $f_X(x)$ 可以取大于1的值(但是 $f_X(x)$ 在 \mathbb{R} 的任何子集上的积分最多为1)。

性质：

- $f_X(x) \geq 0$
- $\int_{-\infty}^{\infty} f_X(x) dx = 1$
- $\int_{x \in A} f_X(x) dx = P(X \in A)$

2.4 期望

假设 X 是一个离散随机变量，其**PMF**为 $p_X(x)$ ， $g: \mathbb{R} \rightarrow \mathbb{R}$ 是一个任意函数。在这种情况下， $g(X)$ 可以被视为随机变量，我们将 $g(X)$ 的期望值定义为：

$$E[g(X)] \triangleq \sum_{x \in Val(X)} g(x)p_X(x)$$

如果 X 是一个连续的随机变量，其**PDF**为 $f_X(x)$ ，那么 $g(X)$ 的期望值被定义为：

$$E[g(X)] \triangleq \int_{-\infty}^{\infty} g(x)f_X(x)dx$$

直觉上， $g(X)$ 的期望值可以被认为是 $g(x)$ 对于不同的 x 值可以取的值的“加权平均值”，其中权重由 $p_X(x)$ 或 $f_X(x)$ 给出。作为上述情况的特例，请注意，随机变量本身的期望值，是通过令 $g(x) = x$ 得到的，这也被称为随机变量的平均值。

性质：

- 对于任意常数 $a \in \mathbb{R}$ ， $E[a] = a$
- 对于任意常数 $a \in \mathbb{R}$ ， $E[af(X)] = aE[f(X)]$
- (线性期望)： $E[f(X) + g(X)] = E[f(X)] + E[g(X)]$
- 对于一个离散随机变量 X ， $E[1\{X = k\}] = P(X = k)$

2.5 方差

随机变量 X 的**方差**是随机变量 X 的分布围绕其平均值集中程度的度量。形式上，随机变量 X 的方差定义为：

$$\text{Var}[X] \triangleq E[(X - E(X))^2]$$

使用上一节中的性质，我们可以导出方差的替代表达式：

$$\begin{aligned} E[(X - E[X])^2] &= E[X^2 - 2E[X]X + E[X]^2] \\ &= E[X^2] - 2E[X]E[X] + E[X]^2 \\ &= E[X^2] - E[X]^2 \end{aligned}$$

其中第二个等式来自期望的线性，以及 $E[X]$ 相对于外层期望实际上是常数的事实。

性质：

- 对于任意常数 $a \in \mathbb{R}$, $\text{Var}[a] = 0$
- 对于任意常数 $a \in \mathbb{R}$, $\text{Var}[af(X)] = a^2 \text{Var}[f(X)]$

举例：

计算均匀随机变量 X 的平均值和方差，任意 $x \in [0, 1]$ ，其**PDF**为 $p_X(x) = 1$ ，其他地方为0。

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} x f_X(x) dx = \int_0^1 x dx = \frac{1}{2} \\ E[X^2] &= \int_{-\infty}^{\infty} x^2 f_X(x) dx = \int_0^1 x^2 dx = \frac{1}{3} \\ \text{Var}[X] &= E[X^2] - E[X]^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12} \end{aligned}$$

举例：

假设对于一些子集 $A \subseteq \Omega$ ，有 $g(x) = 1\{x \in A\}$ ，计算 $E[g(X)]$ ？

离散情况：

$$E[g(X)] = \sum_{x \in \text{Val}(X)} 1\{x \in A\} P_X(x) dx = \sum_{x \in A} P_X(x) dx = P(x \in A)$$

连续情况：

$$E[g(X)] = \int_{-\infty}^{\infty} 1\{x \in A\} f_X(x) dx = \int_{x \in A} f_X(x) dx = P(x \in A)$$

