第三章 概率与信息论

1 概率分布

- 随机变量：随机变量可以是离散的或者连续的。离散随机变量拥有有限或者可数无限多的状态；连续随机变量伴随着实数值。
- **1.1 离散型变量和概率质量函数**

➢ 单个变量分布：$\mathrm{x} = x$ 的概率用 $P(x)$ 来表示

➢ 联合概率分布：$P(\mathrm{x} = x, \mathrm{y} = y)$ 表示 $\mathrm{x} = x$ 和 $\mathrm{y} = y$ 同时发生的概率

- 离散型变量和概率质量函数满足条件：

➢ P 的定义域必须是 x 所有可能状态的集合。

➢ $\forall x \in X, 0 \leq P(x) \leq 1.$

➢ $\sum_{x \in X} P(x) = 1.$ （归一化）

1 概率分布

- **1.2 连续型变量和概率密度函数**
- 连续型变量和概率密度函数满足条件:
  - P 的定义域必须是 x 所有可能状态的集合。
  - $\forall x \in x, p(x) \geq 0$.并不要求 $p(x) \leq 1$。
  - $\int p(x)dx = 1.$
- 概率密度函数 p(x) 并没有直接对特定的状态给出概率，相对的，它给出了落在面积为 δx 的无限小的区域内的概率为 $p(x)\delta x$
- 在单变量的例子中，x 落 在区间 [a,b] 的概率是$\int_{[a,b]} p(x)dx$。

2 边缘概率

- 对于一组变量的联合概率分布，想要了解其中一个子集的概率分布。这种定义在子集上的概率分布被称为边缘概率分布（marginal probability distribution）。
- 离散型随机变量
- 已知 P(x,y)，可以依据下面的求和法则（sum rule）来计算 P(x):
  - $\forall x \in \mathrm{x}, P(\mathrm{x} = x) = \sum_y P(\mathrm{x} = x, \mathrm{y} = y)$
- 连续型变量
  - $p(x) = \int p(x, y)dy$ （模拟二维空间）

3 条件概率

- 某个事件x，在给定其他事件y发生时出现的概率，叫做条件概率。将给定 x = x，y = y 发生的条件概率记为 P(y = y |x = x)。
- 这个条件概率可以通过下面的公式计算（P(x = x) > 0）：
  - $P(\mathrm{y} = y \,|\mathrm{x} = x) = \frac{P(\mathrm{y} = y, \mathrm{x} = x)}{P(\mathrm{x} = x)}$.

4 条件概率的链式法则

- 任何多维随机变量的联合概率分布，都可以分解成只有一个变量的条件概率相 乘的形式:

$$P(x^{(1)}, \ldots, x^{(n)}) = P(x^{(1)}) \prod_{i=2}^{n} P(x^{(i)} | x^{(1)}, \ldots, x^{(i-1)}).$$

- 这个规则被称为概率的链式法则（chain rule）或者乘法法则（product rule）。它可以直接从式(3.5)条件概率的定义中得到。例如，使用两次定义可以得到：

  - $P(a, b, c) = P(a|b, c)P(b, c)$
  - $P(b, c) = P(b|c)P(c)$
  - $P(a, b, c) = P(a|b, c)P(b|c)P(c).$

5 独立性和条件独立性

- **相互独立**
- x 和 y的概率分布可以表示成两个因子的乘积形式
  - $\forall x \in \mathrm{x}, y \in \mathrm{y}, p(\mathrm{x} = x, \mathrm{y} = y) = p(\mathrm{x} = x)p(\mathrm{y} = y).$
- **条件独立**
- x 和 y 的条件概率分布对于 z 的每一个值都可以写成乘积的形式， 那么这两个随机变量 x 和 y 在给定随机变量 z 时是条件独立的（conditionally independent）：
  - $\forall x \in \mathrm{x}, y \in \mathrm{y}, z \in \mathrm{z}, p(\mathrm{x} = x, \mathrm{y} = y \,|\mathrm{z} = z) = p(\mathrm{x} = x \,|\mathrm{z} = z)p(\mathrm{y} = y \,|\mathrm{z} = z).$

6 期望、方差和协方差

- **期望：**函数 f(x) 关于某分布 P(x) 的期望（expectation）或者期望值（expected value）是指，当 x 由 P 产生，f 作用于 x 时，f(x) 的平均值
- 离散型随机变量期望值：
  - $\mathbb{E}_{x\sim P}[f(x)] = \sum_x P(x)f(x)$
- 连续型随机变量期望值：
  - $\mathbb{E}_{x\sim P}[f(x)] = \int p(x)f(x)dx$
- **方差**（variance）衡量的是当我们对 x 依据它的概率分布进行采样时，随机变 量 x 的函数值会呈现多大的差异：
  - $Var(f(x)) = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2]$

6 期望、方差和协方差

- **协方差**（covariance）在某种意义上给出了两个变量线性相关性的强度以及这些变量的尺度：
  - $Cov(f(x), g(y)) = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])(g(y) - \mathbb{E}[g(y)])]$.
- 随机向量 x∈Rn 的协方差矩阵（covariance matrix）是一个 n×n 的矩阵，并 且满足
  - $Cov(x)_{i,j} = Cov(x_i, x_j)$.
- 协方差矩阵的对角元是方差：
  - $Cov(x_i, x_i) = Var(x_i)$.

7 常用概率分布

- **1 Bernoulli 分布**
- Bernoulli 分布（Bernoulli distribution）是单个二值随机变量的分布，它由单 个参数 $\phi \in [0,1]$ 控制，ϕ 给出了随机变量等于 1 的概率。
- 性质:

$$P(\mathrm{x} = 1) = \phi$$

$$P(\mathrm{x} = 0) = 1 - \phi$$

$$P(\mathrm{x} = x) = \phi^{x}(1-\phi)^{1-x}$$

$$\mathbb{E}_{x}[\mathrm{x}] = \phi$$

$$Var_{x}(\mathrm{x}) = \phi(1-\phi)$$

7 常用概率分布

- **2 Multinoulli 分布**
- Multinoulli 分布（multinoulli distribution）或者范畴分布（categorical distribution）是指在具有 k 个不同状态的单个离散型随机变量上的分布，其中 k 是一 个有限值。Multinoulli 分布由向量 $p \in [0,1]^{k-1}$ 参数化，其中每一个分量 pi 表示 第 i 个状态的概率。最后的第 k 个状态的概率可以通过 1−1⊤p 给出，其中，限制 1⊤p≤ 1
- $P\{X = k\} = C_n^k p_x^k (1 - p_x)^{n-k}$