2.6 一些常见的随机变量

离散随机变量

- 伯努利分布：硬币掷出正面的概率为 p （其中： $0 \leq p \leq 1$ ），如果正面发生，则为1，否则为0。

$$p(x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$$

- 二项式分布：掷出正面概率为 p （其中： $0 \leq p \leq 1$ ）的硬币 n 次独立投掷中正面的数量。

$$p(x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

- 几何分布：掷出正面概率为 p （其中： $p > 0$ ）的硬币第一次掷出正面所需要的次数。

- 泊松分布：用于模拟罕见事件频率的非负整数的概率分布（其中： $\lambda > 0$ ）。

$$p(x) = e^{-\lambda} \frac{\lambda^x}{x!}$$

连续随机变量

- 均匀分布：在 a 和 b 之间每个点概率密度相等的分布（其中： $a < b$ ）。

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

- 指数分布：在非负实数上有衰减的概率密度（其中： $\lambda > 0$ ）。

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

- 正态分布：又被称为高斯分布。

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

一些随机变量的概率密度函数和累积分布函数的形状如图2所示。

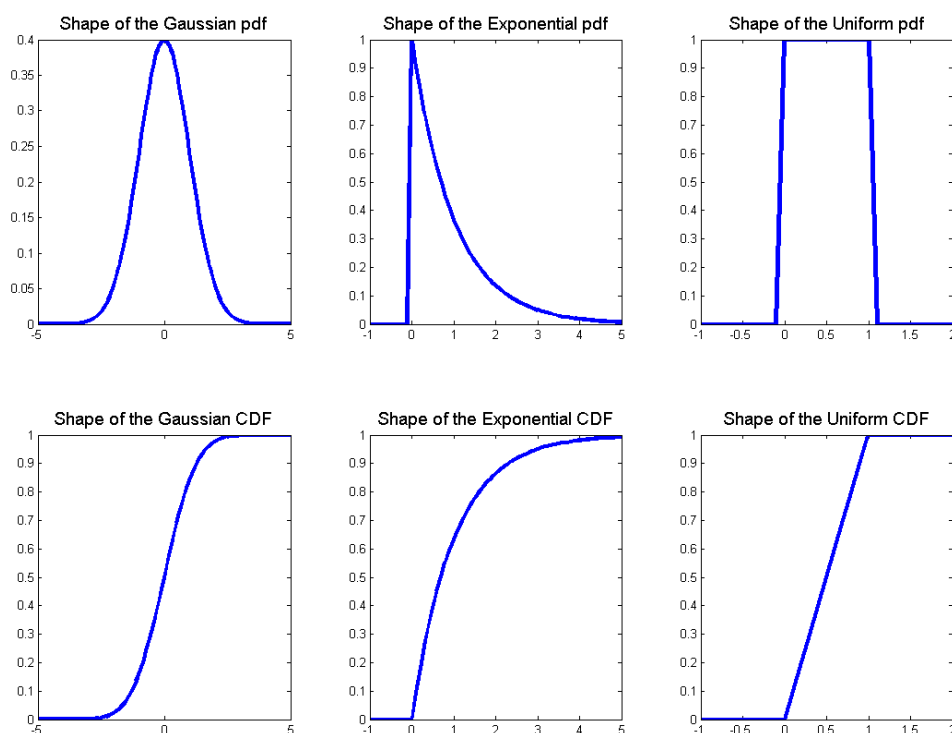


图2：一些随机变量的概率密度函数(PDF)和累积分布函数(CDF)

下表总结了这些分布的一些特性：

分布	概率密度函数(PDF)或者概率质量函数(PMF)	均值	方差
$Bernoulli(p)$ (伯努利分布)	$\begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$	p	$p(1 - p)$
$Binomial(n, p)$ (二项式分布)	$\binom{n}{k} p^k (1 - p)^{n-k} \text{ 其中: } 0 \leq k \leq n$	np	npq
$Geometric(p)$ (几何分布)	$p(1 - p)^{k-1} \text{ 其中: } k = 1, 2, \dots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$

分布	概率密度函数(PDF)或者概率质量函数(PMF)	均值	方差
$Poisson(\lambda)$ (泊松分布)	$e^{-\lambda} \frac{\lambda^k}{k!}$, 其中: $k = 1, 2, \dots$	λ	λ
$Uniform(a, b)$ (均匀分布)	$\frac{1}{b-a}$ 存在 $x \in (a, b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$Gaussian(\mu, \sigma^2)$ (高斯分布)	$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$	μ	σ^2
$Exponential(\lambda)$ (指数分布)	$\lambda e^{-\lambda x}$ $x \geq 0, \lambda > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

3. 两个随机变量

到目前为止，我们已经考虑了单个随机变量。然而，在许多情况下，在随机实验中，我们可能不止一个感兴趣的量。例如，在一个我们掷硬币十次的实验中，我们可能既关心 $X(\omega)$ = 出现的正面数量，也关心 $Y(\omega)$ = 连续最长出现正面的长度。在本节中，我们考虑两个随机变量的设置。

3.1 联合分布和边缘分布

假设我们有两个随机变量，一个方法是分别考虑它们。如果我们这样做，我们只需要 $F_X(x)$ 和 $F_Y(y)$ 。但是如果我们想知道在随机实验的结果中， X 和 Y 同时假设的值，我们需要一个更复杂的结构，称为 X 和 Y 的**联合累积分布函数**，定义如下：

$$F_{XY}(x, y) = P(X \leq x, Y \leq y)$$

可以证明，通过了解联合累积分布函数，可以计算出任何涉及到 X 和 Y 的事件的概率。

联合CDF: $F_{XY}(x, y)$ 和每个变量的联合分布函数 $F_X(x)$ 和 $F_Y(y)$ 分别由下式关联：

$$F_X(x) = \lim_{y \rightarrow \infty} F_{XY}(x, y)$$

$$F_Y(y) = \lim_{x \rightarrow \infty} F_{XY}(x, y)$$

这里我们称 $F_X(x)$ 和 $F_Y(y)$ 为 $F_{XY}(x, y)$ 的**边缘累积概率分布函数**。

性质:

- $0 \leq F_{XY}(x, y) \leq 1$
- $\lim_{x, y \rightarrow \infty} F_{XY}(x, y) = 1$
- $\lim_{x, y \rightarrow -\infty} F_{XY}(x, y) = 0$
- $F_X(x) = \lim_{y \rightarrow \infty} F_{XY}(x, y)$

3.2 联合概率和边缘概率质量函数

如果 X 和 Y 是离散随机变量，那么**联合概率质量函数** $p_{XY} : \mathbb{R} \times \mathbb{R} \rightarrow [0, 1]$ 由下式定义：

$$p_{XY}(x, y) = P(X = x, Y = y)$$

这里, 对于任意 x, y , $0 \leq P_{XY}(x, y) \leq 1$, 并且 $\sum_{x \in \text{Val}(X)} \sum_{y \in \text{Val}(Y)} P_{XY}(x, y) = 1$

两个变量上的**联合PMF**分别与每个变量的概率质量函数有什么关系？事实上：

$$p_X(x) = \sum_y p_{XY}(x, y)$$

对于 $p_Y(y)$ 类似。在这种情况下，我们称 $p_X(x)$ 为 X 的**边际概率质量函数**。在统计学中，将一个变量相加形成另一个变量的边缘分布的过程通常称为“边缘化”。

3.3 联合概率和边缘概率密度函数

假设 X 和 Y 是两个连续的随机变量，具有联合分布函数 F_{XY} 。在 $F_{XY}(x, y)$ 在 x 和 y 中处处可微的情况下，我们可以定义**联合概率密度函数**：

$$f_{XY}(x, y) = \frac{\partial^2 F_{XY}(x, y)}{\partial x \partial y}$$

如同在一维情况下， $f_{XY}(x, y) \neq P(X = x, Y = y)$ ，而是：

$$\iint_{x \in A} f_{XY}(x, y) dx dy = P((X, Y) \in A)$$

请注意，概率密度函数 $f_{XY}(x, y)$ 的值总是非负的，但它们可能大于1。尽管如此，可以肯定的是 $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y) = 1$