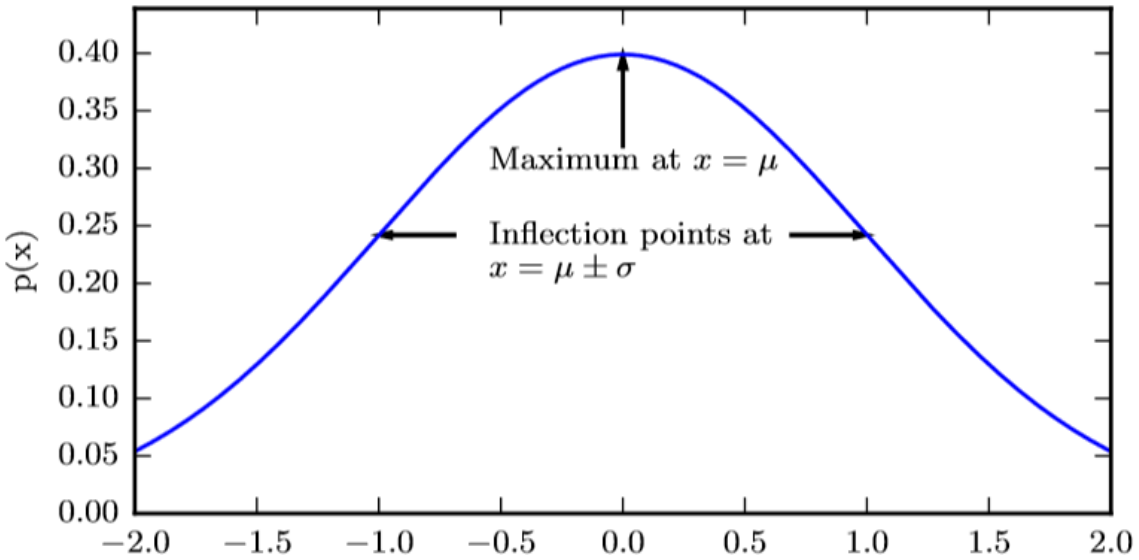
7 常用概率分布

- **3 高斯分布**
- 正态分布（normal distribution），也称为高斯分布 （Gaussian distribution）：

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

- 正态分布由两个参数控制，μ ∈ R 和 σ ∈ (0, ∞)。

参数 μ 给出了中心峰值的坐 标，这也是分布的均值：E[x] = μ。分布的标准差用 σ 表示，方差用 σ2 表示



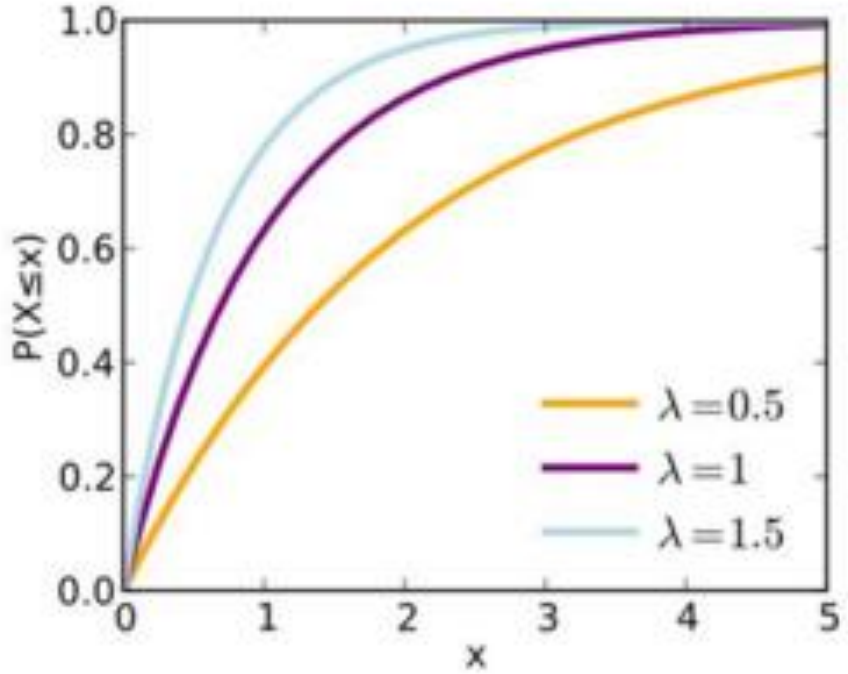
7 常用概率分布

- **3 高斯分布**
- 采用正态分布在很多应用中都是一个明智的选择:
- 中心极限定理（central limit theorem）说明很多独立随机变量的和近似服从正态分布。
- 在具有相同方差的所有可能的概率分布中，正态分布在实数上具有最大的不确定性。因此，可以认为正态分布是对模型加入的先验知识量最少的分布。
- 正态分布可以推广到 Rn 空间，这种情况下被称为多维正态分布（multivariate normal distribution）。它的参数是一个正定对称矩阵 Σ:
  - $N(x;\mu,\Sigma) = \sqrt{\frac{1}{(2\pi)^n \det(\Sigma)}}\exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))$.

7 常用概率分布

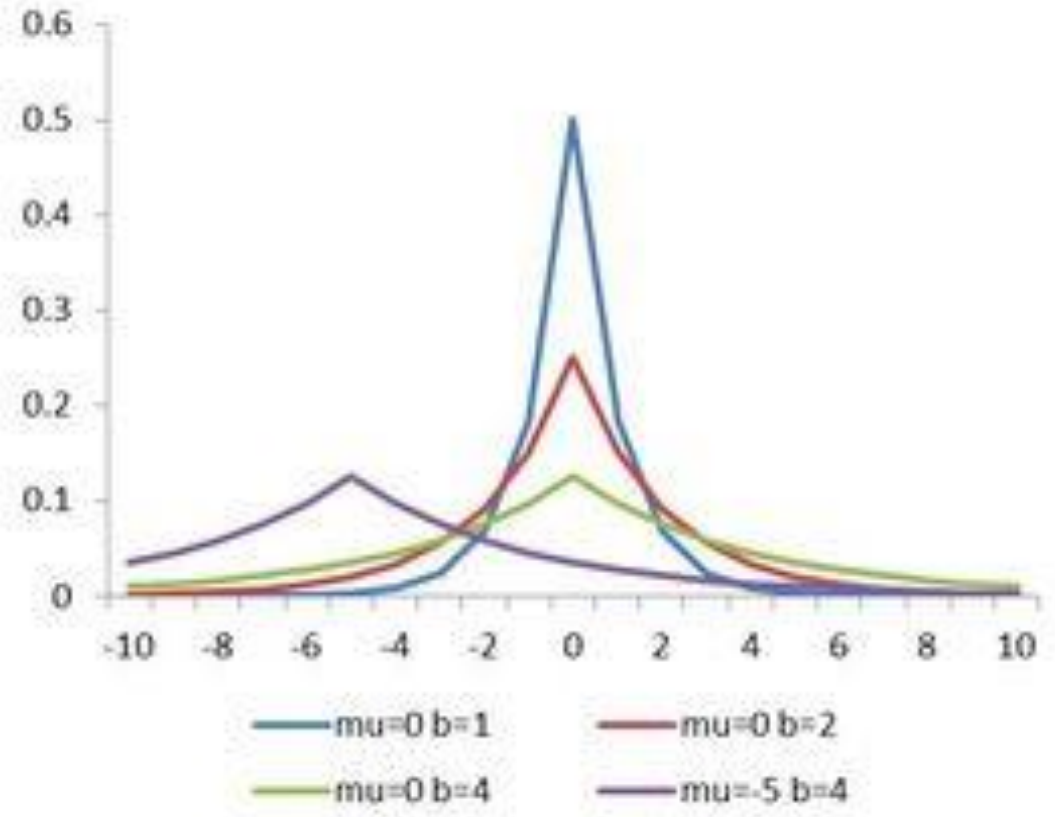
- **4 指数分布和 Laplace 分布**
- 指数分布：在 x = 0 点处取得边界点 (sharp point) 的 分布。
  - $p(x;\lambda) = \lambda 1_{x\geq 0}\exp(-\lambda x)$.