与离散情况相似，我们定义：

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) dy$$

作为 X 的**边际概率密度函数**(或**边际密度**)，对于 $f_Y(y)$ 也类似。

3.4 条件概率分布

条件分布试图回答这样一个问题，当我们知道 X 必须取某个值 x 时， Y 上的概率分布是什么？在离散情况下，给定 Y 的条件概率质量函数是简单的：

$$p_{Y|X}(y|x) = \frac{p_{XY}(x, y)}{p_X(x)}$$

假设分母不等于0。

在连续的情况下，在技术上要复杂一点，因为连续随机变量的概率等于零。忽略这一技术点，我们通过类比离散情况，简单地定义给定 $X = x$ 的条件概率密度为：

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

假设分母不等于0。

3.5 贝叶斯定理

当试图推导一个变量给定另一个变量的条件概率表达式时，经常出现的一个有用公式是**贝叶斯定理**。

对于离散随机变量 X 和 Y ：

$$P_{Y|X}(y|x) = \frac{P_{XY}(x, y)}{P_X(x)} = \frac{P_{X|Y}(x|y) P_Y(y)}{\sum_{y' \in \text{Val}(Y)} P_{X|Y}(x|y') P_Y(y')}$$

对于连续随机变量 X 和 Y ：

$$f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)} = \frac{f_{X|Y}(x|y) f_Y(y)}{\int_{-\infty}^{\infty} f_{X|Y}(x|y') f_Y(y') dy'}$$

3.6 独立性

如果对于 X 和 Y 的所有值， $F_{XY}(x, y) = F_X(x) F_Y(y)$ ，则两个随机变量 X 和 Y 是独立的。等价地，

- 对于离散随机变量，对于任意 $x \in \text{Val}(X)$ ， $y \in \text{Val}(Y)$ ， $p_{XY}(x, y) = p_X(x) p_Y(y)$ 。
- 对于离散随机变量， $p_{Y|X}(y|x) = p_Y(y)$ 当对于任意 $y \in \text{Val}(Y)$ 且 $p_X(x) \neq 0$ 。
- 对于连续随机变量， $f_{XY}(x, y) = f_X(x) f_Y(y)$ 对于任意 $x, y \in \mathbb{R}$ 。

- 对于连续随机变量, $f_{Y|X}(y|x) = f_Y(y)$, 当 $f_X(x) \neq 0$ 对于任意 $y \in \mathbb{R}$ 。

非正式地说, 如果“知道”一个变量的值永远不会对另一个变量的条件概率分布有任何影响, 那么两个随机变量 X 和 Y 是独立的, 也就是说, 你只要知道 $f(x)$ 和 $f(y)$ 就知道关于这对变量 (X, Y) 的所有信息。以下引理将这一观察形式化:

引理3.1

如果 X 和 Y 是独立的, 那么对于任何 $A, B \subseteq \mathbb{R}$, 我们有:

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

利用上述引理, 我们可以证明如果 X 与 Y 无关, 那么 X 的任何函数都与 Y 的任何函数无关。

3.7 期望和协方差

假设我们有两个离散的随机变量 X, Y 并且 $g: \mathbf{R}^2 \rightarrow \mathbf{R}$ 是这两个随机变量的函数。那么 g 的期望值以如下方式定义:

$$E[g(X, Y)] \triangleq \sum_{x \in \text{Val}(X)} \sum_{y \in \text{Val}(Y)} g(x, y) p_{XY}(x, y)$$

对于连续随机变量 X, Y , 类似的表达式是:

$$E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{XY}(x, y) dx dy$$

我们可以用期望的概念来研究两个随机变量之间的关系。特别地, 两个随机变量的**协方差**定义为:

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

使用类似于方差的推导, 我们可以将它重写为:

$$\begin{aligned} \text{Cov}[X, Y] &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY - XE[Y] - YE[X] + E[X]E[Y]] \\ &= E[XY] - E[X]E[Y] - E[Y]E[X] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

在这里, 说明两种协方差形式相等的关键步骤是第三个等号, 在这里我们使用了这样一个事实, 即 $E[X]$ 和 $E[Y]$ 实际上是常数, 可以被提出来。当 $\text{cov}[X, Y] = 0$ 时, 我们说 X 和 Y 不相关。

性质:

- (期望线性) $E[f(X, Y) + g(X, Y)] = E[f(X, Y)] + E[g(X, Y)]$
- $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] + 2\text{Cov}[X, Y]$
- 如果 X 和 Y 相互独立, 那么 $\text{Cov}[X, Y] = 0$
- 如果 X 和 Y 相互独立, 那么 $E[f(X)g(Y)] = E[f(X)]E[g(Y)]$ 。

4. 多个随机变量

上一节介绍的概念和想法可以推广到两个以上的随机变量。特别是, 假设我们有 n 个连续随机变量, $X_1(\omega), X_2(\omega), \dots, X_n(\omega)$ 。在本节中, 为了表示简单, 我们只关注连续的情况, 对离散随机变量的推广工作类似。

4.1 基本性质

我们可以定义 X_1, X_2, \dots, X_n 的**联合累积分布函数**、**联合概率密度函数**, 以及给定 X_2, \dots, X_n 时 X_1 的**边缘概率密度函数**为:

$$F_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n)$$

$$f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) = \frac{\partial^n F_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n)}{\partial x_1 \dots \partial x_n}$$

$$f_{X_1}(X_1) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) dx_2 \dots dx_n$$

$$f_{X_1|X_2, \dots, X_n}(x_1|x_2, \dots, x_n) = \frac{f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n)}{f_{X_2, \dots, X_n}(x_2, \dots, x_n)}$$

为了计算事件 $A \subseteq \mathbb{R}^n$ 的概率, 我们有:

$$P((x_1, x_2, \dots, x_n) \in A) = \int_{(x_1, x_2, \dots, x_n) \in A} f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n$$

链式法则:

从多个随机变量的条件概率的定义中, 可以看出:

$$\begin{aligned} f(x_1, x_2, \dots, x_n) &= f(x_n|x_1, x_2, \dots, x_{n-1}) f(x_1, x_2, \dots, x_{n-1}) \\ &= f(x_n|x_1, x_2, \dots, x_{n-1}) f(x_{n-1}|x_1, x_2, \dots, x_{n-2}) f(x_1, x_2, \dots, x_{n-2}) \\ &= \dots = f(x_1) \prod_{i=2}^n f(x_i|x_1, \dots, x_{i-1}) \end{aligned}$$

独立性: 对于多个事件, A_1, \dots, A_k , 我们说 A_1, \dots, A_k 是相互独立的, 当对于任何子集 $S \subseteq \{1, 2, \dots, k\}$, 我们有:

$$P(\cap_{i \in S} A_i) = \prod_{i \in S} P(A_i)$$

同样, 我们说随机变量 X_1, X_2, \dots, X_n 是独立的, 如果:

$$f(x_1, \dots, x_n) = f(x_1)f(x_2)\dots f(x_n)$$

这里, 相互独立性的定义只是两个随机变量独立性到多个随机变量的自然推广。

独立随机变量经常出现在机器学习算法中, 其中我们假设属于训练集的训练样本代表来自某个未知概率分布的独立样本。为了明确独立性的重要性, 考虑一个“坏的”训练集, 我们首先从某个未知分布中抽取一个训练样本 $(x^{(1)}, y^{(1)})$, 然后将完全相同的训练样本的 $m - 1$ 个副本添加到训练集中。在这种情况下, 我们有:

$$P\left(\left(x^{(1)}, y^{(1)}\right), \dots, \left(x^{(m)}, y^{(m)}\right)\right) \neq \prod_{i=1}^m P\left(x^{(i)}, y^{(i)}\right)$$