$$F(x;\lambda) = \begin{cases} 1-e^{-\lambda x} & , x \geq 0, \\ 0 & , x < 0. \end{cases}$$



7 常用概率分布

- **4 指数分布和 Laplace 分布**
- 一个联系紧密的概率分布是Laplace 分布（Laplace distribution）， 它允许我们 在任意一点 μ 处设置概率质量的峰值 。
- $Laplace(x;\mu,\gamma) = \frac{1}{2\gamma}\exp(-\frac{|x-\mu|}{\gamma})$.



7 常用概率分布

- **5 Dirac 分布和经验分布**
- Dirac delta 函数（Dirac delta function）δ(x) 定义概率密度函数：
  - $p(x) = \delta(x - \mu).$
- Dirac delta 函数被定义成在除了 0 以外的所有点的值都为 0，但是积分为 1，也被称为广义函数（generalized function）。通过把 p(x) 定义成 δ 函数左移 −μ 个单位，我们得到了一个在 x = μ 处具有 无限窄也无限高的峰值的概率质量。
- Dirac 分布经常作为经验分布（empirical distribution）的一个组成部分出现：
  - $\hat{p}(x) = \frac{1}{m}\sum_{i=1}^{m}\delta(x - x^{(i)})$
- 经验分布将概率密度 1 m 赋给 m 个点$x^{(1)}, \ldots, x^{(m)}$中的每一个。对于离散型随机变量，经验分布可以被定义成一 个 Multinoulli 分布

7 常用概率分布

- **6 混合分布**
- 混合分布：组合一些简单的概率分布来定义新的概率分布
- 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 Multinoulli 分 布中采样的结果：
- $$P(x) = \sum_i P(\mathrm{c} = i) P(\mathrm{x} | \mathrm{c} = i)$$
- 这里 P(c) 是对各组件的一个 Multinoulli 分布

7 常用概率分布

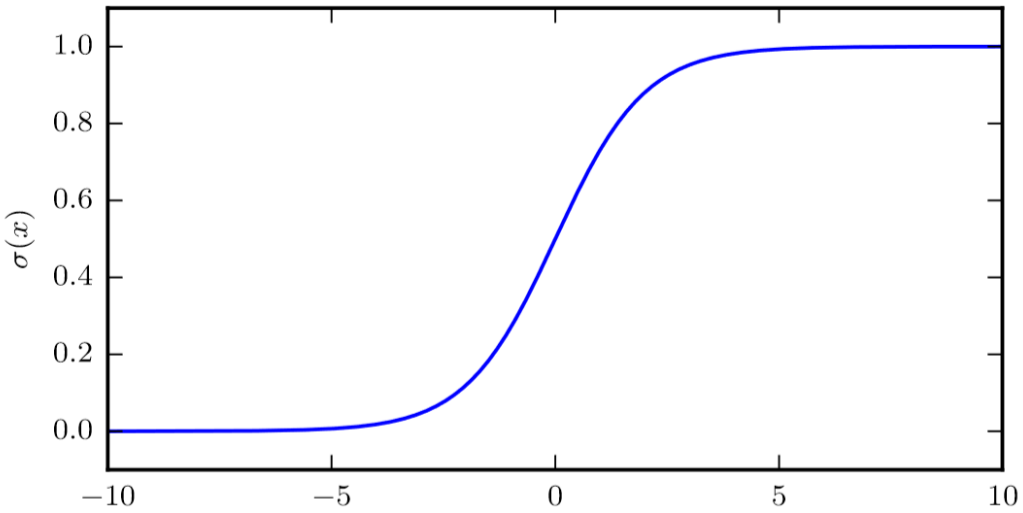
- **6 混合分布**
- 一个非常强大且常见的混合模型是**高斯混合模型**（Gaussian Mixture Model），它的组件$P(\mathrm{x}|\mathrm{c} = i)$是高斯分布。每个组件都有各自的参数，均值 $\mu^{(i)}$ 和协方差矩 阵$\Sigma^{(i)}$ 。除了均值和协方差以外，高斯混合模型的参数指明了给每个组件 i 的先验概率 （prior probability）$\alpha_i = P(\mathrm{c} = i)$。

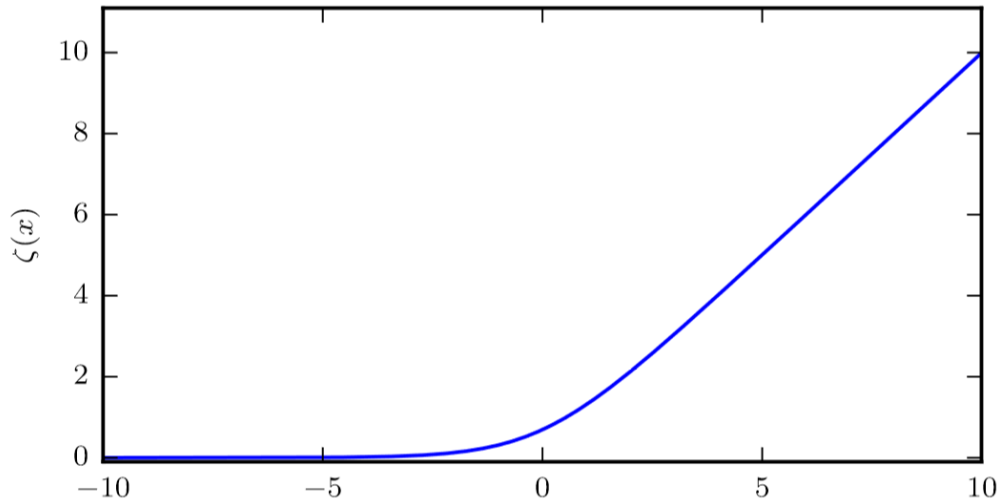
$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$p(x) = \sum_{i=1}^{K} \phi_i \frac{1}{\sqrt{2\sigma_i^2\pi}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}}$$

8 常用函数的有用性质

- **1 logistic sigmoid 函数**
- $\sigma(x) = \frac{1}{1+\exp(-x)}$
- 取值范围为（0，1）
- **2 softplus 函数**
- $\zeta(x) = \log(1 + \exp(x))$.
- 来源于另外一个函数的平滑（软化）形式:
- $x^+ = \max(0, x)$.
- 取值范围为$(0, \infty)$





8 常用函数的有用性质

- **拓展公式**

- $\sigma(x) = \frac{\exp(x)}{\exp(x)+\exp(0)}$
- $\frac{d}{dx}\sigma(x) = \sigma(x)(1-\sigma(x))$
- $1-\sigma(x) = \sigma(-x)$
- $\log\sigma(x) = -\zeta(-x)$
- $\frac{d}{dx}\zeta(x) = \sigma(x)$
- $\forall x \in (0,1), \sigma - 1(x) = \log(\frac{x}{1-x})$
- $\forall x > 0, \zeta^{-1}(x) = \log(\exp(x)-1)$
- $\zeta(x) = \int_{-\infty}^{x} \sigma(y)dy$
- $\zeta(x) - \zeta(-x) = x$

9 贝叶斯规则

- **目标**
- 在已知 $P(\mathrm{y}|\mathrm{x})$ 时计算 $P(\mathrm{x}|\mathrm{y})$，同时还知道$P(\mathrm{x})$。
- **公式：**
- $P(\mathrm{x}|\mathrm{y}) = \frac{P(\mathrm{x})P(\mathrm{y}|\mathrm{x})}{P(\mathrm{y})}$
- $P(\mathrm{y})$ 出现在上面的公式中，它通常使用 $P(\mathrm{y}) = \sum_x P(\mathrm{y} \mid \mathrm{x})P(\mathrm{x})$ 来计算， 所以我们并不需要事先知道 $P(\mathrm{y})$ 的信息。