尽管训练集的大小为 m , 但这些例子并不独立! 虽然这里描述的过程显然不是为机器学习算法建立训练集的明智方法, 但是事实证明, 在实践中, 样本的不独立性确实经常出现, 并且它具有减小训练集的“有效大小”的效果。

4.2 随机向量

假设我们有 n 个随机变量。当把所有这些随机变量放在一起工作时, 我们经常会发现把它们放在一个向量中是很方便的...我们称结果向量为随机向量(更正式地说, 随机向量是从 Ω 到 \mathbb{R}^n 的映射)。应该清楚的是, 随机向量只是处理 n 个随机变量的一种替代符号, 因此联合概率密度函数和综合密度函数的概念也将适用于随机向量。

期望:

考虑 $g: \mathbb{R}^n \rightarrow \mathbb{R}$ 中的任意函数。这个函数的期望值 被定义为

$$\begin{aligned} E[g(X)] &= \int_{\mathbb{R}^n} g(x_1, x_2, \dots, x_n) f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n E[g(X)] \\ &= \int_{\mathbb{R}^n} g(x_1, x_2, \dots, x_n) f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n \end{aligned}$$

其中, $\int_{\mathbb{R}^n}$ 是从 $-\infty$ 到 ∞ 的 n 个连续积分。如果 g 是从 \mathbb{R}^n 到 \mathbb{R}^m 的函数, 那么 g 的期望值是输出向量的元素期望值, 即, 如果 g 是:

$$g(x) = \begin{bmatrix} g_1(x) \\ g_2(x) \\ \vdots \\ g_m(x) \end{bmatrix}$$

那么,

$$E[g(X)] = \begin{bmatrix} E[g_1(X)] \\ E[g_2(X)] \\ \vdots \\ E[g_m(X)] \end{bmatrix}$$

协方差矩阵: 对于给定的随机向量 $X: \Omega \rightarrow \mathbb{R}^n$, 其协方差矩阵 Σ 是 $n \times n$ 平方矩阵, 其输入由 $\Sigma_{ij} = \text{Cov}[X_i, X_j]$ 给出。从协方差的定义来看, 我们有:

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

其中矩阵期望以明显的方式定义。

协方差矩阵有许多有用的属性:

- $\Sigma \succeq 0$; 也就是说, Σ 是正半定的。
- $\Sigma = \Sigma^T$; 也就是说, Σ 是对称的。

4.3 多元高斯分布

随机向量上概率分布的一个特别重要的例子叫做多元高斯或多元正态分布。随机向量 $X \in \mathbb{R}^n$ 被认为具有多元正态(或高斯)分布, 当其具有均值 $\mu \in \mathbb{R}^n$ 和协方差矩阵 $\Sigma \in \mathbb{S}_{++}^n$ (其中 \mathbb{S}_{++}^n 指对称正定 $n \times n$ 矩阵的空间)

$$f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right)$$

我们把它写成 $X \sim \mathcal{N}(\mu, \Sigma)$ 。请注意, 在 $n = 1$ 的情况下, 它降维成普通正态分布, 其中均值参数为 μ_1 , 方差为 Σ_{11} 。

一般来说，高斯随机变量在机器学习和统计中非常有用，主要有两个原因：

首先，在统计算法中对“噪声”建模时，它们非常常见。通常，噪声可以被认为是影响测量过程的大量小的独立随机扰动的累积；根据中心极限定理，独立随机变量的总和将趋向于“看起来像高斯”。

其次，高斯随机变量便于许多分析操作，因为实际中出现的许多涉及高斯分布的积分都有简单的封闭形式解。我们将在本课程稍后遇到这种情况。

5. 其他资源

一本关于**CS229**所需概率水平的好教科书是谢尔顿·罗斯的《概率第一课》(*A First Course on Probability* by Sheldon Ross)。