10 连续型变量的技术细节

- **1 测度论**
- 对于连续性随机变量的分布，可能会有以下悖论：
- 连续型向量值随机变量 x 落在某个集合 S 中的 概率是通过 p(x) 对集合 S 积分得到的。对于集合 S 的一些选择可能会引起悖论。例 如，构造两个集合 S1 和 S2 使得 $p(x \in S1) + p(x \in S2) > 1$ 并且 $S1 \cap S2 = \varnothing$ 是可能 的。这些集合通常是大量使用了实数的无限精度来构造的，例如通过构造分形形状 (fractal-shaped) 的集合（非整数维）
- 测度论用一种比较严格的定义描述那些非常微小的点集。这种集合被称为 “零测度（measure zero）’’ 的。直观地理解这个概念是有用的，我们可以认为零测度集在我们的度量空间中不占有任何的体积

10 连续型变量的技术细节

- **1 非可逆**
- 连续型随机变量的另一技术细节，涉及到处理那种相互之间有确定性函数关系 的连续型变量。假设我们有两个随机变量 x 和 y 满足 y = g(x)，其中 g 是可逆的、 连续可微的函数。可能有人会想 py(y) = px(g−1(y))。但实际上这并不对。
- 举一个简单的例子，假设我们有两个标量值随机变量 x 和 y，并且满足 $y = \frac{x}{2}$ 以及 x~ U(0,1)。

  如果我们使用 $py(y) = px(2y)$，那么 py 除了区间 $[0, \frac{1}{2}]$ 以外都为 0，并且在这个区间上的值为 1。这意味着
- $\int py(y)dy = \frac{1}{2}$
- 而这违背了概率密度的定义 (积分为 1)。这个常见错误之所以错是因为它没有考虑 到引入函数 g 后造成的空间变形。回忆一下，x 落在无穷小的体积为 δx 的区域内的 概率为 p(x)δx。因为 g 可能会扩展或者压缩空间，在 x 空间内的包围着 x 的无穷小体积在 y 空间中可能有不同的体积。

11 信息论

- 信息论是应用数学的一个分支，主要研究的是对一个信号包含信息的多少进行量化。在本书中，我们主要使用信息论的一些关键思想来描述概率分布或者量化概 率分布之间的相似性
- **性质**

➢ 非常可能发生的事件信息量要比较少，并且极端情况下，确保能够发生的事件 应该没有信息量。

➢ 较不可能发生的事件具有更高的信息量。

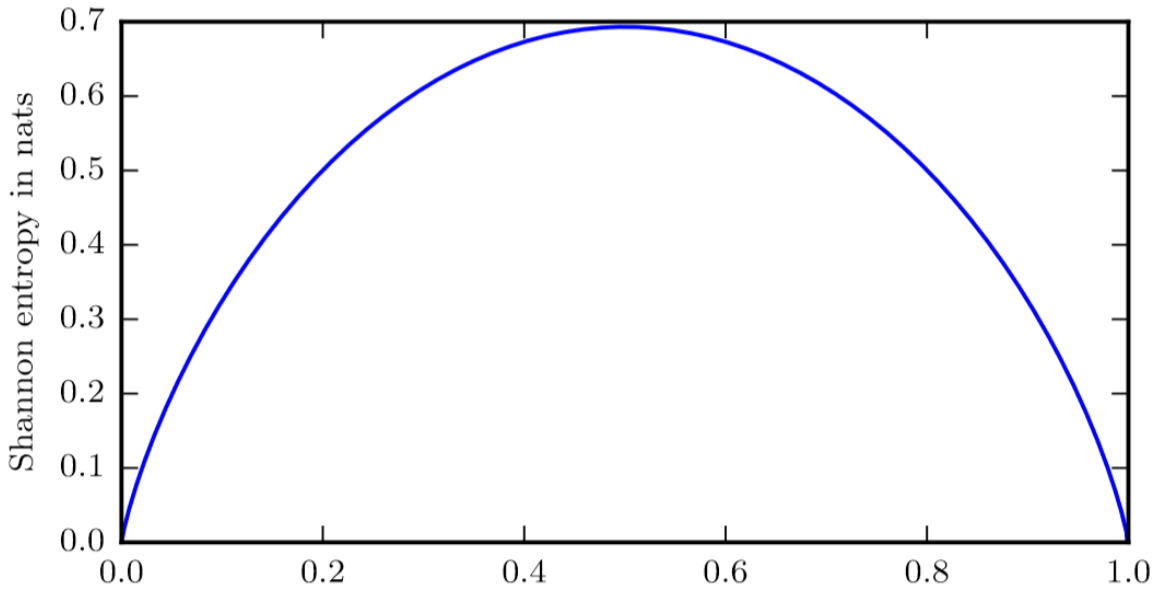
➢ 独立事件应具有增量的信息。例如，投掷的硬币两次正面朝上传递的信息量， 应该是投掷一次硬币正面朝上的信息量的两倍。

11 信息论

- 为了满足上述三个性质，我们定义一个事件 x = x 的**自信息**（self-information）为
- $I(x) = -logP(x).$ 单位是奈特（nats）
- 自信息只处理单个的输出。我们可以用**香农熵（Shannon entropy）**来对整个概率分布中的不确定性总量进行量化
- $H(x) = \mathbb{E}_{x\sim P}[I(x)] = -\mathbb{E}_{x\sim P}[logP(x)]$
- 换言之，一个分布的香农熵是指遵循这个分布的事件所产生的期望信息总量。
- 当 x 是连续的，**香农熵**被称为**微分熵**（differential entropy）。

11 信息论

- 二值随机变量的香农熵。该图说明了更接近确定性的分布是如何具有较低的香农熵，而更接近均匀分布的分布是如何具有较高的香农熵。
- 水平轴是 p，表示二值随机变量等于
- 1 的概率。熵由 $(p-1)\log(1-p)-p\log p$
- 给出。当 p 接近 0 时，分布几乎是确定
- 的，因为随机变量几乎总是 0。当 p 接
- 近 1 时，分布也几乎是确定的，因为随
- 机变量几乎总是 1。当 p = 0.5 时，熵是
- 最大的， 因为分布在两个结果（0 和 1）上是均匀的。

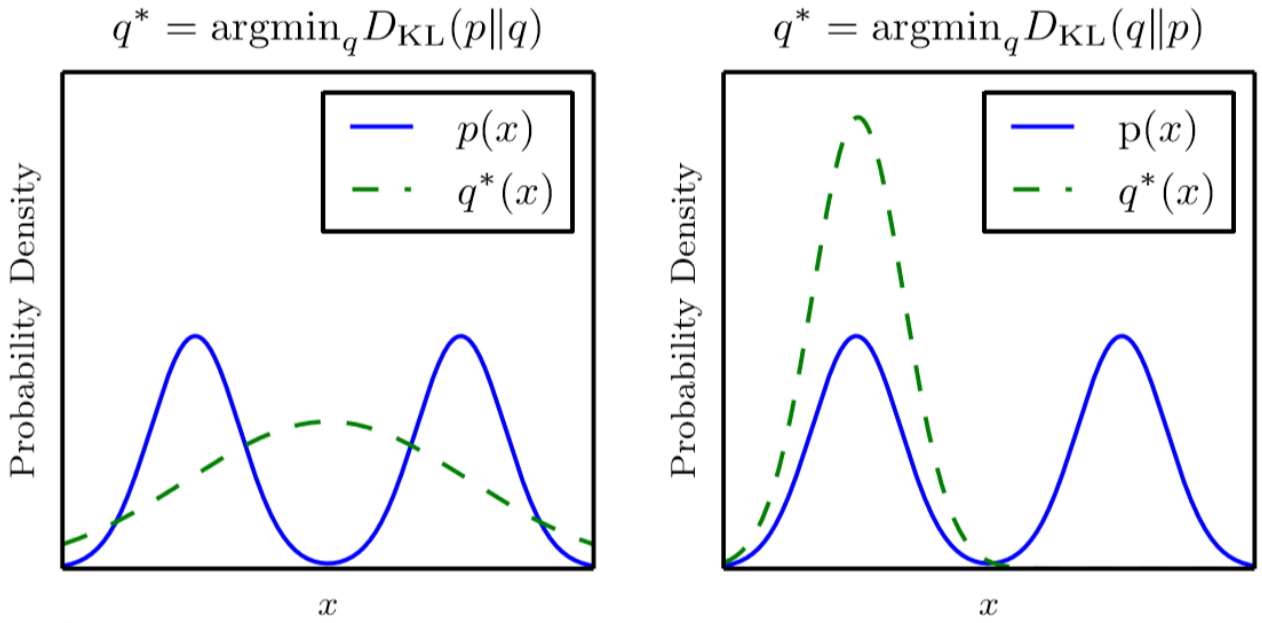


11 信息论

- KL散度
- 如果我们对于同一个随机变量 x 有两个单独的概率分布 P(x) 和 Q(x)，我们可 以使用**KL 散度**（Kullback-Leibler (KL) divergence）来衡量这两个分布的差异:
- $D_{KL}\left(P||Q\right) = \mathbb{E}_{x\sim P}[\log\frac{P(x)}{Q(x)}] = \mathbb{E}_{x\sim P}[\log P(x) - \log Q(x)]$
- 性质:
- 非负，KL散度为0当且仅当P和Q在离散型变量的情况下是相同的分布
- 它经常被用作分布之间的某种距离。但它并不是真的距离因为它不是对称的：对于某些 P 和 Q，$D_{KL}(P||Q) \neq D_{KL}(Q||P)$。

11 信息论

- KL散度



- 一个和 KL 散度密切联系的量是交叉熵（cross-entropy）$H(P,Q) = H(P) + D_{KL}(P||Q)$，它和 KL 散度很像但是缺少左边一项：

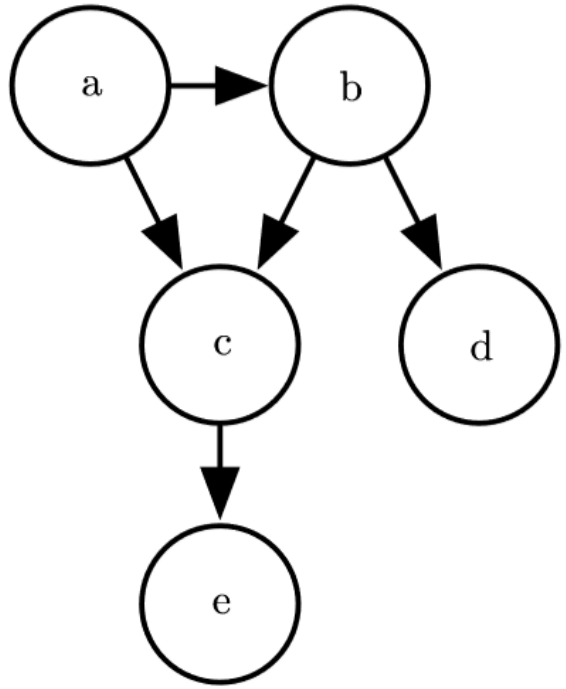
- $H(P,Q) = -\mathbb{E}_{x\sim P}\log Q(x)$.

12 结构化概率模型

- 使用单个函数来描述整个联合概率分布是非常低效的 (无论是计算上还是统计上)，我们可以把概率分布分解成许多因子的乘积形式，而不是使用单一的函数来表示概率分布。
- 例如，假设我们有三个随机变量 a,b 和 c，并且 a 影响 b 的取值，b 影 响 c 的取值，但是 a 和 c 在给定 b 时是条件独立的。我们可以把全部三个变量的概 率分布重新表示为两个变量的概率分布的连乘形式:
  - $p(a, b, c) = p(a)p(b|a)p(c|b)$.
- 这种分解可以极大地减少用来描述一个分布的参数数量。
- 我们可以用图来描述这种分解。这里我们使用的是图论中的 ‘‘图’’ 的概念：由 一些可以通过边互相连接的顶点的集合构成。当我们用图来表示这种概率分布的分 解，我们把它称为结构化概率模型（structured probabilistic model）或者图模型 （graphical model）。

12 结构化概率模型

- 有向（directed）模型
- 有向模型对于分布中的每一个随机变量 x_i 都包含着一个 影响因子，这个组成 x_i条件概率的影响因子被称为 x_i的父节点，记为 $Pa_G(x_i)$:
- $p(x) = \prod_i p(x_i \mid Pa_G(x_i))$
- 关于随机变量 a,b,c,d 和 e 的有向图模型。这幅图对应的
- 概率分布可以分解为
- $p(a,b,c,d,e) = p(a)p(b|a)p(c|a,b)p(d|b)p(e|c)$.



12 结构化概率模型

- 无向（undirected）模型
- 它们将分解表示成一组函数；不 像有向模型那样，这些函数通常不是任何类型的概率分布。G 中任何满足两两之 间有边连接的顶点的集合被称为团。无向模型中的每个团 $C^{(i)}$ 都伴随着一个因子 $\phi^{(i)}(C^{(i)})$。这些因子仅仅是函数，并不是概率分布。
- 随机变量的联合概率与所有这些因子的乘积成比例（proportional）——意味着 因子的值越大则可能性越大。当然，不能保证这种乘积的求和为 1
- 所以我们需要除 以一个归一化常数 Z 来得到归一化的概率分布，归一化常数 Z 被定义为 ϕ 函数乘 积的所有状态的求和或积分。概率分布为：
- $p(x) = \frac{1}{Z} \prod_i \phi^{(i)}(C^{(i)})$

12 结构化概率模型

- 无向（undirected）模型
- 关于随机变量 a,b,c,d 和 e 的无向图模型。这幅图对应的概率分布可以分解为
- $p(a,b,c,d,e) = \frac{1}{Z}\phi^{(1)}(a,b,c)\phi^{(2)}(b,d)\phi^{(3)}(c,e)$.